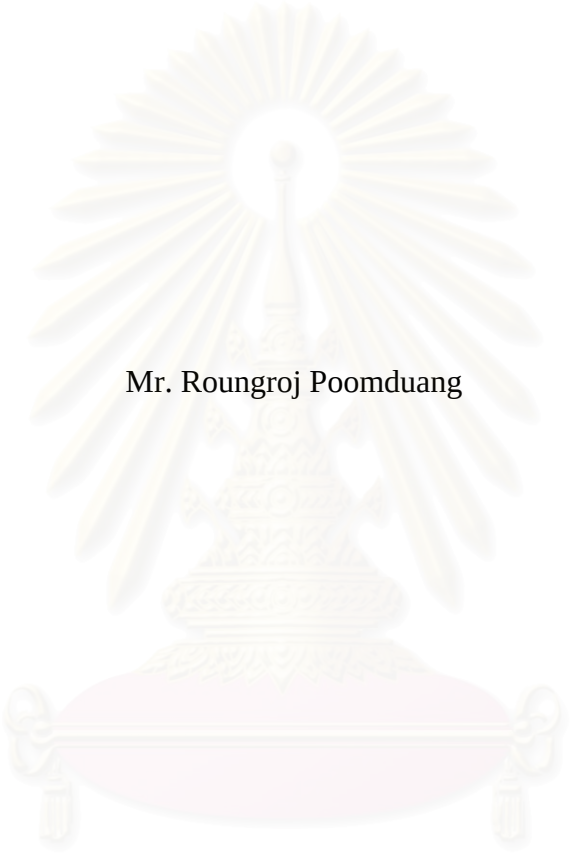
การปรับปรุงขั้นตอนวิธีแบ่งส่วนรอบเมดอยด์

นายรุ่งโรจน์ พุ่มดวง

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาการคณนา ภาควิชาคณิตศาสตร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2550

IMPROVEMENT OF PARTITIONING AROUND MEDOID ALGORITHM

Mr. Roungroj Poomduang



A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Computational Science
Department of Mathematics
Faculty of Science
Chulalongkorn University
Academic Year 2007

หัวข้อวิทยานิพนธ์	การปรับปรุงขั้นตอนวิธีแบ่งส่วนรอบเมดอยด์
โดย	นายรุ่งโรจน์ พุ่มดวง
สาขาวิชา	วิทยาการคณนา
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร. กรุง สินอภิรมย์สราญ

คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะวิทยาศาสตร์
(ศาสตราจารย์ ดร. สุพจน์ หารหนองบัว)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร. จารุโลจน์ จงสถิตย์วัฒนา)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์
(ผู้ช่วยศาสตราจารย์ ดร. กรุง สินอภิรมย์สราญ)

..... กรรมการ
(อาจารย์ ดร. สิริพันธุ์ สงวนสินธุกุล)

รุ่งโรจน์ พุ่มดวง: การปรับปรุงขั้นตอนวิธีแบ่งส่วนรอบเมดอยด์. (IMPROVEMENT PARTITIONING AROUND MEDOID ALGORITHM) อ. ที่ปรึกษา: ผศ. ดร. กรุง สินอภิรมย์สราญ, 49 หน้า.

การวิเคราะห์การเกาะกลุ่มเป็นหนึ่งในหลักการทำเหมืองข้อมูลซึ่งผู้วิจัยใช้แบ่งกั้นเซตข้อมูลออกเป็นกลุ่มย่อยขนาดเล็ก ในวิทยานิพนธ์นี้ เราสนใจการปรับปรุงเฉพาะขั้นตอนวิธีการเกาะกลุ่มที่เรียกว่า การแบ่งกั้นรอบเมดอยด์ (PAM) ขั้นตอนวิธีแพมค้นหาตัวแทนของแต่ละกลุ่มที่เรียก เมดอยด์ ซึ่งใช้คำนวณระยะทางรวมที่น้อยที่สุดจากเมดอยด์ไปยังจุดข้อมูลในกลุ่ม ในแต่ละรอบ วิธีแพมเลือกคู่สลับที่ดีที่สุดระหว่างเมดอยด์กับจุดข้อมูลโดยเปรียบเทียบระยะระหว่างทุกคู่ของการสลับ ทำให้ขั้นตอนวิธีแพมใช้เวลาประมวลผลนานต่อหนึ่งรอบการคำนวณ เราเสนอแนะการปรับปรุงขั้นตอนวิธีแพมผ่านขั้นตอนวิธี 4 แบบ ขั้นตอนวิธีแรกลดเวลาที่ใช้ในการเลือกคู่สลับที่ดีที่สุดโดยยอมรับคู่สลับคู่แรกที่ทำให้ผลรวมของระยะทางทั้งหมดดีขึ้น อย่างไรก็ดีวิธีการดังกล่าวทำให้จำนวนรอบของการประมวลผลมาก จึงทำให้ขั้นตอนวิธีแรกประมวลผลช้ากว่าขั้นตอนวิธีแพม ขั้นตอนวิธีที่สองปรับปรุงจากขั้นตอนวิธีแรกโดยการจัดเรียงคู่ของการสลับที่เหมาะสมก่อนการวนซ้ำ โดยเรียงลำดับเมดอยด์จาก 4 กลยุทธ์ วิธีการดังกล่าวเพิ่มประสิทธิภาพในการทำงานเพียงเล็กน้อย เพราะเมดอยด์ที่เหมาะสมยังไม่ถูกพบในตอนต้นของการทำงาน ขั้นตอนวิธีที่สามปรับปรุงจากขั้นตอนวิธีที่สองโดยเลือกคู่สลับคู่แรกที่ดีที่สุดก่อน แล้วจึงใช้วิธีปรับปรุงการวนซ้ำของรอบที่เหลือ วิธีการนี้ใช้ได้ดี อย่างไรก็ตามผลรวมของเวลาที่ใช้ในการประมวลผลยังคงช้ากว่าขั้นตอนวิธีแพม ดังนั้นเราเสนอขั้นตอนวิธีสุดท้ายซึ่งปรับปรุงมาจากขั้นตอนวิธีที่สาม โดยพิจารณาการเลือกคู่สลับในกลุ่มก่อน หลังจากในรอบแรกผ่านไป วิธีดังกล่าวแสดงเวลาประมวลผลที่ดีที่สุดในขั้นตอนวิธีทั้งหมดรวมทั้งแพม ในการทดลองของเรากับข้อมูลที่จำลองมาจาก 2 ถึง 20 มิติ ค่าเฉลี่ยของเวลาที่ใช้ในการประมวลผล 100 ตัวอย่างแสดงเวลาที่ดีกว่าในกลุ่มของขั้นตอนวิธีเหล่านี้ นอกจากนี้ สำหรับจำนวนจุดข้อมูลที่คงที่ มิติที่มากขึ้นนำไปสู่การใช้เวลาในการประมวลผลที่ลดลง

ภาควิชา **คณิตศาสตร์** ลายมือชื่อ.....รุ่งโรจน์ พุ่มดวง.....

สาขาวิชา **วิทยาการคณนา** ลายมือชื่ออาจารย์ที่ปรึกษา.....

ปีการศึกษา 2550

4772440023: MAJOR COMPUTATIONAL SCIENCE
KEY WORD: DATA MINING/ CLUSTERING/ PARTITIONING AROUND MEDOID

ROUNGROJ POOMDUANG: IMPROVEMENT PARTITIONING AROUND MEDOID ALGORITHM. THESIS ADVISOR: ASST. PROF. KRUNG SINAPIROMSARAN, Ph.D., 49 pp.

Cluster analysis is one of the data mining methodology which researchers use to partition data set into smaller groups. In this thesis, we are interested in improving one special clustering algorithm called Partitioning Around Medoid (PAM). PAM algorithm searches for the representatives for each group named medoid which constitutes the total minimal distance from a medoid to data points within a group. In each iteration, PAM determines the best swapping pair between a medoid and a data point by comparing all possible pairs. This causes PAM algorithm to run slowly per iteration. We suggest improving PAM algorithm via four successive algorithms. The first algorithm reduces the time to determine the best swapping pair by accepting the first swapping pair that improves the total distance. However, this increases the number of iterations that causes this first algorithm to run slower than PAM algorithm. The second algorithm improves from the first algorithm by ordering the most appropriate pair to be considered at the beginning of iterations by sorting medoids via 4 strategies. This gives a little improvement of the total running since the appropriate medoids are not found earlier. The third algorithm improves the second algorithm by choosing the best swapping pair for the first iteration, then it performs the iterative improvement for other iterations. This works well. However, the total running is still slower than PAM algorithm. Therefore, we suggest the last algorithm which improves the third algorithm by first considering the swapping pair within a medoid's group after the first iteration has been performed. This shows the best running time among all algorithms including PAM. In our experiment with the simulated data varying dimensions from 2 to 20, the average running time of 100 runs show the superior running time among other algorithms. Moreover, for a fixed number of data point, the larger the dimension, the lower the running time.

Department Mathematics Student's signature .....
Field of study Computational Science Advisor's signature .....
Academic year 2007

กิตติกรรมประกาศ

ผู้วิจัยขอกราบขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร. กรุง สินอภิรมย์สราญ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ได้กรุณาให้ความรู้ คำแนะนำ และคำปรึกษา ทำให้วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยดี

ขอกราบขอบพระคุณผู้ช่วยศาสตราจารย์ ดร. จารุโลจน์ จงสถิตย์วัฒนา ประธานกรรมการ อาจารย์ ดร. สิริพันธุ์ สงวนสินธุกุล กรรมการที่ได้ให้คำปรึกษา คำแนะนำและแก้ไขข้อบกพร่องต่างๆ ในงานวิจัยนี้ ซึ่งทำให้วิทยานิพนธ์ฉบับนี้มีความสมบูรณ์มากยิ่งขึ้น

ขอกราบขอบพระคุณบิดา มารดา ตลอดจนพี่น้องในครอบครัวและเพื่อนๆ ทุกคนที่คอยเป็นกำลังใจและช่วยเหลือผู้วิจัยมาโดยตลอด

สารบัญ

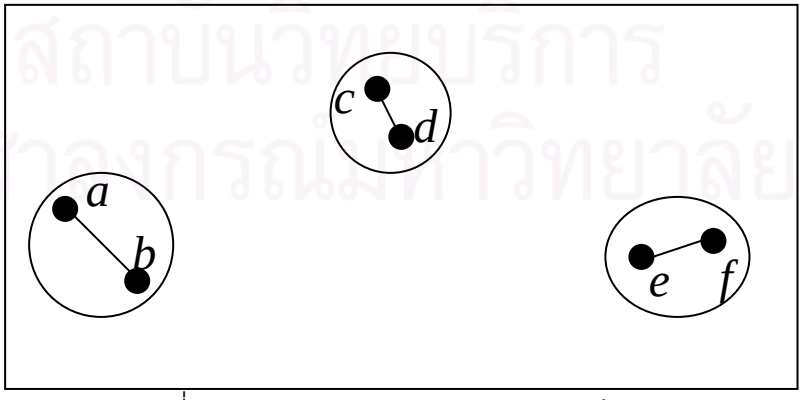
บทที่ 1
บทนำ

ความเป็นมา ความสำคัญของปัญหา และแนวคิดของงานวิจัย

ข้อมูลเป็นสิ่งสำคัญในงานด้านการตัดสินใจ การบริหารองค์กร โดยข้อมูลจะถูกนำมาวิเคราะห์และประมวลผลเพื่อให้ได้ประโยชน์สูงสุดขององค์กร ในอดีตการวิเคราะห์ข้อมูลจะอาศัยผู้ที่มีความเชี่ยวชาญและมีประสบการณ์สูง ทำให้เมื่อมีข้อมูลมาก การทำงานของผู้เชี่ยวชาญจะใช้เวลาในการวิเคราะห์ข้อมูลนาน ซึ่งปัจจุบันมนุษย์ได้คิดค้นวิธีการวิเคราะห์ข้อมูลที่เรียกว่าการทำเหมืองข้อมูล (data mining) [1, 2, 3, 4, 5] เพื่อทำการวิเคราะห์ข้อมูลได้รวดเร็วและสะดวกยิ่งขึ้น

การทำเหมืองข้อมูลเป็นการวิเคราะห์ข้อมูลและประมวลผลข้อมูลขนาดใหญ่เพื่อให้ได้ความรู้ ความสัมพันธ์ที่ซ่อนอยู่ในข้อมูล โดยการทำเหมืองข้อมูลอาศัยความรู้ และความเข้าใจหลายแขนง เช่น ระบบฐานข้อมูล การเรียนรู้ด้วยเครื่อง สถิติ คณิตศาสตร์ คอมพิวเตอร์ ส่งผลให้การทำเหมืองข้อมูลสามารถปรับประยุกต์และตอบคำถามในเชิงธุรกิจและเป็นที่ได้รับความนิยม และเกิดการพัฒนาขั้นตอนวิธีและเทคนิคขึ้นมากมาย ตัวอย่างเช่น การใช้เทคนิคการทำเหมืองข้อมูลสำหรับเว็บ (web mining) [6] การกำหนดแผนผังการวางตู้เอทีเอ็มโดยใช้เทคนิคการเกาะกลุ่มข้อมูล [7]

การเกาะกลุ่มข้อมูล คือวิธีการแยกกลุ่มข้อมูลออกเป็นกลุ่ม โดยดูลักษณะความคล้ายกันของข้อมูล ซึ่งลักษณะความคล้ายกันของข้อมูลจะถูกนิยามด้วยระยะทางระหว่างข้อมูล หมายถึงข้อมูลที่อยู่ใกล้กันจะมีความคล้ายกันกว่าข้อมูลที่อยู่ไกลกันดังรูปที่ 1.1



รูปที่ 1.1 การแบ่งโดยดูระยะทางระหว่างข้อมูล

a คล้ายกับ b , c คล้ายกับ d , e คล้ายกับ f แต่ a,b ไม่คล้ายกับ c,d หรือ e,f

วิธีการเกาะกลุ่มข้อมูลมีหลายวิธี เช่น Partitioning methods, Hierarchical methods, Density based method และอื่นๆ แต่ละวิธีจะมีขั้นตอนการดำเนินการที่แตกต่างกัน วิทยานิพนธ์นี้จะสนใจวิธี Partitioning methods [8, 9] โดยสนใจการปรับปรุงขั้นตอนวิธีแบ่งส่วนรอบเมดอยด์ (Partitioning Around Medoids) ที่เรียกว่าวิธีแพม (PAM) ซึ่งพัฒนาโดย Kaufman และ Housseeuw [10, 11, 12]

งานวิจัยเกี่ยวการปรับปรุงวิธีแพมมีอีกหลายงานวิจัย [13, 14, 15] ซึ่งส่วนใหญ่จะปรับปรุงวิธีแพมเพื่อให้สามารถจัดการกับข้อมูลที่มีขนาดใหญ่ แต่ยังคงหลักการทำงานของวิธีแพมที่รับประกันผลรวมระยะทางที่ได้น้อยที่สุด

ดังนั้นงานวิจัยนี้จะเสนอแนวทางการปรับปรุงวิธีแพม 4 ขั้นตอนวิธี โดยขั้นตอนวิธีแรกจะเป็นการปรับปรุงวิธีแพมโดยการลดเวลาในการประมวลผลใน 1 รอบของการสลับเมดอยด์ เนื่องจากวิธีแพมใช้เวลาในการประมวลผลใน 1 รอบของการสลับเมดอยด์นาน เพราะวิธีแพมจะพิจารณาคู่เมดอยด์กับจุดข้อมูลทุกจุด และเลือกคู่ที่ให้ผลรวมระยะทางลดลงมากที่สุด ดังนั้นขั้นตอนแรกจะลดเวลาการสลับใน 1 รอบด้วยการเลือกคู่เมดอยด์กับจุดข้อมูลที่ให้ผลรวมระยะทางลดลงค่าแรก ซึ่งวิธีการนี้จะทำให้จำนวนรอบของการสลับเพิ่มมากขึ้น จึงทำให้เวลาในการประมวลผลรวมของขั้นตอนวิธีแรกมากกว่าวิธีการแพม ขั้นตอนที่ 2 จึงเสนอวิธีการลดจำนวนรอบของการสลับเมดอยด์เพื่อจะลดเวลารวมของขั้นตอนวิธีแรก โดยการเรียงเมดอยด์ในการพิจารณาก่อน 4 กลยุทธ์ คือ การเรียงเมดอยด์ที่มีสมาชิกในกลุ่มมากก่อน การเรียงเมดอยด์ที่มีจำนวนสมาชิกในกลุ่มน้อยก่อน การเรียงเมดอยด์ที่มีผลรวมระยะทางระหว่างเมดอยด์กับสมาชิกภายในกลุ่มมากก่อน และการเรียงเมดอยด์ที่มีผลรวมระยะทางระหว่างเมดอยด์กับสมาชิกภายในกลุ่มน้อยก่อน ซึ่งผลของขั้นตอนวิธีที่ 2 สามารถลดจำนวนรอบของการสลับเมดอยด์ของขั้นตอนวิธีแรกได้เพียงเล็กน้อย และยังใช้เวลารวมมากกว่าวิธีแพม ขั้นตอนวิธีที่ 3 เป็นขั้นตอนวิธีเพื่อลดจำนวนรอบของการสลับเมดอยด์ของขั้นตอนวิธีที่ 2 ด้วยการเลือกคู่เมดอยด์กับจุดข้อมูลที่ให้ผลรวมระยะทางลดลงมากที่สุดในรอบแรกก่อน ซึ่งขั้นตอนวิธีที่ 3 สามารถลดจำนวนรอบของการสลับเมดอยด์ของขั้นตอนวิธีที่ 2 แต่ใช้เวลารวมมากกว่าวิธีแพม เนื่องจากขั้นตอนวิธีที่ 3 พิจารณาจุดข้อมูลกับเมดอยด์ที่ไม่เหมาะสม ขั้นตอนวิธีที่ 4 เป็นขั้นตอนการลดจำนวนการพิจารณาจุดข้อมูลโดยพิจารณาจุดข้อมูลที่อยู่ในกลุ่มของเมดอยด์ปัจจุบันก่อน เนื่องจากขั้นตอนวิธีที่ 3 จะเลือกคู่ที่ให้ผลรวมระยะทางลดลงมากที่สุดในรอบแรก ซึ่งทำให้ข้อมูลมีการเกาะกลุ่มกันชัดเจนมากขึ้น เพราะจุดข้อมูลที่อยู่ในกลุ่มมีโอกาสเป็นเมดอยด์มากกว่าข้อมูลที่อยู่นอกกลุ่ม จากผลการ

ทดลองของการประมวลผลของขั้นตอนวิธีทั้ง 4 ขั้นตอนวิธีเปรียบเทียบเวลากับวิธีแพม แสดงให้เห็นว่า ขั้นตอนวิธีที่ 4 ใช้เวลาการประมวลผลเร็วกว่าขั้นตอนวิธีแพมเมื่อข้อมูลมีมิติมากขึ้น และขั้นตอนวิธีที่ 4 เหมาะกับข้อมูลที่มีการเกาะกลุ่มชัดเจน และมีมิติมากกว่า 4 มิติ

ในบทที่ 2 จะอธิบายถึงความรู้เบื้องต้น แนวคิดของการทำเหมืองข้อมูล ระยะทางและความคล้ายของข้อมูล วิธีการเกาะกลุ่ม และวิธีเกาะกลุ่มรอบเมดอยด์

บทที่ 3 จะอธิบายถึงวิธีการปรับปรุงวิธีแพม และขั้นตอนการได้มาของขั้นตอนวิธีปรับปรุงแพม 4 ขั้นตอนวิธี

บทที่ 4 ข้อมูลที่ใช้ในการทดลอง ผลการทดลอง และแสดงการเปรียบเทียบโดยกราฟ

บทที่ 5 สรุปผล และกล่าวถึงงานวิจัยในอนาคต

บทที่ 2
นิยาม และทฤษฎีบทที่เกี่ยวข้อง

การวิเคราะห์ข้อมูลในอดีตส่วนใหญ่จะใช้วิธีทางสถิติ โดยการวิเคราะห์ความถี่ การกระจายทางสถิติ เช่น การเลือกตั้งสมาชิกสภาผู้แทนราษฎร ส่วนมากจะมีการทำโพลก่อนการเลือกตั้ง โพลจะเก็บข้อมูลจำนวนคะแนนที่คาดว่าจะได้ของผู้แทนนั้น ซึ่งได้จากการเก็บข้อมูลโดยการไปสอบถามประชาชนอย่างสุ่มในที่ต่าง ซึ่งการทำโพลการเลือกตั้งจะถูกแบ่งออกไปตามภาค และแต่ละภาคจะแบ่งออกเป็นจังหวัด เพื่อที่จะวิเคราะห์ข้อมูลที่ได้ว่าแต่ละภาค หรือแต่ละจังหวัดมีความนิยมพรรคของแต่ละพรรคต่างกันอย่างไร เป็นต้น

ผลที่ได้จากการวิเคราะห์ข้อมูลเราเรียกว่าความรู้จากข้อมูล การได้ความรู้จากข้อมูลนั้นส่วนใหญ่จะถูกวิเคราะห์โดยผู้เชี่ยวชาญที่มีประสบการณ์ ซึ่งจะใช้เวลานานเพราะฉะนั้นนักวิทยาการคอมพิวเตอร์จึงคิดวิธีการค้นความรู้จากข้อมูลโดยอัตโนมัติ ซึ่งเรียกว่าการทำเหมืองข้อมูล (data mining)

2.1 การทำเหมืองข้อมูล

การทำเหมืองข้อมูลหรือการสืบค้นความรู้ในฐานข้อมูล เป็นกระบวนการ วิธี หรือรูปแบบในการสกัดหรือสืบค้นความรู้จากฐานข้อมูลขนาดใหญ่โดยอัตโนมัติ โดยความรู้ที่ได้อาจอยู่ในรูปแบบของความสัมพันธ์ ข้อสรุป แบบจำลอง หรือลักษณะเฉพาะของข้อมูลในฐานข้อมูลนั้น การทำเหมืองข้อมูลทุกครั้งต้องกำหนดวัตถุประสงค์หรือเป้าหมายในการทำเหมืองข้อมูลให้ชัดเจน เพื่อให้สามารถกำหนดขั้นตอนการทำงานและการประเมินผลลัพธ์สุดท้ายได้ตรงตามความต้องการจริง โดยมีเทคนิคในการทำเหมืองข้อมูลดังนี้

1. Association analysis เป็นการวิเคราะห์ความสัมพันธ์เชื่อมโยงภายในข้อมูล ซึ่งเขียนอยู่ในรูปแบบ ถ้าแล้ว เช่น $A \to C$ เมื่อ A และ C เป็นเซตของกลุ่มข้อมูลที่ต้องการ การวิเคราะห์หลักเกณฑ์เชื่อมโยง นิยมใช้กับข้อมูลการขายสินค้าของร้านขายปลีก ซึ่งมีวัตถุประสงค์เพื่อค้นหา สินค้าที่ลูกค้าซื้อพร้อมกันในแต่ละครั้ง

โดยผลลัพธ์ที่ได้จะใช้ในการวางแผนการตลาด และจัดทำรายการส่งเสริมการขาย เพื่อเพิ่มยอดขายของสินค้าแต่ละชนิด

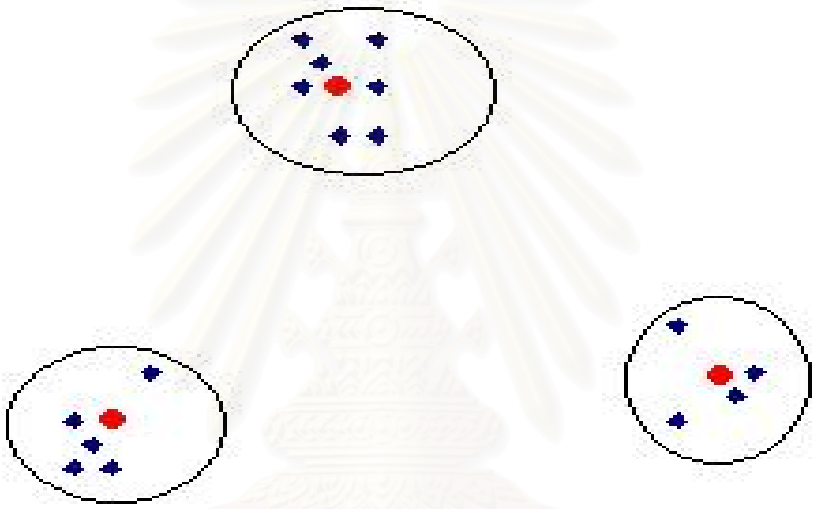
2. Classification and Prediction คือการหาลักษณะของตัวแบบ หรือฟังก์ชัน ที่อธิบาย หรือแยกแยะ ข้อมูลที่มีการบ่งบอกกลุ่มและค่าที่ชัดเจน เทคนิคในการจำแนกประเภทที่ใช้บ่อยได้แก่ การสร้างต้นไม้ตัดสินใจ (decision tree) การจำแนกประเภทแบบเบย์ (bayes classification) เป็นต้น การจำแนกประเภทนิยมใช้ในการวิเคราะห์งานทางด้านธุรกิจ และวิทยาศาสตร์ เช่น การวิเคราะห์ความเสี่ยงของลูกหนี้ การวินิจฉัยทางการแพทย์ และการวิเคราะห์กลุ่มดาวบนท้องฟ้า

3. Cluster Analysis เป็นการวิเคราะห์การเกาะกลุ่มของข้อมูลตามลักษณะที่เหมือนกันและต่างกัน และสามารถแยกประเภทของกลุ่มข้อมูลออกจากกันได้ โดยไม่อาศัยการกำหนดกลุ่มจากผู้วิเคราะห์ ซึ่งการเกาะกลุ่มจะนำไปใช้เพื่อวิเคราะห์การกระจายข้อมูล หรือเป็นขั้นตอนหนึ่งในการเตรียมข้อมูลเพื่อนำไปสู่การวิเคราะห์ข้อมูลแบบอื่น นอกจากนี้ยังสามารถประยุกต์ใช้กับงานได้หลากหลาย เช่นการตลาดใช้ในการแบ่งกลุ่มเพื่อหากลุ่มของลูกค้าที่มีลักษณะเหมือนกัน ทำให้ผู้ผลิตสามารถวางแผนเพื่อเพิ่มความพึงพอใจ หรือเพิ่มยอดขายลูกค้าแต่ละกลุ่มที่แตกต่างกัน

เทคนิคในการทำเหมืองข้อมูลมีมากมาย และมีแนวโน้มที่จะเพิ่มมากขึ้น เพราะข้อมูลที่ต้องการวิเคราะห์มีมาก และหลากหลาย จึงก่อให้เกิดแนวคิด และวิธีใหม่ เพื่อตอบสนองความต้องการทางการวิเคราะห์ในแต่ละสายงาน ในวิทยานิพนธ์นี้ เราสนใจปรับปรุงขั้นตอนวิธีการวิเคราะห์การเกาะกลุ่ม เพื่อให้ได้ขั้นตอนวิธีที่ประมวลผลได้เร็วขึ้น กับข้อมูลที่มีการกระจายของข้อมูลเป็นกลุ่มๆ ที่ชัดเจน

2.2 ระยะและความคล้ายคลึงของข้อมูล

การวิเคราะห์การเกาะกลุ่มใช้ระยะการบ่งบอกความคล้าย และความแตกต่างของจุดข้อมูล (data point) จุดข้อมูลที่คล้ายกันจะมีระยะระหว่างจุดข้อมูลน้อย จุดข้อมูลที่ต่างกันจะมีระยะระหว่างจุดข้อมูลมาก การทำหมืองข้อมูลด้วยวิธีการเกาะกลุ่มเป็น unsupervised classification คือ การแบ่งกลุ่มที่ไม่มีการกำหนดเป้าหมายที่ต้องการ ผลลัพธ์ที่ได้จากวิธีการนี้จะขึ้นอยู่กับลักษณะประจำของข้อมูล และระยะห่างระหว่างข้อมูล ดังรูปที่ 2.1



รูปที่ 2.1 การวิเคราะห์การเกาะกลุ่ม

ลักษณะของข้อมูลที่จะใช้ในการเกาะกลุ่ม จะถูกเก็บอยู่ในรูปของตาราง โดยตารางประกอบไปด้วย แถวและหลัก แต่ละแถวคือจุดข้อมูล และหลักจะเป็นลักษณะประจำหรือสมบัติของจุดข้อมูลนั้น

เพศ	วัย	ส่วนสูง(ซม.)	น้ำหนัก(กก.)
ญ	เด็ก	135	30
ญ	ชรา	167	58
ญ	ผู้ใหญ่	176	90
ช	ชรา	174	65
ช	ผู้ใหญ่	189	89
ญ	เด็ก	150	40
ช	เด็ก	137	45
ช	ผู้ใหญ่	200	109
ญ	เด็ก	120	32

ตาราง 2.1 ลักษณะข้อมูลแบบตาราง

จากตารางข้อมูล 2.1 เป็นข้อมูลลักษณะประจำของบุคคล ต่างๆ โดยมีทั้งหมด 9 จุดข้อมูล โดยแต่ละจุดข้อมูลจะแบ่งลักษณะประจำเป็น 4 ลักษณะ คือ เพศ, วัย, ส่วนสูง, และน้ำหนัก เช่น จุดตัวอย่างแรกคือ เด็กผู้หญิงที่สูง 135 เซนติเมตร และหนัก 30 กิโลกรัม

การทำเหมืองข้อมูลโดยวิธีการเกาะกลุ่มนั้น เราจะต้องใช้ระยะทางในการวัดความคล้ายของข้อมูลในแต่ละจุดข้อมูล ที่เรียกว่า ฟังก์ชันระยะทาง (Distance function) ที่สอดคล้องกับสมบัติเมตริกซ์ (Metric) ดังนี้

กำหนด $\mathbf{X}=(x_1,x_2,...,x_r)$, $\mathbf{Y}=(y_1,y_2,...,y_r)$ และ $\mathbf{Z}=(z_1,z_2,...,z_r)$ แทนจุดข้อมูล r มิติ

1. $d(\mathbf{X},\mathbf{Y})\geq 0$
2. $d(\mathbf{X},\mathbf{X})=0$
3. $d(\mathbf{X},\mathbf{Y})=d(\mathbf{Y},\mathbf{X})$
4. $d(\mathbf{X},\mathbf{Y})\leq d(\mathbf{X},\mathbf{Z})+d(\mathbf{Z},\mathbf{Y})$

จากสมบัติของเมตริกซ์สามารถเขียนเป็นตารางเมทริกซ์ของระยะทางระหว่างคู่ของจุดข้อมูลดังนี้

$$\begin{bmatrix} d(X_1,X_1) & d(X_1,X_2) & \dots & d(X_1,X_n) \\ d(X_2,X_1) & d(X_2,X_2) & \dots & d(X_2,X_n) \\ \vdots & \vdots & & \\ d(X_n,X_1) & d(X_n,X_2) & \dots & d(X_n,X_n) \end{bmatrix}$$

ตาราง 2.2 ตารางเมทริกซ์

จากตาราง 2.2 ด้านของแถวและหลักเป็นจุดข้อมูล $X_1, X_2, \dots, X_n$ และค่าในตารางแสดงค่าของระยะทางระหว่างจุดข้อมูลในแถวกับจุดข้อมูลในหลัก

ฟังก์ชันระยะทางที่นิยมใช้มีดังนี้

1. Manhattan distance เป็นระยะทางระหว่างจุดข้อมูลสองจุดที่ถูกประยุกต์มาจากรูปแบบขนาดของเวกเตอร์ 1-norm จะคำนวณโดยการวัดระยะห่างจากจุดข้อมูลในรูปผลต่างของทุกแกนรวมกัน มีสูตรดังนี้

$$d(\mathbf{X},\mathbf{Y}) = |x_1 - y_1| + |x_2 - y_2| + \dots + |x_r - y_r|$$

2. Euclidean distance เป็นระยะทางระหว่างจุดสองจุดที่ถูกประยุกต์มาจากรูปแบบขนาดของเวกเตอร์ 2-norm จะคำนวณโดยการวัดระยะทางของเส้นตรงที่สั้นที่สุดที่เชื่อมระหว่างจุดข้อมูลสองจุด มีสูตรดังนี้

$$d(\mathbf{X},\mathbf{Y}) = \sqrt{(|x_1 - y_1|)^2 + (|x_2 - y_2|)^2 + \dots + (|x_r - y_r|)^2}$$

3. Minkowski distance เป็นระยะทางระหว่างจุดสองจุดที่ถูกประยุกต์มาจากรูปแบบขนาดของเวกเตอร์ p-norm มีสูตรดังนี้

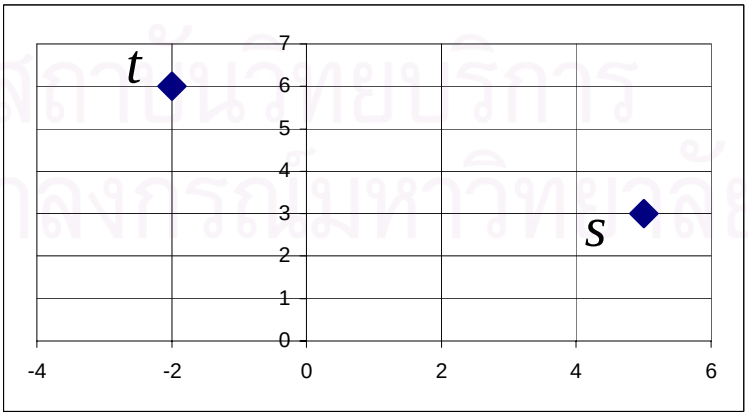
$$d(\mathbf{X},\mathbf{Y}) = \sqrt[p]{\left(|x_1 - y_1|\right)^p + \left(|x_2 - y_2|\right)^p + \ldots + \left(|x_r - y_r|\right)^p}$$

4. Infinity norm distance เป็นระยะทางระหว่างจุดสองจุดที่ถูกประยุกต์มาจากรูปแบบขนาดของเวกเตอร์ ∞-norm จะคำนวณระยะทางโดยการวัดระยะห่างจากจุดข้อมูลในรูปผลต่างของทุกแกนที่มากที่สุด

$$d(\mathbf{X},\mathbf{Y}) = \max\left(|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_r - y_r|\right)$$

จากฟังก์ชันระยะทางที่กล่าวมา ในวิทยานิพนธ์เลือก Manhattan distance เป็นเครื่องมือในการวัดระยะทางของข้อมูล

ตัวอย่าง 2.1 หาระยะทางระหว่าง $t = (-2, 6)$ กับ $s = (5, 3)$ โดยใช้วิธี Manhattan distance และ วิธี Euclidean distance ดังรูปที่ 2.2



รูปที่ 2.2 แสดงตำแหน่งของจุดข้อมูล t และ s

จากสูตร Manhattan distance

$$d(\mathbf{X},\mathbf{Y}) = |x_1 - y_1| + |x_2 - y_2| + \ldots + |x_p - y_p|$$

จะได้

$$\begin{aligned} d(t,s) &= |-2-5| + |6-3| \\ &= 10 \end{aligned}$$

จากสูตร Euclidean distance

$$d(\mathbf{X},\mathbf{Y}) = \sqrt{(|x_1 - y_1|)^2 + (|x_2 - y_2|)^2 + \ldots + (|x_p - y_p|)^2}$$

จะได้

$$\begin{aligned} d(t,s) &= \sqrt{(|-2-5|)^2 + (|6-3|)^2)} \\ &= \sqrt{(|-7|)^2 + (|3|)^2} \\ &\approx 7.616 \end{aligned}$$

2.3 วิธีการเกาะกลุ่ม

เทคนิคการแบ่งกลุ่มข้อมูลมีหลายวิธีโดยจะแบ่งออกเป็นดังนี้

1. Partitioning Method เป็นวิธีการเกาะกลุ่มข้อมูลออกเป็น *k* กลุ่ม โดยแต่ละกลุ่มจะมีสมาชิกอย่างน้อย 1 ตัว และสมาชิกแต่ละตัวต้องอยู่เพียงกลุ่มเดียวเท่านั้น วิธีแบ่งกลุ่มนี้จะเริ่มโดยการกำหนดตัวแทนกลุ่มเริ่มต้น แล้วทำการวนเปลี่ยนตัวแทนกลุ่ม เพื่อลดผลรวมระยะทางระหว่างตัวแทนกลุ่มกับจุดข้อมูลภายในกลุ่ม ให้มีค่าน้อยที่สุด ซึ่งวิธีการที่นิยมใช้มีอยู่ 2 วิธี คือ ***k-mean*** และ ***k-medoid*** ซึ่งในงานวิจัยนี้ สนใจปรับปรุงวิธีการ ***k-medoid***

2. Hierarchical Method วิธีการนี้จะสร้างลำดับชั้นของการเกาะกลุ่ม โดยมี 2 วิธีการหลักคือการรวมข้อมูล ***bottom-up*** เริ่มโดยพิจารณาข้อมูลที่คล้ายกันจับรวมกลุ่มกันเป็นลำดับชั้นขึ้นไป จะหยุดการทำงานก็ต่อเมื่อข้อมูลรวมกัน

เป็นกลุ่มเดียว หรือตามเงื่อนไขการหยุดที่ถูกกำหนด วิธีการแยกข้อมูลแบบ *top-down* จะเริ่มจากการกำหนดระดับบนสุดรวมข้อมูลทั้งหมดเป็นหนึ่งกลุ่มแล้ว พิจารณาการแยกข้อมูลที่แตกต่างกันลงมาเป็นลำดับชั้น จะหยุดการทำงานก็ต่อเมื่อข้อมูลทุกตัวแยกออกจากกันหมด หรือตามเงื่อนไขการหยุดที่ถูกกำหนด วิธีที่นิยมใช้มี 2 วิธี คือ วิธี CURE และ วิธี BIRCH

3. Density-based methods เป็นวิธีที่จะพิจารณาความหนาแน่นของข้อมูลในกลุ่ม โดยการเลือกจุดข้อมูลเพื่อเป็นตัวแทน กำหนดรัศมีและจำนวนข้อมูลในรัศมีที่เหมาะสม ถ้าข้อมูลที่เป็นตัวแทนมีข้อมูลที่มาเกาะอยู่ในกลุ่มน้อยกว่าที่กำหนดจะทำการเปลี่ยนตัวแทนใหม่ วิธีที่ใช้มี DBSCAN และ OPTICS

2.4 วิธีแบ่งกั้นกลุ่มรอบเมดอยด์ (Partitioning Around Medoids)

วิธีแบ่งกั้นกลุ่มรอบเมดอยด์ (PAM: Partitioning Around Medoids) หรือวิธีแพม ออกแบบโดย *Kaufman* และ *Rouseeuw* เป็นวิธีการแบ่งกลุ่มข้อมูลออกเป็น k กลุ่ม จาก n จุดข้อมูล โดยการหาตัวแทนที่อยู่ในแต่ละกลุ่ม ซึ่งเรียกว่าเมดอยด์ (medoid) เมดอยด์ที่เป็นตัวแทนของกลุ่มที่เหมาะสมคือ จุดข้อมูลที่อยู่กึ่งกลางของกลุ่ม ที่ให้ผลรวมระยะทางระหว่างเมดอยด์กับจุดข้อมูลน้อยที่สุด ถ้า X_j เป็นจุดข้อมูลตัวที่ j ที่ไม่ใช่เมดอยด์และ M_i เป็นเมดอยด์ของกลุ่ม i แล้ว X_j จะอยู่กลุ่มเดียวกับ M_i ถ้า $d(X_j, M_i) = min_{M_e}(X_j, M_e)$ เมื่อ min_{M_e} คือค่าน้อยสุดระหว่าง X_j กับเมดอยด์ M_e ทุกตัว

การวัดคุณภาพของกลุ่ม คำนวณได้จากผลรวมระยะทางทั้งหมดระหว่างจุดข้อมูลในกลุ่มกับเมดอยด์

$$\sum_{i=1}^{k}\sum_{j=1}^{N_i} d(M_i, X_{ij})$$

เมื่อ X_{ij} คือ ข้อมูลที่อยู่กลุ่มเดียวกับ M_i

k คือ จำนวนกลุ่ม

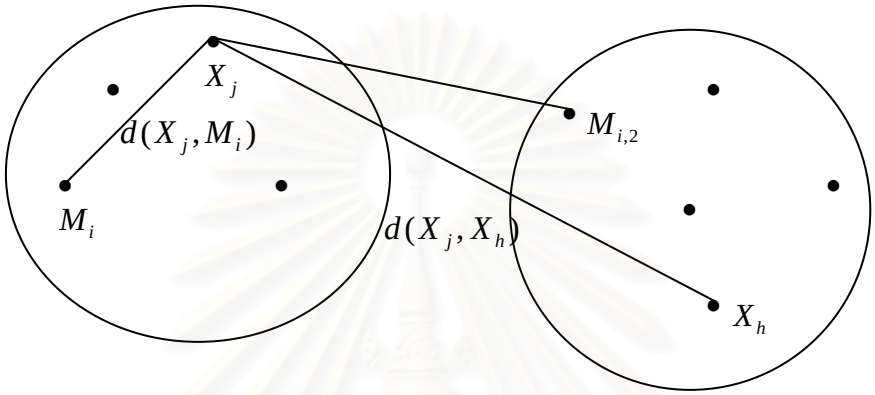
N_i คือ จำนวน X_{ij} ที่อยู่ ในกลุ่ม M_i

ผลรวมค่าระยะทางที่น้อยแสดงการเกาะกลุ่มโดยใช้ชุดของเมดอยด์ $M_1, M_2, ..., M_k$ เป็นชุดที่ดี ในขณะที่ค่าผลรวมระยะทางที่มากแสดงว่าชุดของเมดอยด์ $M_1, M_2, ..., M_k$ ไม่เป็นชุดที่เหมาะสม โดยทฤษฎีแล้ว สำหรับจำนวนจุดข้อมูลที่จำกัดและค่าพารามิเตอร์ k ที่กำหนด จะมีชุดของเมดอยด์ $M_1, M_2, ..., M_k$ ที่ให้ผลรวมระยะทางที่น้อยที่สุด เมื่อเทียบกับชุดของเมดอยด์ k ตัวอื่น

2.5 ขั้นตอนการทำงานของวิธีแบ่งกั้นกลุ่มรอบเมดอยด์

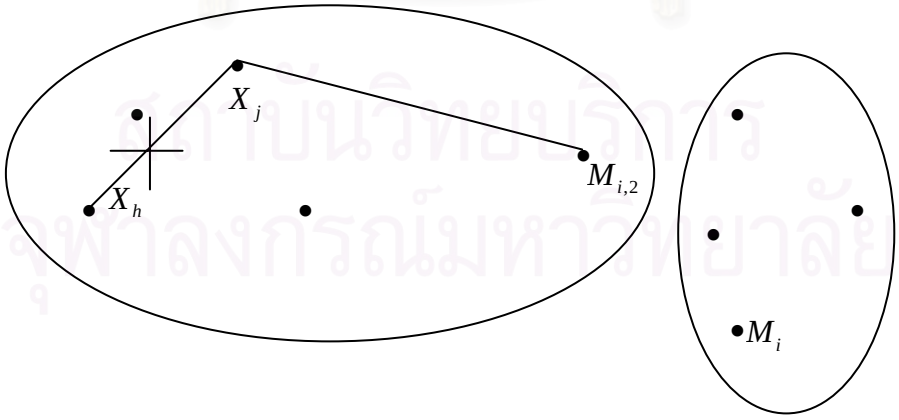
วิธีแพมเป็นขั้นตอนวิธีการวิเคราะห์การเกาะกลุ่ม ในรูปผลแบ่งกั้นของเมดอยด์ k ชุดที่ให้ผลรวมระยะทางต่ำที่สุด เริ่มจากการเลือกเมดอยด์ k ตัว แล้วแต่ละรอบการทำงานจะสลับระหว่างเมดอยด์ M_i กับ จุดข้อมูลที่ไม่ใช่เมดอยด์ X_h โดยกำหนดให้การสลับแต่ละรอบมีผลรวมระยะทางที่ลดลงมากที่สุด การคำนวณค่าผลกระทบของการสลับระหว่าง M_i กับ X_h สำหรับทุกจุดข้อมูล X_j เรียก C_{jih} จะถูกแบ่งออกเป็น 4 กรณี ดังนี้

กรณีที่ 1 : เมื่อ X_j อยู่กลุ่มเดียวกับเมดอยด์ M_i และกำหนดให้ $M_{i,2}$ คือเมดอยด์ที่ไม่ใช่ M_i ที่อยู่ใกล้ X_j ถ้า X_j อยู่ใกล้กับเมดอยด์ $M_{i,2}$ กว่า X_h กล่าวคือ $d(X_j, X_h) \geq d(X_j, M_{i,2})$ ดังรูปที่ 2.3



รูปที่ 2.3 แสดงรูปแบบการแบ่งกั้นกลุ่มรอบเมดอยด์ของกรณีที่ 1

ถ้าแทน M_i ด้วย X_h ดังรูปที่ 2.4

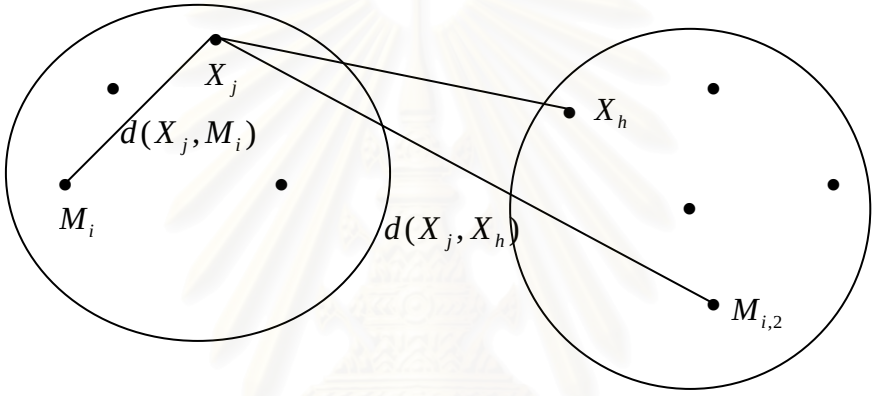


รูปที่ 2.4 แสดงรูปแบบการแบ่งกั้นกลุ่มรอบเมดอยด์ของกรณีที่ 1
เมื่อมีการสลับที่ระหว่าง M_i กับ X_h

จากรูปเมื่อมีการสลับที่ จากเมดอยด์ M_i ไปเป็น X_h จะทำให้ X_j เปลี่ยนกลุ่มไปอยู่กลุ่มเดียวกับ เมดอยด์ $M_{i,2}$ ทำให้ค่าผลรวมระยะทางที่ปลี่ยนไปได้จากการลบระยะทางระหว่าง X_j กับ M_i บวกกับระยะทางระหว่าง X_j กับ $M_{i,2}$ ได้สมการดังนี้

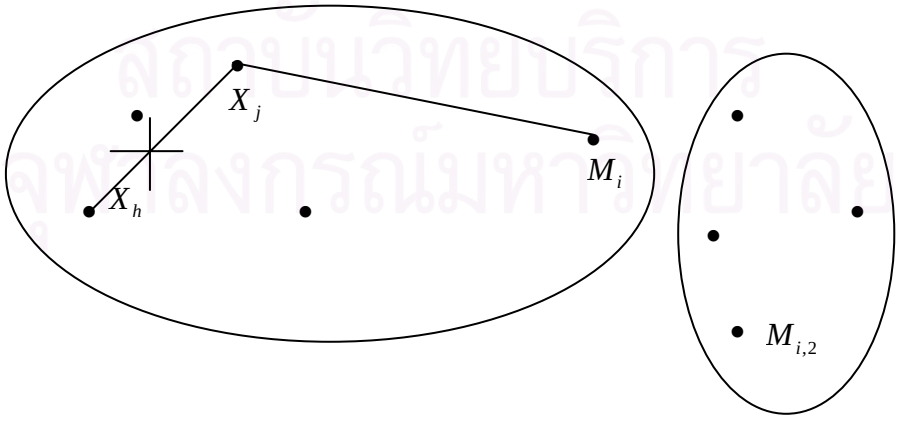
$$C_{jih} = d(X_j, M_{i,2}) - d(X_j, M_i)$$

กรณีที่ 2: เมื่อ X_j อยู่กลุ่มเดียวกับเมดอยด์ M_i ถ้า X_j อยู่ไกลกับ $M_{i,2}$ กว่า X_h กล่าวคือ $d(X_j, X_h) < d(X_j, M_{i,2})$ ดังรูปที่ 2.5



รูปที่ 2.5 แสดงรูปแบบการแบ่งกั้นกลุ่มรอบเมดอยด์ของกรณีที่ 2

ถ้าแทน M_i ด้วย X_h ดังรูปที่ 2.6

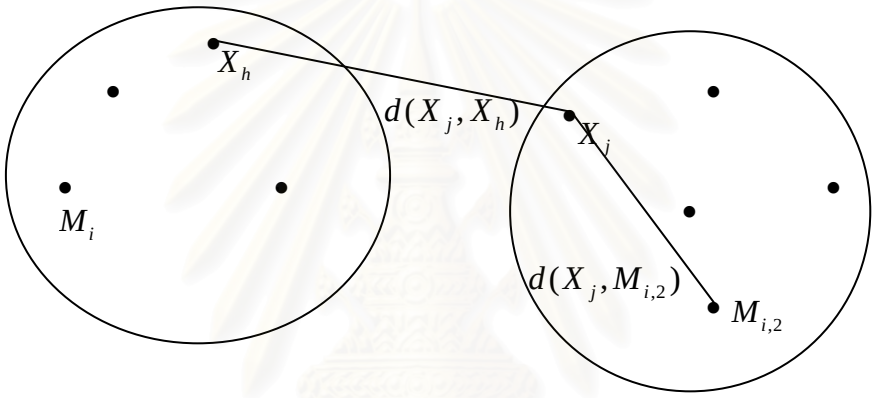


รูปที่ 2.6 แสดงรูปแบบการแบ่งกั้นกลุ่มรอบเมดอยด์ของกรณีที่ 2
เมื่อมีการสลับที่ระหว่าง M_i กับ X_h

จากรูปเมื่อมีการสลับที่ จากเมดอยด์ M_i ไปเป็น X_h จะทำให้ X_j เปลี่ยนกลุ่มไปอยู่กลุ่มเดียวกับ เมดอยด์ X_h ทำให้ค่าผลรวมระยะทางที่เปลี่ยนไปคำนวณได้จากการลบระยะทางระหว่าง X_j กับ M_i บวกกับระยะทางระหว่าง X_j กับ X_h ได้สมการดังนี้

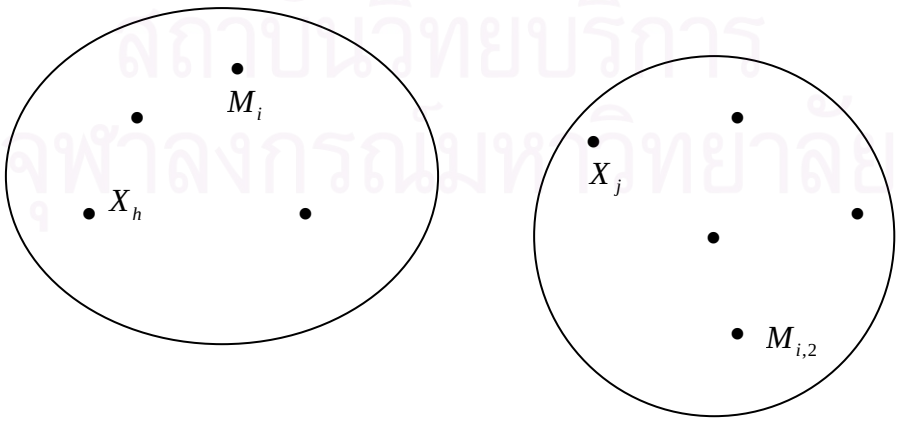
$$C_{jih} = d(X_j, X_h) - d(X_j, M_i)$$

กรณีที่ 3: เมื่อ X_j ไม่ได้อยู่กลุ่มเดียวกับเมดอยด์ M_i แต่อยู่กลุ่มเดียวกับ $M_{i,2}$ ให้ X_j อยู่ใกล้กับ $M_{i,2}$ กว่า X_h กล่าวคือ $d(X_j, X_h) \geq d(X_j, M_{i,2})$ ดังรูปที่ 2.7



รูปที่ 2.7 แสดงรูปแบบการแบ่งกั้นกลุ่มรอบเมดอยด์ของกรณีที่ 3

ถ้าแทน M_i ด้วย X_h ดังรูปที่ 2.8

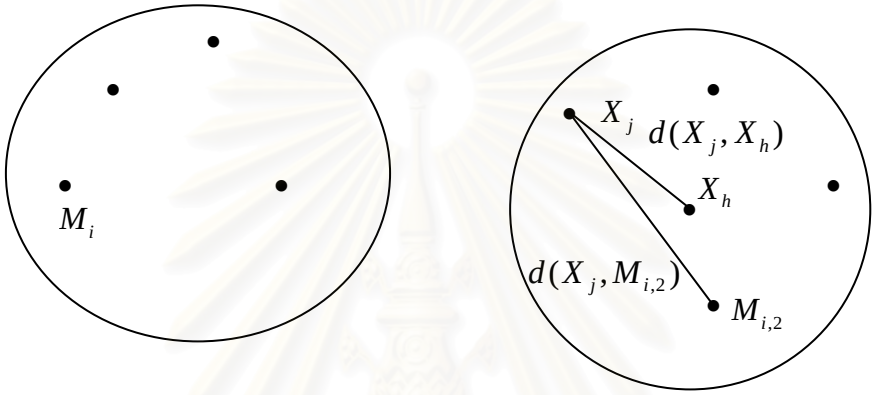


รูปที่ 2.8 แสดงรูปแบบการแบ่งกั้นกลุ่มรอบเมดอยด์ของกรณีที่ 3
เมื่อมีการสลับที่ระหว่าง M_i กับ X_h

จากรูปเมื่อมีการสลับที่จากเมดอยด์ M_i ไปเป็น X_h จะเห็นว่า X_j ยังอยู่กลุ่มเดียวกับ $M_{i,2}$ ทำให้ไม่มีการเปลี่ยนแปลงระยะทาง ค่าผลกระทบจึงไม่เปลี่ยนแปลง ดังสมการ

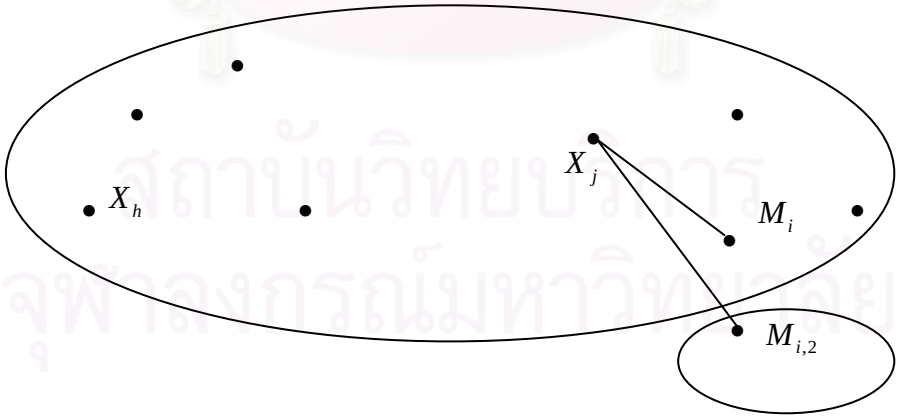
$$C_{jih} = 0$$

กรณีที่ 4: เมื่อ X_j ไม่ได้อยู่กลุ่มเดียวกับเมดอยด์ M_i แต่อยู่กลุ่มเดียวกับเมดอยด์ $M_{i,2}$ ถ้า X_j อยู่ไกล $M_{i,2}$ กว่า X_h กล่าวคือ $d(X_j, X_h) < d(X_j, M_j)$ ดังรูปที่ 2.9



รูปที่ 2.9 แสดงรูปแบบการแบ่งกั้นกลุ่มรอบเมดอยด์ของกรณีที่ 4

ถ้าแทน M_i ด้วย X_h ดังรูปที่ 2.10



รูปที่ 2.10 แสดงรูปแบบการแบ่งกั้นกลุ่มรอบเมดอยด์ของกรณีที่ 4
เมื่อมีการสลับที่ระหว่าง M_i กับ X_h

จากรูปเมื่อมีการสลับที่จากเมดอยด์ M_i ไปเป็น X_h จะทำให้ X_j เปลี่ยนกลุ่มไปอยู่กลุ่มเดียวกับเมดอยด์ X_h ทำให้ค่าผลรวมระยะทางที่เปลี่ยนไปคำนวณได้จากการลบระยะทางระหว่าง X_j กับ $M_{i,2}$ บวกกับระยะทางระหว่าง X_j กับ X_h ได้สมการดังนี้

$$C_{jih} = d(X_j, X_h) - d(X_j, M_{i,2})$$

จากทั้ง 4 กรณีเราจะสามารถคำนวณผลรวมระยะทางของการสลับคู่ระหว่างเมดอยด์ i กับจุดข้อมูล h ได้คือ

$$TC_{ih} = \sum_{j \neq i,\, j \neq h} C_{jih}$$

ค่าที่เป็นค่าลบของ TC_{ih} แสดงระยะทางรวมที่ลดลงของการแบ่งกลุ่มเมดอยด์ใหม่ โดยจะเลือกสลับคู่ M_i กับ X_h ที่ให้ค่าลบมากที่สุด

ขั้นตอนวิธีแพมมีรายละเอียดดังนี้ กำหนดจุดข้อมูล $X_j, j = 1,2,3,...,n$ และ k คือจำนวนกลุ่ม

1. เลือกตัวแทนกลุ่มเริ่มต้นมา k ตัว
2. สำหรับ X_j ที่ไม่ใช่เมดอยด์ เมื่อ $j = 1,2,\ldots,n$
3. สำหรับ M_i เมื่อ $i = 1,2,\ldots,k$
4. คำนวณระยะทางระหว่าง X_j กับ M_i
5. ถ้า $i = 1$ แล้ว
6. $Minimum =$ ระยะทางระหว่าง X_j กับ M_1
7. $group = 1$
8. กรณีอื่น
9. ถ้า $Minimum >$ ระยะทางระหว่าง X_j กับ M_1
10. $Minimum =$ ระยะทางระหว่าง X_j กับ M_1
11. $group = i$

12. กำหนดกลุ่มให้ X_j ให้อยู่กลุ่มเดียวกับเมดอยด์ M_i ที่ใกล้ที่สุด

13. สำหรับ M_i เมื่อ $i = 1, 2, ..., k$

14. สำหรับ X_h เมื่อ $h = 1, 2, ..., n$ และ $X_h \neq M_i$

15. $TC_{ih} = 0$

16. สำหรับ X_h เมื่อ $j = 1, 2, ..., n$ และ $X_j \neq X_h \neq M_i$

17. คำนวณค่า C_{jih} และรวมเก็บไว้ใน TC_{ih}

18. หาค่าต่ำที่สุดของ TC_{ih}

19. ถ้าค่าต่ำสุดของ TC_{ih} มีค่าเป็นลบแล้ว จะสลับ M_i ด้วย X_h แล้วกลับไปทำขั้นที่ 2

20. ถ้าค่าต่ำสุดของ TC_{ih} มีค่าไม่เป็นลบแล้วจะหยุดการทำงาน โดยกลุ่มเมดอยด์ที่ได้เป็นกลุ่มเมดอยด์ที่ให้ผลรวมของระยะทางน้อยที่สุด

จากขั้นตอนที่ 17 นั้นจะเห็นว่าวิธีแพมจะใช้เวลาในการประมวลผลเป็น $k(n-k)$ เมื่อ k คือจำนวนกลุ่ม และ n คือจำนวนจุดข้อมูล ถ้าเราต้องการปรับระยะของการสลับระหว่าง M_i กับ X_h จะใช้เวลาในการทำงาน $k(n-k)$ สำหรับคู่ในการคำนวณ TC_{ih} จะใช้ $(n-k)$ จุดข้อมูล เพราะฉะนั้น เวลาในการทำงานทั้งหมด คือ $O(k(n-k)^2)$ จะเห็นว่าเวลาในการทำงานของวิธีแพม จะขึ้นอยู่กับจำนวนจุดข้อมูล เพราะถ้าข้อมูลมีจำนวนมาก ก็จะใช้เวลาในการคำนวณมากขึ้น แต่สำหรับปริมาณข้อมูลที่มีขนาดเล็ก ถึงขนาดกลาง จะใช้เวลาในการคำนวณที่เป็นที่ยอมรับได้ วิธีการนี้จึงเหมาะสมกับข้อมูลที่มีขนาดไม่ใหญ่มากนัก

งานวิจัยที่พัฒนาต่อจากวิธีแพม เพื่อจัดการกับข้อมูลที่มีปริมาณมากได้แก่ วิธี CLARA เป็นวิธีการที่ใช้วิธีแพมกับข้อมูลตัวแทนขนาดเล็ก โดยวิธีนี้จะสุ่มข้อมูลออกมาจากข้อมูลทั้งหมดหลายชุด นำข้อมูลที่สุ่มแต่ละชุดมาประมวลผลด้วยวิธีแพม แล้วเปรียบเทียบผลรวมระยะทางที่ได้จากการประมวลผลของแต่ละชุด เลือกผลลัพธ์จากชุดข้อมูลที่ให้ผลรวมระยะทางน้อยที่สุด แต่ผลลัพธ์ที่ได้จากวิธี CLARA อาจไม่ได้ชุดเมดอยด์ที่ให้ผลรวมระยะทางที่น้อยที่สุดของข้อมูลทั้งหมด

อีกวิธีหนึ่งคือ CLARAN จะแบ่งกลุ่มข้อมูลโดยการสุ่มเซตของเมดอยด์เริ่มต้นแล้วจะสุ่มเลือกเซตเมดอยด์ที่มีสมาชิกต่างกันกับเซตคำตอบ 1 ตัว ซึ่งเรียกว่า *neighbor* โดยจะ

พิจารณาค่าผลรวมระยะทางระหว่าง 2 เซตนี้ ถ้าค่าผลรวมระยะทางของ *neighbor* ให้ค่าลดลงก็จะกำหนดให้ *neighbor* เป็นเซตคำตอบ ทำซ้ำจนกว่าเซตคำตอบไม่มีการเปลี่ยนโดยแปรค่า *neighbor* ไปถึงค่า *maxneighbor* ที่ถูกกำหนด จึงหยุดการทำงาน

วิธีทั้งสองวิธีพบว่าเป็นวิธีการเกาะกลุ่มข้อมูลขนาดใหญ่ที่ยังคงใช้วิธีแพม ผู้วิจัยจึงเสนอวิธีการที่จะปรับเปลี่ยนวิธีแพมเพื่อให้มีประสิทธิภาพขึ้นที่แตกต่างจากวิธี CLARA และ CLARAN ในบทที่ 3

บทที่ 3
ขั้นตอนวิธีสำหรับการเกาะกลุ่ม

ในวิทยานิพนธ์นี้ ผู้วิจัยใช้แนวความคิดการปรับปรุงวิธีแพม 4 ขั้นตอนวิธีด้วยวิธีการทำซ้ำ (Iterative improvement) โดยแนวคิดนี้มาจากการจะลดเวลาใน 1 รอบของการสลับเมดอยด์ เพราะวิธีแพมจะพิจารณาคู่เมดอยด์กับจุดข้อมูลทุกคู่ก่อนแล้วสลับจึงใช้เวลาประมวลผลนาน ขั้นตอนแรก ปรับปรุงวิธีแพมโดยการเลือกสลับคู่เมดอยด์กับจุดข้อมูลที่ให้ผลรวมระยะทางลดลงคู่แรก ซึ่งวิธีการนี้จำนวนรอบของการสลับจะมากขึ้นทำให้เวลาในการประมวลผลรวมของขั้นตอนวิธีแรกมากกว่าวิธีแพม ขั้นตอนที่ 2 เรียงเมดอยด์เพื่อพิจารณาก่อน 4 รูปแบบเพื่อลดจำนวนรอบของการสลับเมดอยด์ ในขั้นตอนที่ 3 ปรับปรุงขั้นตอนที่ 2 โดยการลดจำนวนรอบของการสลับเมดอยด์เพิ่มเติมด้วยการเลือกคู่ที่ให้ผลรวมระยะทางลดลงมากที่สุดก่อนในรอบแรก และขั้นตอนสุดท้าย ลดจำนวนการพิจารณาเมดอยด์กับจุดข้อมูลที่ไม่ทำให้ผลรวมระยะทางลดลง โดยการเลือกพิจารณาคู่เมดอยด์กับจุดข้อมูลที่อยู่ในกลุ่มเมดอยด์ปัจจุบันก่อน ซึ่งแบ่งขั้นตอนได้ดังนี้

3.1 วิธีปรับปรุงวิธีแพมโดยการทำซ้ำ

ขั้นตอนวิธีแรกเรียกว่า (FIPAM: First-in First-out Iterative Improvement Partitioning Around Medoids) แนวคิดของขั้นตอนวิธีนี้ได้มาจากการลดเวลาในการประมวลผล 1 รอบของการสลับเมดอยด์ของวิธีแพม โดยการเลือกคู่เมดอยด์กับจุดข้อมูลที่ให้ผลรวมระยะทางลดลงค่าแรก ซึ่งมีขั้นตอนดังนี้

ขั้นตอนวิธี FIPAM เมื่อจุดข้อมูล $X_j, j = 1,2,3,...,n$ และ k คือจำนวนกลุ่ม

1. เลือกตัวแทนกลุ่มเริ่มต้นมา k กลุ่ม
2. ให้ตัวแปร $i = 0$
3. วนรอบจนกว่า $i \leq 0$
4. $repeat = false$
5. สำหรับ X_h เมื่อ $h = 1,2,...,n$ และ $X_h \neq M_i$

6. $TC_{ih} = 0$

7. สำหรับ ทุกๆ $j = 1, 2, ..., n$ เมื่อ $X_j \neq X_h$ และ $X_j \neq M_i$

8. คำนวณค่า C_{jih} และรวมเก็บไว้ใน TC_{ih}

9. ถ้า $TC_{ih} < 0$ แล้ว

10. จะสลับ M_i กับ X_h

11. $i = 0$

12. $repeat = true$

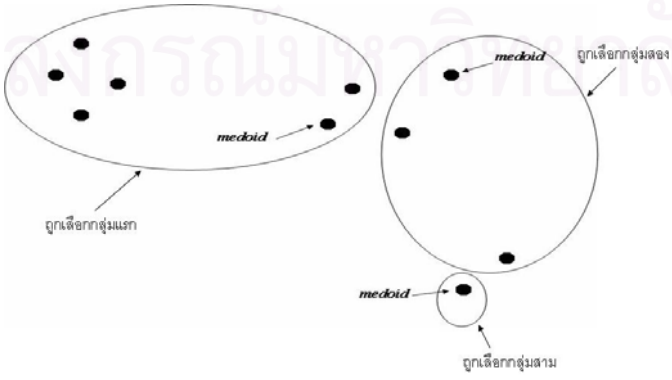
13. ถ้าไม่มีการวนรอบ กำหนดค่า $i = i + 1$

14. จบการทำงาน

ขั้นตอนวิธีนี้ใช้เวลาการประมวลผล 1 รอบของการสลับเมดอยด์น้อยกว่าวิธีแพม แต่จำนวนรอบของการสลับเมดอยด์มากกว่าวิธีแพม ซึ่งทำให้เวลาการประมวลผลรวมมากกว่าวิธีแพม ดูผลได้ในบทที่ 4

ขั้นตอนวิธีที่ 2 เรียกว่า (OFIPAM: Ordered First-in First-out Iterative improvement of the Partition Around Medoids) ขั้นตอนวิธีนี้จะใช้ขั้นตอนวิธีแรก และเพิ่มการเรียงเมดอยด์โดยการพิจารณาก่อน 4 กลยุทธ์ เพื่อลดจำนวนรอบของการสลับเมดอยด์ ซึ่งมีรูปแบบการเรียงดังนี้

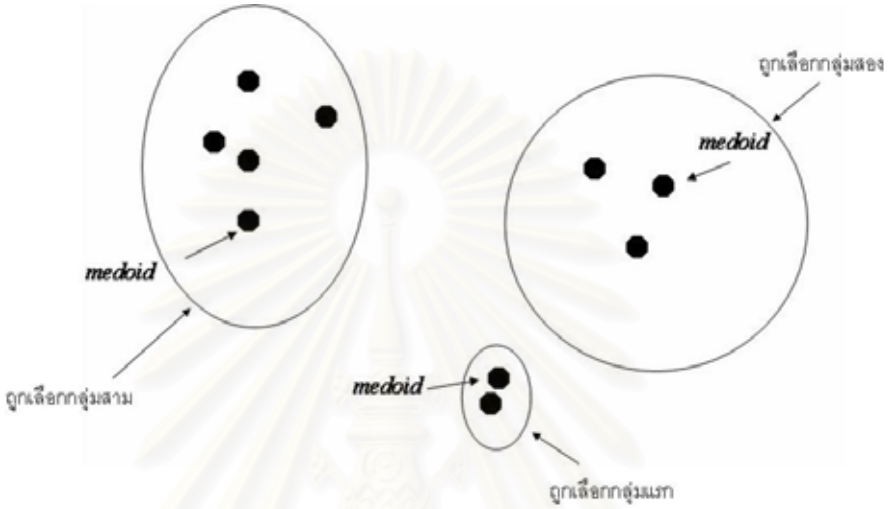
การเรียงเมดอยด์ที่มีจำนวนสมาชิกในกลุ่มมากก่อน OFIPAM-LE ดังรูปที่ 3.1



รูปที่ 3.1 การเลือกเมดอยด์จากกลุ่มที่มีจำนวนสมาชิกมากก่อน

จากรูปที่ 3.2 จะเห็นว่ากลุ่มที่จำนวนสมาชิกมากมีข้อมูลที่อยู่ไกลจากเมดอยด์เมื่อมีการสลับเมดอย์ที่กลุ่มนี้ก่อน พบว่าผลรวมระยะทางจะลดลงมาก

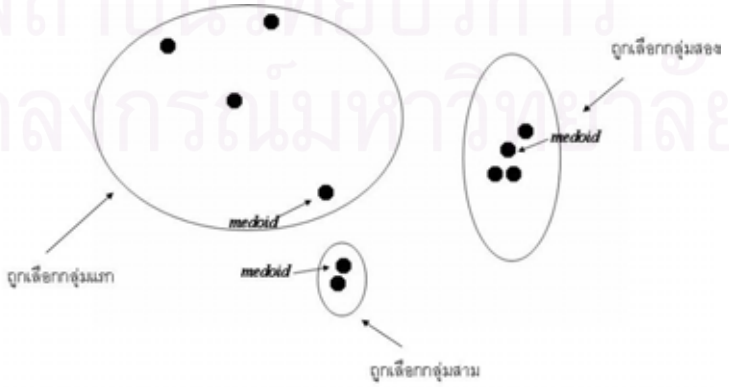
การเรียงเมดอยด์ที่มีจำนวนสมาชิกน้อยก่อน OFIPAM-SE ดังรูปที่ 3.2



รูปที่ 3.2 การเลือกเมดอยด์จากกลุ่มที่มีจำนวนสมาชิกน้อยก่อน

จากรูปที่ 3.2 จะเห็นว่ากลุ่มที่มีสมาชิกน้อยจะมีจุดให้พิจารณาน้อย เมื่อพิจารณากลุ่มที่มีสมาชิกน้อยก่อนทำให้ลดการพิจารณาข้อมูลที่ไม่เป็นคำตอบ ดังนั้นผู้วิจัยจึงคาดว่าจำนวนการทำซ้ำจะลดลง

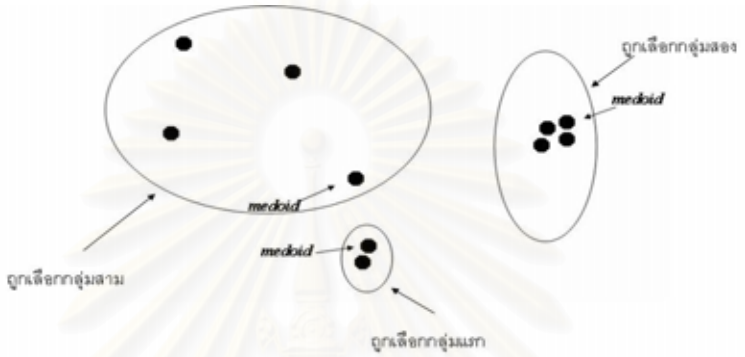
การเรียงเมดอยด์ที่มีผลรวมระยะทางระหว่างเมดอยด์กับสมาชิกภายในกลุ่มมากก่อน ดังรูปที่ 3.3



รูปที่ 3.3 การเลือกเมดอยด์ที่มีผลรวมระยะทางระหว่างเมดอยด์กับสมาชิกภายในกลุ่มมากก่อน

จากรูปที่ 3.3 จะเห็นได้ว่ากลุ่มที่มีผลรวมระยะทางภายในกลุ่มมากนั้นเมื่อมีการสลับเมดอยด์ก่อนจะทำให้ผลรวมระยะทางลดลงมาก

การเลือกเมดอยด์ที่มีผลรวมระยะทางระหว่างเมดอยด์กับสมาชิกภายในกลุ่มน้อยก่อน ดังรูปที่ 3.4



รูปที่ 3.4 การเลือกเมดอยด์ที่มีผลรวมระยะทางระหว่างเมดอยด์กับสมาชิกภายในกลุ่มน้อยก่อน

จากรูปที่ 3.4 จะเห็นว่ากลุ่มที่มีผลรวมระยะทางภายในกลุ่มน้อยจะมีการเกาะกลุ่มของข้อมูลอยู่แล้ว ดังนั้นเมื่อพิจารณากลุ่มนี้ก็จะลดการสลับเมดอยด์ลงได้

ขั้นตอนวิธีOFIPAM ตามการเลือกเมดอยด์ทั้ง 4 แบบ กำหนดจุดข้อมูล $X_j, j = 1,2,3,...,n$ และ k คือจำนวนกลุ่ม

1. สุ่มเลือกตัวแทนกลุ่มเริ่มต้นมา k กลุ่ม
2. ให้ตัวแปร $i = 0$
3. วนรอบจนกว่า $i \le 0$
4. สำหรับ X_j ที่จัดกลุ่มเข้ากับ M_i เมื่อ $d(M_i, X_j) = \min d(M_e, X_j)$ เมื่อ $e = 1,2,...,k$
5. เลือกเมดอยด์ ตัวที่ i จากวิธีเลือกทั้ง 4 แบบ
6. $repeat = false$
7. สำหรับ X_h เมื่อ $h = 1,2,...,n$ และ $X_h \neq M_i$
8. $TC_{ih} = 0$

9. สำหรับ ทุก $j = 1, 2, ..., n$ เมื่อ $X_j \neq X_h$ และ $X_j \neq M_i$

10. คำนวณค่า C_{jih} และรวมเก็บไว้ใน TC_{ih}

11. ถ้า $TC_{ih} < 0$ แล้ว

12. จะสลับ M_i กับ X_h

13. $i = 0$

14. $repeat = true$

15. ถ้า ไม่มีการวนรอบ จะกำหนดค่า $i = i + 1$

16. จบการทำงาน

จากขั้นตอนวิธี OFIPAM ทั้ง 4 กลยุทธ์ พบว่าสามารถลดจำนวนการสลับเมดอยด์ของวิธีการ FIPAM ได้เพียงเล็กน้อย

ขั้นตอนวิธีที่ 3 เรียกว่า (OFIPAM-FA: Ordered First-in First-out Iterative improvement of the Partition Around Medoids with first iteration - all pairs comparison) ขั้นตอนวิธีนี้ลดจำนวนรอบของการสลับเมดอยด์เพิ่มเติมจากขั้นตอนวิธีที่ 2 โดยการเลือกคู่เมดอยด์กับจุดข้อมูลที่ให้ผลรวมระยะทางลดลงมากที่สุดในรอบแรกก่อน แล้วใช้ขั้นตอนวิธีแรกต่อไป โดยมีขั้นตอนดังนี้

ขั้นตอนวิธีOFIPAM-FA กำหนดจุดข้อมูล $X_j, j = 1, 2, 3, ..., n$ และ k คือ จำนวนกลุ่ม

1. สุ่มเลือกตัวแทนกลุ่ม k กลุ่ม

2. สำหรับทุก M_i เมื่อ $i = 0, 1, .., k$

3. $TC_{ih} = 0$

4. สำหรับทุก X_h เมื่อ $h = 1, 2, ..., n$ และ $X_h \neq M_i$

5. สำหรับ $j = 1, 2, ..., n$ เมื่อ $X_j \neq X_h$ และ $X_j \neq M_i$

6. คำนวณค่า C_{jih} เก็บไว้ใน TC_{ih}

7. $Minimum = TC_{ih}$

8. ถ้า $Minimum < 0$ แล้ว สลับ M_i กับ X_h

9. ให้ตัวแปร $i = 0$

10. วนรอบจนกว่า $i \leq 0$

11. $repeat = false$

12. สำหรับทุก X_h เมื่อ $h = 1, 2, ..., n$ และ $X_h \neq M_i$

13. $TC_{ih} = 0$

14. สำหรับ $j = 1, 2, ..., n$ เมื่อ $X_j \neq X_h$ และ $X_j \neq M_i$

15. คำนวณค่า C_{jih} และรวมเก็บไว้ใน TC_{ih}

16. ถ้า $TC_{ih} < 0$ แล้ว

17. จะสลับ M_i กับ X_h

18. $i = 0$

19. $repeat = true$

20. ถ้า ไม่มีการวนรอบ จะกำหนดค่า $i = i + 1$

21. จบการทำงาน

จากขั้นตอนวิธี OFIPAM-FA พบว่าขั้นตอนวิธีนี้สามารถลดจำนวนรอบของการสลับเมดอยด์ของวิธี FIPAM ได้มาก แต่เวลาในการประมวลผลมากกว่าวิธีแพม ซึ่งแสดงผลในบทที่ 4 เนื่องจากขั้นตอนวิธีนี้ยังพิจารณาเมดอยด์กับจุดข้อมูลที่ไม่สามารถลดระยะทางรวมมากเกิน ดังนั้นจึงเสนอขั้นตอนการปรับปรุงเพื่อลดจำนวนการพิจารณาคู่เมดอยด์กับจุดข้อมูลที่ไม่ลดระยะทางรวม

ขั้นตอนวิธี 4 เรียกว่า (OFIPAM-FAW: Ordered First-in First-out Iterative improvement of the Partition Around Medoids with first iteration – all pairs comparison within group) ขั้นตอนวิธีนี้จะใช้ขั้นตอนวิธีที่ 3 ก่อน แล้วจะลดจำนวนการพิจารณาข้อมูลที่ไม่ลดระยะทางรวม โดยการพิจารณาคู่เมดอยด์กับจุดข้อมูลที่อยู่ในกลุ่มปัจจุบันก่อนในรอบถัดไป เพราะในขั้นตอนที่ 3 ทำให้ข้อมูลมีการเกาะกลุ่มที่ชัดเจนขึ้นซึ่งเป็นจุดข้อมูลที่มีโอกาสเป็นเมดอยด์สูงกว่าจุดข้อมูลที่อยู่นอกกลุ่ม โดยมีขั้นตอนดังนี้

ขั้นตอนวิธีการของขั้นตอนวิธี OFIPAM-FAW โดยใช้ $P_j, j = 1,2,3,...,n$ และ k คือจำนวนกลุ่ม

1. สุ่มเลือกตัวแทนกลุ่ม k กลุ่ม
2. สำหรับทุก M_i เมื่อ $i = 0,1,..,k$
3. $TC_{ih} = 0$
4. สำหรับทุก P_h เมื่อ $h = 1,2,...,n$ และ $X_h \neq M_i$
5. สำหรับ $j = 1,2,...,n$ และ $X_j \neq X_h$ เมื่อ $X_j \neq M_i$
6. คำนวณค่า C_{jih} เก็บไว้ใน TC_{ih}
7. $Minimum = TC_{ih}$
8. ถ้า $Minimum < 0$ แล้ว สลับ M_i กับ X_h
9. ให้ตัวแปร $i = 0$
10. วนรอบจนกว่า $i \leq 0$
11. เลือกเมดอยด์ ตัวที่ i จาก
12. $repeat = false$
13. สำหรับบาง X_h เมื่อ h เรียงจากดัชนีภายในกลุ่ม M_i และ $P_h \neq M_i$
14. $TC_{ih} = 0$
15. สำหรับ $j = 1,2,...,n$ เมื่อ $X_j \neq X_h$ และ $P_j \neq M_i$
16. คำนวณค่า C_{jih} และรวมเก็บไว้ใน TC_{ih}
17. ถ้า $TC_{ih} < 0$ แล้ว
18. จะสลับ M_i กับ P_i
19. $i = 0$

20. $repeat = true$

21. ถ้า ไม่มีการวนรอบ จะกำหนดค่า $i = i + 1$

22. จบการทำงาน

จากขั้นตอนที่ 2 กับ 8 นั้นจะเห็นว่าขั้นตอนวิธี OFIPAM-FAW ใช้เวลาในการประมวลผลเป็น $k(n-k)^2$ เมื่อ k จำนวนเมดอยด์ n คือจำนวนจุดข้อมูล และในขั้นตอนที่ 14 ถึง 21 ใช้เวลาในการประมวลผลเป็น $k(n-k)^2$ เช่นกัน ดังนั้นเวลาที่ใช้ในการประมวลผลของขั้นตอนวิธี OFIPAM-FAW คือ $O(k(n-k)^2)$

ในบทต่อไปจะเป็นผลการทดลอง โดยการประมวลผลเพื่อเทียบเวลาระหว่างขั้นตอนวิธีแพม กับขั้นตอนวิธีการปรับปรุงทั้ง 4 ขั้นตอนวิธี

บทที่ 4

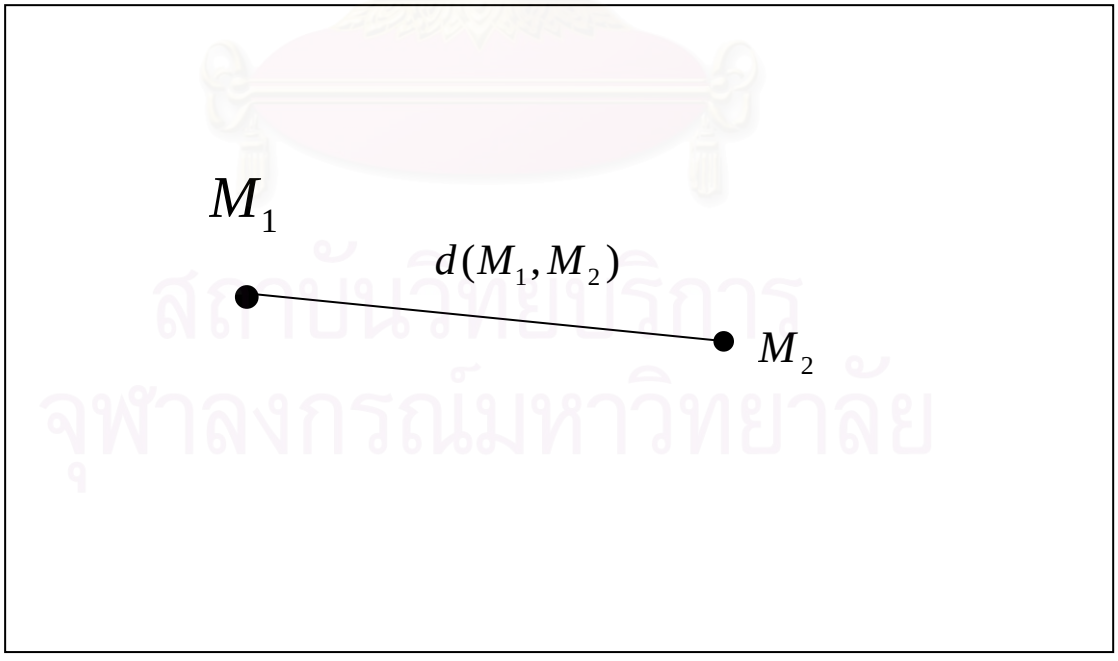
การวิเคราะห์ผลการทดลอง

ในบทที่ 4 จะกล่าวถึงลักษณะของข้อมูลที่ใช้ในการทดลอง และขั้นตอนการจำลองข้อมูล ผลการทดลอง แสดงการเปรียบเทียบเวลาประมวลผลของขั้นตอนวิธีทั้ง 4 ขั้นตอนวิธีกับวิธีแพมโดยกราฟ และวิเคราะห์ผลการทดลอง

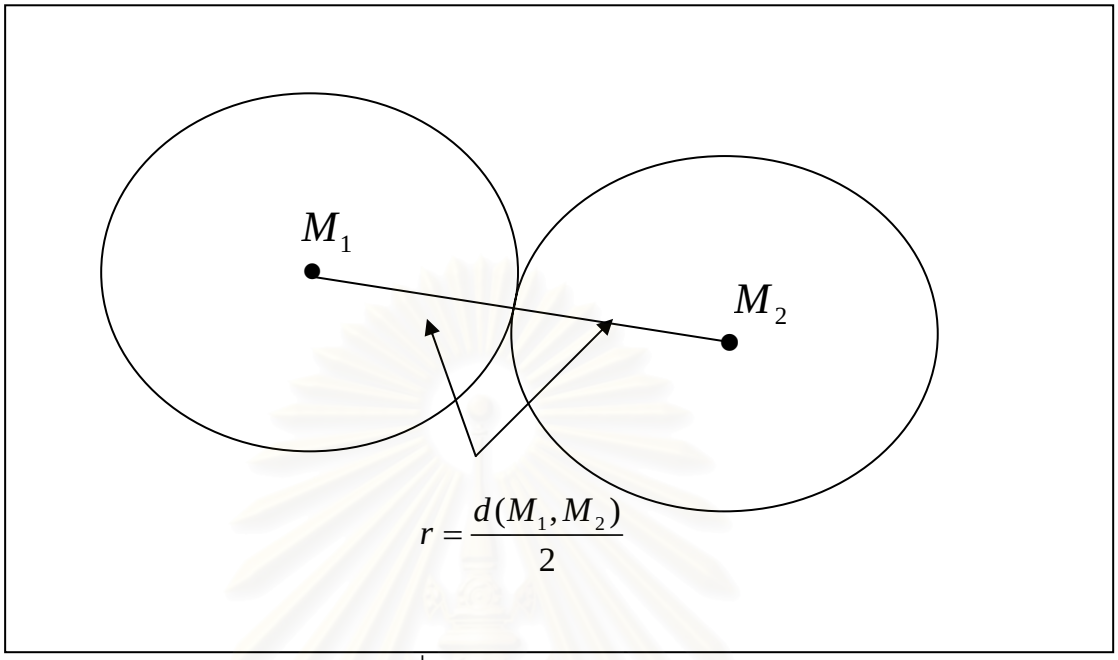
4.1 ข้อมูลนำเข้า

การทดสอบเวลาในการประมวลผลระหว่างขั้นตอนวิธีที่ปรับปรุงในวิทยานิพนธ์นี้กับวิธีแพม ผู้วิจัยใช้ข้อมูลจำลองตั้งแต่ 2 ถึง 20 มิติ ขนาด 1000 จุดข้อมูล ข้อมูลแต่ละขนาดถูกจำลอง โดยการกำหนดเมดอยด์ก่อน ในวิทยานิพนธ์นี้ กำหนดเมดอยด์เริ่มต้น k ตัว แล้วทำการสุ่มจุดข้อมูลรอบเมดอยด์ โดยจุดข้อมูลที่สุ่มระยะห่างกับเมดอยด์เริ่มต้นไม่เกินรัศมีที่ถูกคำนวณด้วยผลรวมระยะทางระหว่างเมดอยด์ทุกคู่หารด้วยสอง ข้อมูลจำลองแต่ละขนาดจะมี 100 แฟ้ม

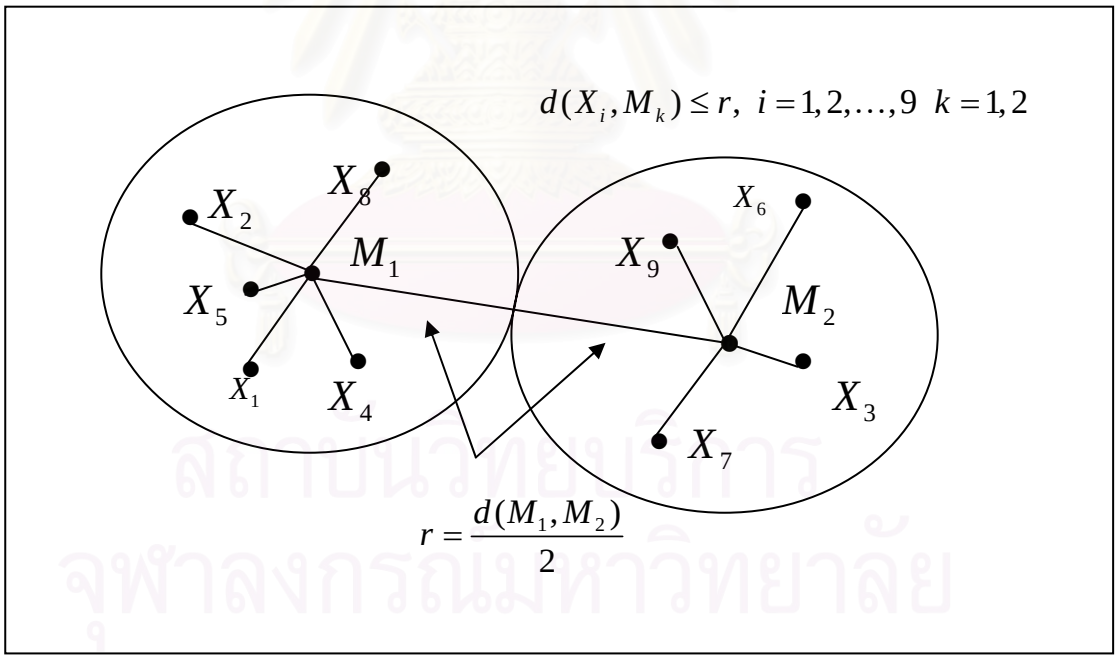
ให้ M_1 และ M_2 เป็นเมดอยด์เริ่มต้น



รูปที่ 4.1 การกำหนดเมดอยด์เริ่มต้น



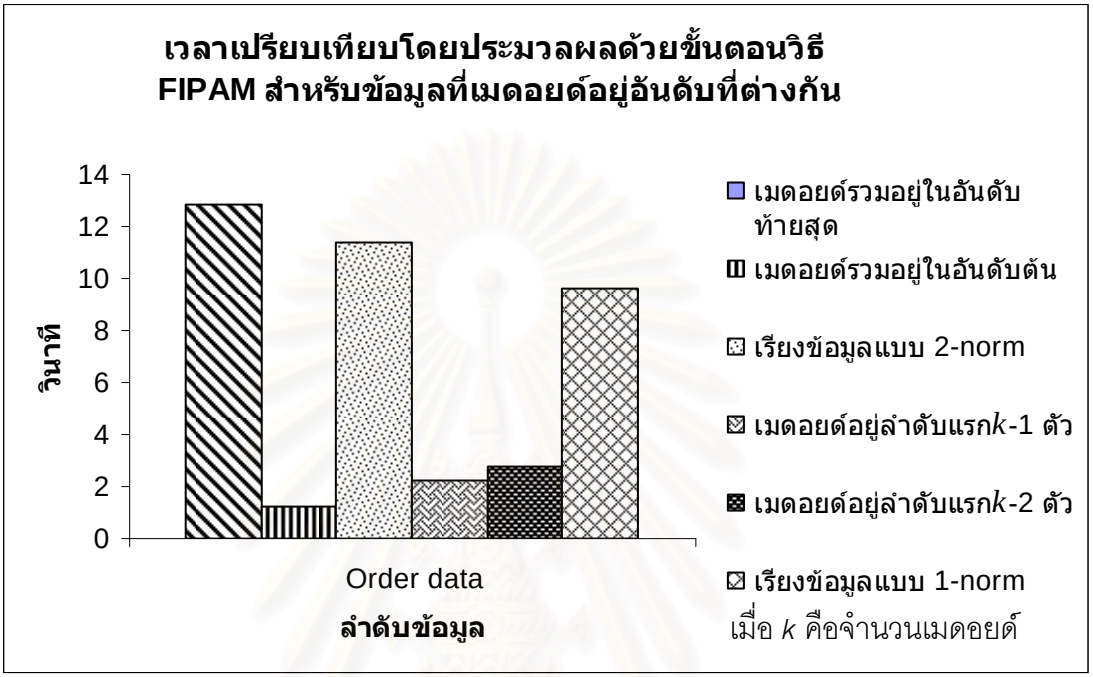
รูปที่ 4.2 คำนวณรัศมีของกลุ่ม



รูปที่ 4.3 สุ่มจุดข้อมูลภายในกลุ่ม

ในรูปที่ 4.2 คำนวณรัศมีของกลุ่ม โดยการหาระยะทางระหว่างสองเมดอยด์ แล้วหารด้วย 2 และรูปที่ 4.3 สุ่มข้อมูลให้มีระยะทางระหว่างเมดอยด์กับข้อมูลที่สุ่มไม่เกินค่ารัศมีที่กำหนด

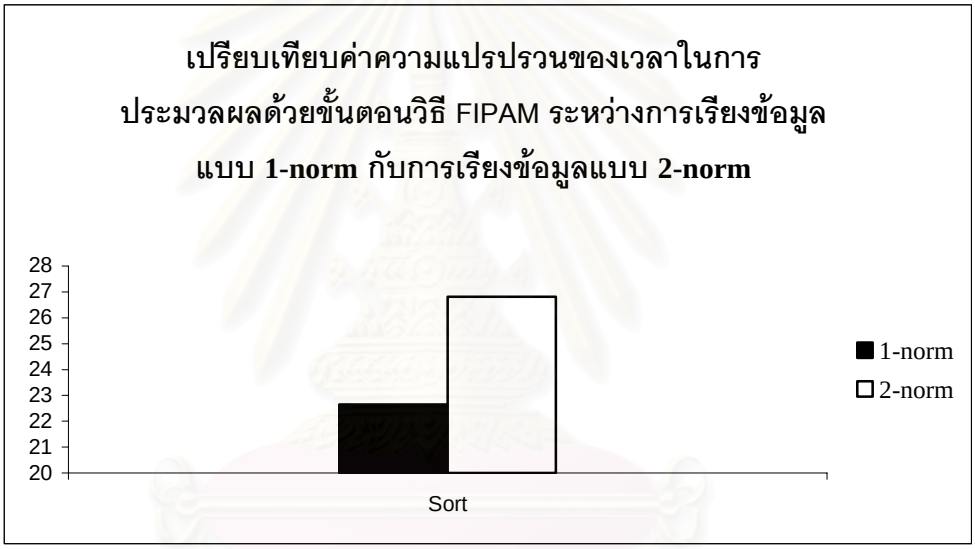
ข้อมูลที่จะใช้วัดเวลาเฉลี่ยจะผ่านกระบวนการเรียงลำดับ เพราะผู้วิจัยไม่ต้องการให้ลำดับของข้อมูลมีผลกระทบต่อเวลาประมวลผลที่ใช้ของแต่ละวิธี ดังรูปที่ 4.4



รูปที่ 4.4 เวลาเปรียบเทียบสำหรับข้อมูลที่ลำดับต่างกัน

จากรูปที่ 4.4 เมื่อผู้วิจัยได้ทดลองการนำเข้าข้อมูลชุดเดียวกัน แต่มีการสลับลำดับการเรียงข้อมูล 6 รูปแบบ รูปแบบที่ 1 คือการเรียงข้อมูลโดยให้ผลเฉลยเมดอยด์ของข้อมูลอยู่ลำดับสุดท้ายจากผลเห็นว่าใช้เวลาประมวลผลนานที่สุด เนื่องจากขั้นตอนวิธี FIPAM จะเลือกคู่เมดอยด์กับจุดข้อมูลที่ให้ผลรวมระยะทางลดลงคู่แรก ดังนั้นเมื่อผลเฉลยอยู่ลำดับสุดท้ายทำให้ขั้นตอนวิธี FIPAM ใช้เวลานาน รูปแบบที่ 2 คือการเรียงข้อมูลโดยให้ผลเฉลยเมดอยด์ของข้อมูลอยู่ลำดับแรก พบว่าใช้เวลาในการประมวลผลเร็วที่สุด เนื่องจากการเรียงข้อมูลโดยให้ผลเฉลยอยู่ลำดับแรกทำให้ขั้นตอนวิธี FIPAM พบผลเฉลยเร็วจึงใช้เวลาประมวลผลน้อย รูปแบบที่ 3 คือการเรียงข้อมูลโดยใช้การเรียงแบบ 2-norm พบว่าใช้เวลาประมวลผลนาน เนื่องจากการเรียงรูปแบบนี้มีการกระจายเมดอยด์ไปในลำดับต่างๆ กัน รูปแบบที่ 4 คือการเรียงข้อมูลโดยให้ผลเฉลยเมดอยด์อยู่ลำดับแรก k-1 ตัว พบว่าใช้เวลาประมวลผลเร็วแต่ช้ากว่ารูปแบบที่ 2 เนื่องจากผลเฉลยอยู่ในลำดับแรก k-1 ทำให้ขั้นตอนวิธี FIPAM พบผลเฉลย k-1 แรกก่อน รูปแบบที่ 5 คือการเรียงเมดอยด์โดยให้ผลเฉลยเมดอยด์อยู่ในลำดับแรก k-2 ตัว พบว่าใช้เวลาประมวลผลช้ากว่ารูปแบบที่ 2 และรูปแบบที่ 4 เนื่องจากผลเฉลยอยู่ในลำดับแรก k-2 ทำให้ขั้นตอนวิธี FIPAM พบผลเฉลย

k-2 แรกก่อน และรูปแบบสุดท้าย คือการเรียงเมดอยด์การเรียงแบบ 1-norm พบว่าใช้เวลาประมวลผลนาน เนื่องจากการเรียงรูปแบบนี้มีการกระจายผลเฉลยไปในลำดับต่างๆ กัน เหมือนรูปแบบที่ 3 จากผลการทดลองแสดงให้เห็นว่าการเรียงลำดับการนำเข้าข้อมูลที่แตกต่างกันมีผลกระทบต่อเวลาที่ใช้ของขั้นตอนวิธี FIPAM ดังนั้นผู้วิจัยเสนอการแก้ปัญหาด้วยการเรียงลำดับข้อมูลก่อนนำเข้าสู่การประมวลผล โดยเลือกระหว่างการเรียงลำดับข้อมูลด้วยค่า 1-norm กับการเรียงลำดับข้อมูลด้วย 2-norm โดยใช้ความแปรปรวนของเวลาเป็นตัววัดความเหมาะสม ซึ่งความแปรปรวนของเวลาคำนวณได้จาก เวลาประมวลผลของขั้นตอนวิธี FIPAM จากจำนวน 100 แฟ้มข้อมูล

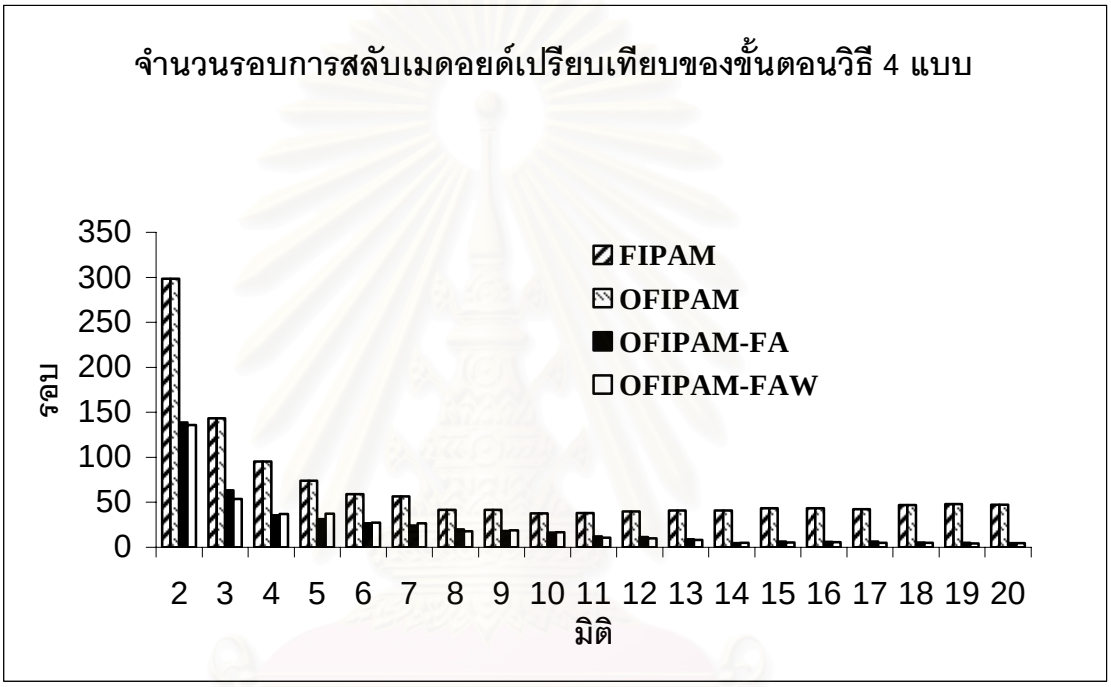


รูปที่ 4.5 เปรียบเทียบความแปรปรวนระหว่างการเรียงข้อมูลที่เรียงโดยวิธี 1-norm กับข้อมูลที่เรียงโดยวิธี 2-norm

พบว่าการเรียงข้อมูลแบบ 1-norm ใช้เวลาประมวลผลที่เสถียรกว่าการเรียงข้อมูลแบบ 2-norm โดยมีค่าความแปรปรวนที่ต่ำกว่า ดังรูปที่ 4.5 ผู้วิจัยจึงเลือกการเรียงแบบ 1-norm ในการเรียงลำดับข้อมูลก่อนนำไปประมวลผล

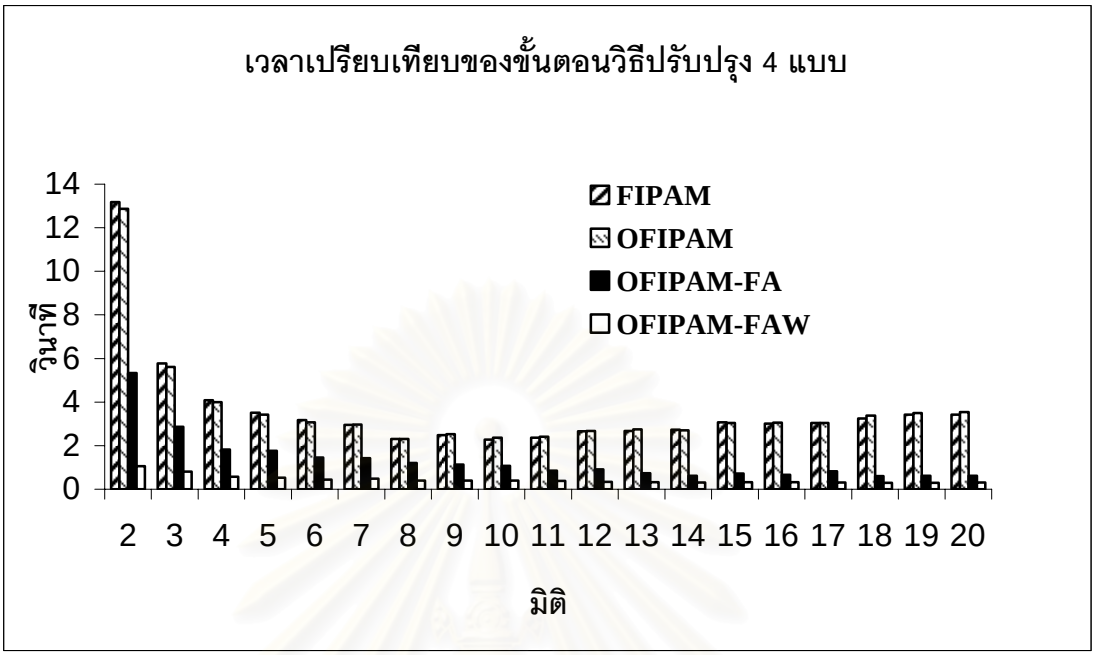
4.2 ผลการทดลองขั้นตอนวิธีปรับปรุงทั้ง 4 ขั้นตอนวิธี

จากขั้นตอนวิธีปรับปรุงทั้ง 4 ขั้นตอนวิธีที่กล่าวมาแล้วในบทที่ 3 เราเลือกแสดงจำนวนรอบของการสลับเมดอยด์ และเวลาที่ใช้ประมวลผลเปรียบเทียบจากข้อมูลที่ถูกจำลอง 100 รอบ ได้ผลดังรูปที่ 4.6 และ 4.7



รูปที่ 4.6 จำนวนรอบของการสลับเมดอยด์เปรียบเทียบของ 4 ขั้นตอนวิธี ของข้อมูลขนาด 2 ถึง 20 มิติ

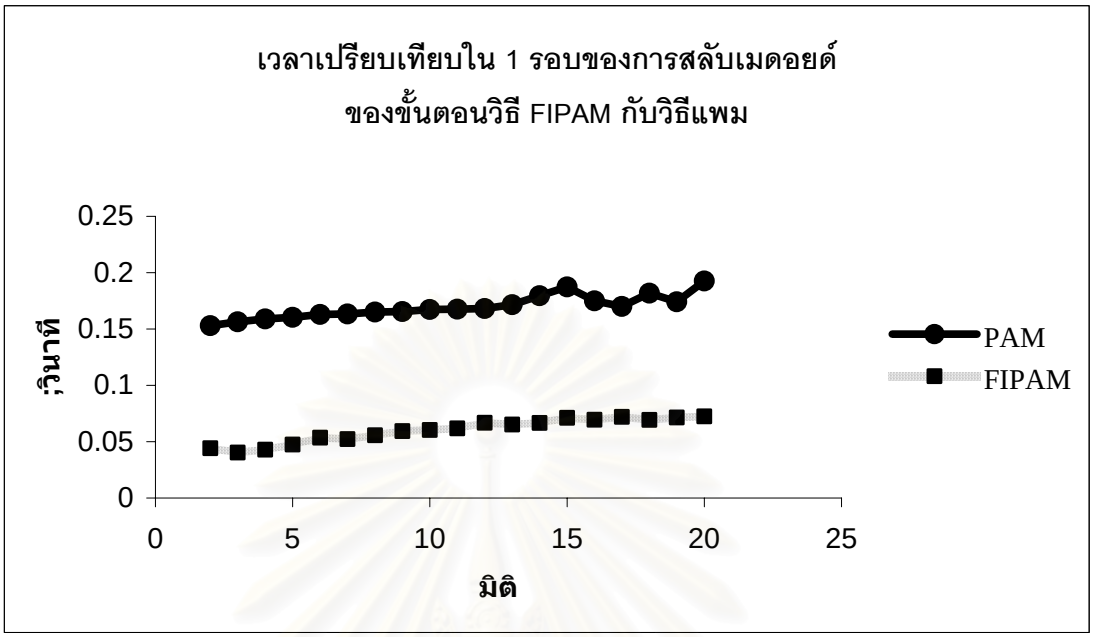
จากรูปที่ 4.6 พบว่าจำนวนรอบของการสลับเมดอยด์ของขั้นตอนวิธีปรับปรุงทั้ง 4 ขั้นตอนวิธีจะลดลงจนลู่เข้าสู่ค่าหนึ่งเมื่อข้อมูลมีมิติมากขึ้น ขั้นตอนวิธี OFIPAM-FA และขั้นตอนวิธี OFIPAM-FAW ใช้จำนวนรอบของการสลับน้อยที่สุดเมื่อเทียบกับขั้นตอนวิธีปรับปรุงทั้ง 4 ขั้นตอนวิธี



รูปที่ 4.7 เวลาเปรียบเทียบของขั้นตอนวิธี 4 ขั้นตอนวิธีของข้อมูลขนาด 2 ถึง 20 มิติ

จากรูปที่ 4.7 พบว่าเวลาประมวลผลของขั้นตอนวิธีปรับปรุงทั้ง 4 ขั้นตอนวิธีจะลดลงเมื่อมิติของข้อมูลเพิ่มขึ้น และมีแนวโน้มคงที่ ขั้นตอนวิธี OFIPAM-FAW ใช้เวลาในการประมวลผลน้อยที่สุดเมื่อเทียบกับขั้นตอนวิธีปรับปรุงทั้ง 4 ขั้นตอนวิธี

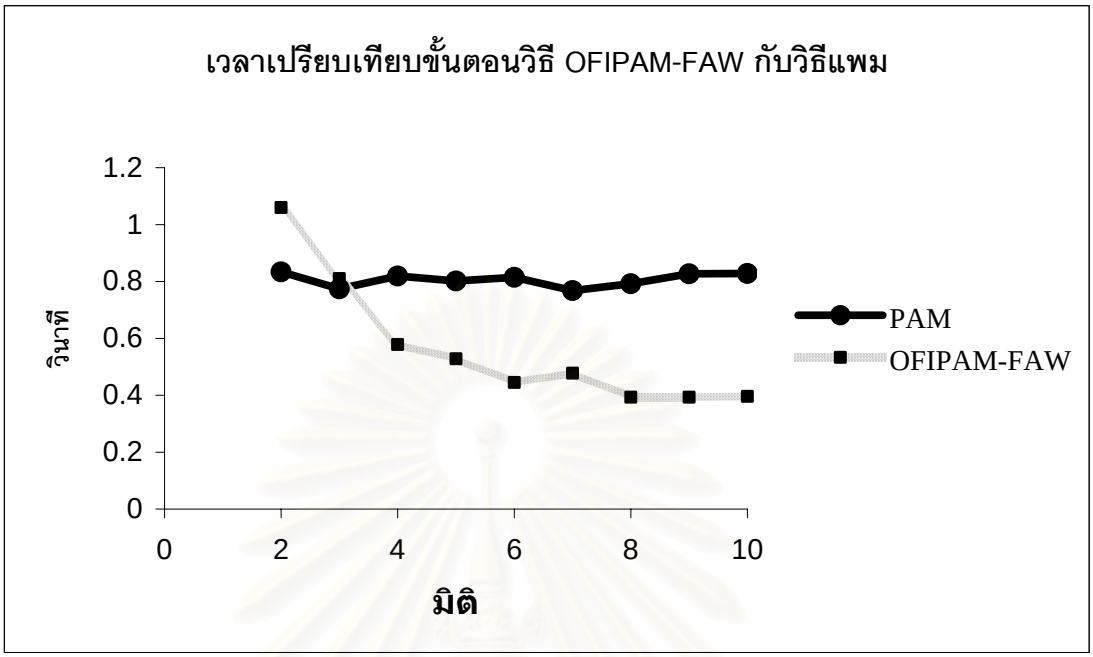
จากผลการทดลองข้างต้นพบว่าขั้นตอนวิธี OFIPAM-FAW ใช้เวลาในการประมวลผล และจำนวนรอบของการสลับเมดอยด์น้อยที่สุดเมื่อเปรียบเทียบกับขั้นตอนวิธีทั้ง 4 ขั้นตอนวิธี เนื่องจากขั้นตอนวิธี OFIPAM-FAW เป็นขั้นตอนวิธีที่ปรับปรุงมาจากขั้นตอนวิธีทั้ง 3 ขั้นตอนวิธี โดยได้แนวคิดเพื่อจะลดเวลาในการประมวลผลใน 1 รอบของการสลับเมดอยด์มาจากขั้นตอนวิธี FIPAM ด้วยการเลือกคู่เมดอยด์ที่ให้ผลรวมระยะทางลดลงค่าแรกโดยแสดงผลได้ดังรูปที่ 4.8



รูปที่ 4.8 เวลาเปรียบเทียบใน 1 รอบของขั้นตอนวิธีแพม กับขั้นตอนวิธี FIPAM

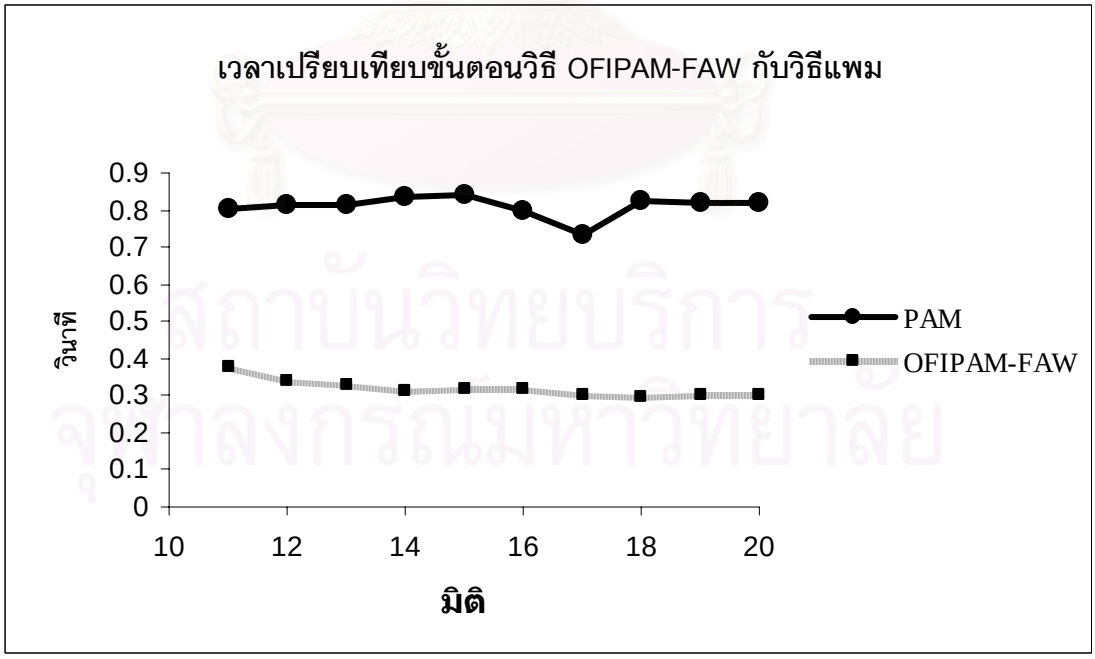
จากรูปที่ 4.8 พบว่าขั้นตอนวิธี FIPAM ใช้เวลาประมวลผลใน 1 รอบของการสลับเมดอยด์เร็วกว่าวิธีแพม และใช้เวลาใกล้เคียงกันในทุกมิติ

วิธีการนี้จะเพิ่มจำนวนรอบของการสลับเมดอยด์ดังรูปที่ 4.6 ดังนั้นขั้นตอนวิธี OFIPAM-FAW จะลดจำนวนรอบของการสลับเมดอยด์โดยใช้แนวคิดมาจากขั้นตอนวิธี OFIPAM-FA ด้วยการเลือกคู่เมดอยด์ที่ให้ผลรวมระยะทางลดลงมากที่สุดในรอบแรก และยังลดจำนวนการพิจารณาคู่เมดอยด์กับจุดข้อมูลด้วยการเลือกพิจารณาคู่เมดอยด์กับข้อมูลที่อยู่ในกลุ่มปัจจุบันก่อน เพราะการเลือกคู่เมดอยด์กับจุดข้อมูลที่ให้ผลรวมระยะทางลดลงมากที่สุดในรอบแรกทำให้ข้อมูลถูกกั้นออกเป็นกลุ่มชัดเจน กราฟในรูปที่ 4.9, 4.10, 4.11 และ 4.12 แสดงเวลาและจำนวนรอบของการสลับเมดอยด์เปรียบเทียบระหว่างขั้นตอนวิธี OFIPAM-FAW กับวิธีแพม



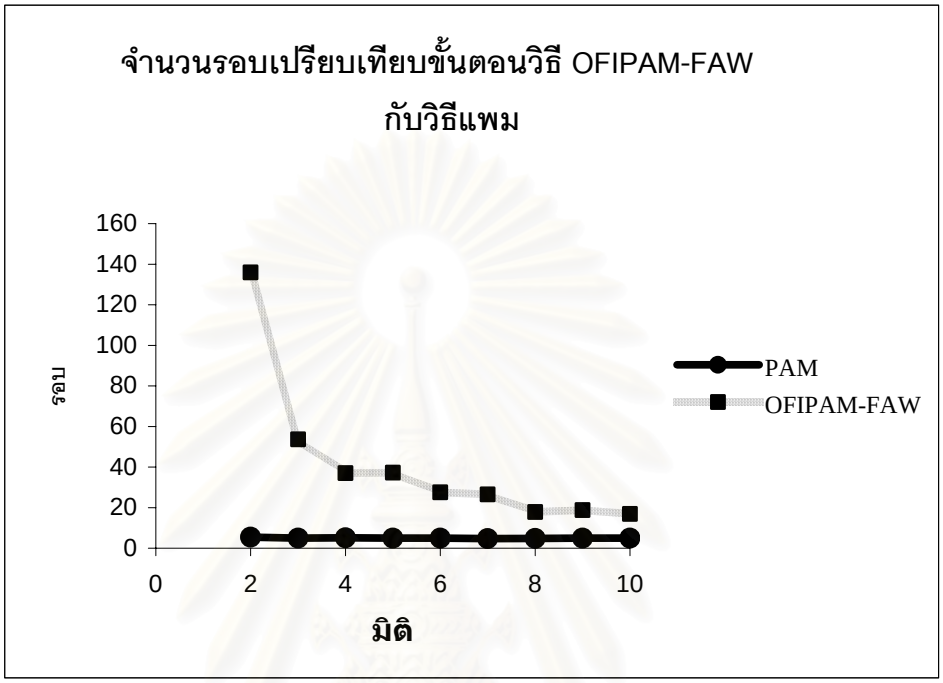
รูปที่ 4.9 เวลาเปรียบเทียบของขั้นตอนวิธี OFIPAM-FAW กับวิธีแพม

จากรูปที่ 4.9 พบว่าเวลาประมวลผลของขั้นตอนวิธี OFIPAM-FAW ใช้เวลาเร็วกว่าวิธีแพมเมื่อข้อมูลมีมิติ ตั้งแต่ 4 มิติไป และมีแนวโน้มที่จะคงที่



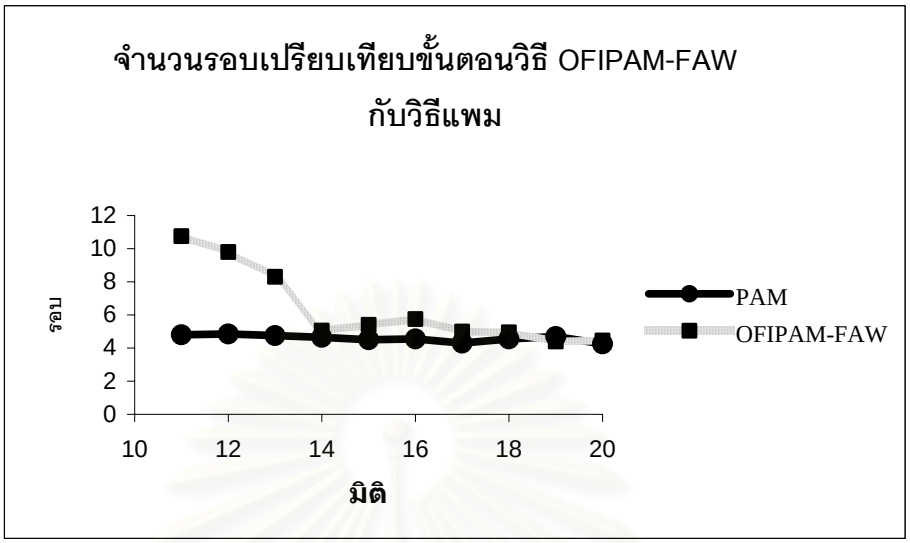
รูปที่ 4.10 เวลาเปรียบเทียบของขั้นตอนวิธี OFIPAM-FAW กับวิธีแพม

จากรูปที่ 4.10 พบว่าขั้นตอนวิธี OFIPAM-FAW เมื่อประมวลผลกับข้อมูลขนาด 10 มิติไป เวลาประมวลผลของขั้นตอนวิธี OFIPAM-FAW มีลักษณะคงที่



รูปที่ 4.11 จำนวนรอบของการสลับเมดอยด์เปรียบเทียบ
ของขั้นตอนวิธี OFIPAM-FAW กับวิธีแพม

จากรูปที่ 4.11 พบว่าจำนวนรอบของการสลับเมดอยด์ของขั้นตอนวิธี OFIPAM-FAW จะลดลงเร็วเมื่อข้อมูลมากกว่า 2 มิติ และจะลดลงช้าๆ เมื่อข้อมูลมีขนาด 8 มิติขึ้นไป



รูปที่ 4.12 จำนวนรอบของการสลับเมดอยด์เปรียบเทียบ
ของขั้นตอนวิธี OFIPAM-FAW กับวิธีแพม

จากรูปที่ 4.12 พบว่าจำนวนรอบของการสลับเมดอยด์ของขั้นตอนวิธี OFIPAM-FAW จะลู่เข้าเทียบเท่ากับจำนวนรอบในการสลับเมดอยด์ของวิธีแพมเมื่อข้อมูลมีขนาด 14 มิติไป เนื่องจากเมื่อมิติของข้อมูลมากขึ้นระยะทางระหว่างเมดอยด์กับจุดข้อมูลภายในกลุ่มไกลขึ้นทำให้ระยะระหว่างเมดอยด์กับจุดข้อมูล และระยะระหว่างเมดอยด์กับเมดอยด์อื่นห่างกันมากการสลับจึงไม่เกิดขึ้นมาก ทำให้ขั้นตอนวิธี OFIPAM-FAW ใช้เวลาต่ำกว่าในมิติที่สูงกว่า

ผลการทดลองจากการประมวลผลกับข้อมูล UCI ชื่อ *"Water treatment"* ซึ่งเป็นข้อมูลขนาด 38 มิติ 527 จุดข้อมูล โดยจะทำการ normalize ข้อมูลให้มีค่าระหว่าง 0 ถึง 1 ข้อมูลที่ได้ไม่มีการเกาะกลุ่มกันชัดเจนเหมือนข้อมูลที่ถูกจำลองขึ้น แต่ละจุดข้อมูลมีการขาดหายไปของลักษณะประจำ ในวิทยานิพนธ์นี้เราจะเติมข้อมูลที่ขาดหายไปด้วยเลข 9 เราจะพิจารณาคำตอบของการประมวลผลด้วยวิธีแพมกับขั้นตอนวิธี OFIPAM-FAW ซึ่งผลที่ได้พบว่าวิธีแพมใช้เวลาประมวลผลเท่ากับ 0.732 วินาที และผลรวมระยะทางเท่ากับ 6754.1919 ขั้นตอนวิธี OFIPAM-FAW ใช้เวลาประมวลผลเท่ากับ 0.368 วินาที และให้ผลรวมระยะทางเท่ากับ 6761.9146 เพราะฉะนั้นจะเห็นได้ว่าขั้นตอนวิธี OFIPAM-FAW ใช้เวลาประมวลผลน้อยกว่าวิธีแพมเมื่อข้อมูลมีมิติมาก แต่ให้ผลรวมระยะทางมากกว่าวิธีแพม เนื่องจากขั้นตอนวิธี OFIPAM-FAW พิจารณาจุดข้อมูลที่อยู่ในกลุ่มก่อน แต่ข้อมูล *"Water treatment"* เป็นข้อมูลที่ไม่มีการเกาะกลุ่มชัดเจนดังนั้น การพิจารณาจุดข้อมูลที่อยู่ในกลุ่มจึงไม่เพียงพอกับการหาผลเฉลยที่ได้ระยะทางรวมน้อยที่สุด

บทที่ 5
สรุปผลการทดลอง

งานวิจัยนี้เสนอขั้นตอนวิธีการปรับปรุงวิธีการแบ่งกลุ่มที่เรียกว่าวิธีแพมโดยมีขั้นตอนการปรับปรุง 4 ขั้นตอน ขั้นตอนที่ 1 คือขั้นตอนการลดเวลาในการประมวลผลใน 1 รอบของการสลับเมดอยด์ ด้วยการเลือกคู่เมดอยด์กับจุดข้อมูลที่ให้ผลรวมระยะทางลดลงค่าแรก (FIPAM) ขั้นตอนที่ 2 คือขั้นตอนการลดจำนวนรอบของการสลับเมดอยด์โดยการเรียงเมดอยด์ (OFIPAM) ขั้นตอนที่ 3 คือขั้นตอนการลดจำนวนรอบของการสลับเมดอยด์เพิ่มจากขั้นตอนที่ 2 โดยการเลือกคู่ที่ให้ผลรวมระยะทางลดลงค่าแรกก่อน (OFIPAM-FA) ขั้นตอนสุดท้ายคือขั้นตอนการพิจารณาเฉพาะจุดข้อมูลที่อยู่ในกลุ่มก่อน (OFIPAM-FAW)

จากผลการทดลองพบว่าวิธีการปรับปรุงวิธีแพมใช้จำนวนรอบของการสลับเมดอยด์ใกล้เคียงกับจำนวนรอบของการสลับเมดอยด์ของวิธีแพมเมื่อข้อมูลมีมิติมากขึ้น เนื่องจากข้อมูลที่มีมิติมากทำให้ระยะห่างของจุดข้อมูลมีความห่างมากขึ้นดังนั้นข้อมูลที่มีมิติน้อยกว่าจะมีความใกล้ชิดของข้อมูลที่เป็นเมดอยด์ทำให้มีการสลับมากขึ้น

เวลาในการประมวลผลใน 1 รอบของวิธีการปรับปรุงใช้เวลาน้อยกว่าวิธีแพม เนื่องจากขั้นตอนวิธี OFIPAM-FAW ไม่พิจารณาทุกคู่เมดอยด์กับจุดข้อมูล แต่เลือกสลับเมดอยด์เมื่อเจอคู่เมดอยด์กับจุดข้อมูลที่ให้ผลรวมระยะทางลดลงทันที ดังนั้นเวลาในการประมวลผลรวมของวิธีการปรับปรุงจึงใช้เวลาน้อยการเวลาในการประมวลผลรวมของวิธีการแพมดังรูปที่ 4.9 และ 4.10

ผลจากการประมวลผลกับข้อมูล *“Water treatment”* ที่ได้จาก UCI พบว่าขั้นตอนวิธี OFIPAM –FAW ใช้เวลาประมวลผลน้อยกว่าวิธีแพม แต่ผลรวมระยะทางมากกว่าวิธีแพมโดยคิดเป็น 0.11% ซึ่งสามารถสรุปได้ว่าขั้นตอนวิธี OFIPAM-FAW สามารถแบ่งกลุ่มข้อมูลกับข้อมูลที่เกาะกลุ่มชัดเจน และข้อมูลที่มีมิติมาก

จากข้อมูลที่ใช้ในการทดลองของงานวิจัยนี้เป็นข้อมูลที่ถูกจำลองขึ้น โดยการกำหนดให้เป็นตัวเลขเท่านั้นเพื่อให้สามารถคำนวณระยะทางได้ง่ายขึ้น แต่ในความเป็นจริงลักษณะของข้อมูลอาจไม่เป็นตัวเลข ดังนั้นงานในอนาคตคือ การพัฒนาขั้นตอนวิธี OFIPAM-

FAW ให้สามารถใช้กับข้อมูลที่ไม่เป็นตัวเลขได้ นอกจากนี้ข้อมูลที่ถูกจำลองขึ้น ผู้วิจัยได้มีการกำหนดเมดอยด์และกลุ่มขึ้นมาก่อนทำให้ขั้นตอนวิธี OFIPAM-FAW สามารถหาคำตอบได้ แต่ในข้อมูลจริงนั้นไม่จำเป็นที่จะมีการเกาะกลุ่มที่ชัดเจน งานต่อไปคือการนำข้อมูลจริงมาประมวลผลผ่าน Train-Validate-Test แล้วนำไปเปรียบเทียบกับขั้นตอนวิธี CLARA และ CLARAN

รายการอ้างอิง

[1] Han, J. and Kamber, M. *Data mining: Concepts and techniques*. San Fransisco: Morgan Kaufmann Publishers, 2001.

[2] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. From data mining to knowledge discovery in databases. *AI Magazine* (1996): 37-54.

[3] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. The kdd process for extract in useful knowledge from volumes data. *Communications of the ACM* 11 (November 1996): 27-31.

[4] Two Crows corporations. *Introduction to data mining and knowledge discovery* [online]. (1999). Available from: http://www.twocrows.com [cited 18 January 2007].

[5] Kuonen, D. A statistical perspective of data mining. *CRM Zine* (December 2004): 1-6.

[6] Srivastava, J., Cooley, R., Deshpande, M. and Tan, P. N. Web usage mining: Discovery and applications of usage patterns from web data, *SIGKDD Explorations* (2000):12–23.

[7] Tung, K. H., Hou, A. J. and Han, J. COE: Clustering with Obstacles Entities, a preliminary study. In *Proceedings of the 4th Pacific-Asia Conference on Knowledges Discovery and Data Mining (PAKDD) (*April 2000).

[8] Ng, R. T. and Han, J. CLARANS: A method for clustering objects for spatial Data Mining. *IEEE Transactions on Knowledge and Data Engineering.* (September/October 2002): 1003 – 1016.

[9] Zhange, T., Ramakrishnan, R. and Livny, M., An efficient data clustering medoid for very large databases, In *Proceeding ACM-SIGMOD International Conference Management* (June 1996):103-114.

[10] Ng, R. T. and Han, J. Efficient and effective clustering methods for spatial data mining. In *Proceeding of the 20th International Conference on Very Large Data Bases*, (September 1994) : 144-155.

[11] Handl, J. and Knowles, J., Multiobjective clustering around medoids, *Evolutionary Computation, 2005. The 2005 IEEE Congress Vol.1* (September 2005): 632 – 639.

[12] Kaufman, L. and Rouseeuw, P. J., *Finding groups in data: an introduction to cluster analysis*, John Wiley & Sons, 1990.

[13] Van der Laan, M. J., Pollard, K. S. and Bryan, J. A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation (*2003): 575-584.

[14] Ioannidis, Y. and Kang, Y., Randomized algorithms for optimizing large join queries. In *Proceeding of SIGMOD* (1990): 312 – 321.

[15] Zhang, Q. and Couloigner, I., A new and efficient k-medoid algorithm for spatial clustering. In *International Conference on Computational Science and Its Applications* (2005): 181- 189.

ประวัติผู้เขียนวิทยานิพนธ์

ชื่อ-นามสกุล : นายรุ่งโรจน์ พุ่มดวง

วัน-เดือน-ปีเกิด : 29 กันยายน พ.ศ.2525

ภูมิลำเนา : แขวงบางซื่อ เขตบางซื่อ จ.กรุงเทพมหานคร

สำเร็จการศึกษา : วิทยาศาสตรบัณฑิต (คณิตศาสตร์)
จากสถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ