รายงานวิจัยฉบับสมบูรณ์

โครงการขั้นตอนวิธีการแบ่งกลุ่มแบบยืดหยุ่นที่ใช้ความซับซ้อนด้านเวลาและพื้นที่เกือบเชิงเส้นสำหรับข้อมูลสตรีมมิ่งที่มีแนวคิดดริฟท์

โดย ผู้ช่วยศาสตราจารย์ ดร.อุรีรัฐ สุขสวัสดิ์ชน

พฤษภาคม 2563

สัญญาเลขที่ TRG5880092

รายงานวิจัยฉบับสมบูรณ์

โครงการขั้นตอนวิธีการแบ่งกลุ่มแบบยืดหยุ่นที่ใช้ความซับซ้อนด้านเวลาและพื้นที่เกือบเชิงเส้นสำหรับข้อมูลสตรีมมิ่งที่มีแนวคิดดริฟท์

โดย ผู้ช่วยศาสตราจารย์ ดร.อุรีรัฐ สุขสวัสดิ์ชน
คณะวิทยาการสารสนเทศ
มหาวิทยาลัยบูรพา

สนับสนุนโดยสำนักงานกองทุนสนับสนุนการวิจัยและ
มหาวิทยาลัยบูรพา

กิตติกรรมประกาศ

งานวิจัยฉบับนี้เสร็จสมบูรณ์ได้ด้วยดีด้วยการสนับสนุนจากศาสตราจารย์ ดร.ชิดชนก เหลือสินทรัพย์ นักวิจัยที่ปรึกษา ที่ช่วยแนะนำให้คำปรึกษา และแนะแนวทางในการดำเนินงานวิจัยให้สำเร็จลุล่วง รวมถึงผู้ช่วยศาสตราจารย์ ดร.จักริน สุขสวัสดิ์ชน ที่ช่วยในการดำเนินงานวิจัย และยังช่วยแก้ไขข้อบกพร่อง แนะนำในงานเขียนงานวิจัยให้มีความสมบูรณ์มากยิ่งขึ้น

ขอขอบคุณนายนรงค์ฤทธิ์ ตั้งปฐมพงศ์ นิสิตระดับปริญญาโท ผู้ช่วยวิจัย ที่เป็นกำลังสำคัญในการร่วมทำงานวิจัยนี้ให้สำเร็จลุล่วง

ขอขอบคุณพื้นที่ทำงานจาก อาคารสิรินธร คณะวิทยาศาสตร์ และห้องปฏิบัติการวิจัย Mobile Application Developers Incubation คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา อาคารใหม่

งานวิจัยนี้ได้รับทุนอุดหนุนจากสำนักงานกองทุนสนับสนุนการวิจัยและมหาวิทยาลัยบูรพา สัญญาเลขที่ TRG5880092

ผู้ช่วยศาสตราจารย์ ดร.อุรีรัฐ สุขสวัสดิ์ชน

บทคัดย่อ

ความก้าวหน้าทางเทคโนโลยีทั้งด้านฮาร์ดแวร์และซอฟต์แวร์ส่งผลให้โปรแกรมหรือแอปพลิเคชัน ต้องคำนึงถึงข้อมูลที่จะประมวลผลมากขึ้น เนื่องจากข้อมูลที่ใช้ประมวลผลมักเป็นข้อมูลที่มีขนาดใหญ่ มีการเปลี่ยนแปลงอย่างรวดเร็วมาก และมีข้อมูลที่เกิดขึ้นใหม่อย่างต่อเนื่องไม่สิ้นสุด อีกทั้งยังมีปริมาณที่มากเกินกว่าจะประมวลผลข้อมูลทั้งหมดในหน่วยความจำหลักได้ เรียกข้อมูลลักษณะเช่นนี้ว่า ข้อมูลสตรีมมิ่งหรือข้อมูลกระแส แต่ขั้นตอนวิธีการในปัจจุบันที่จัดกลุ่มข้อมูลแบบกระแสโดยพิจารณาจากความหนาแน่นยังคงมีข้อผิดพลาด เนื่องจากกลุ่มข้อมูลขนาดเล็กสำหรับเก็บข้อสรุปเชิงสถิติของข้อมูลกระแสที่เข้ามา เพื่อลดหน่วยความจำหลักที่ใช้ประมวลผล ยังไม่เหมาะสมกับลักษณะของข้อมูล และการจัดกลุ่มเพื่อให้ได้กลุ่มข้อมูลสำหรับไปใช้งานยังคงมีความผิดพลาด อีกทั้งยังไม่เหมาะสมกับข้อมูลที่ให้ความสำคัญกับข้อมูลเท่ากันแม้เวลาผ่านไป ดังนั้นงานวิจัยนี้จึงนำเสนอ “ขั้นตอนวิธีการจัดกลุ่มข้อมูลที่มีความยืดหยุ่นสำหรับข้อมูลสตรีมมิ่ง โดยใช้เพียงข้อมูลที่เข้ามาใหม่ร่วมกับโครงสร้างกลุ่มย่อยไฮเพอร์อัลลิปโซดอยล์” เรียกโดยย่อว่า “DyCluStream” โดยแบ่งการทำงานออกเป็นสองส่วน คือส่วนออนไลน์ที่จะใช้โครงสร้างไดนามิคไฮเพอร์อัลลิปโซดอยล์เป็นกลุ่มข้อมูลขนาดเล็กสำหรับเก็บข้อสรุปเชิงสถิติสำหรับข้อมูลที่เข้ามาด้วยมาตรวัดระยะทางที่มีความเหมาะสมกับโครงสร้างกลุ่มข้อมูลขนาดเล็ก โดยรูปทรงของกลุ่มข้อมูลขนาดเล็กถูกคำนวณจากข้อมูลทั้งหมดที่อยู่ในกลุ่ม และส่วนออฟไลน์จะใช้การเชื่อมต่อด้วยการซ้อนทับกันของกลุ่มข้อมูลขนาดเล็กเพื่อให้ได้ผลลัพธ์สำหรับนำไปใช้งาน และ “ขั้นตอนวิธีการจัดกลุ่มข้อมูลที่มีความยืดหยุ่นสำหรับข้อมูลสตรีมมิ่ง โดยใช้เพียงข้อมูลที่เข้ามาใหม่ร่วมกับโครงสร้างกลุ่มย่อยไฮเพอร์อัลลิปโซดอยล์พลัส” หรือเรียกว่า DyCluStream+ เป็นขั้นตอนวิธีการที่พัฒนาต่อยอดมาจาก DyCluStream ที่มีความสามารถในการจัดกลุ่มข้อมูลขนาดเล็กต่อจากการจัดกลุ่มก่อนหน้าได้ทันทีโดยไม่จำเป็นต้องจัดกลุ่มใหม่ทั้งหมด เพิ่มความสามารถในการจัดการกับข้อมูลจริงได้ และมีการรวมกลุ่มข้อมูลขนาดเล็กที่มีคุณลักษณะที่คล้ายกัน เพื่อลดหน่วยความจำที่ใช้และช่วยเพิ่มความเร็วในการประมวลผล จากการทดลองทั้งข้อมูลจริงและข้อมูลสังเคราะห์โดยใช้ตัววัดประสิทธิภาพ Purity, Jaccard Index และ Rand Index แสดงให้เห็นว่าวิธีการ DyCluStream และ DyCluStream+ มีประสิทธิภาพที่ดีกว่า DenStream และ HECES รวมทั้งยังใช้เวลาในการคำนวณที่รวดเร็วกว่า

คำสำคัญ: การจัดกลุ่มโดยพิจารณาความหนาแน่น/ข้อมูลกระแส/โครงสร้างกลุ่มย่อยไฮเพอร์อัลลิปโซดอยล์

Abstract

Recent advances in both hardware and software induced the current programs or applications considering in more processing. Because the data used to process are massive and fast-changing. Those applications potentially generate infinite data which are very large to process in memory. We called the data characteristic as mentioned as “data stream”. However, recent density-based clustering algorithms still have drawbacks, since the micro-clusters which kept the statistic summary of incoming data used to reduce space in main memory were not suitable to describe the characteristic of data. In addition, the clustering results had some mistakes and were not suitable for the data which has all sample having the same important even though when time passed. In this work, we proposed “ the Dynamic Clustering Algorithm for Streaming Data Using One- Pass- Throw- Away Datum and Hyper- ellipsoidal Micro-Clustering” or “DyCluStream”. The DyCluStream consists of online and offline phases. In an online phase, the DyCluStream used dynamic hyper- ellipsoidal micro- cluster to capture the incoming data and calculated shape from that data by suitable measurements. In the offline phase, the final clusters were generated by expanding the hyper-ellipsoidal micro-clusters. Moreover, we improved our method called “the Dynamic Clustering Algorithm for Streaming Data Using One- Pass- Throw- Away Datum and Hyper- ellipsoidal Micro- Clustering Plus” or “ DyCluStream+ ” from DyCluStream which has the ability to continue group the micro-cluster from the previous clustering. Beside, DyCluStream+ can utilize with the real datasets and the micro- clusters with the same characteristic were merged to reduce the memory usage and speed up the time in processing. From the experimental result with synthetic and real data shown that DyCluStream and The DyCluStream+ outperform than DenStream and HECES in Purity, Jaccard index and Rand index measure. Additionally, efficient in computational time.

KEYWORD: DENSITY-BASED CLUSTERING/STREAMING DATA/HYPER-ELLIPSOIDAL MICROCLUSTER

สารบัญ

หน้า
กิตติกรรมประกาศ ..... ค
บทคัดย่อ ..... ง
Abstract ..... จ
สารบัญ ..... ฉ
สารบัญตาราง ..... ช
สารบัญตาราง (ต่อ) ..... ช
สารบัญภาพ ..... ฌ
สารบัญภาพ (ต่อ) ..... ญ
สารบัญภาพ (ต่อ) ..... ฎ
สารบัญภาพ (ต่อ) ..... ฎ
บทที่ 1 1
บทนำ 1
1.1 ที่มาและความสำคัญปัญหา 1
1.2 แนวทางการแก้ปัญหา 7
1.3 วัตถุประสงค์ของงานวิจัย 8
1.4 ขอบเขตของงานวิจัย 8
1.5 ประโยชน์ที่คาดว่าจะได้รับ 9
บทที่ 2 10
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง 10
2.1 ทฤษฎีที่เกี่ยวข้อง 10
2.1.1 ลักษณะของข้อมูลสตรีมมิ่งหรือข้อมูลกระแส (Streaming Data) 10
2.1.2 การจัดกลุ่มกับข้อมูลกระแส (Stream Clustering) 10
2.1.3 Recursive Computation 11
2.1.4 วิธีการหาค่าความคล้ายคลึงแบบโคไซน์ (Cosine Similarity) 12
2.1.5 Bhatthacharyya Distance 12
2.1.6 Principal Component Analysis 12
2.2 งานวิจัยที่เกี่ยวข้อง 13
บทที่ 3 19

วิธีการที่นำเสนอ 19
3.1 การนำเข้าข้อมูล 22
3.2 โครงสร้างสำหรับกลุ่มข้อมูลขนาดเล็ก 22
3.2.1 Amc - Actual micro-clusters 22
3.3.2 Smc - Small micro-clusters 22
3.3 ขั้นตอนวิธีการในการจัดกลุ่มข้อมูล DyCluStream 27
3.3.1 ขั้นตอนส่วนออนไลน์ 28
3.3.2 ขั้นตอนส่วนออฟไลน์ 36
3.4 ขั้นตอนวิธีการ DyCluStream+ 40
3.4.1 การปรับปรุงขั้นตอนส่วนออนไลน์ 41
3.4.2 การปรับปรุงขั้นตอนส่วนออฟไลน์ 48
บทที่ 4 54
ผลการดำเนินงาน 54
4.1 ข้อมูลที่ใช้ในการทดลอง 54
4.1.1 ข้อมูลสังเคราะห์ 54
4.1.2 ข้อมูลจริง 57
4.2 วิธีการวัดประสิทธิภาพความถูกต้องและหน่วยความจำที่ใช้ของวิธีการจัดกลุ่ม 58
4.2.1 Purity 58
4.2.2 Rand statistic หรือ Rand Index (RI) 58
4.2.3 Jaccard Index (JI) 59
4.2.4 หน่วยความจำที่ใช้ในการจัดกลุ่ม 59
4.3 การออกแบบการทดลอง 61
4.3.1 พารามิเตอร์ของขั้นตอนวิธีการต่าง ๆ 61
4.3.2 การออกแบบการทดลองของข้อมูลสังเคราะห์ 62
4.3.3 การออกแบบการทดลองของข้อมูลจริง 80
บทที่ 5 108
สรุปผลการดำเนินงานและข้อเสนอแนะ 108
5.1 สรุปผลการดำเนินงาน 108
5.2 วิจารณ์ผลการดำเนินงาน 109

5.2.1 ข้อดีของงานวิจัย 109
5.2.2 ข้อจำกัดของงานวิจัย 110
5.2.3 ข้อเสนอแนะของงานวิจัย 110
บรรณานุกรม 111
ภาคผนวก 125
ภาคผนวก ก 126
เอกสารเผยแพร่ผลงานวิจัย 126

สารบัญตาราง

ตารางที่ หน้า

2-1 เปรียบเทียบงานวิจัยที่เกี่ยวข้องในประเด็นต่าง ๆ 18
4-1 คุณลักษณะของข้อมูล 58
4-2 ตัวแปรที่เก็บในหน่วยความจำหลักของแต่ละขั้นตอนวิธีการ 60
4-3 หน่วยความจำที่ใช้ต่อหนึ่งกลุ่มข้อมูลขนาดเล็กในขั้นตอนวิธีการ **DyCluStream, DyCluStream+** 60
4-4 หน่วยความจำที่ใช้ต่อหนึ่งกลุ่มข้อมูลขนาดเล็กในขั้นตอนวิธีการ **HECES** โดยจำเป็นต้องเก็บข้อมูลทั้งหมดไว้ 61
4-5 หน่วยความจำที่ใช้ต่อหนึ่งกลุ่มข้อมูลขนาเล็กในขั้นตอนวิธีการ **DenStream** 61
4-6 พารามิเตอร์ของขั้นตอนวิธีการต่าง ๆ 62
4-7 การกำหนดพารามิเตอร์ของขั้นตอนวิธีการต่าง ๆ 63
4-8 พารามิเตอร์ที่ให้ผลดีที่สุดหลังการจัดกลุ่ม 64
4-9 ผลการจัดกลุ่มข้อมูล S1, S2, S3 70
4-10 ผลการจัดกลุ่มข้อมูล AbitaryHCM, AbitaryHCM2 74
4-11 ผลการจัดกลุ่มข้อมูล t4.8k, t4.10k, t8.8k 80
4-12 ผลการทดลองของตัววัด Purity สำหรับการปรับเปลี่ยนตัวแปร Stream Speed 82
4-13 ผลการทดลองของตัววัด Rand statistic สำหรับการปรับเปลี่ยนตัวแปร Streamimg speed 83
4-14 ผลการทดลองของตัววัด Jaccard Index สำหรับการปรับเปลี่ยนตัวแปร Streaming Speed 84
4-15 ผลลัพธ์ของกลุ่มข้อมูลขนาดเล็กที่ผ่านการจัดกลุ่ม สำหรับการปรับเปลี่ยนตัวแปร Streaming Speed 85
4-16 ผลลัพธ์ของกลุ่มข้อมูลที่ผ่านการจัดกลุ่ม สำหรับการปรับเปลี่ยนตัวแปร Streaming Speed 86

สารบัญตาราง (ต่อ)

ตารางที่ หน้า

4-17 หน่วยความจำที่ใช้งาน สำหรับการปรับเปลี่ยนตัวแปร Streaming Speed 88

4-18 ผลการทดลองของตัววัด Purity สำหรับการปรับเปลี่ยนตัวแปร Radius 90

4-19 ผลการทดลองของตัววัด Rand statistic สำหรับการปรับเปลี่ยนตัวแปร Radius 91

4-20 ผลการทดลองของตัววัด Jaccard Index สำหรับการปรับเปลี่ยนตัวแปร Radius..... 92

4-21 ผลลัพธ์ของกลุ่มข้อมูลขนาดเล็กที่ผ่านการจัดกลุ่ม สำหรับการปรับเปลี่ยนตัวแปร Radius 93

4-22 ผลลัพธ์ของกลุ่มข้อมูลที่ผ่านการจัดกลุ่ม สำหรับการปรับเปลี่ยนตัวแปร Radius 94

4-23 หน่วยความจำที่ใช้งาน สำหรับการปรับเปลี่ยนตัวแปร Radius 95

4-23 การเปรียบเทียบประสิทธิภาพเชิงเวลาของ DenStream, HECES, DyCluStream และ DyCluStream+..... 107

สารบัญภาพ

ภาพที่ หน้า

1-1 การแทนรูปร่างตัวแทนกลุ่มของข้อมูลเป็น spherical 2
1-2 รูปร่าง micro-cluster ที่มีลักษณะเป็นแบบ hyper-elliptical ของ CompactStream 3
1-3 การกำหนดรูปทรงของตัวแทนกลุ่มข้อมูลจากขั้นตอนวิธีการ CompactStream 3
1-4 กลุ่มข้อมูลจากข้อมูลบางส่วน 4
1-5 ผลลัพธ์การจัดกลุ่มที่เกิดจากการรวมกันและแทนที่ด้วยกลุ่มข้อมูลใหม่ 4
1-6 ผลลัพธ์การจัดกลุ่มที่เกิดจากการซ้อนทับกันของกลุ่มข้อมูล 5
1-7 ขั้นตอนวิธีการ DyCluStream 7
2-1 การแทนรูปร่างกลุ่มของข้อมูลขนาดเล็กเป็น spherical 13
2-2 ขั้นตอนวิธีการ DenStream ส่วนออนไลน์ 15
3-1 ขั้นตอนวิธีการ DyCluStream ส่วนออนไลน์ 20
3-2 ขั้นตอนวิธีการ DyCluStream ส่วนออฟไลน์ 21
3-3 โครงสร้าง dynamic hyper-ellipsoidal micro-cluster 23
3-4 โครงสร้าง hyper-ellipsoidal Micro-cluster ที่มีจุดศูนย์กลางอยู่ที่จุดกำเนิด 24
3-5 โครงสร้าง hyper-ellipsoidal Micro-cluster ที่สามารถหมุนไปตามฐานหลักเชิงตั้งฉากปรกติที่มีจุดศูนย์กลางอยู่ที่จุดกำเนิด 24
3-6 โครงสร้าง dynamic hyper-ellipsoidal micro-cluster 25
3-7 ข้อมูลที่นำเข้าอยู่ในรัศมีของกลุ่มข้อมูล Amc_w ที่ปรับปรุงโครงสร้าง (เส้นประสีแดง) 26
3-8 การปรับปรุงโครงสร้างกลุ่มข้อมูล Amc_w จากรูปร่างเดิมเป็นรูปร่างใหม่ 26
3-9 ขั้นตอนวิธีการ DyCluStream ส่วนออนไลน์ 28
3-10 การสร้างตัวแทนกลุ่มของข้อมูลเพื่อใช้ในการจัดเก็บข้อมูลที่ถูกนำเข้าเป็นครั้งแรก 29
3-11 การปรับปรุงค่า เฉลี่ย จำนวนสมาชิก และ Covariance matrix ของกลุ่มข้อมูลโครงสร้าง Smc_w เมื่อข้อมูลที่เข้ามาอยู่ภายในรัศมี 30
3-12 กลุ่มข้อมูล Smc_w มีความหนาแน่นถึงระดับที่ผู้ใช้กำหนด (สังเกตุได้จากจำนวนจุดสีเทา) . 32
3-13 โครงสร้าง $Amc_{|Amc|+1}$ ที่ถูกสร้างขึ้นจาก Smc_w 32

สารบัญภาพ (ต่อ)

ภาพที่ หน้า
3-14 ข้อมูลที่นำเข้าอยู่ในรัศมีของ $\mathbf{Amc_w}$ ที่ทดลองปรับปรุงโครงสร้าง (เส้นประสีแดง).....33
3-15 การปรับปรุงโครงสร้าง $\mathbf{Amc_w}$เมื่อรูปร่างที่ถูกทดลองตามข้อมูลที่เข้ามาสามารถครอบคลุมข้อมูลได้.....33
3-16 ขั้นตอนวิธีการ DyCluStream ส่วนออนไลน์ 35
3-17 ขั้นตอนวิธีการ DyCluStream ส่วนออฟไลน์ 36
3-18 ขั้นตอนเริ่มต้นการแผ่กระจายกลุ่มข้อมูล Amc.....37
3-19 กลุ่มข้อมูล Amc_w ไม่สามารถแผ่ขยายต่อไปได้.....38
3-20 เริ่มการแผ่ขยายจากกลุ่มข้อมูล Amc_w ที่มีความหนาแน่นมากที่สุดที่ยังไม่ได้ถูกกำหนดหมายเลขกลุ่มเป็นลำดับถัดไป.....39
3-21 ผลลัพธ์ของการจัดกลุ่มข้อมูลในส่วนออฟไลน์.....39
3-22 ขั้นตอนส่วนออฟไลน์.....40
3-23 ขั้นตอน DyCluStream+ ส่วนออนไลน์ 41
3-24 ตัวอย่างของข้อมูลที่ตัวแปรใด ๆ มีความสัมพันธ์เชิงเส้น.....42
3-25 ตัวอย่างของข้อมูลที่ตัวแปรมีลักษณะเป็นค่าเดียวกัน.....42
3-26 ขั้นตอนวิธีการประมาณค่า Covariance Matrix.....44
3-27 มาตรวัดระยะทาง $mahaDist(xi, Amc_w)$ 45
3-28 มาตรวัดระยะทาง $pointToEllipse(xi, Amc_w)$ 45
3-29 ขั้นตอน DyCluStrem+ ส่วนออนไลน์ 47
3-30 ขั้นตอน DyCluStrem+ ส่วนออฟไลน์ 48
3-31 ขั้นตอน DyCluStrem+ ส่วนออฟไลน์ (ExpandedProcess).....49
3-32 การผสานกลุ่มข้อมูลขนาดเล็ก.....50
3-33 การจัดกลุ่มแบบต่อยอด 52
3-34 ขั้นตอนวิธีการในการจัดกลุ่มข้อมูลแบบต่อยอด.....53
3-35 ขั้นตอนวิธีการในการแผ่ขยายกลุ่มข้อมูลแบบต่อยอด.....53

สารบัญภาพ (ต่อ)

ภาพที่ หน้า

4-1 คุณลักษณะข้อมูล S3 55
4-2 คุณลักษณะข้อมูล AbitaryHCM1, AbitaryHCM2 56
4-3 คุณลักษณะของข้อมูล t4.8k, t7.10k, t8.8k 57
4-4 ผลลัพธ์การจัดกลุ่มของข้อมูล S1 66
4-5 ผลลัพธ์การจัดกลุ่มของข้อมูล S2 67
4-6 ผลลัพธ์การจัดกลุ่มของข้อมูล S3 69
4-7 ผลลัพธ์การจัดกลุ่มของข้อมูล AbitaryHCM 71
4-8 ผลลัพธ์การจัดกลุ่มของข้อมูล AbitaryHCM2 73
4-9 ผลลัพธ์การจัดกลุ่มของข้อมูล t4.8k 75
4-10 ผลลัพธ์การจัดกลุ่มของข้อมูล t7.10k 76
4-11 ผลลัพธ์การจัดกลุ่มของข้อมูล t8.8k 78
4-12 แผนภูมิแท่งเปรียบเทียบผลการทดลองของตัววัด Purity สำหรับการปรับเปลี่ยนตัวแปร Stream Speed 82
4-13 แผนภูมิแท่งเปรียบเทียบผลการทดลองของตัววัด Rand statistic สำหรับการปรับเปลี่ยนตัวแปร Streamimg Speed 83
4-14 แผนภูมิแท่งเปรียบเทียบผลการทดลองของตัววัด Jaccard Index สำหรับการปรับเปลี่ยนตัวแปร Streaming Speed 84
4-15 แผนภูมิแท่งเปรียบเทียบผลลัพธ์ของกลุ่มข้อมูลขนาดเล็กที่ผ่านการจัดกลุ่ม สำหรับการปรับเปลี่ยนตัวแปร Streaming Speed 85
4-16 แผนภูมิแท่งเปรียบเทียบผลลัพธ์ของกลุ่มข้อมูลที่ผ่านการจัดกลุ่ม สำหรับการปรับเปลี่ยนตัวแปร Streaming Speed 87
4-17 แผนภูมิแท่งเปรียบเทียบหน่วยความจำที่ใช้งาน สำหรับการปรับเปลี่ยนตัวแปร Streaming Speed 88

สารบัญภาพ (ต่อ)

ภาพที่ หน้า

4-18 แผนภูมิแท่งเปรียบเทียบผลการทดลองของตัววัด Purity สำหรับการปรับเปลี่ยนตัวแปร Radius 90
4-19 แผนภูมิแท่งเปรียบเทียบผลการทดลองของตัววัด Rand statistic สำหรับการปรับเปลี่ยนตัวแปร Radius 91
4-20 แผนภูมิแท่งเปรียบเทียบผลการทดลองของตัววัด Jaccard Index สำหรับการปรับเปลี่ยนตัวแปร Radius 92
4-21 แผนภูมิแท่งเปรียบเทียบผลลัพธ์ของกลุ่มข้อมูลขนาดเล็กที่ผ่านการจัดกลุ่ม สำหรับการปรับเปลี่ยนตัวแปร Radius 93
4-22 แผนภูมิแท่งเปรียบเทียบผลลัพธ์ของกลุ่มข้อมูลที่ผ่านการจัดกลุ่ม สำหรับการปรับเปลี่ยนตัวแปร Radius 94
4-23 แผนภูมิแท่งเปรียบเทียบหน่วยความจำที่ใช้งาน สำหรับการปรับเปลี่ยนตัวแปร Radius 95
4-24 แผนภูมิกราฟเปรียบเทียบผลกระทบระหว่างหน่วยความจำที่ใช้งานกับความแม่นยำในแต่ละขั้นตอนวิธีการของชุดข้อมูล Statlog Shutter 96
4-25 แผนภูมิกราฟเปรียบเทียบผลกระทบระหว่างหน่วยความจำที่ใช้งานกับความแม่นยำในแต่ละขั้นตอนวิธีการของชุดข้อมูล Letter Recognition 97
4-26 แผนภูมิกราฟเปรียบเทียบผลกระทบระหว่างหน่วยความจำที่ใช้งานกับความแม่นยำในแต่ละขั้นตอนวิธีการของชุดข้อมูล KDD Cup 1999 97
4-27 ขั้นตอนวิธีการ DBSCAN 99
4-28 ขั้นตอนวิธีการ DenStream ส่วนออนไลน์ 100
4-29 ขั้นตอนวิธีการ HECES 102
4-30 ขั้นตอนวิธีการ DyCluStream ส่วนออนไลน์ 102
4-31 ขั้นตอนวิธีการ DyCluStream ส่วนออฟไลน์ 103
4-32 ขั้นตอนวิธีการ DyCluStream+ ส่วนออนไลน์ 105
4-33 ขั้นตอนวิธีการ DyCluStream+ ส่วนออฟไลน์ 106

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญปัญหา

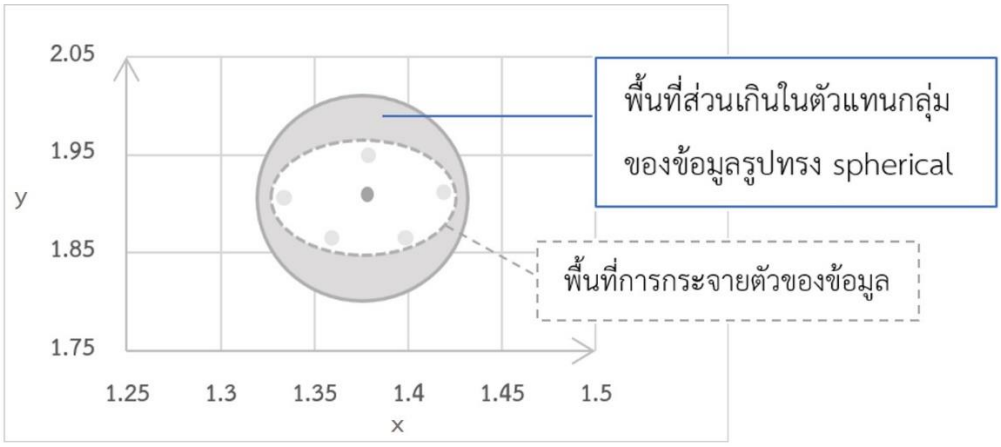
ความก้าวหน้าทางเทคโนโลยีทั้งด้านฮาร์ดแวร์และซอฟต์แวร์ส่งผลให้โปรแกรมหรือแอปพลิเคชันในปัจจุบันอย่างเช่น ระบบเฝ้าระวังแบบทันท่วงที การซื้อขายหุ้น การตรวจสอบการทุจริตบัตรเครดิต หรือแม้กระทั่งการวิเคราะห์ข้อมูลเครือข่ายทางสังคม เป็นต้น ต้องคำนึงถึงข้อมูลที่ต้องประมวลผลมากขึ้น เนื่องจากข้อมูลที่โปรแกรมหรือแอปพลิเคชันใช้ประมวลผลมักเป็นข้อมูลขนาดใหญ่ (Large-scale data) มีการเปลี่ยนแปลงอย่างรวดเร็วมาก และมีข้อมูลเกิดขึ้นใหม่ที่เป็นข้อมูลกระแสอย่างต่อเนื่องไม่มีที่สิ้นสุด อีกทั้งยังมีปริมาณที่มากเกินกว่าจะประมวลผลข้อมูลทั้งหมดในหน่วยความจำหลักได้ ซึ่งเราเรียกข้อมูลลักษณะเช่นนี้ว่า ข้อมูลสตรีมมิ่ง (Streaming data) หรือข้อมูลกระแส ในปัจจุบันหลายงานวิจัยที่เกี่ยวกับการทำเหมืองข้อมูลได้เห็นความสำคัญของข้อมูลสตรีมมิ่งมากขึ้น และการสกัดข้อมูลที่สำคัญหรือการค้นหาข้อมูลที่ซ่อนอยู่จากข้อมูลสตรีมมิ่งยังคงเป็นงานที่มีความท้าทาย (J. Han & M. Kamber, 2006; S.-S. Ho & H. Wechsler, 2010; Y. Li, D. Li, S. Wang, Y. Zhai, 2014)

เทคนิคหนึ่งในการสกัดหาองค์ความรู้จากข้อมูลเหล่านี้คือ การจัดกลุ่มข้อมูล โดยจะแบ่งชุดข้อมูลออกเป็นกลุ่ม ๆ โดยข้อมูลที่มีคุณลักษณะเหมือนกันหรือคล้ายกันจะถูกจัดไว้ในกลุ่มเดียวกัน และข้อมูลที่มีคุณลักษณะที่ต่างกันจะถูกจัดไว้ในกลุ่มที่ต่างกัน แต่อย่างไรก็ตามเทคนิคการจัดกลุ่มในลักษณะดังกล่าวไม่เหมาะสมที่จะนำมาใช้กับข้อมูลที่เป็นแบบกระแส ทั้งนี้เพราะลักษณะของข้อมูลสตรีมมิ่งที่ถูกผลิตขึ้นมาอย่างต่อเนื่องและรวดเร็ว และมีปริมาณของข้อมูลมากจนไม่สามารถเก็บไว้ในหน่วยความจำได้ ดังนั้นขั้นตอนวิธีในการจัดกลุ่มข้อมูลจะต้องมีการเก็บข้อมูลในรูปแบบทางคณิตศาสตร์และข้อมูลสรุปทางสถิติแทนที่การเก็บข้อมูลทั้งหมด เพื่อให้สามารถจัดเก็บและประมวลผลได้ในหน่วยความจำที่จำกัด และเมื่อมีข้อมูลใหม่เข้ามา ฟังก์ชันทางคณิตศาสตร์และสถิติจะถูกปรับปรุงได้ในเวลาอันรวดเร็ว

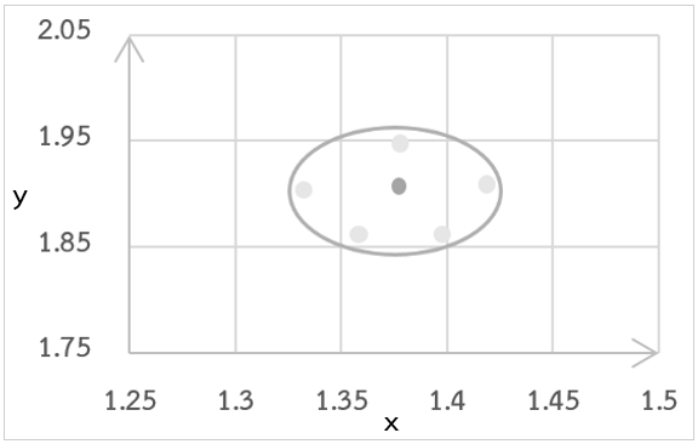
ปัจจุบันมีหลายงานวิจัยที่ศึกษาเกี่ยวกับการจัดกลุ่มของข้อมูลสตรีมมิ่งอย่างมากมาย โดยวิธีที่นิยมใช้คือ การนำวิธีแบบ k-means มาดัดแปลงให้สามารถแบ่งกลุ่มกับข้อมูลสตรีมมิ่งได้ ตัวอย่างเช่น ขั้นตอนวิธี StreamKM++ (M. R. Ackermann et al.,2012) ขั้นตอนวิธี CluStream (C. Aggarwal, J. Han, J. Wang, & P. Yu, 2003) ขั้นตอนวิธี Single-pass-k-means ขั้นตอนวิธี STREAM (S. Guha, A. Meyerson, N. Mishra, R. Motwani, & L. O'Callaghan, 2003) เป็นต้น ซึ่งวิธีการเหล่านี้ก็ยังมีข้อเสียที่เกิดจากวิธีจัดกลุ่มแบบ k-means คือต้องมีการรู้จำนวนกลุ่มของข้อมูล

ล่วงหน้า ไม่เหมาะสมกับข้อมูลที่มีการกระจายตัวและความหนาแน่นที่แตกต่างกัน และผลการจัดกลุ่มจะจัดกลุ่มได้ดีสำหรับข้อมูลที่มีลักษณะรูปร่างหรือการกระจายตัวของข้อมูลเป็น spherical เนื่องจากใช้ระยะทางแบบยูคลิดในการวัดความเหมือนกันของข้อมูลที่จะพิจารณาข้อมูลที่อยู่ในกลุ่มเดียวกันจากข้อมูลและค่าเฉลี่ยของกลุ่มข้อมูลเท่านั้น จึงส่งผลให้กลุ่มข้อมูลที่ถูกแบ่งด้วยวิธีการแบบ k-means ไม่สามารถมีรูปร่างตามลักษณะการกระจายตัวของข้อมูลได้

ขั้นตอนวิธีการจัดกลุ่มที่นิยมอีกวิธีหนึ่งคือ การนำวิธีการแบบ DBSCAN มาประยุกต์ใช้ ซึ่งเป็นวิธีการจัดกลุ่มที่พิจารณาจากความหนาแน่นของกลุ่มข้อมูล อาทิเช่น DenStream (Cao, Feng, Martin Ester, Weining Qian, & Aoying Zhou, 2006) โดย DenStream ประกอบไปด้วยการทำงานสองส่วนคือ ส่วนออนไลน์และออฟไลน์ โดยส่วนออนไลน์ทำหน้าที่จัดกลุ่มข้อมูลที่เข้ามาให้อยู่ในกลุ่มของข้อมูลขนาดเล็กเรียกว่า micro-clusters ซึ่งเป็นโครงสร้างสำหรับเก็บข้อมูลเชิงสถิติ ที่ประกอบไปด้วย ค่าเฉลี่ย ความแปรปรวน และจำนวนของข้อมูล มีลักษณะโครงสร้างเป็น spherical โดยกระบวนการในการจัดกลุ่มส่วนนี้จะเรียกว่า micro-clustering ซึ่งเมื่อข้อมูลที่เข้ามาถูกจัดให้อยู่ใน micro-cluster ใด ๆ เพื่อเก็บข้อมูลเชิงสถิติแล้ว หลังจากนั้นข้อมูลที่เข้ามาจะถูกนำออกไปจากหน่วยความจำเพื่อเป็นการประหยัดพื้นที่ในหน่วยความจำ และในส่วนของการจัดกลุ่มสำหรับนำไปใช้งานจะอยู่ในส่วนของออฟไลน์ ซึ่งจะถูกเรียกใช้เมื่อมีความต้องการนำไปใช้งาน โดยกลุ่มข้อมูลจะได้จากการเชื่อมกันด้วยรูปแบบการซ้อนทับกัน (Intersecting) ของ micro-clusters ดังภาพที่ 1-6 ก) จึงทำให้รูปร่างของกลุ่มข้อมูลสามารถเป็นรูปแบบใด ๆ ก็ได้ ดังภาพที่ 1-6 ข) อีกทั้งกระบวนการในการจัดกลุ่มข้อมูลสามารถดำเนินการได้โดยไม่จำเป็นต้องรู้จำนวนกลุ่มของข้อมูลมาก่อน แต่อย่างไรก็ตามการแทนรูปร่างกลุ่มข้อมูลขนาดเล็กเป็น spherical ตัวอย่างดังภาพที่ 1-1 นั้นยังคงไม่เหมาะสมกับลักษณะการกระจายตัวของข้อมูล เนื่องจากกลุ่มข้อมูลขนาดเล็กรูปทรง spherical ครอบคลุมพื้นที่ส่วนเกิน (บริเวณสีเทา) ที่เกินจากการกระจายตัวของข้อมูล

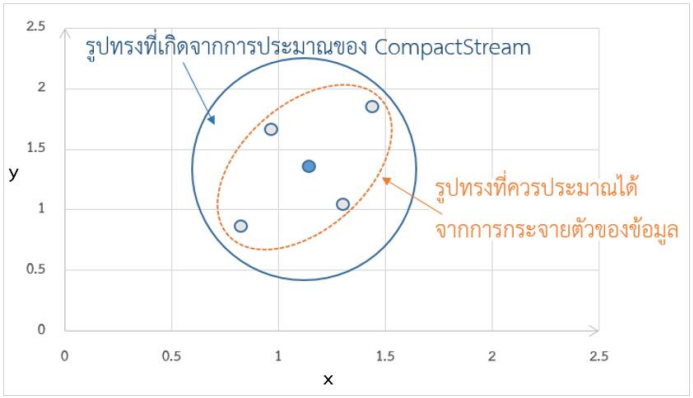


ภาพที่ 1-1 การแทนรูปร่างกลุ่มของข้อมูลขนาดเล็กเป็น spherical



ภาพที่ 1-2 รูปร่าง micro-cluster ที่มีลักษณะเป็นแบบ hyper-elliptical ของ CompactStream

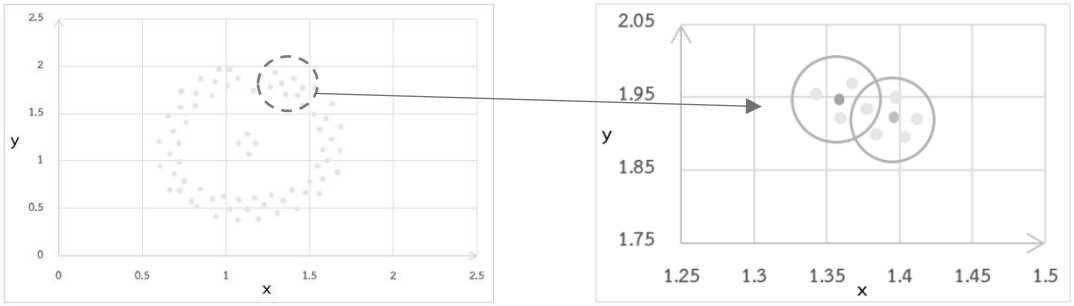
CompactStream (N. Wattanakitrungroj & C. Lursinsap, 2011) เป็นขั้นตอนวิธีการที่นำเสนอแนวคิดในแทนกลุ่มข้อมูลขนาดเล็กด้วยรูปทรง hyper-elliptical ดังภาพที่ 1-2 ที่เหมาะสมกับรูปแบบการกระจายตัวของข้อมูลที่มีความหลากหลายมากกว่า และในส่วนของการจัดกลุ่มสำหรับนำไปใช้งาน วิธีการนี้ได้ใช้การเชื่อมกันด้วยรูปแบบการซ้อนทับกัน (Intersecting) ของกลุ่มข้อมูลขนาดเล็กที่มีลักษณะเป็นรูปทรง hyper-elliptical จึงทำให้รูปร่างของกลุ่มข้อมูลสำหรับนำไปใช้งานสามารถเป็นรูปแบบใด ๆ ก็ได้ ดังภาพที่ 1-6 ข) แต่อย่างไรก็ตาม โครงสร้างของกลุ่มข้อมูลขนาดเล็กที่งานวิจัยนี้ได้นำเสนอนั้นไม่ได้เกิดจาก Covariance matrix ของข้อมูลที่อยู่ในกลุ่มข้อมูลขนาดเล็กทั้งหมด เพราะในช่วงแรกที่กลุ่มข้อมูลขนาดเล็กยังมีความหนาแน่นไม่มากพอ โดยยังเป็นรูปทรง spherical อยู่ ไม่ได้เก็บค่า Covariance matrix ที่บอกถึงลักษณะการกระจายตัวของข้อมูล แต่เกิดจากการกำหนดด้วยค่าคงที่ดังภาพที่ 1-3 เพื่อหลีกเลี่ยงการเกิดปัญหา Not Positive Definite ที่มีผลกระทบกับการคำนวณ Covariance matrix จึงทำให้รูปทรง hyper-elliptical ที่เกิดจาก spherical มีรูปร่างคลาดเคลื่อนจากความเป็นจริง



ภาพที่ 1-3 การกำหนดรูปทรงของกลุ่มข้อมูลขนาดเล็กจากขั้นตอนวิธีการ CompactStream

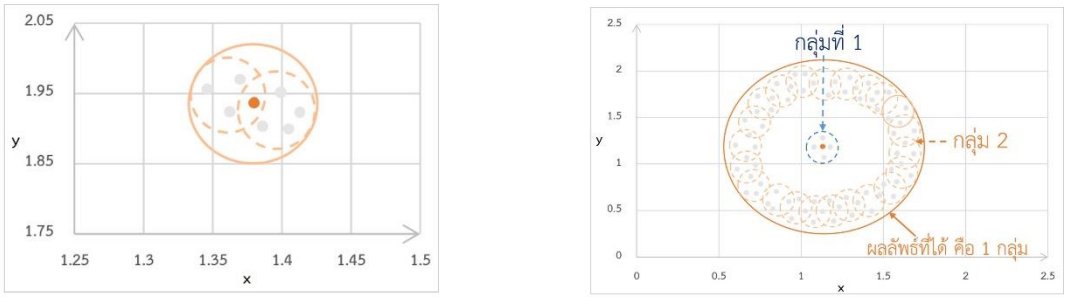
HECES (M.Z. R., T. Li, Y. Yang, & H. Wang, 2014) เป็นขั้นตอนวิธีการที่นำเสนอแนวคิดเช่นเดียวกับงานวิจัย CompactStream ในเรื่องการแทนกลุ่มข้อมูลด้วยโครงสร้าง hyper-ellipsoid

ที่เหมาะสมกับรูปแบบการกระจายตัวของข้อมูลที่มีความหลากหลายมากกว่า spherical แต่อย่างไรก็ตาม ขั้นตอนวิธีการนี้มีการรวมกลุ่มข้อมูล (Merging) ตัวอย่างดังภาพที่ 1-4 ข) ที่สามารถรวมกันได้และแทนด้วยกลุ่มข้อมูลใหม่ทุกครั้งเมื่อจัดกลุ่ม ซึ่งใช้ค่าเฉลี่ยของกลุ่มที่สามารถรวมกันได้เป็นจุดศูนย์กลาง ดังภาพที่ 1-5 ก) จึงทำให้เกิดกรณีของกลุ่มข้อมูลที่มีขนาดเล็กกว่าอยู่ในกลุ่มข้อมูลที่มีขนาดใหญ่กว่าแล้วได้ผลลัพธ์ที่มีเพียงหนึ่งกลุ่มข้อมูล ดังภาพที่ 1-5 ข) จึงไม่เหมาะสมกับข้อมูลที่มีรูปร่างไม่แน่นอน เมื่อเปรียบเทียบกับวิธีการเชื่อมกันด้วยรูปแบบการซ้อนทับกันของกลุ่มข้อมูลดังภาพที่ 1-6 ก) ที่ให้ผลลัพธ์ในการจัดกลุ่มของข้อมูลเป็นรูปแบบใด ๆ ก็ได้ ดังภาพที่ 1-6 ข)



ก) ข้อมูลทั้งหมด ข) กลุ่มข้อมูล

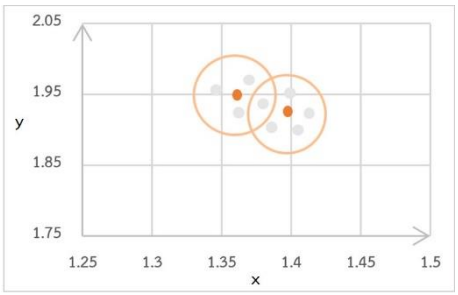
ภาพที่ 1-4 กลุ่มข้อมูลที่ได้จากข้อมูลบางส่วน



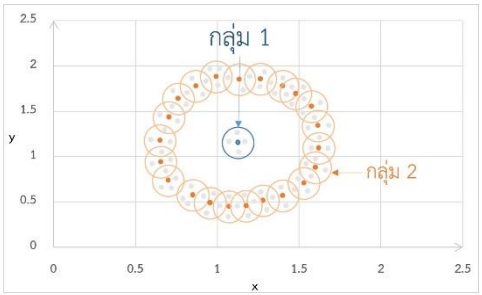
ก) การจัดกลุ่มที่เกิดจากการรวมกันและแทนด้วยกลุ่มข้อมูลใหม่

ข) ผลลัพธ์การจัดกลุ่มที่เกิดจากการรวมกันและแทนที่ด้วยกลุ่มข้อมูลใหม่

ภาพที่ 1-5 ผลลัพธ์การจัดกลุ่มที่เกิดจากการรวมกันและแทนที่ด้วยกลุ่มข้อมูลใหม่ของวิธีการ HECES และ eVQ-AMS



ก) การจัดกลุ่มที่เกิดจากการเชื่อมกันด้วยรูปแบบการซ้อนทับกันของกลุ่มข้อมูลขนาดเล็ก



ข) ผลลัพธ์การจัดกลุ่มที่เกิดจากการซ้อนทับกันของกลุ่มข้อมูลขนาดเล็ก

ภาพที่ 1-6 ผลลัพธ์การจัดกลุ่มที่เกิดจากการซ้อนทับกันของกลุ่มข้อมูลขนาดเล็กของวิธีการ DenStream, CompactStream, DBSTREAM และ VHEC

eVQ-AMS (E. Lughofer & M. S. Mouchaweh, 2015) เป็นขั้นตอนวิธีการที่ได้นำเสนอแนวคิดในการแทนกลุ่มของข้อมูลเป็นรูป ellipsoids และมีขั้นตอนวิธีการที่มีส่วนออนไลน์เพียงอย่างเดียว คือเมื่อมีข้อมูลเข้ามา จะนำข้อมูลไปปรับปรุงกับกลุ่มข้อมูลที่ใกล้สุด หลังจากนั้นจึงดำเนินการตรวจสอบกับกลุ่มข้อมูลรอบข้างว่าสามารถรวมกลุ่มได้หรือไม่ คล้ายกับ HECES และเพิ่มส่วนของการตรวจสอบว่าสามารถแยกกลุ่มข้อมูลได้หรือไม่เมื่อกลุ่มข้อมูลมีการเปลี่ยนแปลง แต่อย่างไรก็ตามการรวมกลุ่มที่คล้ายกับ HECES ดังภาพที่ 1-5 ก) จึงยังทำให้เกิดปัญหากับข้อมูลที่มีรูปร่างไม่แน่นอน ดังภาพที่ 1-5 ข) และขั้นตอนวิธีการนี้จำเป็นต้องเก็บข้อมูลตัวอย่างสำหรับพิจารณาว่าข้อมูลควรรวมกลุ่มหรือแยกกลุ่มออกจากกัน ทำให้มีปัญหาในเรื่องของหน่วยความจำ

DBSTREAM (M. Hahsler & M. Bolaños, 2016) ได้นำเสนอแนวคิดในการจัดกลุ่มโดยแบ่งออกเป็นสองส่วนคือ ส่วนออนไลน์ทำหน้าที่จัดกลุ่มข้อมูลที่เข้ามาให้อยู่ในกลุ่มข้อมูลขนาดเล็กที่เรียกว่า hyper-sphere และส่วนออฟไลน์ที่จะเชื่อมกลุ่มข้อมูลขนาดเล็กด้วยรูปแบบการซ้อนทับกันโดยจะพิจารณาถึงความหนาแน่นที่ถูกเก็บในรูปแบบของกราฟกลุ่มข้อมูลขนาดเล็กที่เป็น hyper-sphere แต่อย่างไรก็ตามขั้นตอนวิธีการนี้ยังคงแทนกลุ่มข้อมูลขนาดเล็กด้วยรูปแบบ hypers-phere ซึ่งไม่เหมาะสมกับลักษณะการกระจายตัวของกลุ่มข้อมูลและให้ความสำคัญกับข้อมูลในแต่ละช่วงเวลาที่ไม่เท่ากัน

VHEC (N. Wattanakitrungroj, S. Maneeroj & C. Lursinsap, 2017) ได้นำเสนอแนวคิดในการแทนกลุ่มข้อมูลขนาดเล็กด้วย versatile hyper-ellipsoidal ซึ่งเป็นรูปทรงที่มีลักษณะใกล้เคียงกับการกระจายตัวของข้อมูลมากกว่า spherical และในส่วนของการจัดกลุ่มสำหรับนำไปใช้งาน วิธีการนี้ได้ใช้การเชื่อมกันด้วยรูปแบบการซ้อนทับกัน (Intersecting) ของกลุ่มข้อมูลขนาดเล็กที่มีลักษณะเป็นรูปทรง versatile hyper-ellipsoidal จึงทำให้รูปร่างของกลุ่มข้อมูลสำหรับนำไปใช้งานสามารถเป็นรูปแบบใด ๆ ก็ได้ ดังภาพที่ 1-6 ข) นอกจากนี้ยังได้เพิ่มส่วนของการผสานกลุ่มข้อมูลขนาดเล็กที่มีลักษณะใกล้เคียงกัน เพื่อลดหน่วยความจำที่ใช้ในการจัดเก็บข้อมูล แต่อย่างไรก็ตามโครงสร้างกลุ่มข้อมูลขนาดเล็กที่เป็น versatile hyper-ellipsoidal ที่งานวิจัยนี้ได้นำเสนอในส่วนของลักษณะของ versatile hyper-ellipsoidal นั้น มีลักษณะคล้ายกับ hyper-elliptical ของ

CompactStream คือไม่ได้เกิดจาก Covariance matrix ของข้อมูลที่อยู่ในกลุ่มข้อมูลขนาดเล็กทั้งหมด จึงทำให้รูปร่างกลุ่มข้อมูลขนาดเล็กคลาดเคลื่อนจากความเป็นจริง

ขั้นตอนวิธีในการจัดกลุ่มข้อมูลสตรีมมิ่งที่มีอยู่นั้นส่วนใหญ่มุ่งที่จะตรวจจับความเปลี่ยนแปลงของกลุ่มข้อมูลขนาดเล็ก ซึ่งมีวิธีการที่หลากหลายในการบ่งบอกหรือจัดการกับการเปลี่ยนแปลงของกลุ่มข้อมูลขนาดเล็ก โดยใช้ Landmark Window model, Sliding Window Model, Fading Window Model, Damped Window model (A. Amini & Y. Teh, H. Saboohi, 2013) เป็นต้น วิธีการเหล่านี้จะมีการกำหนดค่าน้ำหนักที่ขึ้นอยู่กับเวลาให้กับข้อมูลที่เข้ามา ก่อนจะถูกนำไปใช้ในกระบวนการในการจัดกลุ่มกับกลุ่มข้อมูลขนาดเล็กต่อไป ดังนั้นกลุ่มข้อมูลขนาดเล็กที่ไม่มีข้อมูลใหม่เข้ามา จะถูกลดค่าน้ำหนักไปเรื่อย ๆ จนกระทั่งถูกนำออกไปจากหน่วยความจำด้วยค่าน้ำหนักที่ลดลง ซึ่งเกิดจากฟังก์ชันที่เกี่ยวข้องกับเวลา โดย DenStream ได้มีการใช้ Damped Window Model ซึ่งจะมีการใช้ฟังก์ชันที่เป็นเอกซ์โพเนนเชียลเป็นค่าถ่วงน้ำหนักกับทุกกลุ่มข้อมูลขนาดเล็ก ซึ่งกลุ่มข้อมูลขนาดเล็กที่มีข้อมูลเข้ามาจะถูกปรับปรุงให้มีค่าน้ำหนักมากที่สุด ส่วนกลุ่มข้อมูลขนาดเล็กที่ไม่มีข้อมูลเข้าจะถูกลดค่าน้ำหนักลงไปเรื่อย ๆ จนหายไปในที่สุด ซึ่งแสดงให้เห็นว่ามีผลกระทบกับทุกข้อมูลที่อยู่ในกลุ่มข้อมูลขนาดเล็กนั้น แต่คุณลักษณะของข้อมูลในงานด้านอื่น ๆ นั้นขั้นตอนวิธีในการจัดกลุ่มไม่ควรมุ่งเน้นไปยังการตรวจจับความเปลี่ยนแปลงรูปร่างลักษณะของกลุ่มข้อมูลขนาดเล็ก หรือการลดความสำคัญของข้อมูลก่อนหน้าลง ยกตัวอย่างเช่นข้อมูลบัญชีธนาคาร ที่ข้อมูลเก่านั้นไม่ควรถูกมองว่ามีความสำคัญน้อยกว่าข้อมูลใหม่ เพราะทุก ๆ ข้อมูลของลูกค้าที่เปิดบัญชีอยู่ควรมีความสำคัญเท่ากันหมด ดังนั้นจึงต้องการขั้นตอนวิธีการจัดกลุ่มที่ไม่ได้ลดความสำคัญของกลุ่มข้อมูลขนาดเล็กลงแม้เวลาเปลี่ยนแปลงไปก็ตาม

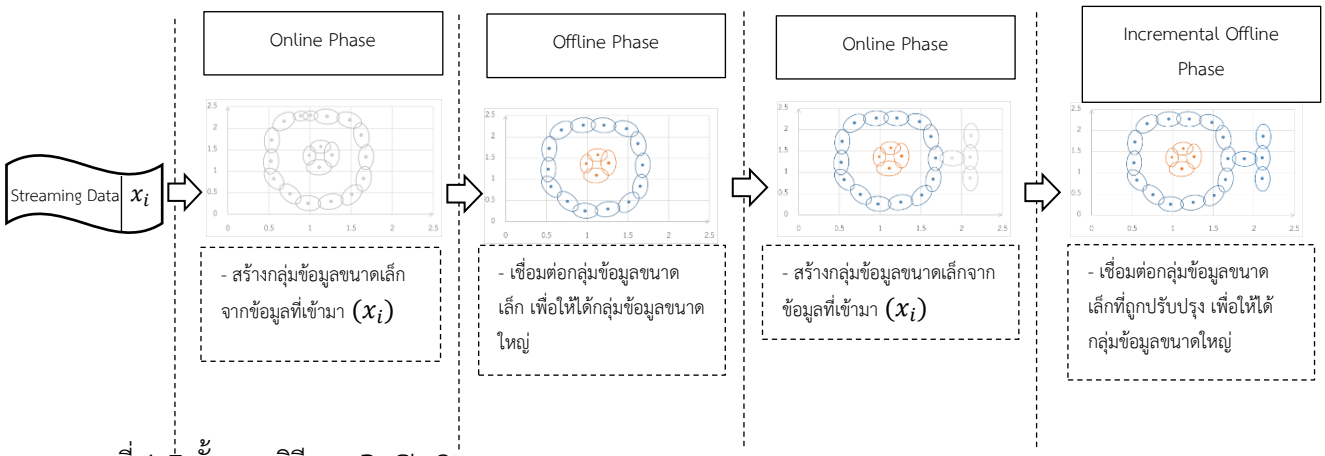
จากงานวิจัยที่กล่าวมาข้างต้น สามารถสรุปเป็นประเด็นปัญหาดังนี้

- โครงสร้างกลุ่มข้อมูลขนาดเล็กที่มีลักษณะเป็น spherical ไม่เหมาะสมกับลักษณะการกระจายตัวของข้อมูลที่มีรูปร่างไม่แน่นอน และโครงสร้างรูปทรง hyper-ellipsoidal ที่นำเสนอโดยงานวิจัยที่กล่าวถึงยังจำเป็นต้องเก็บข้อมูลตัวอย่างเพื่อใช้ในการสร้างกลุ่มข้อมูลขนาดเล็ก หรือโครงสร้างที่เป็นรูปทรง hyper-ellipsoidal ไม่ได้เกิดจาก Covariance matrix ของข้อมูลทั้งหมดที่อยู่ในกลุ่มข้อมูลขนาดเล็ก แต่เกิดจากการกำหนดค่าคงที่ ซึ่งส่งผลให้ผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูลมีความคลาดเคลื่อน ดังภาพที่ 1-3
- การจัดการกับการเปลี่ยนแปลงของกลุ่มข้อมูลขนาดเล็ก ที่เกิดจากการใช้ Landmark Window model, Sliding Window Model, Fading Window Model ซึ่งมีผลทำให้ลดความสำคัญของกลุ่มข้อมูลขนาดเล็ก เมื่อเวลาเปลี่ยนแปลงไป แต่ในกรณีของข้อมูลบัญชีธนาคาร ลูกค้าทุกคนควรมีความสำคัญเท่ากันแม้เวลาจะเปลี่ยนแปลงไปก็ตาม แต่การใช้โมเดลที่ลดความสำคัญของข้อมูลเก่าส่งผลให้ลูกค้าทุกคนมีความสำคัญไม่เท่ากันเมื่อเวลาเปลี่ยนแปลงไป จึงเกิดผลกระทบต่อกลุ่มข้อมูลขนาดเล็กไม่เคยมีลูกค้าใหม่เข้ามา
- วิธีการรวมกลุ่มข้อมูลขนาดเล็กที่สามารถรวมกันได้ แล้วแทนที่ด้วยกลุ่มข้อมูลขนาดเล็กใหม่ โดยใช้ค่าเฉลี่ยของกลุ่มข้อมูลขนาดเล็กที่สามารถรวมกันได้เป็นจุดศูนย์กลางเพื่อทำให้ได้กลุ่มข้อมูลขนาดใหญ่สำหรับนำไปใช้งาน มีผลทำให้เกิดกรณีของกลุ่มข้อมูลที่มีขนาดเล็กกว่าอยู่ในกลุ่ม

ข้อมูลที่มีขนาดใหญ่กว่าแล้วได้ผลลัพธ์ที่มีเพียงหนึ่งกลุ่มข้อมูล ซึ่งส่งผลให้ไม่สามารถจัดการกับกลุ่มของข้อมูลที่มีรูปร่างไม่แน่นอนได้อย่างถูกต้อง ดังภาพที่ 1-5

1.2 แนวทางการแก้ปัญหา

จากปัญหาดังกล่าวข้างต้นงานวิจัยนี้จึงได้นำเสนอขั้นตอนวิธีการจัดกลุ่มข้อมูลที่มีความยืดหยุ่นสำหรับข้อมูลสตรีมมิ่ง โดยใช้เพียงข้อมูลที่เข้ามาใหม่ร่วมกับโครงสร้างกลุ่มย่อยไฮเพอร์อัลลิปโซดอยล์พลัส ที่เรียกว่า **The Dynamic Clustering Algorithm for Streaming Data Using One-Pass-Throw-Away Datum and Hyper-ellipsoidal Micro-Clustering Plus** หรือเรียกว่า **DyCluStream+** ดังภาพที่ 1-7 ซึ่งเป็นขั้นตอนวิธีการที่แบ่งการทำงานออกเป็นสองส่วน คือ ส่วนออนไลน์สำหรับเก็บข้อมูลที่เข้ามาให้อยู่ในรูปแบบทางคณิตศาสตร์และข้อมูลสรุปในเชิงสถิติที่ใช้เป็นกลุ่มข้อมูลขนาดเล็ก และส่วนออฟไลน์ทำหน้าที่ในการจัดกลุ่มจากกลุ่มข้อมูลขนาดเล็ก โดยมีรายละเอียดดังนี้



ภาพที่ 1-7 ขั้นตอนวิธีการ DyCluStream+

- **ขั้นตอนออนไลน์ (Online Phase)** เป็นขั้นตอนสำหรับเก็บสถิติของข้อมูลนำเข้าที่นำเข้าทีละ 1 ตัวอย่าง ให้อยู่ในรูปกลุ่มข้อมูลขนาดเล็กที่มีรูปทรง spherical โดยจะมีการเก็บลักษณะการกระจายตัวของข้อมูลที่อยู่ในกลุ่ม หลังจากกลุ่มข้อมูลขนาดเล็กรูปทรง spherical มีความหนาแน่นมากพอ จะดำเนินการเปลี่ยนเป็นรูปทรง dynamic hyper-ellipsoidal และพิจารณาเก็บข้อมูลเชิงสถิติกับ dynamic hyper-ellipsoidal เป็นลำดับแรก ซึ่งทุกครั้งที่ข้อมูลเข้ามาเมื่อถูกดำเนินการเก็บข้อมูลเชิงสถิติให้อยู่ในรูปแบบทางคณิตศาสตร์และข้อสรุปเชิงสถิติแล้ว ข้อมูลที่เข้ามาก็จะถูกนำออกไปจากระบบแล้วดำเนินการรับข้อมูลเข้ามาใหม่
- **ขั้นตอนออฟไลน์ (Offline Phase)** เป็นขั้นตอนสำหรับการเชื่อมต่อกลุ่มข้อมูลขนาดเล็กรูปทรง dynamic hyper-ellipsoidal เพื่อให้ได้กลุ่มข้อมูลขนาดใหญ่ และจะรวมกลุ่มข้อมูลขนาดเล็กสองกลุ่มที่มีคุณลักษณะเหมือนกันเข้าด้วยกันในระหว่างการเชื่อมต่อ โดยเริ่มต้นจากกลุ่มข้อมูลขนาดเล็กรูปทรง dynamic hyper-ellipsoidal ที่มีความหนาแน่นมากที่สุดเป็นตัวตั้งต้นในการแผ่ขยายเพื่อเชื่อมต่อกับกลุ่มข้อมูลขนาดเล็กรูปทรง dynamic hyper-ellipsoidal กลุ่มอื่น โดยถ้าสามารถแผ่ขยายไปยังกลุ่มข้อมูลขนาดเล็กรูปทรง dynamic hyper-ellipsoidal กลุ่มอื่นได้ ก็จะนำ

กลุ่มข้อมูลขนาดเล็กรูปทรง dynamic hyper-ellipsoidal ที่ถูกแผ่ขยายเป็นตัวดำเนินแผ่ขยายไปยังกลุ่มข้อมูลขนาดเล็กรูปทรง dynamic hyper-ellipsoidal กลุ่มอื่นต่อไป ซึ่งถ้าไม่สามารถแผ่ขยายไปยังกลุ่มข้อมูลขนาดเล็กรูปทรง dynamic hyper-ellipsoidal ใด ๆ ได้อีก ก็จะหากลุ่มข้อมูลขนาดเล็กรูปทรง dynamic hyper-ellipsoidal ที่ยังไม่ถูกแผ่ขยายเป็นตัวดำเนินการแผ่ขยาย โดยในระหว่างการแผ่ขยาย จะมีการตรวจสอบด้วยว่าถ้ามีกลุ่มข้อมูลขนาดเล็กสองกลุ่มที่มีคุณลักษณะใกล้เคียงกันมากก็จะรวมกลุ่มข้อมูลขนาดเล็กสองกลุ่ม เพื่อประหยัดหน่วยความจำ

- **ขั้นตอนออฟไลน์แบบต่อยอด (Incremental Offline Phase)** เป็นขั้นตอนสำหรับการเชื่อมต่อกลุ่มข้อมูลขนาดเล็กรูปทรง dynamic hyper-ellipsoidal เพื่อให้ได้กลุ่มข้อมูลขนาดใหญ่ที่เหมือนกับ **ขั้นตอนออฟไลน์ (Offline Phase)** แต่จะเชื่อมต่อกลุ่มข้อมูลขนาดเล็กรูปทรง dynamic hyper-ellipsoidal ที่ได้รับการปรับปรุงจากขั้นตอนในส่วนออนไลน์ เพื่อให้ได้กลุ่มข้อมูลขนาดใหญ่

จากวิธีการข้างต้นทำให้ **DyCluStream+** มีความสามารถดังนี้

- เป็นขั้นตอนวิธีการที่ใช้ dynamic hyper-ellipsoidal micro-clusters เป็นกลุ่มข้อมูลขนาดเล็ก ซึ่งเป็นรูปทรงที่เข้ากับรูปร่างของข้อมูลมากกว่า spherical โดยข้อมูลจะถูกนำเข้าสู่หน่วยความจำและเก็บข้อมูลทางสถิติ หลังจากนั้นจะถูกนำออกไปจากหน่วยความจำซึ่งช่วยลดหน่วยความจำที่ใช้ในการเก็บข้อมูล
- โครงสร้าง dynamic hyper-ellipsoidal micro-clusters ที่เป็นกลุ่มข้อมูลขนาดเล็กจะถูกคำนวณมาจากข้อมูลที่นำเข้าเท่านั้น ซึ่งแตกต่างจากโครงสร้างของงานวิจัยอื่นที่เกิดจากการประมาณ ทำให้โครงสร้างที่ได้มีความถูกต้องแม่นยำ
- เป็นขั้นตอนวิธีการที่ไม่ลดความสำคัญของข้อมูลเมื่อเวลาผ่านไป
- เป็นขั้นตอนวิธีที่จะใช้วิธีการเชื่อมต่อกันไปเรื่อย ๆ ของการซ้อนทับกันของกลุ่มข้อมูลขนาดเล็ก เพื่อให้ได้กลุ่มข้อมูลสำหรับนำไปใช้งาน
- เป็นขั้นตอนวิธีการที่ไม่จำเป็นต้องเริ่มจัดกลุ่มข้อมูลขนาดเล็กใหม่ทั้งหมด เมื่อเคยมีการจัดกลุ่มมาแล้ว โดยจะจัดกลุ่มข้อมูลขนาดเล็กที่ถูกปรับปรุงจากส่วนออนไลน์เท่านั้น

1.3 วัตถุประสงค์ของงานวิจัย

1.3.1 เพื่อนำเสนอขั้นตอนวิธีแบ่งกลุ่มแบบใหม่สำหรับข้อมูลสตรีมมิ่งที่ไม่สามารถทราบลักษณะการกระจายตัวและจำนวนกลุ่มได้ล่วงหน้า

1.3.2 เพื่อนำเสนอโครงสร้างที่มีความยืดหยุ่นสำหรับเก็บข้อสรุปแต่ละกลุ่มข้อมูล โดยโครงสร้างดังกล่าวจะใช้เพียงข้อมูลที่เข้ามาใหม่เพื่อใช้ในการปรับปรุงโครงสร้างทางสถิติ

1.4 ขอบเขตของงานวิจัย

1.4.1 ข้อมูลที่ใช้ในงานวิจัยเป็นข้อมูลสังเคราะห์ขึ้นเองและข้อมูลจากปัญหาจริง โดยข้อมูลจะถูกนำเข้าทีละหนึ่งตัวอย่างอย่างต่อเนื่อง ไม่สามารถทราบการกระจายตัวหรือความหนาแน่นของข้อมูลและจำนวนกลุ่มล่วงหน้า

1.4.2 ข้อมูลแต่ละตัวอย่างจะถูกนำเข้าสู่หน่วยความจำเพียงครั้งเดียว เพื่อนำมาสร้างโครงสร้างสำหรับเก็บข้อมูลทางสถิติ จากนั้นข้อมูลที่เข้ามาก็จะถูกนำออกไปจากหน่วยความจำ

1.4.3 การทดลองและการประเมินผลในงานวิจัยนี้ได้เปรียบเทียบวิธีการ DyCluStream+, DyCluStream กับวิธีการ DenStream และวิธีการ HECES สำหรับการออกแบบการทดลองเพื่อเปรียบเทียบประสิทธิภาพได้แบ่งเป็นสองส่วนคือ ส่วนแรกใช้ข้อมูลสังเคราะห์โดยการวัดจากความสามารถในการแบ่งกลุ่มของข้อมูลที่ดีที่สุด และส่วนที่สองใช้ข้อมูลจริงซึ่งจะทดสอบทั้งความสามารถในการจัดการกับข้อมูลกระแสที่นำเข้าและการมีผลของระยะทางที่ใช้ในการจัดกลุ่มข้อมูล โดยงานวิจัยนี้ได้ใช้วิธีการวัดประสิทธิภาพความถูกต้องคือ Purity , Rand Statistic และ Jaccard Index รวมทั้งการวัดประสิทธิภาพเชิงเวลา

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1.5.1 ได้ขั้นตอนวิธีที่สามารถนำไปประยุกต์ใช้กับงานที่ข้อมูลทั้งหมดมีความสำคัญเท่ากัน

1.5.2 ได้ขั้นตอนวิธีจัดกลุ่มข้อมูลที่เป็นแบบกระแส ไม่สามารถทราบการกระจายตัวของข้อมูลจำนวนกลุ่มได้ล่วงหน้า

1.5.3 ได้ขั้นตอนวิธีจัดกลุ่มที่มีความถูกต้องในการจัดกลุ่ม โดยใช้เวลาและหน่วยความจำที่น้อยลง

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้องที่นำมาใช้ในการพัฒนาขั้นตอนวิธีที่นำเสนอในงานวิจัย โดยจะแบ่งเป็นส่วนต่าง ๆ ดังนี้

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 ลักษณะของข้อมูลสตรีมมิ่งหรือข้อมูลกระแส (Streaming Data)

2.1.2 การจัดกลุ่มกับข้อมูลกระแส (Stream Clustering)

2.1.3 Recursive Computation

2.1.4 Cosine Similarity

2.1.5 Bhatthacharyya Distance

2.1.6 Principal Component Analysis

2.2 งานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 ลักษณะของข้อมูลสตรีมมิ่งหรือข้อมูลกระแส (Streaming Data)

ข้อมูลกระแส คือ ข้อมูลที่ถูกผลิตขึ้นมาจากโปรแกรมหรือแอปพลิเคชันต่าง ๆ อย่างต่อเนื่อง มีจำนวนไม่สิ้นสุด จึงทำให้ไม่สามารถประมวลผลในหน่วยความจำหลัก (Main memory) ได้และเมื่อเวลาผ่านไปจะมีการเปลี่ยนแปลงลักษณะการกระจายตัวหรือความหนาแน่นของข้อมูล ตัวอย่างเช่น ข้อมูลเครือข่ายทางสังคม ข้อมูลการซื้อขายหุ้น และข้อมูลจากอุปกรณ์ที่ใช้ในการรับรู้การเคลื่อนที่ของโทรศัพท์

นิยามที่ 2.1 ข้อมูลกระแส นิยามได้โดย Streaming Data D คือข้อมูลที่มีลำดับตามเวลา $x_1, x_2, \dots, x_n$ นั่นคือ $D = \{x_i\}_{i=1}^{n}$ ที่มีจำนวนมากไม่สิ้นสุด $(n \to \infty)$ ซึ่งแต่ละ x_i มีมิติ (Dimension) ของข้อมูลเป็น $[x_{i1}, x_{i2}, \dots, x_{id}]$ ทั้งหมด d มิติ นั่นคือ $x_i = [x_{ij}]_{j=1}^{d}$

2.1.2 การจัดกลุ่มกับข้อมูลกระแส (Stream Clustering)

การจัดกลุ่มข้อมูลที่ดำเนินการกับข้อมูลกระแสนั้นต้องคำนึงถึงข้อจำกัดในด้านหน่วยความจำหลักที่จำกัด และความรวดเร็วในการประมวลผลของข้อมูลกระแสที่เข้ามา เนื่องจากข้อมูลถูกผลิตขึ้นมาอย่างต่อเนื่องและไม่มีที่สิ้นสุด ดังนั้นการจัดกลุ่มของข้อมูลที่เป็นแบบกระแสนี้จึงนิยมเก็บข้อมูล

ที่เข้ามาให้อยู่ในรูปแบบทางคณิตศาสตร์หรือข้อมูลสรุปในเชิงสถิติ โดยการแทนด้วยกลุ่มข้อมูลขนาดเล็กเพื่อลดหน่วยความจำที่ใช้สำหรับเก็บข้อมูลทั้งหมด และในส่วนของการจัดกลุ่มข้อมูลจะใช้กลุ่มข้อมูลขนาดเล็กจัดกลุ่ม เพื่อเพิ่มความเร็วในการจัดกลุ่มข้อมูลเพราะไม่ต้องประมวลผลจากข้อมูลทั้งหมด ดังนั้นการจัดกลุ่มกับข้อมูลกระแสคือการจัดกลุ่มกับกลุ่มข้อมูลขนาดเล็กที่มีลักษณะที่คล้ายกันให้อยู่ในกลุ่มเดียวกัน และกลุ่มข้อมูลขนาดเล็กที่มีคุณลักษณะที่อยู่ต่างกลุ่มกันจะถูกจัดให้อยู่ต่างกลุ่ม

2.1.3 Recursive Computation

เนื่องจากการจัดกลุ่มของข้อมูลกระแสนั้นจำเป็นต้องคำนึงถึงหน่วยความจำที่ใช้ และความรวดเร็วในการประมวลผล ดังนั้นข้อมูลที่เข้ามาจะถูกเก็บไว้ในรูปแบบทางคณิตศาสตร์และข้อสรุปเชิงสถิติจึงทำให้ประหยัดหน่วยความจำ และเมื่อมีข้อมูลกระแสเข้ามาใหม่ ข้อมูลดังกล่าวจะถูกนำไปปรับปรุงกับรูปแบบทางคณิตศาสตร์และข้อสรุปเชิงสถิติที่อยู่ในหน่วยความจำ ซึ่งเป็นการเพิ่มความรวดเร็วในการประมวลผล จึงทำให้มีความรวดเร็วกว่าการคำนวณกับข้อมูลก่อนหน้าทั้งหมด ซึ่งวิธีการที่ใช้ในการปรับปรุงรูปแบบทางคณิตศาสตร์และข้อสรุปเชิงสถิติจะใช้การคำนวณแบบ Recursive function แสดงได้ดังสมการ ที่ (2.1), (2.2), (2.3) และ (2.4)

$$N_i^{new} = N_i^{curr} + 1 \tag{2.1}$$

$$c_i^{new} = \frac{N_i^{curr} c_i^{curr} + x_i}{N_i^{new}} \tag{2.2}$$

$$S_i^{new} = \frac{N_i^{curr}}{N_i^{new}} S_i^{new} + L \tag{2.3}$$

$$L = \left(\frac{x_i x_i^T}{N_i^{new}}\right) - c_i^{new} (c_i^{new})^T + c_i^{cur} (c_i^{cur})^T - \frac{c_i^{cur} (c_i^{cur})^T}{N_i^{new}} \tag{2.4}$$

โดยที่ x_i คือ ข้อมูลที่เข้ามาใหม่

c_i^{cur} คือ ค่าเฉลี่ยของข้อมูลที่เคยนำเข้ามาทั้งหมดจนถึงเวลาปัจจุบัน

N_i^{cur} คือ จำนวนสมาชิกของข้อมูลที่เคยนำเข้ามาทั้งหมดจนถึงเวลาปัจจุบัน

S_i^{cur} คือ Covariance matrix ของข้อมูลที่เคยนำเข้ามาทั้งหมดจนถึงเวลาปัจจุบัน

และ c_i^{new} คือ ค่าเฉลี่ยของข้อมูลที่เคยนำเข้ามาทั้งหมดจนถึงเวลาปัจจุบันที่ต้องปรับปรุง

N_i^{new} คือ จำนวนสมาชิกของข้อมูลที่เคยนำเข้ามาทั้งหมดจนถึงเวลาปัจจุบันที่ต้องปรับปรุง

S_i^{new} คือ Covariance matrix ของข้อมูลที่เคยนำเข้ามาทั้งหมดจนถึงเวลาปัจจุบันที่ต้องปรับปรุง

2.1.4 วิธีการหาค่าความคล้ายคลึงแบบโคไซน์ (Cosine Similarity)

การหาค่าความคล้ายคลึงแบบโคไซน์ เป็นมาตรวัดความคล้ายของการทำมุมระหว่างสองเวกเตอร์ใด ๆ ซึ่งมีค่าอยู่ในช่วง [-1, 1] โดยค่าเข้าใกล้ 1 คือ สองเวกเตอร์มีทิศทางเดียวกัน และค่าเข้าใกล้ -1 คือ สองเวกเตอร์มีทิศทางต่างกัน สามารถแสดงได้ดังสมการ (2.5)

$$\cos(u_i, u_j) = \frac{u_i \cdot u_j}{||u_i||_2 \, ||u_j||_2} \tag{2.5}$$

โดยที่ u_i คือ เวกเตอร์ที่ i^{th}

u_j คือ เวกเตอร์ที่ j^{th}

$||u_i||_2$, $||u_j||_2$ คือ Euclidean norm ของ u_i และ u_j

2.1.5 Bhatthacharyya Distance

มาตรวัดระยะทาง Bhattacharyya Distance (A. Bhattacharyya, 1946) เป็นมาตรวัดระยะทางที่คำนึงถึงระดับการซ้อนทับกันของสองกลุ่มประชากรใด ๆ ซึ่งมีช่วงอยู่ระหว่าง [0, ∞] โดยค่าที่คำนวณได้มีค่าใกล้เคียง 0 มากเท่าใด หมายถึงสองกลุ่มประชากรมีลักษณะคล้ายกันมากเท่านั้น ซึ่งสามารถแสดงได้ดังสมการ (2.7)

$$S = \frac{S_i + S_j}{2} \tag{2.6}$$

$$B(mc_i, mc_j) = \frac{1}{8}(c_i - c_j)^T S^{-1}(c_i - c_j) + \frac{1}{2} ln \left(\frac{det\, S}{\sqrt{det\, S_i \; det\, S_j}}\right) \tag{2.7}$$

โดยที่ mc_i, mc_j คือ กลุ่มประชากรที่ i^{th} และ j^{th} ตามลำดับ

c_i, c_j คือ ค่าเฉลี่ยของข้อมูลที่อยู่ในกลุ่มประชากร mc_i และ mc_j ตามลำดับ

S_i, S_j คือ Covariance Matrix ของ mc_i และ mc_j ตามลำดับ

S^{-1} คือ Inverse Covariance matrix

2.1.6 Principal Component Analysis

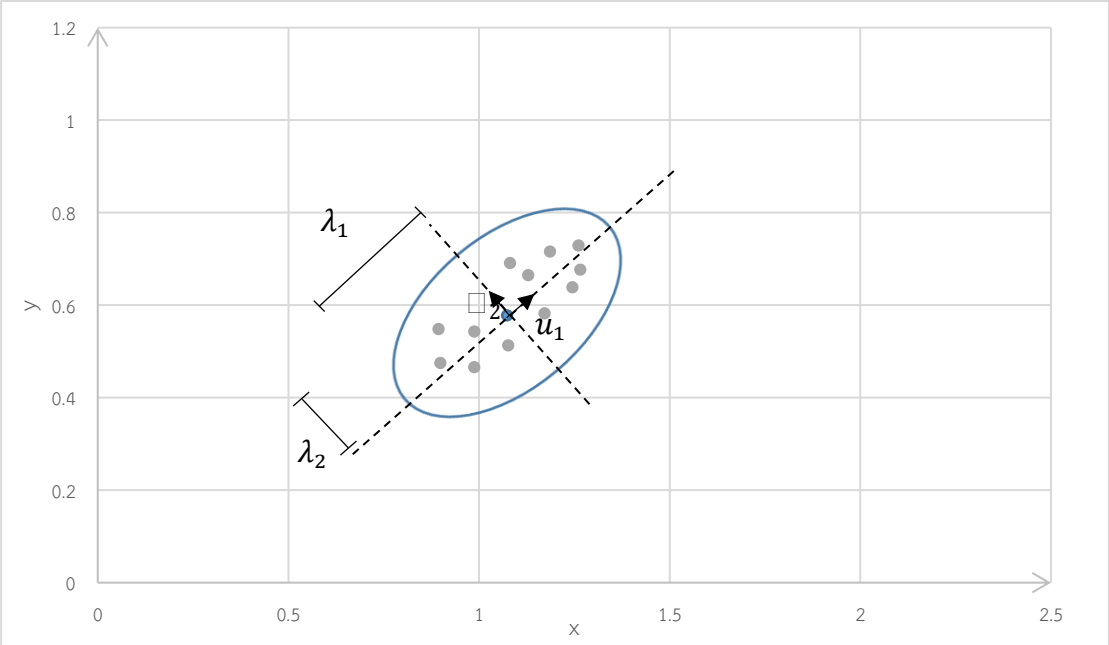
Principal Component Analysis หรือ PCA เป็นวิธีการทางสถิติที่ใช้สำหรับการวิเคราะห์ส่วนประกอบที่สำคัญของข้อมูล เพื่อใช้อธิบายคุณลักษณะของข้อมูลและหาความคล้ายคลึงกันของข้อมูล มีขั้นตอนวิธีการดังนี้

1) ปรับข้อมูลเข้าสู่ศูนย์กลาง (Standardization)

2) คำนวณหา Covariance matrix

3) คำนวณหา Eigen value และ Eigen vector

4) เลือกคุณลักษณะของข้อมูลจาก Eigen value ตามจำนวนที่ต้องการ โดยที่ $\lambda_1 > \lambda_2 > \ldots > \lambda_d$ คือ Eigen value ที่เรียงลำดับจากค่ามากที่สุดไปยังค่าน้อยที่สุด ซึ่งสอดคล้องกับ Eigen vector $u_1, u_2, \ldots, u_d$ ตามลำดับ ดังภาพที่ 2-1



ภาพที่ 2-1 การแทนรูปร่างกลุ่มของข้อมูลขนาดเล็กเป็น spherical

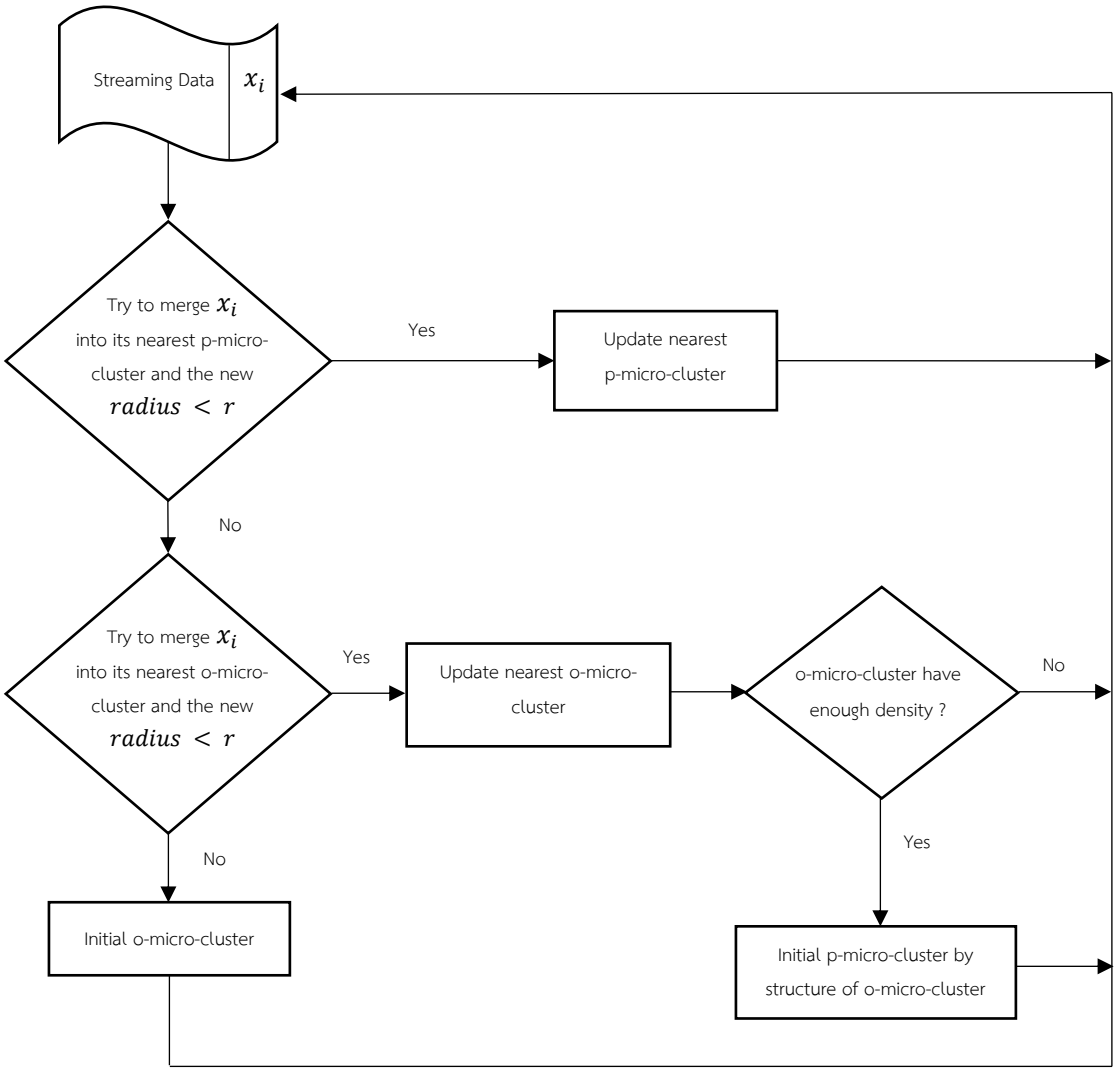
ในงานวิจัยนี้ได้ประยุกต์ใช้ PCA เพื่อหาค่า Eigen value ที่มีค่ามากที่สุดเพื่อใช้ในการหาทิศทางแกนหลักของกลุ่มข้อมูลขนาดเล็ก รวมถึงใช้ในการคำนวณหาปริมาตรของกลุ่มข้อมูลขนาดเล็ก

2.2 งานวิจัยที่เกี่ยวข้อง

Martin Ester และคณะ (1998) ได้นำเสนอขั้นตอนวิธีการจัดกลุ่มที่ชื่อ Incremental DBSCAN ซึ่งเป็นวิธีการในการจัดกลุ่มที่ต่อยอดมาจาก DBSCAN แต่ถูกปรับปรุงให้สามารถจัดกลุ่มข้อมูลที่เข้ามาใหม่ต่อยอดจากกลุ่มข้อมูลที่เคยจัดกลุ่มไว้จากข้อมูลเดิมได้ โดยเมื่อมีข้อมูลใหม่ที่ต้องการเพิ่มเข้ามา จะหาเซตของ core-object ที่ได้รับผลกระทบ ซึ่งหาได้จากลักษณะของการทำ Density-reachable จากข้อมูลที่เข้ามา หลังจากนั้นจึงจัดกลุ่มเพิ่มเติมเฉพาะส่วนที่มีผลกระทบเท่านั้น และยังมีความสามารถในการลบข้อมูลเมื่อเวลาใด ๆ ซึ่งจะทำในลักษณะเดียวกันกับส่วนของการเพิ่มข้อมูล แต่จะนำเซ็ตของ core-objects มาลบสมาชิกที่มีข้อมูลของข้อมูลที่ต้องการลบออกไปแล้วจึงดูว่ากลุ่มข้อมูลต่าง ๆ มีการเปลี่ยนแปลงไปหรือไม่ เช่น กลุ่มข้อมูลเปลี่ยนเป็น noise กลุ่มข้อมูลมีสมาชิกหายไปที่เกิดจากการลบแล้วทำ Density-reachable ไม่ได้ และกลุ่มข้อมูลต้องแยกออกจากกันเป็นคนละกลุ่มเนื่องจากคุณลักษณะเปลี่ยนแปลง แต่ขั้นตอนวิธีการนี้ใช้ข้อมูลทั้งหมดใน

การดำเนินการ ทำให้ไม่สามารถจัดการกับข้อมูลสตรีมมิ่งได้ เนื่องจากข้อมูลสตรีมมิ่งมีขนาดมหาศาล ไม่สามารถเก็บข้อมูลได้ทั้งหมดในหน่วยความจำที่จำกัด

Feng Cao และคณะ (2006) ได้นำเสนอขั้นตอนวิธีการจัดกลุ่มที่ชื่อ DenStream ซึ่งเป็นขั้นตอนวิธีที่ถูกพัฒนามาจาก DBSCAN มีกรอบแนวคิดดังภาพที่ 2-2 โดยข้อมูลที่เข้ามา (x_i) ในช่วงแรกจะถูกเก็บไว้จำนวนหนึ่ง และนำข้อมูลที่เก็บไว้ไปทำในส่วนออฟไลน์ในครั้งแรก เพื่อสร้างเป็นกลุ่มข้อมูลขนาดเล็กสำหรับใช้ในขั้นตอนออนไลน์ต่อไป โดยการทำในส่วนของออฟไลน์นี้จะมีการจัดกลุ่มของข้อมูลที่มีลักษณะแบบ density-based ซึ่งเป็นวิธีการจัดกลุ่มข้อมูลของ DBSCAN เมื่อดำเนินการจัดกลุ่มเสร็จสิ้นแล้ว ก็จะแทนกลุ่มข้อมูลด้วยกลุ่มข้อมูลขนาดเล็กรูปทรง spherical โดยกลุ่มข้อมูลขนาดเล็กนี้จะถูกเรียกว่า potential-micro-cluster หรือเรียกว่า p-micro-cluster หลังจากนั้นก็จะเข้าสู่ส่วนออนไลน์ เมื่อมีข้อมูลใหม่เข้ามาก็จะตรวจสอบลักษณะของข้อมูลที่เข้ามาว่าสามารถอยู่ใน p-micro-cluster ที่มีอยู่หรือไม่ โดยใช้ Euclidean distance เป็นมาตรวัดความใกล้เพื่อหา p-micro-cluster ที่ใกล้สุด และคำนวณรัศมีใหม่ ($radius$) ของ p-micro-cluster ที่ใกล้สุดเมื่อถูกปรับปรุงโดยข้อมูลที่เข้ามาใหม่ ถ้ารัศมีที่ได้มีค่าน้อยกว่าที่กำหนด (r) ก็จะถือว่าข้อมูลที่เข้ามาอยู่ใน p-micro-cluster ข้อมูลใหม่จะก็ถูกรวมเข้ากับ p-micro-cluster ที่อยู่ใกล้สุด แต่ถ้ารัศมีมากกว่าที่กำหนดก็จะทำการสร้าง outlier-micro-cluster หรือเรียกว่า o-micro-cluster ขึ้นมา หลังจากนั้นจึงกลับไปรับข้อมูลใหม่ โดยข้อมูลใหม่จะถูกจัดกลุ่มจาก p-micro-cluster ก่อน ถ้าไม่อยู่ในเงื่อนไขที่กำหนดจึงตรวจสอบจาก outlier-micro-cluster เป็นลำดับถัดมา เมื่อไม่เข้าเงื่อนไขใดเลย จะสร้าง o-micro-cluster ขึ้นมาใหม่ โดยกลุ่มข้อมูลขนาดเล็กเหล่านี้จะมีฟังก์ชันซึ่งเกี่ยวข้องกับเวลาเป็นตัวลดความสำคัญของกลุ่มข้อมูล เมื่อใดที่มีคำสั่งว่าต้องการจัดกลุ่มข้อมูลเพื่อนำไปใช้งาน ก็จะนำ p-



ภาพที่ 2-2 ขั้นตอนวิธีการ DenStream ส่วนออนไลน์

micro-clusters มาจัดกลุ่มโดยถือว่าเป็น Pseudo points ด้วยวิธีแบบ DBSCAN ดั้งเดิม แต่อย่างไรก็ตามรูปทรงของกลุ่มข้อมูลเป็นแบบ spherical ที่งานวิจัยนี้ได้นำเสนอไม่เหมาะสมกับลักษณะการกระจายตัวของข้อมูลที่มีรูปร่างไม่แน่นอน เนื่องจากรูปทรงของกลุ่มข้อมูล spherical ครอบคลุมพื้นที่ส่วนเกินที่ไม่มีข้อมูล ทำให้ผลลัพธ์ในการจัดกลุ่มมีความผิดพลาด

Niwan Wattanakitrungroj และคณะ (2011) ได้นำเสนอขั้นตอนวิธีการจัดกลุ่มที่ชื่อ CompactStream ซึ่งวิธีการจัดกลุ่มแบ่งออกเป็นสองส่วนคือ แบบออนไลน์ และแบบออฟไลน์ นอกจากนี้ งานวิจัยนี้ได้นำเสนอแนวคิดของการหาระยะห่างระหว่างกลุ่มข้อมูลขนาดเล็กแบบ hyper-elliptical โดยใช้อัตราส่วนระยะทางของจุดศูนย์กลางและขอบของกลุ่มข้อมูลขนาดเล็ก รวมทั้งยังนำเสนอให้กลุ่มข้อมูลขนาดเล็กมีระดับของการขยายตัว โดยขั้นตอนส่วนออนไลน์ จะมีการนำเข้าข้อมูลทีละหนึ่งตัวอย่าง เพื่อนำไปเก็บข้อมูลทางสถิติและนำออกจากหน่วยความจำ ซึ่งจะแทนข้อมูลทางสถิติด้วยกลุ่มข้อมูลขนาดเล็กรูปทรง spherical หลังจากนั้นเมื่อกลุ่มข้อมูลขนาดเล็กรูปทรง

spherical มีจำนวนข้อมูลที่อยู่ในกลุ่มตามที่ผู้ใช้กำหนด กลุ่มข้อมูลขนาดเล็กรูปทรง spherical ก็จะถูกแทนที่ด้วยกลุ่มข้อมูลขนาดเล็กรูปทรง hyper-elliptical โดยจะปรับเปลี่ยนรูปทรงของกลุ่มข้อมูลขนาดเล็กไปเรื่อย ๆ ตามข้อมูลที่เข้ามา และเมื่อมีคำสั่งที่ต้องการจัดกลุ่ม จะจัดกลุ่มในส่วนออฟไลน์ด้วยวิธีเชื่อมต่อกันไปเรื่อย ๆ ของกลุ่มข้อมูลขนาดเล็กรูปทรง hyper-elliptical แต่อย่างไรก็ตาม รูปทรง hyper-elliptical ที่ถูกปรับเปลี่ยนมาจากรูปทรง spherical ไม่ได้เกิดจาก Covariance ของข้อมูลที่เข้ามาอยู่ในกลุ่มทั้งหมด ทำให้กลุ่มข้อมูลขนาดเล็กรูปทรง hyper-elliptical มีความคลาดเคลื่อน

Muhammad Zia-ur Rehman และคณะ (2014) ได้นำเสนอขั้นตอนวิธีการจัดกลุ่มที่ชื่อ HECES ซึ่งจะใช้ Grid-cell เพื่อเก็บข้อมูลทางสถิติของข้อมูลที่เข้ามา โดยจะนำเข้าข้อมูลด้วย Sliding Window ที่มีการลดความสำคัญของข้อมูลเมื่อเวลาผ่านไป ซึ่ง Sliding Window ใด ๆ จะถูกเก็บข้อมูลทางสถิติในช่วงนั้นลงใน Grid-cell ซึ่งเมื่อถึงเวลาที่กำหนดก็จะดำเนินการลบ Grid-cell ที่มีค่านัยสำคัญทางสถิติที่ต่ำ ต่อจากนั้นจึงนำ Grid-cell ที่มีค่านัยสำคัญทางสถิติสูงไปสร้างกลุ่มข้อมูลรูปทรง hyper-ellipsoid โดยในขั้นตอนนี้ได้มีการคำนวณค่า Covariance matrix เพื่อหารูปร่างและทิศทางของ hyper-ellipsoid ด้วย shrinkage method เนื่องจากการหา Covariance matrix จากข้อมูลทางสถิติที่อยู่ใน Grid-cell สามารถเกิดปัญหา singular covariance matrix เพราะข้อมูลไม่เพียงพอทำให้ไม่สามารถใช้ในการคำนวณสมการต่าง ๆ ได้ หลังจากนั้นจึงนำกลุ่มของข้อมูลรูปทรง hyper-ellipsoid ที่ใกล้เคียงกันโดยวัดจากมาตรวัด Mahalanobis distance สองกลุ่มมารวมกันและแทนที่ด้วยกลุ่มข้อมูลใหม่ จากนั้นจะลบกลุ่มข้อมูลรูปทรง hyper-ellipsoid ที่มีความซ้ำซ้อนกันเพื่อให้ได้กลุ่มข้อมูลสำหรับนำไปใช้งานต่อไป แต่อย่างไรก็ตาม การนำกลุ่มของข้อมูลรูปทรง hyper-ellipsoid ที่ใกล้เคียงกันมารวมกัน และแทนที่ด้วยกลุ่มข้อมูลใหม่เพื่อให้ได้กลุ่มข้อมูลสำหรับนำไปใช้งาน ทำให้ผลลัพธ์ที่ได้จากการจัดกลุ่มมีความผิดพลาด

Edwin Lughofera และคณะ (2015) ได้นำเสนอขั้นตอนวิธีการจัดกลุ่มที่ชื่อ eVQ-AMS ซึ่งมีการแทนลักษณะของกลุ่มข้อมูลด้วย ellipsoidal ตั้งแต่ต้น ซึ่งเป็นลักษณะของกลุ่มข้อมูลที่มีความคล้ายคลึงกับรูปแบบการกระจายตัวของข้อมูลที่ใกล้เคียงกว่า และใช้ Mahalanobis distance เป็นมาตรวัดระยะทางระหว่าง ellipsoidal ซึ่งเป็นมาตรวัดที่อธิบายคุณลักษณะของรูปทรงข้อมูลได้ดีกว่า Euclidean distance และมีการทำการจัดกลุ่มในลักษณะที่มีส่วนเดียวคือ กระบวนการทั้งหมดถูกรวมไว้เป็นขั้นตอนเดียวคือส่วนออนไลน์ โดยเมื่อข้อมูลเข้ามาจะถูกนำไปหากลุ่มข้อมูลที่ใกล้สุด หลังจากนั้นจึงปรับปรุงกลุ่มข้อมูลที่ใกล้สุดและกลุ่มข้อมูลรอบ ๆ แล้วจึงพิจารณาว่าจะเก็บไว้เป็นตัวแบบสำหรับการแบ่งกลุ่มข้อมูลออกเป็นสองกลุ่มหรือจะทำการรวมกลุ่มข้อมูลที่ใกล้เคียงกันเป็นกลุ่มเดียวกันหรือไม่ ซึ่งกลุ่มข้อมูลที่ได้จากขั้นตอนวิธีการนี้จะถูกพิจารณาเป็นกลุ่มข้อมูลสำหรับนำไปใช้งานโดยไม่ต้องผ่านขั้นตอนออฟไลน์ แต่ทั้งนี้ยังคงมีปัจจัยหลายอย่าง ทั้งในเรื่องของการจัดเก็บข้อมูลจำนวนหนึ่งเพื่อใช้ในการจัดกลุ่ม ซึ่งมีผลกระทบในเรื่องของหน่วยความจำที่ใช้ และปัญหาในเรื่องตัวแทนกลุ่มของข้อมูลที่มีจุดศูนย์กลางร่วมกันที่เกิดจากการรวมกันของกลุ่มของข้อมูลไปเรื่อย ๆ ทำให้ผลลัพธ์ที่ได้จากการจัดกลุ่มมีความผิดพลาด

Michael Hahsler และคณะ (2016) ได้นำเสนอขั้นตอนวิธีการจัดกลุ่มที่มีชื่อ DBSTREAM ซึ่งยังคงแทนลักษณะของกลุ่มข้อมูลขนาดเล็กด้วยรูปทรง hyper-sphere และมีลักษณะการทำงานที่

เหมือน DenStream คือประกอบไปด้วยการทำงานสองส่วน คือส่วนออนไลน์และส่วนออฟไลน์ โดยในส่วนออนไลน์ เมื่อมีข้อมูลเข้ามาจะถูกตรวจสอบว่าสามารถอยู่ในกลุ่มข้อมูลขนาดเล็กใดได้ หลังจากนั้นจึงทำการเก็บข้อมูลเชิงสถิติที่แสดงถึงความหนาแน่นของกลุ่มข้อมูลขนาดเล็กให้อยู่ในรูปแบบกราฟ และพิจารณาว่ากลุ่มข้อมูลขนาดเล็กควรเปลี่ยนแปลงคุณลักษณะทางสถิติหรือไม่เพื่อป้องกันกลุ่มงข้อมูลขนาดเล็กทับซ้อนกัน และได้นำเสนอวิธีการปรับปรุงการจัดกลุ่มของกลุ่มข้อมูลขนาดเล็กที่อยู่ในส่วนออฟไลน์ ว่าควรคำนึงถึงความหนาแน่นของกลุ่มข้อมูลขนาดเล็กสองกลุ่มว่าควรมีการแผ่ขยายเพื่อเชื่อมต่อกันเป็นกลุ่มข้อมูลสำหรับเป็นผลลัพธ์ของการจัดกลุ่มหรือไม่ แต่อย่างไรก็ตามเนื่องจากรูปทรงที่ใช้ในการจัดกลุ่มของข้อมูลยังคงเป็น hyper-sphere ซึ่งไม่เข้ากับลักษณะของข้อมูลและยังให้ค่าน้ำหนักของข้อมูลเมื่อเวลาผ่านไป

Niwan Wattanakitrungroj และคณะ (2017) ได้นำเสนอขั้นตอนวิธีการจัดกลุ่มที่มีชื่อ VHEC ซึ่งเป็นวิธีการจัดกลุ่มที่มีลักษณะใกล้เคียงกับ CompactStream โดยแบ่งออกเป็นสองส่วนคือส่วนออนไลน์สำหรับเก็บสถิติของข้อมูลที่เข้ามาให้อยู่ในรูปแบบกลุ่มข้อมูลขนาดเล็กรูปทรง versatile hyper-ellipsoidal ซึ่งเป็นรูปทรงที่สามารถกำหนดขอบเขตให้ครอบคลุมกับข้อมูลได้ และส่วนออฟไลน์ที่แบ่งออกเป็นส่วนย่อย ๆ คือ การรวมกลุ่มข้อมูลขนาดเล็กที่มีคุณลักษณะใกล้เคียงกันเข้าด้วยกัน และส่วนของการเชื่อมต่อกลุ่มข้อมูลขนาดเล็กเพื่อให้ได้กลุ่มข้อมูลสำหรับนำไปใช้งาน แต่อย่างไรก็ตามเนื่องจากขั้นตอนวิธีการนี้นิยามรูปร่างของกลุ่มข้อมูลขนาดเล็กที่เป็นทรง versatile hyper-ellipsoidal ไม่ได้เกิดจากข้อมูลที่อยู่ในกลุ่ม แต่เกิดจากการประมาณให้เป็นรัศมีของกลุ่มข้อมูลขนาดเล็กที่เป็นรูปทรง hyper-spherical ดังนั้นจึงทำให้เกิดความคลาดเคลื่อนในระยะเริ่มต้นในการจัดกลุ่มของข้อมูลซึ่งอาจส่งผลกระทบต่อการจัดกลุ่มข้อมูลเพื่อให้ได้เป็นผลลัพธ์สำหรับนำไปใช้งาน

ตารางที่ 2-1 เปรียบเทียบงานวิจัยที่เกี่ยวข้องในประเด็นต่าง ๆ

เทคนิคการแบ่งกลุ่มข้อมูล	**รูปร่างของข้อมูล**	**การนำข้อมูลแต่ละตัวเข้าเพียงครั้งเดียว**	**มีการเก็บข้อมูล**	**โครงสร้างถูกคำนวณจากข้อมูลที่เข้ามาทั้งหมด***	**รูปทรงของโครงสร้างกลุ่มข้อมูลขนาดเล็ก****	**สามารถจัดกลุ่มได้ต่อจากรอบก่อนหน้า**	**พารามิเตอร์ที่ต้องกำหนด**
Incremental DBSCAN (1998)	arbitrary	No	Yes	Yes	spherical	Yes	• radius • minimum of points
Den-Stream (2006)	arbitrary	Yes	No	Yes	spherical	No	• radius • minimum of points • fading factor
CompactStream (2011)	arbitrary	Yes	No	No	hyper-elliptical	No	• radius • the number of points • the scale of the boundary of the ellipstical micro-cluster
HECES (2014)	arbitrary	Yes	No	Yes	hyper-ellipsoid	No	• size of sliding window • width of a grid-cell • number of grid-cell
eVQ-AMS (2015)	arbitrary	Yes	Yes	No	ellipsoidal	No	• variation range • fraction
DBSTREAM (2016)	arbitary	YES	No	Yes	spherical	No	• Radius • Minimum of points • Intersection factor • Noise threshold
VHEC (2017)	arbitary	YES	No	No	versatile hyper-ellipsoidal	No	• radius • the number of points • the scale of the boundary of the VHEF micro-cluster • the cosine similarity threshold
DyCluStream+	**arbitrary**	**Yes**	**No**	**Yes**	**dynamic hyper-ellipsoidal**	**Yes**	• **radius** • **the number of points** • **the expanding threshold** • **the merging threshold by bhatthacharriya distance** • **the merging threshold by cosine similarity**

* โครงสร้างของกลุ่มข้อมูลขนาดเล็กเกิดจากการคำนวณด้วยข้อมูลที่เข้ามาทั้งหมดโดยไม่ได้เกิดจากการประมาณ

** รูปทรงของโครงสร้างที่ใช้สำหรับเป็นกลุ่มข้อมูลขนาดเล็ก

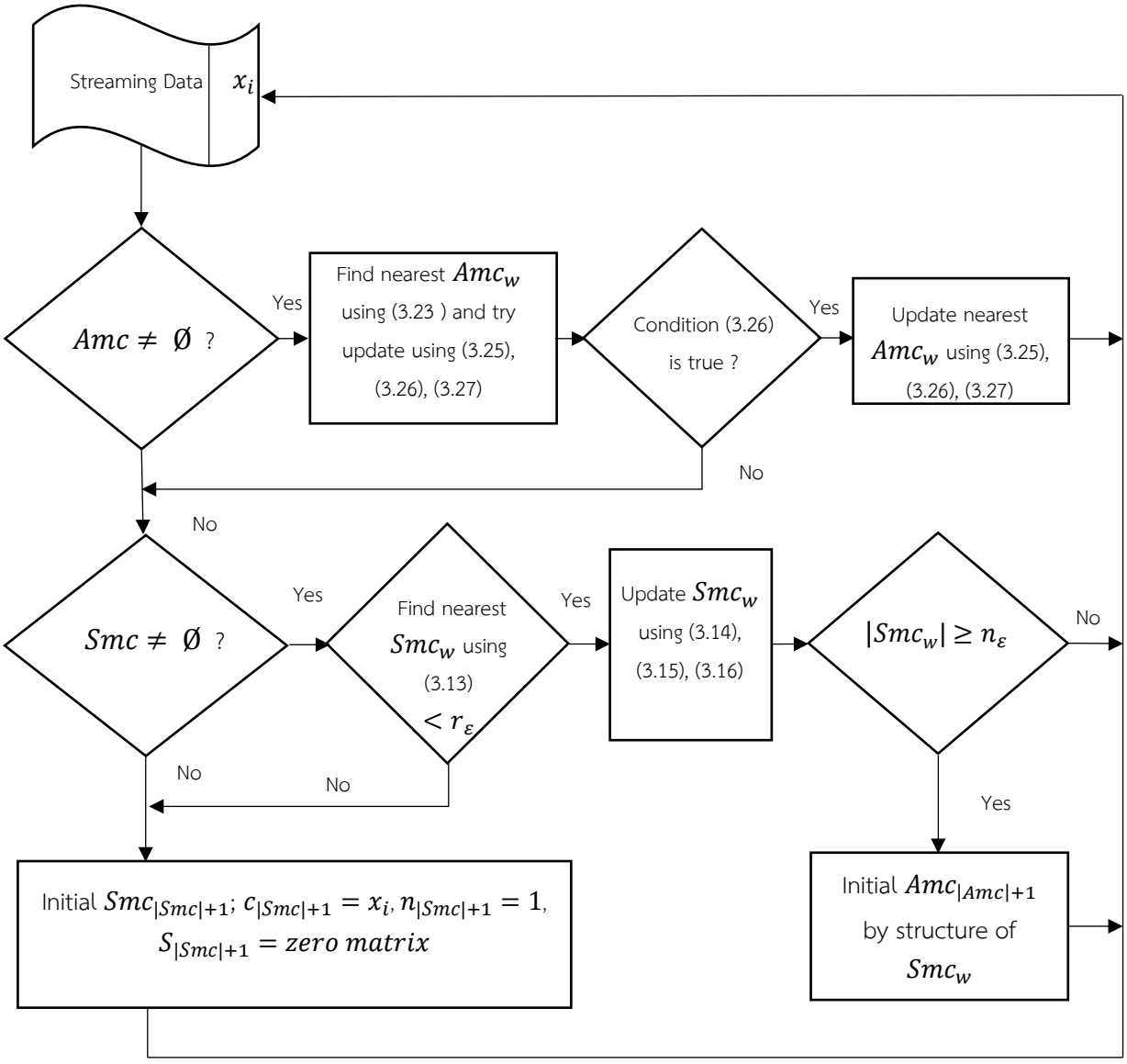
บทที่ 3
วิธีการที่นำเสนอ

งานวิจัยได้นำเสนอวิธีการจัดกลุ่มข้อมูลที่มีชื่อว่า **The Dynamic Clustering Algorithm for Streaming Data Using One-Pass-Throw-Away Datum and Hyper-ellipsoidal Micro-Clustering** หรือเรียกว่า **DyCluStream** ซึ่งนำเสนอโครงสร้างในการจัดเก็บข้อมูลในเชิงสถิติ เพื่อลดหน่วยความจำที่ใช้ และเป็นอัลกอริทึมในการจัดกลุ่มข้อมูลแบบกระแสที่สามารถจัดกลุ่มข้อมูลโดยไม่สามารถทราบลักษณะการกระจายตัวและจำนวนกลุ่มได้ล่วงหน้า โดยอัลกอริทึมนี้จะแบ่งการทำงานออกเป็นสองส่วนคือ ส่วนออนไลน์สำหรับเก็บสถิติของข้อมูลที่นำเข้า และส่วนออฟไลน์สำหรับการจัดกลุ่มข้อมูล

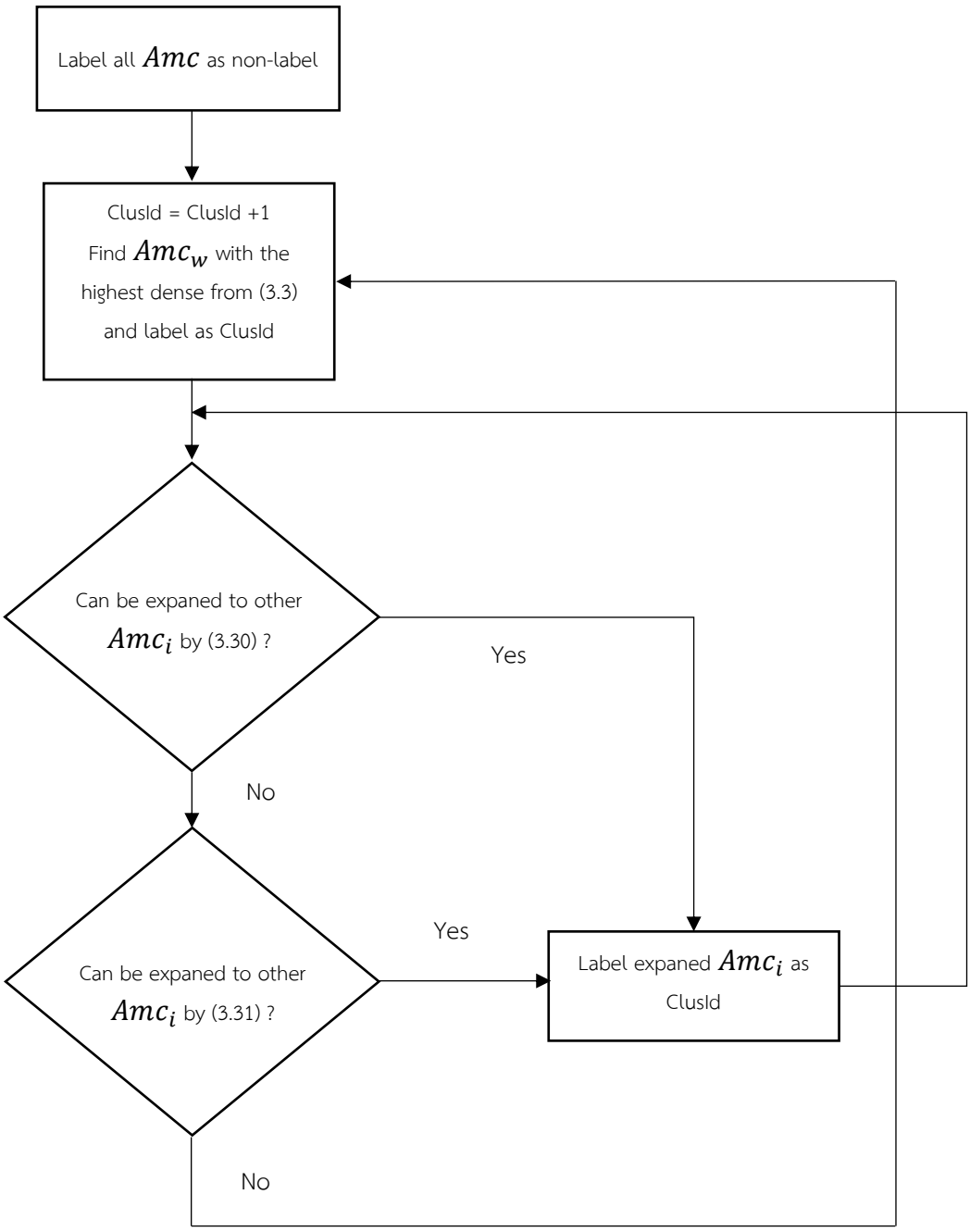
ส่วนออนไลน์ ในตอนต้นข้อมูลที่เข้ามาจะถูกจัดเก็บข้อมูลทางสถิติให้อยู่ในกลุ่มของข้อมูลขนาดเล็กรูปทรง spherical โดยใช้มาตรวัดระยะทางแบบ Euclidean และเมื่อกลุ่มข้อมูลขนาดเล็กมีจำนวนสมาชิกตามที่ผู้ใช้กำหนด ก็จะเปลี่ยนรูปร่างกลุ่มข้อมูลขนาดเล็กให้เป็นรูปทรง dynamic hyper-ellipsoidal หลังจากนั้น ข้อมูลที่เข้ามาใหม่จะถูกตรวจสอบกับกลุ่มข้อมูลขนาดเล็กรูปทรง dynamic hyper-ellipsoidal เป็นลำดับแรก โดยใช้มาตรวัดระยะทางแบบ Mahalanobis เพราะเข้ากับลักษณะรูปทรง dynamic hyper-ellipsoidal เนื่องจากมีการคำนึงถึงลักษณะการกระจายตัวของข้อมูลที่อยู่ในกลุ่มข้อมูลขนาดเล็ก เมื่อข้อมูลถูกเก็บสถิติในตัวแทนกลุ่มข้อมูลขนาดเล็กใด ๆ แล้ว ข้อมูลจะถูกนำออกจากหน่วยความจำ ซึ่งมีกรอบแนวคิดดังภาพที่ 3-1

ส่วนออฟไลน์จะถูกเรียกใช้ก็ต่อเมื่อมีความต้องการกลุ่มข้อมูลเพื่อนำไปใช้งาน โดยการจัดกลุ่มจะเริ่มจากกลุ่มข้อมูลขนาดเล็กรูปทรง dynamic hyper-ellipsoidal ที่มีความหนาแน่นมากที่สุดเป็นตัวแผ่ขยายเพื่อเชื่อมต่อไปยังกลุ่มข้อมูลขนาดเล็กรูปทรง dynamic hyper-ellipsoidal อื่น ๆ โดยใช้ฟังก์ชันในการผสานเป็นเงื่อนไขหลัก และอัตราส่วนความหนาแน่นของข้อมูลเป็นเงื่อนไขรองในการเชื่อมต่อกลุ่มข้อมูลขนาดเล็กรูปทรง dynamic hyper-ellipsoidal เข้าด้วยกันเพื่อจัดเป็นกลุ่มขนาดใหญ่สำหรับนำไปใช้งานดังภาพที่ 3-2 และขั้นตอนวิธีการ **DyCluStream+** ที่ถูกพัฒนาต่อยอดเพื่อปรับปรุงประสิทธิภาพความถูกต้องในการจัดกลุ่มข้อมูลด้วยการปรับเปลี่ยนมาตรวัดระยะทางและเพิ่มการรวมกลุ่มข้อมูลขนาดเล็กสองกลุ่มที่มีคุณลักษณะใกล้เคียงกันเพื่อลดหน่วยความจำที่ใช้งาน อีกทั้งยังมีความสามารถในการจัดกลุ่มกับข้อมูลจริง และความสามารถในการจัดกลุ่มข้อมูลที่อยู่

ในส่วนออฟไลน์ที่สามารถดำเนินการจัดกลุ่มต่อจากส่วนออนไลน์ที่ผ่านมาเพื่อลดเวลาที่ใช้ในการจัดกลุ่มข้อมูล



ภาพที่ 3-1 ขั้นตอนวิธีการ DyCluStream ส่วนออนไลน์



ภาพที่ 3-2 ขั้นตอนวิธีการ DyCluStream ส่วนออฟไลน์

3.1 การนำเข้าข้อมูล

ในงานวิจัยนี้ได้จำลองเหตุการณ์การจัดกลุ่มข้อมูลแบบกระแส โดยการนำเข้าข้อมูลกระแสทีละหนึ่งตัวอย่างจากชุดข้อมูลหนึ่ง ๆ และเมื่อประมวลผลเรียบร้อยแล้ว จะนำข้อมูลนั้นออกจากหน่วยความจำหลัก

กำหนดให้ Streaming Data D คือ เซตของข้อมูลที่มีลำดับเวลา $x_1, x_2, \dots, x_i, \dots, x_n$ โดยที่ $n \to \infty$ นั่นคือ $D = \{x_i\}_{i=1}^{n}$ ซึ่งแต่ละ $x_i = [x_{i1}, x_{i2}, \dots, x_{id}]$ ทั้งหมด d มิติ นั่นคือ $x_i = [x_{ij}]_{j=1}^{d}$

3.2 โครงสร้างสำหรับกลุ่มข้อมูลขนาดเล็ก

โครงสร้างของกลุ่มข้อมูลขนาดเล็กรูปทรง dynamic hyper-ellipsoidal ที่ใช้ในการจัดเก็บข้อมูลทางสถิติของข้อมูลที่เข้ามาประกอบไปด้วย 2 โครงสร้าง ดังนี้

3.2.1 Amc - Actual micro-clusters

Amc เป็นเซตของกลุ่มข้อมูลขนาดเล็ก โดย $Amc = \{Amc_1, Amc_2, \dots, Amc_i, \dots, Amc_m\}$ ซึ่งมีคุณลักษณะเป็น **dynamic hyper-ellipsoidal** สามารถนิยามได้ดัง นิยามที่ 3.1

นิยามที่ 3.1 กำหนดให้ Actual micro-cluster ลำดับที่ i^{th} หรือเรียกว่า Amc_i คือ dynamic hyper-ellipsoidal micro-cluster ลำดับที่ i^{th} ของข้อมูล $x = \{x_1, x_2, \dots, x_n\}$ โดยโครงสร้างของ Amc_i ประกอบไปด้วย ค่าเฉลี่ยของข้อมูล (c_i), Covariance matrix (S_i) และจำนวนสมาชิกที่อยู่ในกลุ่มของข้อมูล (n_i) เขียนแทนด้วย $Amc_i(c_i, S_i, n_i)$

กลุ่มข้อมูลขนาดเล็กจะเป็น Amc_i ก็ต่อเมื่อ $n_i \geq n_\varepsilon$

3.3.2 Smc - Small micro-clusters

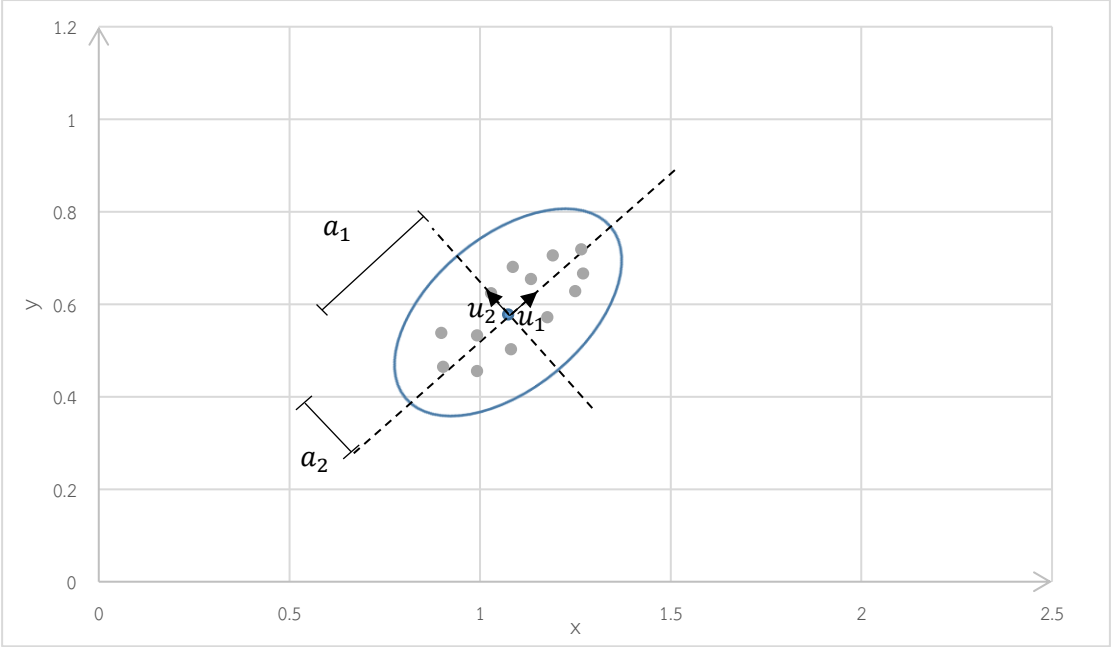
Smc เป็นเซตของกลุ่มข้อมูลขนาดเล็ก โดย $Smc = \{Smc_1, Smc_2, \dots, Smc_i, \dots, \dots, Smc_m\}$ ซึ่งมีคุณลักษณะเป็น **spherical micro-cluster** สามารถนิยามได้ดังนิยามที่ 3.2

นิยามที่ 3.2 กำหนดให้ Small micro-cluster ลำดับที่ i^{th} หรือเรียกว่า Smc_i คือ spherical micro-cluster ลำดับที่ i^{th} สำหรับกลุ่มของข้อมูล $x = \{x_1, x_2, \dots, x_n\}$ โดยโครงสร้างของ Smc_i ประกอบไปด้วย ค่าเฉลี่ยของข้อมูล (c_i), Covariance matrix (S_i) และจำนวนสมาชิกที่อยู่ในตัวแทนกลุ่มของข้อมูล (n_i) เขียนแทนด้วย $Smc_i(c_i, S_i, n_i)$

กลุ่มข้อมูลขนาดเล็กจะเป็น Smc_i ก็ต่อเมื่อ $n_i < n_\varepsilon$

โครงสร้างของกลุ่มข้อมูลขนาดเล็ก Amc_i ในงานวิจัยนี้ได้ประยุกต์จากงานวิจัย A Very Fast Neural Learning for Classification Using Only New Incoming Datum (S. Jaiyen et al. 2010)

ซึ่งงานวิจัยดังกล่าวมีกลุ่มข้อมูลขนาดเล็กที่มีโครงสร้างเป็น dynamic hyper-ellipsoidal micro-cluster ที่สามารถหมุนไปตามฐานหลักเชิงตั้งฉากปรกติ (Orthonormal Basis) และสามารถเคลื่อนย้ายจุดศูนย์กลางไปยังพิกัดใด ๆ ดังภาพที่ 3-3



ภาพที่ 3-3 โครงสร้าง dynamic hyper-ellipsoidal micro-cluster

สมมติให้แต่ละตัวอย่างข้อมูลเป็น $x_i = [\, x_{i1}, x_{i2}, \ldots, x_{id}]^T$

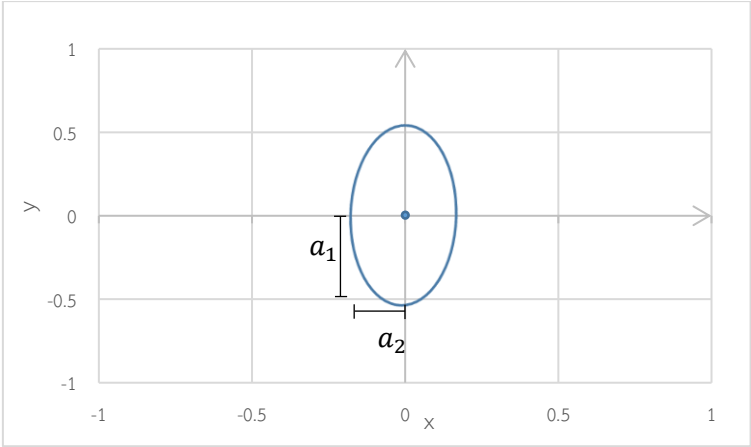
โครงสร้าง dynamic hyper-ellipsoidal micro-cluster ที่มีจำนวนมิติของข้อมูล d โดยมีจุดศูนย์กลางอยู่ที่จุดกำเนิด ตัวอย่างดังภาพที่ 3-4 สามารถนิยามได้ดังสมการที่ (3.1)

$$\frac{x_{i1}^2}{a_1^2} + \frac{x_{i2}^2}{a_2^2} + \cdots + \frac{x_{id}^2}{a_d^2} = \sum_{j=1}^{d} \frac{x_{ij}^2}{a_j^2} = 1 \qquad (3.1)$$

โดย d คือ จำนวนมิติของข้อมูล

x_{ij} คือ คือตัวอย่างข้อมูลที่ i^{th} มิติที่ j^{th}

$a_j;\ j = 1, 2, \ldots, \mathrm{d}$ คือ ความกว้างของแกนที่ j^{th} ในโครงสร้างของ hyper-ellipsoidal-microcluster



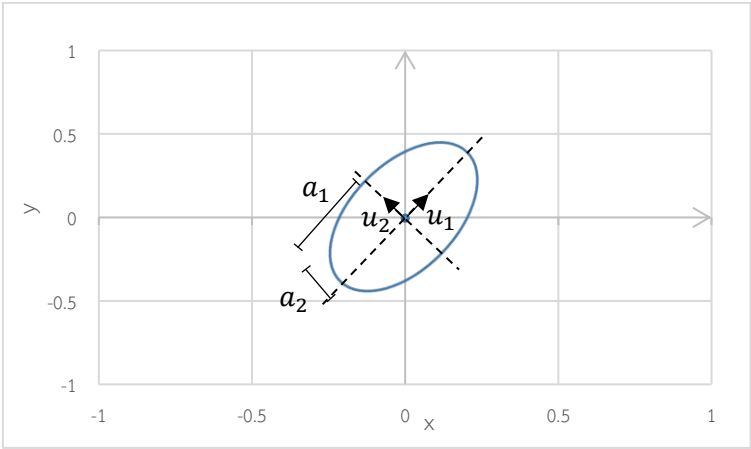
ภาพที่ 3-4 โครงสร้าง hyper-ellipsoidal Micro-cluster ที่มีจุดศูนย์กลางอยู่ที่จุดกำเนิด

ซึ่งแต่ละตัวอย่างข้อมูลสามารถถูกเคลื่อนย้ายโดยคำนึงถึงฐานหลักเชิงตั้งฉากปรกติ ดังสมการที่ (3.2)

$$x_{ij} = x_i{}^T u_j \tag{3.2}$$

โดย u_j คือ Eigenvector มิติที่ j^{th} ของตัวอย่างข้อมูล

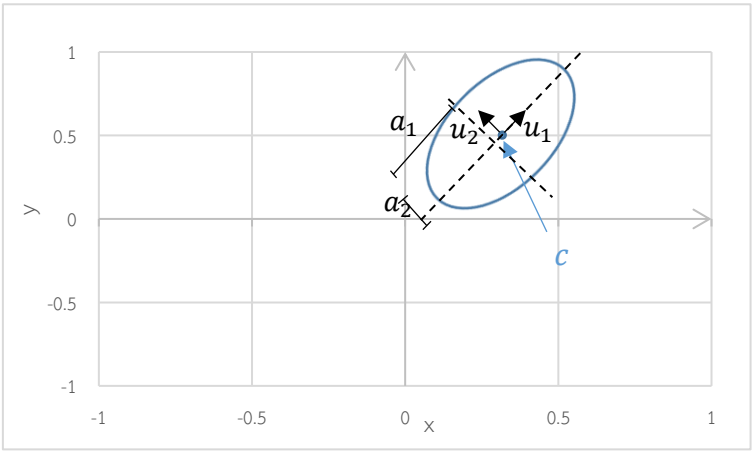
จากสมการที่ (3.1) และ (3.2) ทำให้ได้สมการ hyper-ellipsoidal ที่สามารถหมุนไปตามฐานหลักเชิงตั้งฉากปรกติที่มีจุดศูนย์กลางอยู่ที่จุดกำเนิด ดังสมการ (3.3) สามารถแสดงได้ ดังภาพที่ 3-5



ภาพที่ 3-5 โครงสร้าง hyper-ellipsoidal Micro-cluster ที่สามารถหมุนไปตามฐานหลักเชิงตั้งฉากปรกติที่มีจุดศูนย์กลางอยู่ที่จุดกำเนิด

$$\sum_{j=1}^{d} \frac{\left(x_i{}^T u_j\right)^2}{a_j^2} = 1 \tag{3.3}$$

ถ้าแกนเดิมของ hyper-ellipsoidal micro-cluster ถูกเคลื่อนย้ายจากจุดกำเนิดไปยังจุดศูนย์กลางของจุดอื่น $c = [c_1, c_2, \dots, c_d]^T$ ดังภาพที่ 3-6 จึงสามารถเขียนสมการ **dynamic hyper-ellipsoidal** ได้ดังสมการที่ (3.4)



ภาพที่ 3-6 โครงสร้าง dynamic hyper-ellipsoidal micro-cluster

$$\sum_{j=1}^{d} \frac{\left((x_i - c)^T u_j\right)^2}{a_j^2} = 1 \tag{3.4}$$

ในงานวิจัยนี้เราได้ประยุกต์ใช้ Principle Component Analysis (PCA) เพื่อใช้ในการหาฐานหลักเชิงตั้งฉากปรกติ โดยที่ Eigenvalue และ Eigenvector สามารถคำนวณได้จาก Covariance matrix ซึ่ง Eigenvector จะตั้งฉากซึ่งกันและกัน และมีความยาวเท่ากับ 1 ดังนั้นจึงกล่าวได้ว่า Eigenvector มีลักษณะเป็นฐานหลักเชิงตั้งฉากปรกติ (orthonormal basis) และ Eigenvalue นั้นสามารถปรับเปลี่ยนขนาดความกว้างได้เนื่องด้วยถูกคำนวณมาจากความแปรปรวนของข้อมูล ดังนั้นค่าความกว้าง $a_j, j = 1, 2, \dots, d$ สามารถนิยามได้ว่า $\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots, \sqrt{\lambda_d}$ โดยที่แต่ละ λ_j คือค่า Eigenvalue ที่สอดคล้องกับ Eigenvector u_j

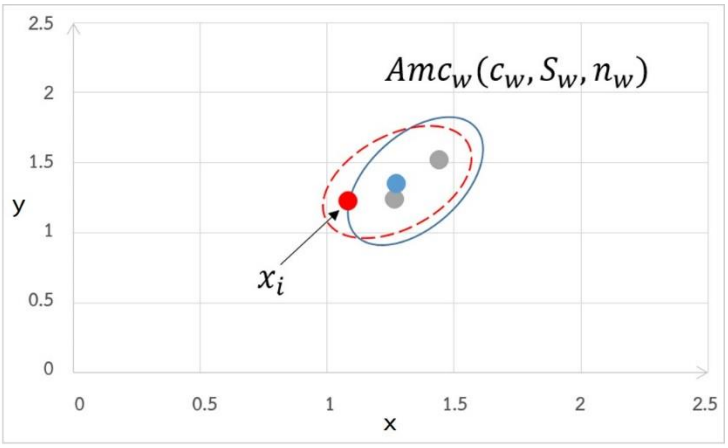
จากสมการที่ (3.4) สามารถนิยามสมการ dynamic hyper-ellipsoidal ได้ดังสมการที่ (3.5) และ (3.6)

$$\sum_{j=1}^{d} \frac{\left((x_i - c)^T u_j\right)^2}{\lambda_j^2} = 1 \tag{3.5}$$

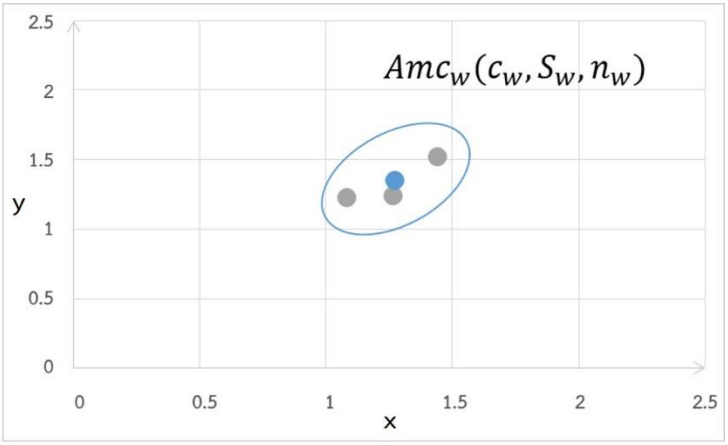
$$\varphi(x) = \sum_{j=1}^{d} \frac{\left((x_i - c)^T u_j\right)^2}{\lambda_j^2} = 1 \tag{3.6}$$

ในงานวิจัยนี้ได้ใช้สมการ (3.6) สำหรับหาค่าความใกล้ชิดระหว่างตัวอย่างข้อมูล x_i และกลุ่มข้อมูลขนาดเล็ก Amc_w โดยคำนึงถึงรูปทรงใหม่ที่เกิดจากการปรับปรุงรูปร่างเดิมกับข้อมูลที่เข้ามา

สามารถครอบคลุมข้อมูลที่เข้ามาใหม่ได้หรือไม่ ดังภาพที่ 3-7 โดยเส้นสีฟ้าคือรูปร่างเดิม และเส้นประสีแดงคือรูปร่างใหม่ หากรูปร่างใหม่สามารถครอบคลุม ก็จะใช้รูปทรงที่ถูกปรับปรุงจากข้อมูลที่เข้ามาใหม่แทนที่รูปทรงเดิม ดังภาพที่ 3-8 แต่ถ้าไม่ครอบคลุมก็จะดำเนินการสร้างกลุ่มข้อมูลขนาดเล็กขึ้นมาใหม่โดยใช้ข้อมูลที่เข้ามาใหม่



ภาพที่ 3-7 ข้อมูลที่นำเข้าอยู่ในรัศมีของกลุ่มข้อมูล Amc_w ที่ปรับปรุงโครงสร้าง (เส้นประสีแดง)



ภาพที่ 3-8 การปรับปรุงโครงสร้างกลุ่มข้อมูล Amc_w จากรูปร่างเดิมเป็นรูปร่างใหม่

จากสมการที่ (3.6) สามารถประยุกต์ให้อยู่ในรูปแบบของระยะทางระหว่างกลุ่มข้อมูลขนาดเล็กที่ $m\ (Amc_m)$ และ $n\ (Amc_n)$ ซึ่งใช้วัดความใกล้เคียงของระยะทางได้ดังสมการที่ (3.7)

$$\delta(Amc_m, Amc_n) = \sum_{j=1}^{d} \frac{\left((c_m - c_n)^T u_j\right)^2}{\lambda_j^2} = 1 \qquad (3.7)$$

โดยที่ c_m, c_n คือ จุดศูนย์กลางของ Amc_m และ Amc_n

u_j คือ องค์ประกอบของ Eigen Vector ที่ j^{th} ของ Amc_n

λ_j คือ องค์ประกอบของ Eigen Value ที่ j^{th} ของ Amc_n

ถ้าพิจารณาแต่ละกลุ่มข้อมูล Amc_i ที่มีความกว้างตามแกนต่าง ๆ เป็น $\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_d}$ โดยที่แต่ละ λ_j คือค่า Eigenvalue ซึ่งสอดคล้องกับ Eigenvector u_j จะสามารถคำนวณปริมาตร (vol) และความหนาแน่น ($dense$) ได้ดังสมการ (3.8) และ (3.9)

$$vol = \lambda_1 \times \lambda_2 \times \ldots \times \lambda_d \tag{3.8}$$

$$dense = \frac{n_i}{vol} \tag{3.9}$$

โดยที่ n_i คือ จำนวนสมาชิกที่อยู่ในกลุ่มข้อมูลขนาดเล็กที่มีโครงสร้างเป็น dynamic hyper-ellipsoidal

3.3 ขั้นตอนวิธีการในการจัดกลุ่มข้อมูล DyCluStream

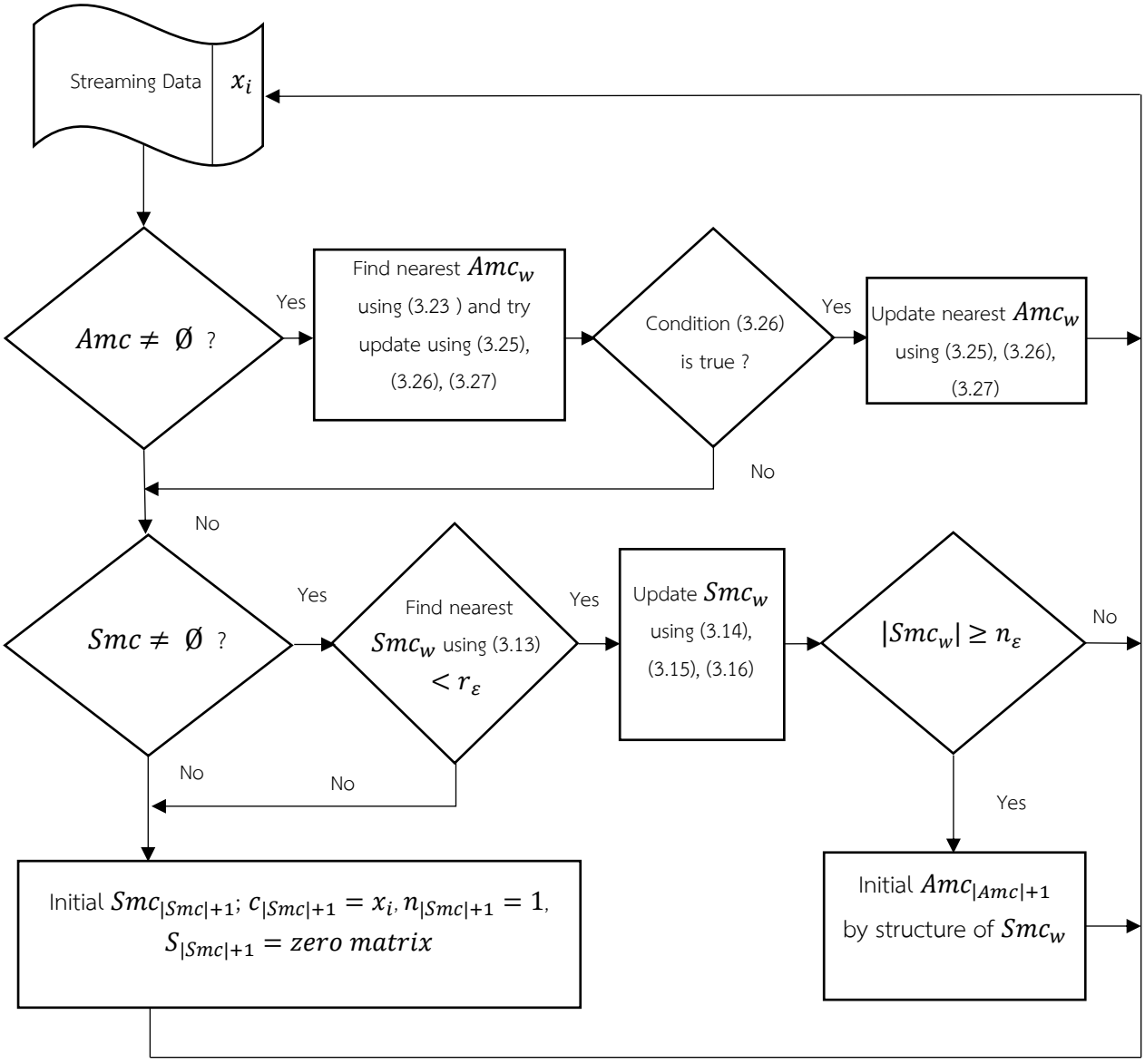
งานวิจัยนี้ได้แบ่งขั้นตอนวิธีการออกเป็น 2 ส่วนคือ

ขั้นตอนส่วนออนไลน์ เป็นขั้นตอนที่ใช้ในการเก็บข้อมูลที่เข้ามาให้อยู่ในรูปโครงสร้างทางสถิติที่นำเสนอคือ กลุ่มข้อมูลขนาดเล็ก Amc และกลุ่มข้อมูลขนาดเล็ก Smc

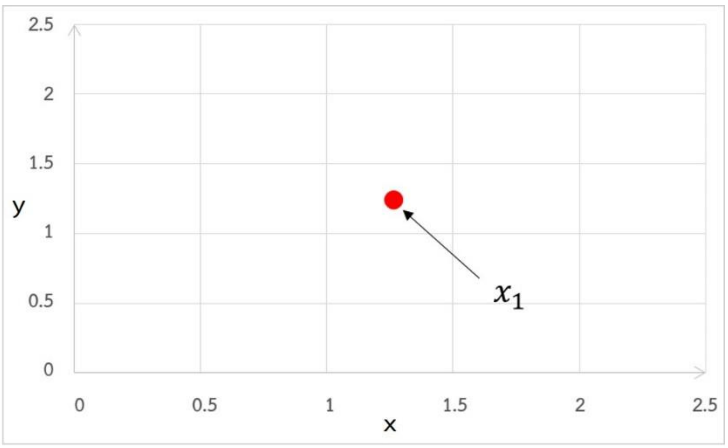
ขั้นตอนส่วนออฟไลน์ เป็นขั้นตอนจัดกลุ่มข้อมูลโดยใช้เทคนิคการเชื่อมต่อกันไปเรื่อย ๆ ของกลุ่มข้อมูลขนาดเล็ก Amc เพื่อให้ได้กลุ่มข้อมูลสำหรับนำไปใช้งาน

3.3.1 ขั้นตอนส่วนออนไลน์

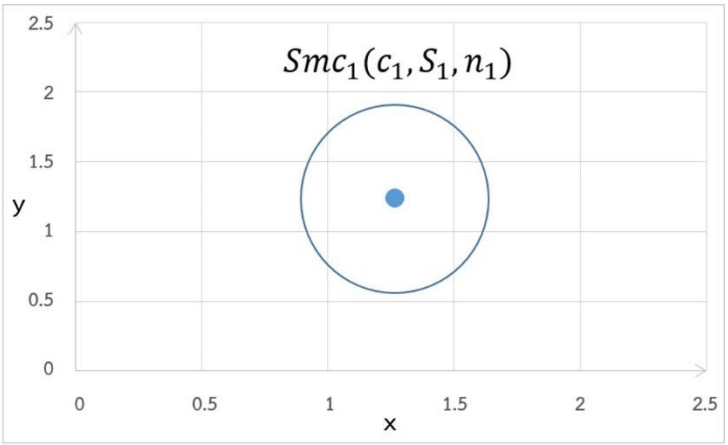
เป็นการดำเนินการจัดเก็บข้อมูลที่เข้ามา ให้อยู่ในรูปแบบโครงสร้างกลุ่มข้อมูลขนาดเล็ก โดยมีขั้นตอนดังภาพที่ 3-9 ขั้นตอนวิธีการ DyCluStream ส่วนออนไลน์



ภาพที่ 3-9 ขั้นตอนวิธีการ DyCluStream ส่วนออนไลน์



ก) ข้อมูลเข้าสู่ระบบครั้งแรก



ข) Smc_1 ถูกสร้างขึ้นเพื่อจัดเก็บข้อมูลที่เข้ามา

ภาพที่ 3-10 การสร้างตัวแทนกลุ่มของข้อมูลเพื่อใช้ในการจัดเก็บข้อมูลที่ถูกนำเข้าเป็นครั้งแรก

การดำเนินการส่วนออนไลน์ในครั้งแรก ข้อมูลที่ถูกนำเข้ามาจะถูกสร้างเป็นกลุ่มข้อมูล Smc_1 เริ่มต้นก่อน ดังภาพที่ 3-10 ก) โดยกำหนดโครงสร้างเป็น $Smc_1(c_1, S_1, n_1)$ ให้จุดศูนย์กลางตัวแทนกลุ่มของข้อมูล (c_1) คือข้อมูลที่ถูกนำเข้า (x_1) จำนวนสมาชิก (n_1) มีค่าเท่ากับหนึ่ง และ Covariance matrix (S_1) มีค่าเป็น zero matrix ซึ่งบ่งบอกถึงรูปร่างที่เป็น spherical ดังภาพที่ 3-10 ข) สามารถกำหนดได้ดังสมการที่ (3.10), (3.11) และ (3.12)

$$c_1 = x_1 \tag{3.10}$$

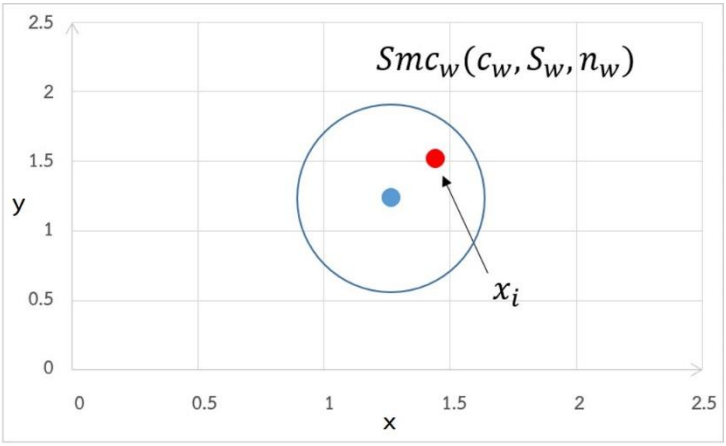
$$n = 1 \tag{3.11}$$

$$S_1 = zero\ matrix \tag{3.12}$$

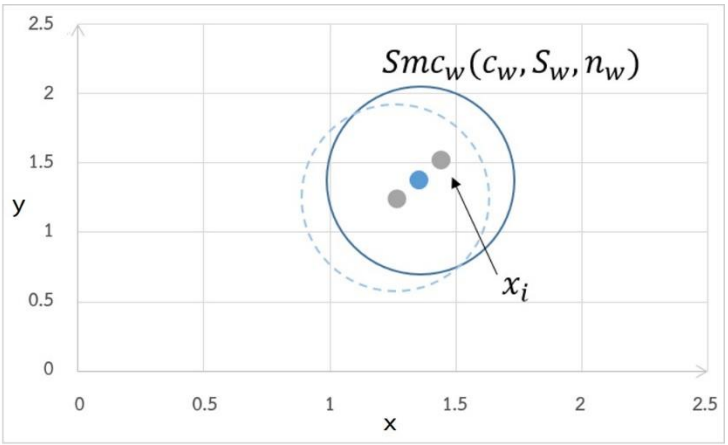
การดำเนินการส่วนออนไลน์ในกรณีที่ยังไม่มีกลุ่มข้อมูล Amc แต่มีกลุ่มข้อมูล Smc แล้ว เมื่อมีข้อมูลเข้ามาใหม่จะถูกนำไปตรวจสอบก่อนว่าอยู่ใกล้กับกลุ่มข้อมูล Smc กลุ่มใดมากที่สุดก่อน โดยใช้สมการที่ (3.14) ว่าสามารถครอบคลุมข้อมูลที่เข้ามาใหม่ได้หรือไม่

$$euclidDist(x_i, Smc_k) = \sqrt{\sum_{j=1}^{d}(x_{ij} - c_{kj})^2} \tag{3.13}$$

$$w = argmin_{k=1,2,\ldots,|Smc|} euclidDist(x_i, Smc_k) \tag{3.14}$$



ก) ข้อมูลที่นำเข้าอยู่ในรัศมีของ Smc_w



ข) การปรับปรุงโครงสร้าง Smc_w

ภาพที่ 3-11 การปรับปรุงค่า เฉลี่ย จำนวนสมาชิก และ Covariance matrix ของกลุ่มข้อมูลโครงสร้าง Smc_w เมื่อข้อมูลที่เข้ามาอยู่ภายในรัศมี

หากค่าที่ได้จากมาตรวัดระยะทาง $euclidDist(x_i, Smc_w)$ ไม่เกินรัศมีที่ผู้ใช้กำหนด (r_ε) ดังภาพที่ 3-11 ก) จะนำข้อมูล x_i ไปปรับปรุงกลุ่มข้อมูล $Smc_w(c_w, S_w, n_w)$ โดย w คือลำดับของกลุ่มข้อมูล Smc ที่ใกล้สุด ดังภาพที่ 3-11 ข) ซึ่งเส้นประคือรูปทรงเก่า และเส้นทึบคือรูปทรงใหม่

หลังการปรับปรุงกลุ่มข้อมูล การปรับปรุงกลุ่มข้อมูล $Smc_w(c_w, S_w, n_w)$ สามารถทำได้ดังสมการที่ (3.15), (3.16) และ (3.17)

$$c_{w(new)} = \frac{n_{w(old)}}{(n_{w(old)}+1)}\, c_{old} + \frac{x_i}{n_{w(old)}+1} \tag{3.15}$$

$$S_{w(new)} = \frac{n_{w(old)}}{n_{w(old)}+1}\, S_{w(old)} + \left(\frac{x_i {x_i}^T}{n_{w(old)}+1}\right) - c_{w(new)} c_{w(new)}^T + c_{w(old)} c_{w(old)}^T - \frac{c_{w(old)} c_{w(old)}^T}{n_{w(old)}+1} \tag{3.16}$$

$$n_{w(new)} = n_{w(old)} + 1 \tag{3.17}$$

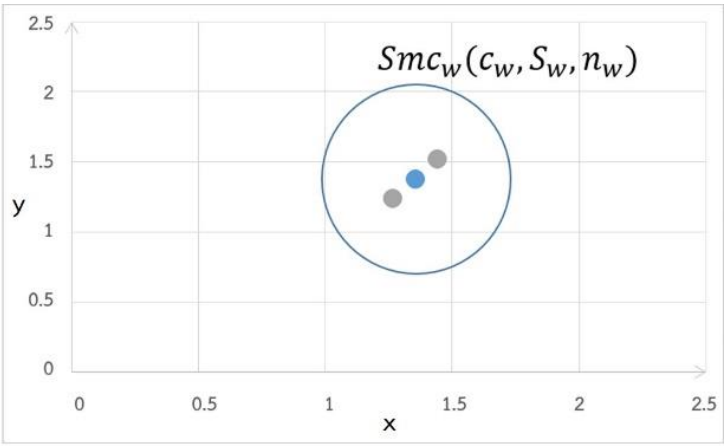
จากนั้นจะพิจารณาต่อไปว่า ถ้า n_{new} มีค่ามากกว่าหรือเท่ากับที่ผู้ใช้กำหนด $(n_w \geq n_\varepsilon)$ ดังภาพที่ 3-12 โดยจุดสีเทาคือข้อมูลที่เคยเข้ามาอยู่ในกลุ่มข้อมูล Smc_w ก็จะเปลี่ยนจาก $Smc_w(c_w, S_w, n_w)$ เป็น $Amc_{|Amc|+1}(c_{|Amc|+1}, S_{|Amc|+1}, n_{|Amc|+1})$ ดังภาพที่ 3-13 โดยใช้สมการ (3.18), (3.19) และ (3.20)

$$c_{|Amc|+1} = c_w \tag{3.18}$$

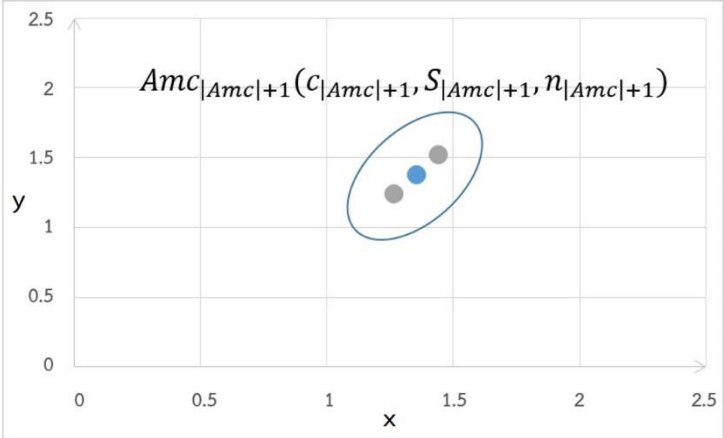
$$S_{|Amc|+1} = S_w \tag{3.19}$$

$$n_{|Amc|+1} = n_w \tag{3.20}$$

โดย $|Amc|$ คือ จำนวนของกลุ่มข้อมูล Amc



ภาพที่ 3-12 กลุ่มข้อมูล Smc_w มีความหนาแน่นถึงระดับที่ผู้ใช้กำหนด (สังเกตุได้จากจำนวนจุดสีเทา)



ภาพที่ 3-13 โครงสร้าง $Amc_{|Amc|+1}$ที่ถูกสร้างขึ้นจาก Smc_w

หากค่าที่ได้จากมาตรวัดระยะทาง $euclidDist(x_i, Smc_w)$ เกินรัศมีที่ผู้ใช้กำหนด (r_ε) ก็จะสร้างกลุ่มข้อมูล $Smc_{|Smc|+1}(c_{|Smc|+1}, S_{|Smc|+1}, n_{|Smc|+1})$ ขึ้นมาใหม่ โดยใช้สมการที่ (3.21), (3.22) และ (3.23) แล้วจึงนำเข้าข้อมูลใหม่

$$c_{|Smc|+1} = x_i \tag{3.21}$$

$$n_{|Smc|+1} = 1 \tag{3.22}$$

$$S_{|Smc|+1} = zero\ matrix \tag{3.23}$$

การดำเนินการส่วนออนไลน์กรณีที่มีกลุ่มข้อมูล Amc แล้ว เมื่อมีข้อมูลเข้ามาใหม่จะตรวจสอบว่าอยู่ใกล้กลุ่มข้อมูล Amc_w กลุ่มใดมากที่สุดจากสมการที่ (3.25) จากมาตรวัดระยะทาง Mahalanobis ซึ่งเป็นมาตรวัดระยะทางของ dynamic hyper-ellipsoidal micro-cluster

$$mahaDist(x_i, Amc_k) = \sqrt{(x_i - c)^T S^{-1}(x_t - c)} \tag{3.24}$$

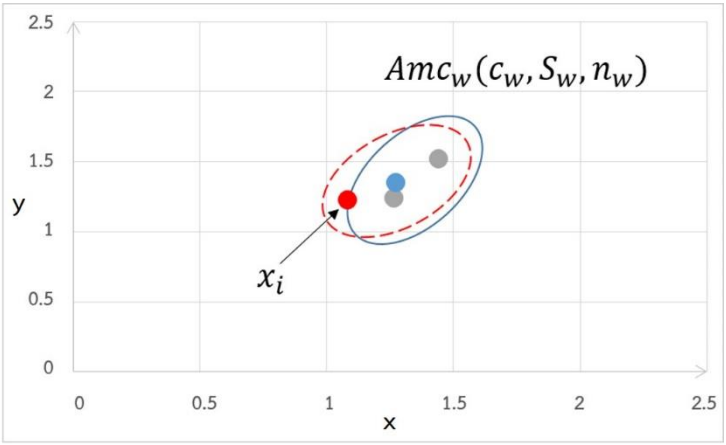
$$w = argmin_{k=1,2,\ldots,|Smc|}\ mahaDist(x_i, Amc_k) \tag{3.25}$$

โดย x_i คือ เวกเตอร์ของข้อมูลนำเข้า

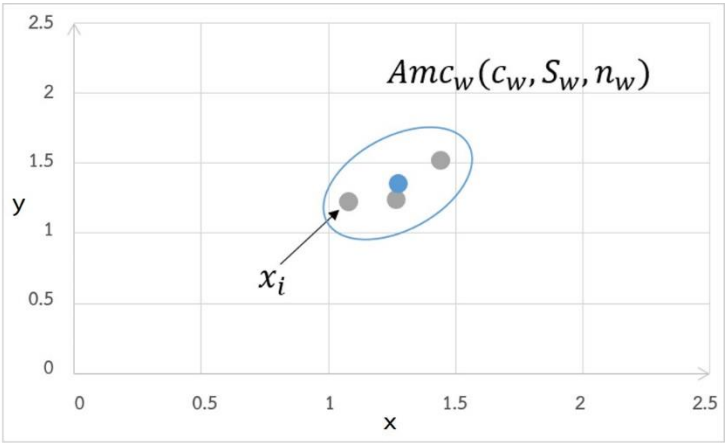
c คือ ค่าเฉลี่ยของข้อมูลของ Amc

S^{-1} คือ Inverse Covariance matrix ของ Amc

และหลังจากที่กลุ่มข้อมูล Amc_w มีข้อมูล (x_i) เข้าไปเป็นสมาชิกแล้ว ดังภาพที่ 3-14 จะตรวจสอบต่อไปว่ารูปร่างของกลุ่มข้อมูล Amc_w ที่ได้สามารถครอบคลุมข้อมูลตัวนั้นได้หรือไม่ โดยตรวจสอบได้จากสมการที่ (3.26)



ภาพที่ 3-14 ข้อมูลที่นำเข้าอยู่ในรัศมีของ Amc_w ที่ทดลองปรับปรุงโครงสร้าง (เส้นประสีแดง)



ภาพที่ 3-15 การปรับปรุงโครงสร้าง Amc_w เมื่อรูปร่างที่ถูกทดลองตามข้อมูลที่เข้ามาสามารถครอบคลุมข้อมูลได้

$$\psi(x_i, Amc_w) = \sum_{j=1}^{d} \frac{((x_i - c_w)^T u_j)^2}{(2 * \lambda_j)^2} - 1 \qquad (3.26)$$

โดย c_w คือ จุดศูนย์กลางของกลุ่มข้อมูล

u_j คือ Eigen Vector ของมิติลำดับที่ j^{th}

λ_j คือ Eigen Value ของมิติลำดับที่ j^{th}

หากค่าจากสมการ (3.26) น้อยกว่าหรือเท่ากับ 0 หมายถึง กลุ่มข้อมูล Amc_w ที่ได้นำข้อมูลเข้าไปปรับปรุงรูปร่างแล้ว รูปร่างใหม่ไม่สามารถครอบคลุมข้อมูลตัวที่เข้ามาได้ ข้อมูลใหม่จะถูกนำไปเปรียบเทียบกับกลุ่มข้อมูล Smc_w ที่ใกล้ที่สุด ว่าสามารถครอบคลุมข้อมูลที่เข้ามาใหม่ได้หรือไม่ ดังกรณีที่ไม่มีกลุ่มข้อมูล Amc เริ่มต้น

แต่ถ้าค่าจากสมการ (3.26) มากกว่า 0 หมายถึง Amc_wที่ได้นำข้อมูลเข้าไปปรับปรุงรูปร่างแล้ว รูปร่างใหม่สามารถครอบคลุมข้อมูลตัวที่เข้ามาได้ ก็จะปรับปรุงกลุ่มข้อมูล Amc_w ดังภาพที่ 3-15 โดยปรับปรุงจุดศูนย์กลางของข้อมูล Covariance matrix และจำนวนสมาชิกในกลุ่ม ดังสมการที่ (3.27), (3.28) และ (3.29) แล้วจึงนำเข้าข้อมูลใหม่

$$c_{w(new)} = \frac{n_{w(old)}}{(n_{w(old)}+1)}\, c_{w(old)} + \frac{x_i}{n_{w(old)}+1} \tag{3.27}$$

$$S_{w(new)} = \frac{n_{w(old)}}{n_{w(old)}+1}\, S_{w(old)} + \left(\frac{x_i x_i{}^T}{n_{w(old)}+1}\right) - c_{w(new)} c_{w(new)}^T$$

$$+ c_{w(old)} c_{w(old)}^T - \frac{c_{w(old)} c_{w(old)}^T}{n_{w(old)}+1} \tag{3.28}$$

$$n_{w(new)} = n_{w(old)} + 1 \tag{3.29}$$

ขั้นตอนวิธีการในส่วนออนไลน์สามารถอธิบายได้ดังภาพที่ 3-16

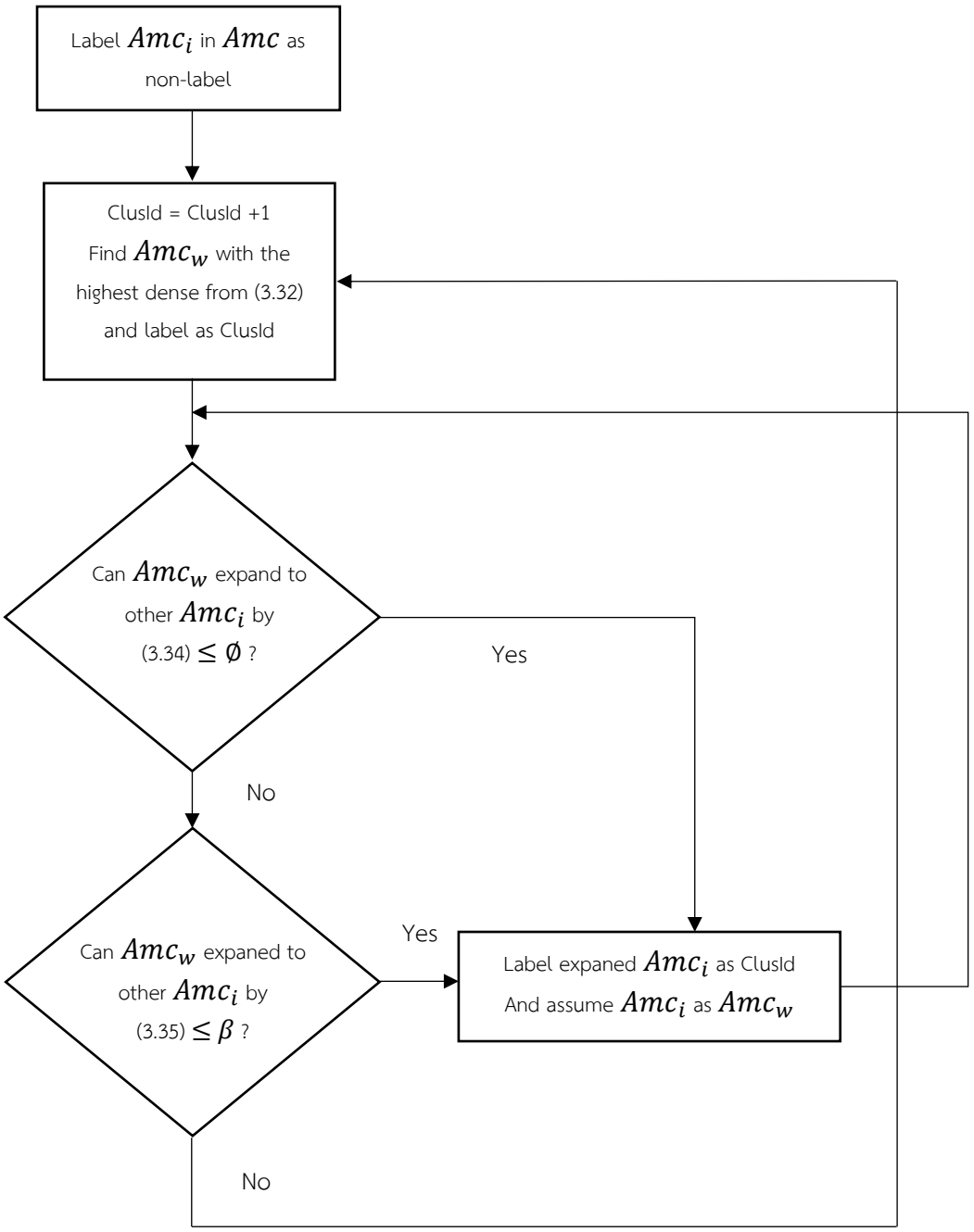
Algorithm Online Phase($x_i, r_\varepsilon, n_\varepsilon$)

Get the arriving data point x_i
if $Amc \neq \emptyset$ **then**
    Find the winning Actual micro-cluster Amc_w using (3.25)
    if x_i is satisfies the conditions in eq. (3.26) with respect Amc_w **then**
        Update the parameters of Amc_w from eq. (3.27), (3.28), (3.29)
        return
    end
end
if $Smc \neq \emptyset$ **then**
    Find the winning Small micro-cluster Smc_w using eq. (3.14)
    if $mahaDist(x_i, Smc_w) \leq r_\varepsilon$ then
        Update the parameters of Smc_w using eq. (3.15), (3.16), (3.17)
        if $n_w > n_\varepsilon$ then
            Create a new $Amc_{|Amc|+1}$ by Smc_w using eq. (3.18), (3.19), (3.20)
            return
        end
        return
    end
end
Create new $Smc_{|Smc|+1}$ by the arriving data point x_i using eq. (3.20), (3.21), (3.22)
return

ภาพที่ 3-16 ขั้นตอนวิธีการ DyCluStream ส่วนออนไลน์

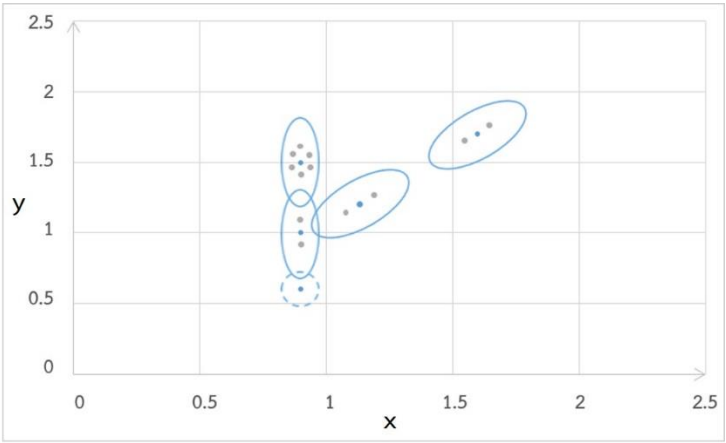
3.3.2 ขั้นตอนส่วนออฟไลน์

ขั้นตอนส่วนออฟไลน์จะทำงานเมื่อมีความต้องการกลุ่มข้อมูลสำหรับไปใช้งาน มีลักษณะการทำงานคือ การเชื่อมต่อกลุ่มข้อมูล Amc ให้กลายเป็นกลุ่มข้อมูลที่มีขนาดใหญ่สำหรับนำไปใช้งาน ดังภาพที่ 3-14 โดยมีกระบวนการดังต่อไปนี้

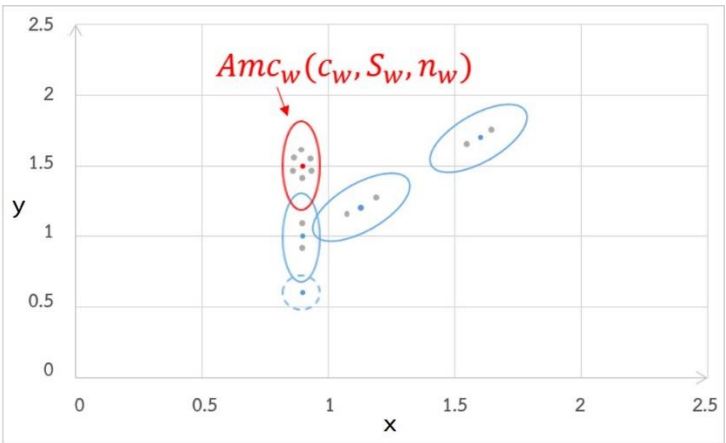


ภาพที่ 3-17 ขั้นตอนวิธีการ DyCluStream ส่วนออฟไลน์

กระบวนการในการจัดกลุ่มจากกลุ่มข้อมูล Amc จะมีรูปแบบการทำงานเหมือนกับกระบวนการในการจัดกลุ่มจากความหนาแน่นของข้อมูล (Density-based clustering) โดยจะเริ่มต้นจากกลุ่มข้อมูล Amc และกลุ่มข้อมูล Smc ทุกตัวจะถูกกำหนดว่ายังไม่มีหมายเลขกลุ่ม ดังภาพที่ 3-18 ก) โดยเส้นทึบคือกลุ่มข้อมูล Amc และเส้นประคือกลุ่มข้อมูล Smc จากนั้นจะเริ่มจัดกลุ่มโดยแผ่กระจายจากกลุ่มข้อมูล Amc_w ที่มีความหนาแน่นมากที่สุดจากสมการที่ (3.32) ที่ยังไม่ถูกกำหนดกลุ่ม ดังภาพที่ 3-18 ข) เป็นตัวเริ่มต้นไปยังกลุ่มข้อมูล Amc_i ใด ๆ ต่อไปโดยเงื่อนไขดังสมการที่ (3.33) เพื่อตรวจสอบว่าสามารถแผ่กระจายได้หรือไม่



ก) กำหนดให้ Amc ทุกกลุ่มยังไม่กำหนดหมายเลขกลุ่ม



ข) กลุ่มข้อมูล Amc_w ที่มีความหนาแน่นมากที่สุด

ภาพที่ 3-18 ขั้นตอนเริ่มต้นการแผ่กระจายกลุ่มข้อมูล Amc

$$volumn(Amc_k) = \lambda_1 \times \lambda_2 \times \ldots \times \lambda_d \tag{3.30}$$

$$density(Amc_k) = \frac{|Amc_k|}{Amc_k.volumn(Amc_k)} \tag{3.31}$$

$$w = argmax_{k=1,2,\ldots,|Smc|}\ density(Amc_k) \tag{3.32}$$

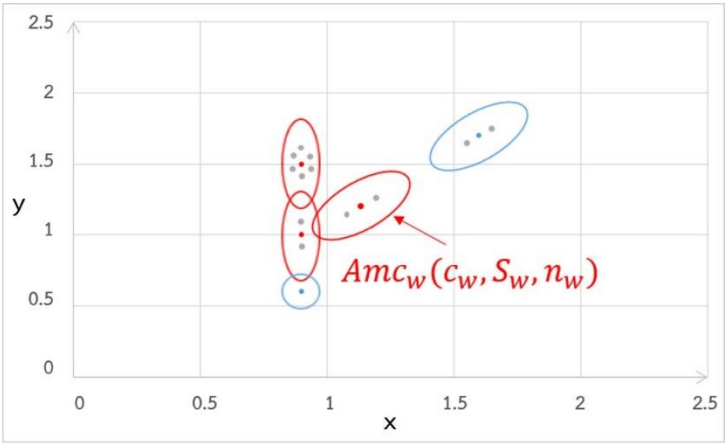
$$\delta(Amc_w, Amc_i) = \sum_{j=1}^{d} \frac{((c_i - c_w)^T \vec{\mu}_{wi})^2}{\lambda_{wi}} - 1 \tag{3.33}$$

$$min\delta(Amc_w, Amc_i) = min\ (\delta(Amc_w, Amc_i), \delta(Amc_i, Amc_w)) \tag{3.34}$$

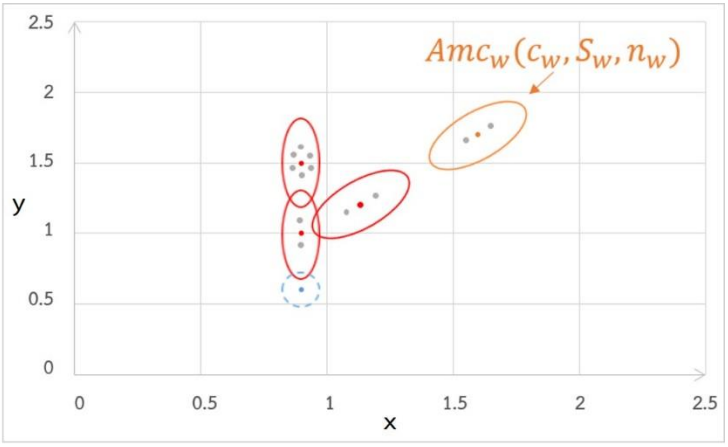
หากค่าที่ได้จากสมการที่ (3.34) น้อยกว่าหรือเท่ากับที่ผู้ใช้กำหนด ก็จะแผ่ขยายกลุ่มข้อมูล Amc_i ต่อไป แต่ถ้าหากค่าที่ได้จากสมการที่ (3.34) มากกว่าที่ผู้ใช้กำหนด ก็จะตรวจสอบจากอัตราส่วนของความหนาแน่นของกลุ่มข้อมูล Amc_w กับกลุ่มข้อมูล Amc_i ต่อไปดังสมการที่ (3.35)

$$densityRatio(Amc_w, Amc_i) = \frac{volumn(Amc_w \cup Amc_i)}{volumn(Amc_w) + volumn(Amc_i)} \tag{3.35}$$

หากค่าที่ได้จากสมการที่ (3.34) น้อยกว่าหรือเท่ากับที่ผู้ใช้กำหนด ก็จะแผ่ขยายกลุ่มข้อมูล Amc_i ต่อไป ดังภาพที่ 3-19 แต่ถ้ามากกว่า กลุ่มข้อมูล Amc_w ก็จะหยุดการแผ่ขยาย เมื่อไม่สามารถแผ่ขยายเพื่อสร้างกลุ่มข้อมูลได้แล้ว ก็จะเริ่มสร้างกลุ่มใหม่โดยมีกลุ่มข้อมูล Amc_w ที่มีความหนาแน่นมากที่สุดและยังไม่ถูกกำหนดหมายเลขกลุ่มเป็นจุดเริ่มต้นในการแผ่ขยายต่อไป ดังภาพที่ 3-20

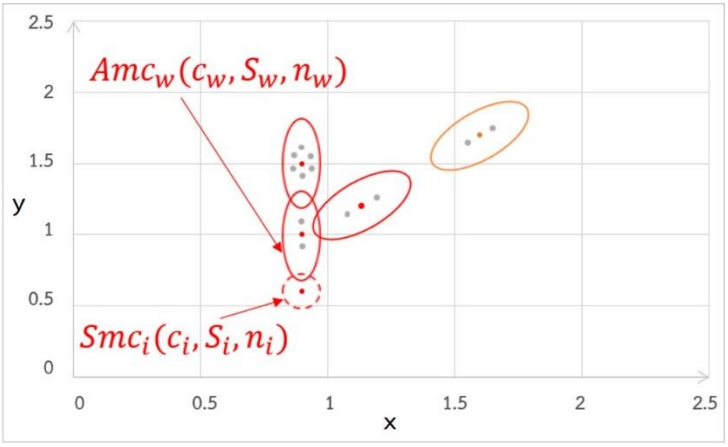


ภาพที่ 3-19 กลุ่มข้อมูล Amc_w ไม่สามารถแผ่ขยายต่อไปได้



ภาพที่ 3-20 เริ่มการแผ่ขยายจากกลุ่มข้อมูล Amc_w ที่มีความหนาแน่นมากที่สุดที่ยังไม่ได้ถูกกำหนดหมายเลขกลุ่มเป็นลำดับถัดไป

โดยกระบวนการข้างต้นจะดำเนินการไปเรื่อย ๆ จนกระทั่งกลุ่มข้อมูล Amc ถูกกำหนดหมายเลขกลุ่มทั้งหมด และจะเริ่มกำหนดกลุ่มให้ Smc โดยพิจารณาจากกลุ่มข้อมูล Amc_w ที่ใกล้ที่สุดจากสมการที่ (3.14) ดังภาพที่ 3-21



ภาพที่ 3-21 ผลลัพธ์ของการจัดกลุ่มข้อมูลในส่วนออฟไลน์

ขั้นตอนวิธีการในส่วนออฟไลน์สามารถอธิบายได้ดังภาพที่ 3-18

Algorithm ClusteringAmc(Amc)

1: Label Amc_i in Amc as non-label
2: Initial $ClusId = 0$
3: **for** each non-labelled Amc_w $with\ most\ dense$ from eq.(3.32) ∈ Amc do
4: Incremental ClusId by 1
5: Label Amc_w as $ClusId$
6: ExpandClusAmc(ClusId, Amc_w)
7: **end**

Algorithm ExpandClusAmc($ClusId, Amc_w, Amc$)

8: **for** each non-labelled $Amc_w \in Amc$ **do**
9: **if** ($min\delta(Amc_w, Amc_i)$ from eq. (3.34) $\leq \emptyset$)) **then**
10: Label Amc_i as $ClusId$
11: ExpandClusAmc($ClusId, Amc_i$)
12: **else if** ($densityRatio(Amc_w, Amc_i)$ from eq. (3.35) $\leq \beta$)
13: Label Amc_i as $ClusId$
14: ExpandClusAmc($ClusId, Amc_i$)
15: **end**
16: **end**

ภาพที่ 3-22 ขั้นตอนส่วนออฟไลน์

3.4 ขั้นตอนวิธีการ DyCluStream+

ขั้นตอนวิธีการ DyCluStream ที่ได้นำเสนอยังคงมีข้อบกพร่องในเรื่องของความสามารถในการจัดการกับข้อมูลจริง ในกรณีที่ Covariance matrix มีคุณลักษณะเป็น Not Positive definite ทำให้มีผลกระทบต่อการคำนวณในสมการต่าง ๆ นอกจากนี้มาตรวัดระยะทาง $mahaDist(x_i, Amc_k)$ ที่ใช้ยังคงไม่เหมาะสมกับลักษณะของข้อมูล และกลุ่มข้อมูล Amc สองกลุ่มที่มีลักษณะคล้ายกันมากจนเป็นกลุ่มเดียวกัน ยังถูกแยกเป็นคนละกลุ่ม รวมไปถึงต้องจัดกลุ่มใหม่ทุกครั้งเมื่อมีความต้องการกลุ่มข้อมูลขนาดใหญ่สำหรับไปใช้งาน ดังนั้นขั้นตอนวิธีการ DyCluStream+ จึงได้ถูกพัฒนาต่อยอดจากขั้นตอนวิธี DyCluStream ดังนี้

- ขั้นตอนวิธีการ DyCluStream ในส่วนออนไลน์ จะมีการปรับปรุงดังนี้

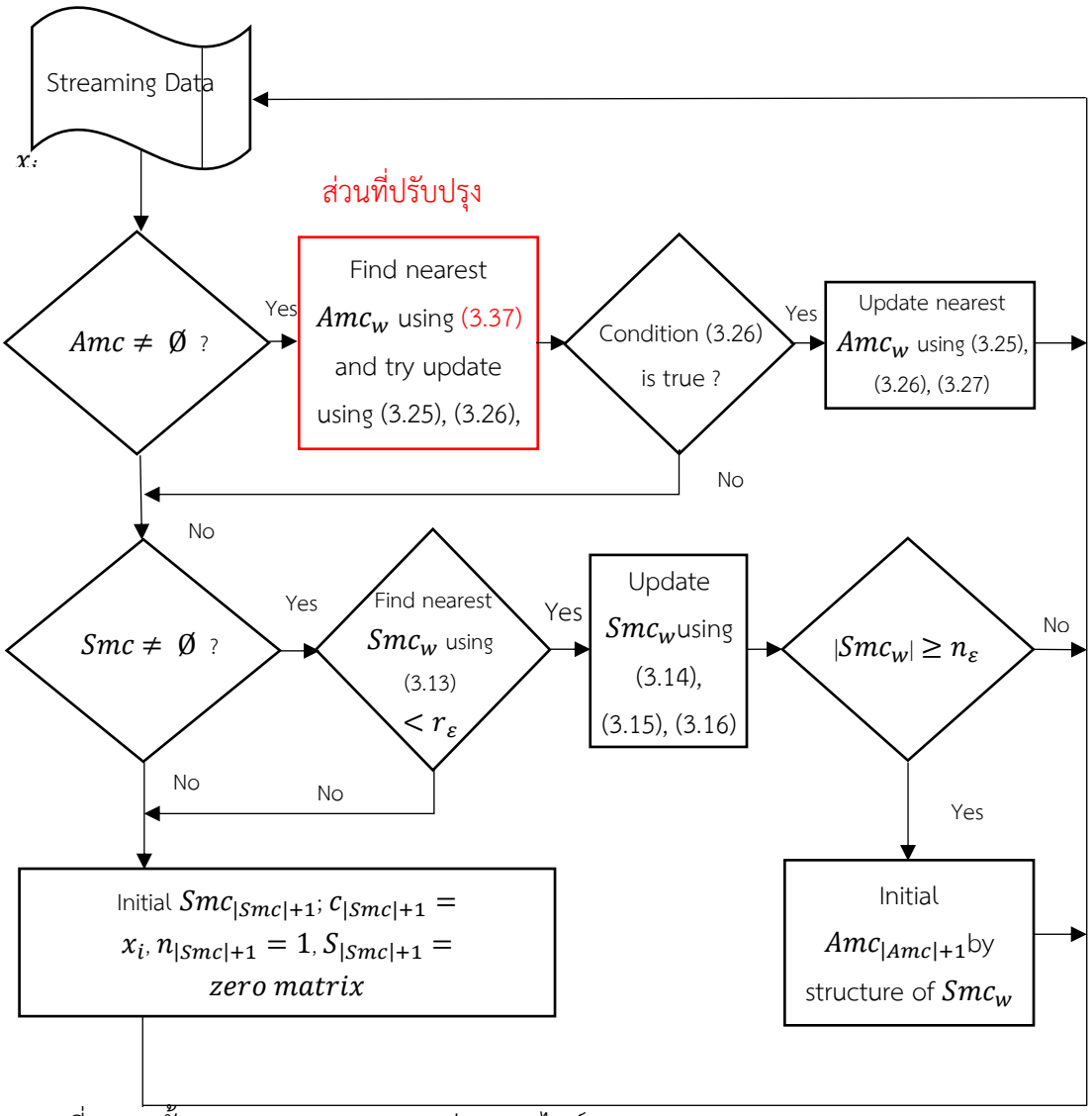
1. เพิ่มวิธีประมาณค่า Covariance Matrix เพื่อใช้สำหรับการคำนวณสมการต่าง ๆ
2. เปลี่ยนแปลงมาตรวัดระยะทางให้เหมาะสมกับลักษณะของข้อมูล

- ขั้นตอนวิธีการ DyCluStream ในส่วนออฟไลน์ จะมีการปรับปรุงดังนี้

1. ปรับเปลี่ยนเงื่อนไขในการเชื่อมต่อกันของ Amc เพื่อเพิ่มความถูกต้องแม่นยำ
2. การรวม Amc ที่มีคุณลักษณะใกล้เคียงกันเพื่อลดหน่วยความจำที่ใช้ในการจัดเก็บ
3. สามารถจัดกลุ่มข้อมูลแบบต่อยอดเพื่อลดเวลาที่ใช้ในการจัดกลุ่มข้อมูล

3.4.1 การปรับปรุงขั้นตอนส่วนออนไลน์

การปรับปรุงขั้นตอนส่วนออนไลน์ สามารถแสดงส่วนที่ปรับปรุงได้ดังภาพที่ 3-23

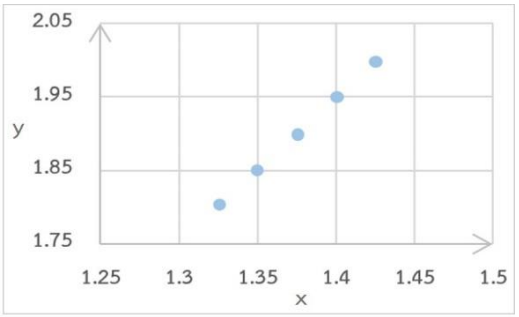


ภาพที่ 3-23 ขั้นตอน DyCluStream+ ส่วนออนไลน์

- **การประมาณค่า Covariance Matrix**

เนื่องจากขั้นตอนวิธีการ DyCluStream ใช้สมการที่เกี่ยวข้องกับ dynamic hyper-ellipsoidal ซึ่งในหลายงานวิจัยที่ใช้ dynamic hyper-ellipsoidal เป็นกลุ่มข้อมูลขนาดเล็กมักจะพบปัญหาที่เรียกว่า Not Positive definite โดยสังเกตได้จากค่า Eigen value น้อยกว่าหรือเท่ากับ 0 ซึ่งเป็นปัญหาที่พบเนื่องมาจาก Covariance Matrix ที่ได้จากกลุ่มข้อมูลที่มีลักษณะดังต่อไปนี้

- ค่าของตัวแปรใด ๆ มีความสัมพันธ์ในลักษณะเชิงเส้นกับตัวแปรอื่น ดังภาพที่ 3-24
- ข้อมูลตัวอย่างมีจำนวนน้อยกว่ามิติของข้อมูล (M.Z. R., T. Li, Y. Yang, & H. Wang, 2014) โดยงานวิจัย VHEC (N. Wattanakitrungroj, S. Maneeroj & C. Lursinsap, 2017) กำหนดให้คำนวณ Covariance matrix เมื่อจำนวนของข้อมูลมีจำนวนมากกว่ามิติของข้อมูล
- ตัวแปรบางตัวมีลักษณะเป็นค่าเดียวกัน ดังภาพที่ 3-25 สามารถสังเกตได้จากคอลัมน์ที่ V4, V5, V7, V8, V9, V10, V11 ของข้อมูลแต่ละตัวอย่างที่มีค่าเป็น 0 ทั้งหมด



ภาพที่ 3-24 ตัวอย่างของข้อมูลที่ตัวแปรใด ๆ มีความสัมพันธ์เชิงเส้น

sample	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
1	0.0003943150	0.000000493239	0.000207935	0	0	0.2	0	0	0	0	0
2	0.0000171441	0.000000956192	0.000063040	0	0	0	0	0	0	0	0
3	0	0.000000046151	0	0	0	0	0	0	0	0	0
4	0	0.000000274022	0.000120649	0	0	0	0	0	0	0	0
5	0	0.000000356228	0.000751823	0	0	0	0	0	0	0	0
6	0	0.000000328826	0.000150132	0	0	0	0	0	0	0	0
7	0	0.000000393726	0.002218421	0	0	0	0	0	0	0	0
8	0	0.000000341806	0.000287268	0	0	0	0	0	0	0	0
9	0	0.000000395168	0.001830872	0	0	0	0	0	0	0	0
10	0	0.000000252388	0.000044613	0	0	0	0	0	0	0	0
11	0	0.000000297097	0.000116769	0	0	0	0	0	0	0	0
12	0	0.000000449973	0.000805746	0	0	0	0	0	0	0	0

ภาพที่ 3-25 ตัวอย่างของข้อมูลที่ตัวแปรมีลักษณะเป็นค่าเดียวกัน

ในงานวิจัยนี้ได้ประยุกต์ใช้วิธีการจากงานวิจัยเรื่อง **Computing the nearest correlation matrix - a problem from finance** (N.J. Higham, 2002) ที่ได้นำเสนอวิธีในการแก้ปัญหาโดยวิธีการประมาณค่า Covariance Matrix ที่มีปัญหา Not Positive Definite ให้มีค่าใกล้เคียงกับ Positive Definite มากที่สุด เพื่อทำให้สามารถใช้ในการคำนวณในสมการต่าง ๆ ของงานวิจัยนี้ได้ โดยใช้ Frobenius norm ดังสมการที่ 3.36 ในการหาค่าระยะทางที่ใกล้สุดเพื่อใช้ในการประมาณค่า Nearest Positive Definite

$$FrobDist(S) = \sum_{i=1}^{m}\sum_{j=1}^{n}|S_{ij}|^2 \tag{3.36}$$

โดยที่ $|S|$ คือ ค่าที่เป็นบวก (Absolute) ของ Covariance Matrix
ลำดับที่ i^{th} มิติที่ j^{th}
m, n คือ ขนาดมิติของ Covariance Matrix $(m \times n)$

การประมาณค่า Nearest Positive Definite จะนำไปใช้ในการแก้ปัญหา Not Positive Definite ที่เกิดจากการคำนวณของสมการที่ (3.24), (3.26), (3.30), (3.33) มีขั้นตอนวิธีการดังภาพที่ 3-26

Algorithm nearPD(S)

$ds = zero\ matrix_{dxd}$; d is the number of dimensions.

$w = ones\ matrix_{dxd}$; d is the number of dimensions.

$tol = d \times 2.220446049250313\mathrm{e}^{-16}$

$max\_iterations = 100$

$X = S, Y = S$

$rel\_diffY = \infty, rel\_diffX = \infty, rel\_diffXY = \infty$

$Whalf = sqrt(w \times w^T)$

$iteration = 0$

while $max(rel\_diffX, rel\_diffY, rel\_diffXY) > tol$ && $iteration < 100$:

  $iteration \mathrel{+}= 1$

  $Xold = X$

  $R = X - ds$

  $R\_wtd = Whalf \times R$

  $X =$ proj_spd(R_wtd)

  $X = X / Whalf$

  $ds = X - R$

  $Yold = Y$

  $Y = X$

  Fill the diagonal element of matrix Y with 1

  $normY = FrobDist(Y)$

  $rel\_diffX = FrobDist(X - Xold)/FrobDist(X)$

  $rel\_diffY = FrobDist(Y - Yold)/ normY$

  $rel\_diffXY = FrobDist(X - Yold) / normY$

  $X = Y$

return X

Algorithm proj_spd(S):

Find the eigen values λ and the eigen vectors u

$S = \left(u \times maximum(\lambda, 0)\right) \times u^T$

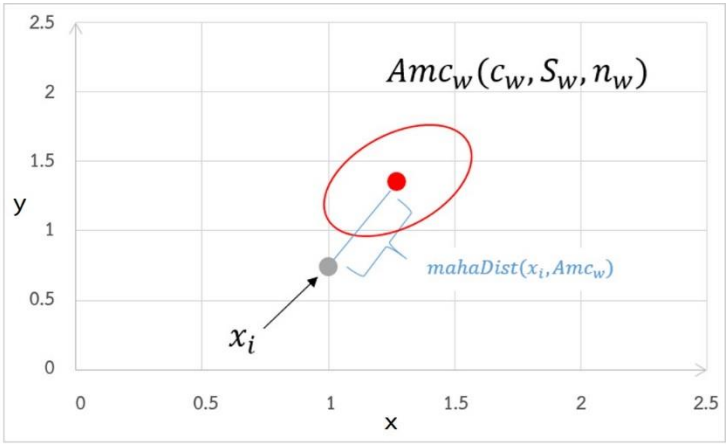
$S = (S + S^T) / 2$

return S

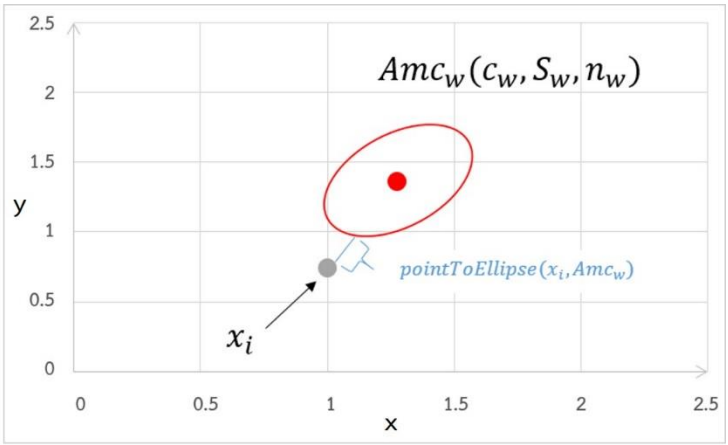
ภาพที่ 3-26 ขั้นตอนวิธีการประมาณค่า Covariance Matrix

- **การปรับปรุงระยะทาง**

เนื่องจากระยะทางที่ใช้เป็นมาตรวัดระหว่างกลุ่มข้อมูลขนาดเล็กที่มีรูปทรงเป็น dynamic hyper-ellipsoidal และข้อมูลที่เข้ามาใช้มาตรวัดระยะทาง $mahaDist(x_i, Amc_w)$ ที่เป็นมาตรวัดที่วัดจากจุดศูนย์กลางของ dynamic hyper-ellipsoidal micro-cluster ดังภาพที่ 3-27 ซึ่งยังเป็นค่าระยะทางที่ไม่เหมาะสม เนื่องจากระยะที่ได้ควรเป็นระยะทางที่วัดจากขอบของ dynamic hyper-ellipsoidal micro-cluster ที่มีค่าแม่นยำ ดังภาพที่ 3-28



ภาพที่ 3-27 มาตรวัดระยะทาง $mahaDist(x_i, Amc_w)$



ภาพที่ 3-28 มาตรวัดระยะทาง $pointToEllipse(x_i, Amc_w)$

งานวิจัยนี้จึงได้ประยุกต์ใช้มาตรวัดระยะทางระหว่างกลุ่มข้อมูลขนาดเล็กรูปทรง dynamic hyper-ellipsoidal กับข้อมูลที่เข้ามาใหม่ จากงานวิจัย Versatile Hyper-Elliptic Clustering

Approach for Streaming Data Based on One-Pass-Thrown-Away Learning สามารถแสดงได้ดังสมการที่ (3.37)

$$pointToEllipse(x_i, Amc_w) = || c_w - x_i|| \times [1 - \frac{1}{((x_i - c_w)^T S_w^{-1}(x_i - c))^{1/2}}] \quad (3.37)$$

โดย x_i คือ ข้อมูลนำเข้า

c_w คือ จุดศูนย์กลางของกลุ่มข้อมูลขนาดเล็ก Amc_w

S_w^{-1} คือ Inverse Covariance matrix ของกลุ่มข้อมูลขนาดเล็ก Amc_w

$|| c_w - x_i||$ คือ Euclidean Distance ระหว่าง c_w และ x_i

การปรับปรุงขั้นตอนส่วนออฟไลน์สามารถแสดงภาพที่ 3-29

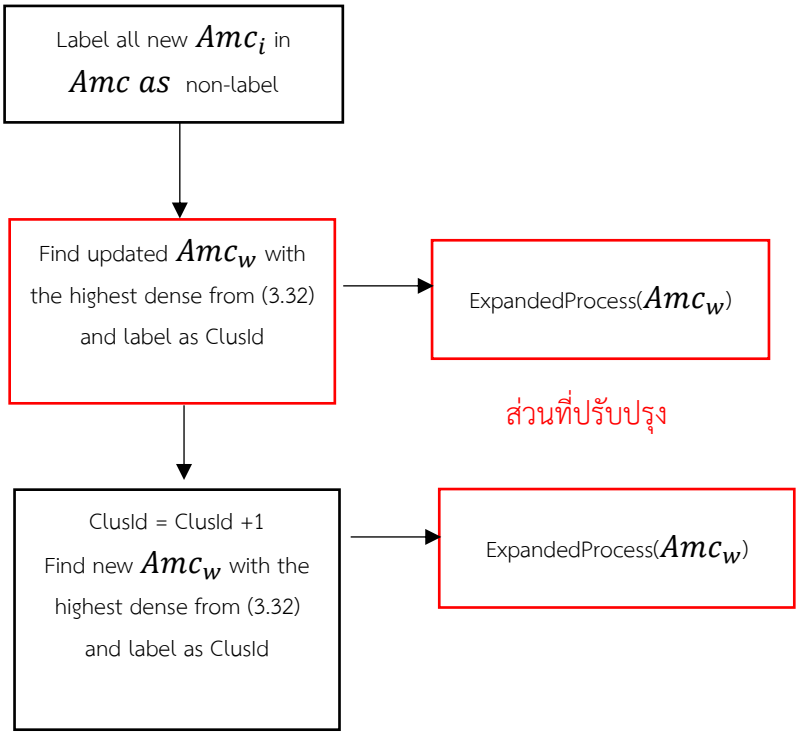
Algorithm Online Phase($x_i, r_\varepsilon, n_\varepsilon$)

1: Get the arriving data point x_i
2: **if** $Amc \neq \emptyset$ **then**
3: Find the winning Actual micro-cluster Amc_w using (3.25)
4: **if** x_i is satisfies the conditions in eq. (3.37) with respect Amc_w **then**
5: Update the parameters of Amc_w from eq. (3.27), (3.28), (3.29)
6: **return**
7: **end**
8: **end**
9: **If** $Smc \neq \emptyset$ **then**
10: Find the winning Small micro-cluster Smc_w using eq. (3.14)
11: **If** $mahaDist(x_i, Smc_w) \leq r_\varepsilon$ then
12: Update the parameters of Smc_w using eq. (3.15), (3.16), (3.17)
13: **If** $n_w > n_\varepsilon$ then
14: Create a new $Amc_{|Amc|+1}$ by Smc_w using eq. (3.18), (3.19), (3.20)
15: **return**
16: **end**
17: **return**
18: **end**
19: **end**
20: Create new $Smc_{|Smc|+1}$ by the arriving data point x_i using eq. (3.20), (3.21), (3.22)
21: **return**

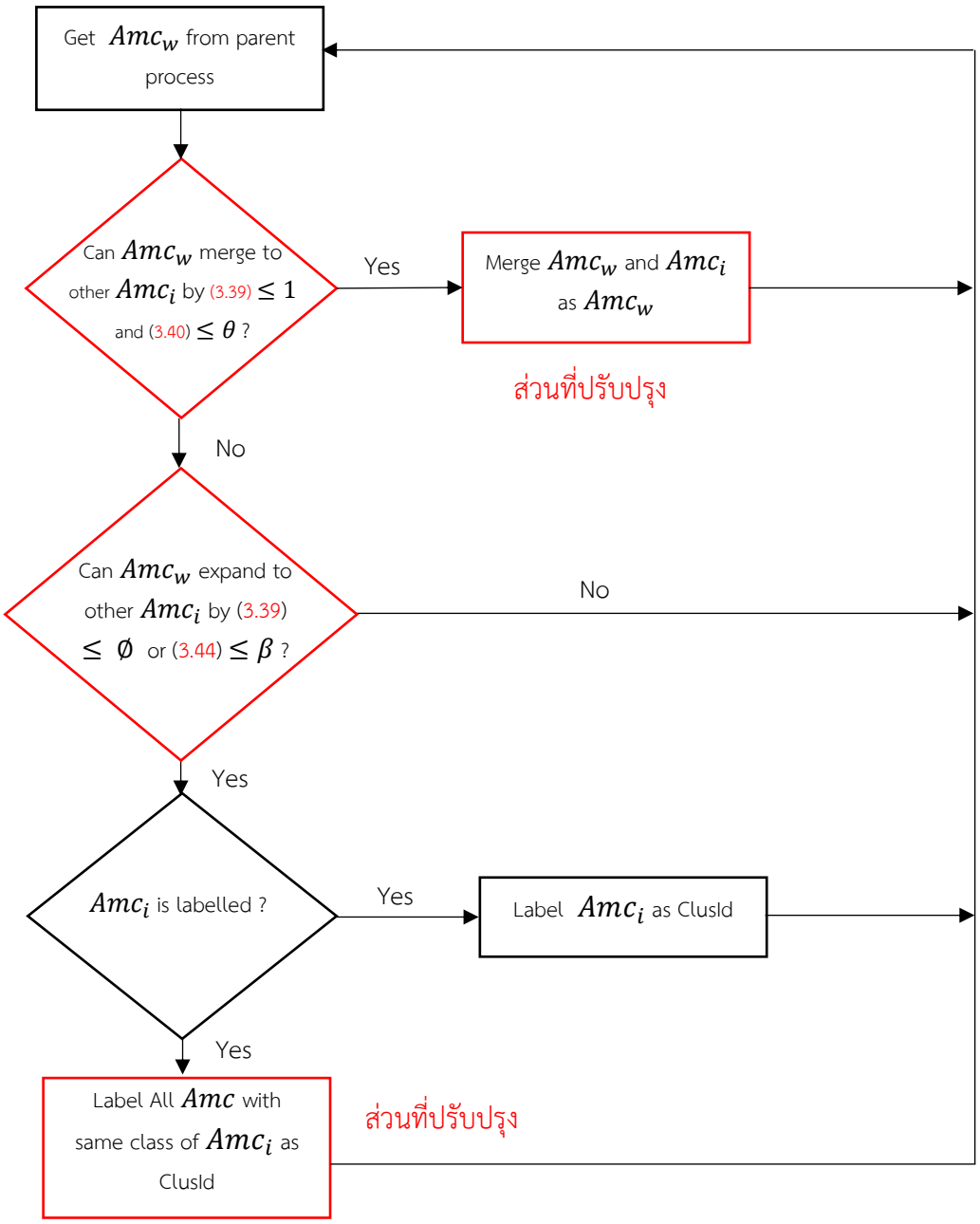
ภาพที่ 3-29 ขั้นตอน DyCluStrem+ ส่วนออนไลน์

3.4.2 การปรับปรุงขั้นตอนส่วนออฟไลน์

การปรับปรุงขั้นตอนส่วนออฟไลน์ สามารถแสดงส่วนที่ปรับปรุงได้ดังภาพที่ 3-30 และ 3-31



ภาพที่ 3-30 ขั้นตอน DyCluStrem+ ส่วนออฟไลน์



ภาพที่ 3-31 ขั้นตอน DyCluStrem+ ส่วนออฟไลน์ (ExpandedProcess)

- **เพิ่มมาตรวัดระยะทาง** เนื่องจากมาตรวัดระยะจากสมการ (3.33) ไม่ได้คำนึงถึงระดับของการซ้อนทับกันของข้อมูล DyCluStream+ จึงได้เพิ่มมาตรวัดระยะทาง Bhatthacharyya ดังสมการที่ (3.39) ซึ่งเป็นมาตรวัดระยะทางที่คำนึงถึงระดับการซ้อนทับกันของข้อมูลจึงช่วยให้การจัดกลุ่มข้อมูลมีความถูกต้องมากยิ่งขึ้น

$$S = \frac{S_w + S_i}{2} \tag{3.38}$$

$$B(Amc_w, Amc_i) = \frac{1}{8}(c_w - c_i)^T S^{-1}(c_w - c_i) + \frac{1}{2} ln \left(\frac{det\, S}{\sqrt{det\, S_w \; det\, S_i}}\right) \quad (3.39)$$

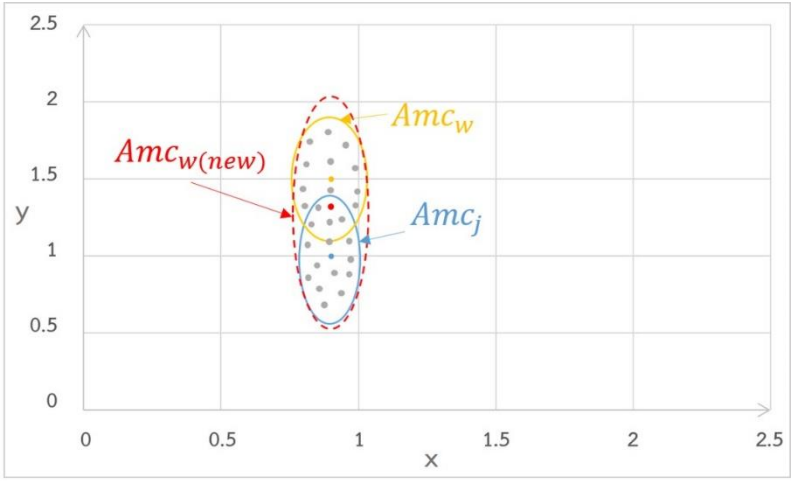
โดยที่ Amc_w, Amc_j คือ โครงสร้างกลุ่มข้อมูลขนาดเล็ก w^{th} และ i^{th}

c_w, c_i คือ เวกเตอร์จุดศูนย์กลางของกลุ่มข้อมูลขนาดเล็กAmc_w และ Amc_i

S_w, S_i คือ Covariance Matrix ของกลุ่มข้อมูลขนาดเล็ก Amc_w และ Amc_i

S^{-1} คือ Inverse Covariance matrix

● **การรวมกลุ่มข้อมูลขนาดเล็ก** การเก็บคุณลักษณะของข้อมูลให้อยู่ในรูปกลุ่มข้อมูลขนาดเล็ก แม้จะช่วยลดหน่วยความจำที่ใช้ในการเก็บข้อมูล แต่ถ้ามีกลุ่มข้อมูลขนาดเล็กที่มีคุณลักษณะที่เหมือนกันหรือคล้ายกันจำนวนมาก การรวมกลุ่มข้อมูลขนาดเล็กที่เหมือนหรือคล้ายกันแล้วแทนด้วยกลุ่มข้อมูลใหม่ ก็จะช่วยลดหน่วยความจำที่ใช้ในการจัดเก็บข้อมูล และยังเพิ่มความเร็วที่ใช้ในการจัดกลุ่มข้อมูลเนื่องจากจำนวนกลุ่มข้อมูลขนาดเล็กมีขนาดลดลงดังภาพที่ 3-32 โดยเงื่อนไขที่ใช้ในการกำหนดความคล้ายคลึงกันของลุ่มข้อมูลขนาดเล็ก Amc_i ใด ๆ คือ cosine similarity ซึ่งเป็นการวัดค่าความคล้ายคลึงกันในเรื่องขององศาของ dynamic hyper-ellipsoidal micro-clusters สองกลุ่มที่ทำมุมกันดังสมการที่ (3.40) และ Bhatthacharyya distance ซึ่งเป็นมาตรวัดระยะทางที่คำนึงถึงความหนาแน่นของกลุ่มข้อมูลขนาดเล็ก



ภาพที่ 3-32 การผสานกลุ่มข้อมูลขนาดเล็ก

$$\cos(Amc_w, Amc_i) = \left|\frac{u_w \cdot u_i}{||u_w||_2 \; ||u_i||_2}\right| \quad (3.40)$$

โดยที่ u_w, u_i คือ Eigen vector ที่สอดคล้องกับ Eigen value ที่มีค่ามากที่สุดของ Amc_w และ Amc_i

$||u_w||_2$ คือ Euclidean norm ของ u_w และ u_i

$\left|\frac{u_w \cdot u_i}{||u_w||_2 \, ||u_i||_2}\right|$ คือ ค่าที่เป็นบวกของ cosine similarity

หากค่าที่ได้จากสมการที่ (3.39) และ (3.40) น้อยกว่าหรือเท่ากับที่ผู้ใช้กำหนด จะรวมกลุ่มข้อมูลขนาดเล็ก Amc_w และ Amc_i เข้าด้วยกัน ดังสมการที่ (3.41), (3.42) และ (3.43)

$$c_{w(new)} = \frac{n_w c_w + n_i c_i}{n_w + n_i} \tag{3.41}$$

$$S_{w(new)} = \frac{n_w}{n_w+n_i} S_w + \frac{n_i}{n_w+n_i} S_j + \frac{n_w n_i}{(n_w+n_i)^2} \times (c_w - c_i)(c_w - c_i)^T \tag{3.42}$$

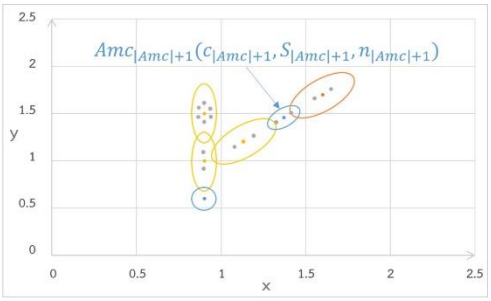
$$n_{w(new)} = n_w + n_i \tag{3.43}$$

แต่ถ้าค่าที่ได้จากสมการที่ (3.38) และ (3.39) มากกว่าที่ผู้ใช้กำหนด จะตรวจสอบว่ากลุ่มข้อมูล Amc_w สามารถแผ่ขยายออกไปได้หรือไม่ ตามเงื่อนไขดังสมการที่ (3.38) และ (3.43)

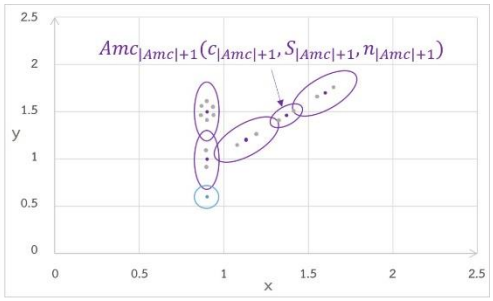
$$\delta(Amc_w, Amc_i) = \sum_{k=1}^{d} \frac{((c_w - c_i)^T \mu_{ik})^2}{(2 * a_{ik})^2} - 1 \tag{3.44}$$

ถ้าค่าที่ได้จากสมการที่ (3.38) และ (3.43) น้อยกว่าหรือเท่ากับที่ผู้ใช้กำหนด ก็จะทำการแผ่ขยายกลุ่มข้อมูล Amc_w ต่อไป แต่ถ้ามากกว่าก็จะหยุดการแผ่ขยาย เมื่อไม่สามารถแผ่ขยายเพื่อสร้างกลุ่มข้อมูลได้แล้ว ก็จะเริ่มสร้างกลุ่มใหม่โดยมี Amc_w ที่มีความหนาแน่นมากที่สุดและยังไม่ถูกกำหนดกลุ่มเป็นจุดเริ่มต้นในการแผ่ขยายต่อไป

- **การปรับปรุงขั้นตอนส่วนออฟไลน์ให้สามารถทำงานแบบต่อยอด** เนื่องจากกระบวนการในการจัดกลุ่มข้อมูลนั้นต้องดำเนินการจัดกลุ่มข้อมูลขนาดเล็กใหม่ทั้งหมดเมื่อต้องการกลุ่มข้อมูลสำหรับนำไปใช้จริง ทำให้เสียเวลาในการจัดกลุ่มข้อมูลขนาดเล็กที่เคยได้จัดกลุ่มไปแล้วทั้งหมด ดังนั้นใน DyCluStream+ จึงได้มีการนำเสนอกระบวนการในการจัดกลุ่มแบบต่อยอด โดยจะทำงานในลักษณะคล้ายคลึงกับ DyCluStream แต่จะดำเนินการจัดกลุ่มกับกลุ่มข้อมูลขนาดเล็กที่ได้รับการปรับปรุงจากกระบวนการในการปรับปรุงกลุ่มข้อมูลขนาดเล็กจากขั้นตอนในส่วนออนไลน์เท่านั้น ดังภาพที่ 3-33 เพื่อประหยัดเวลาในการหากลุ่มข้อมูลที่จะนำไปใช้จริง



ก) $Amc_{|Amc|+1}$ ที่สร้างขึ้นใหม่หรือ Amc_w ถูกปรับปรุงจากกระบวนการในส่วนออนไลน์



ข) การจัดกลุ่มแบบต่อยอดโดยการแผ่ขยาย จาก $Amc_{|Amc|+1}$ที่สร้างขึ้นใหม่หรือ Amc_w ถูกปรับปรุง

ภาพที่ 3-33 การจัดกลุ่มแบบต่อยอด

โดยจะแบ่งกรณีที่เกิดขึ้นจากกระบวนการในการปรับปรุงกลุ่มข้อมูลขนาดเล็กจากขั้นตอนในส่วนออนไลน์ออกเป็น 2 กรณี ดังต่อไปนี้

กรณีที่ 1 Amc_w ถูกปรับปรุง เนื่องจากสามารถครอบคลุมข้อมูลที่เข้ามาได้

กรณีที่ 2 มีการสร้าง $Amc_{|Amc|+1}$ ขึ้นมาใหม่ โดยเกิดจากกรณีของ Smc_w จากกระบวนการในการปรับปรุงกลุ่มข้อมูลขนาดเล็กในรอบก่อนหน้าหรือเกิดจากกระบวนการในการปรับปรุงกลุ่มข้อมูลขนาดเล็กในรอบนี้มีความหนาแน่นตามที่ผู้ใช้ได้กำหนด จึงกลายเป็น $Amc_{|Amc|+1}$

ขั้นตอนส่วนออฟไลน์ที่สามารถทำงานแบบต่อยอดมีวิธีการดังนี้

ขั้นตอนส่วนออฟไลน์จะแผ่กระจายจาก Amc_w ที่ได้รับผลกระทบที่มีความหนาแน่นมากที่สุดตามสมการที่ (3.32) หากสามารถแผ่กระจายกับ Amc_i รอบๆ ได้จะดำเนินการดังนี้

ถ้า Amc_i ที่ถูกแผ่มายังไม่มีการกำหนดกลุ่ม กลุ่มข้อมูล Amc_i นั้นจะถูกกำหนดกลุ่มและดำเนินการแผ่ขยายออกไปดังเช่นกระบวนการในการจัดกลุ่มข้อมูลส่วนออฟไลน์ในหัวข้อที่ **3.3.2**

ถ้า Amc_i ที่ถูกแผ่มาถูกกำหนดกลุ่มไว้แล้ว จะทำการกำหนด Amc_i ที่ถูกแผ่มาและ Amc ที่มีกลุ่มเหมือนกับ Amc_i ที่ถูกแผ่มา ให้เป็นกลุ่มเดียวกันกับกลุ่มข้อมูลที่เป็นตัวแผ่กระจาย

ขั้นตอนวิธีการในส่วนออฟไลน์ที่ได้รับการปรับปรุง สามารถอธิบายได้ดังภาพที่ 3-34 และ 3-35

Algorithm ClusteringAmcbyIncremental(Amc)

1: Label all new Amc as non-labelled

2: **for** each updated Amc_w $with\ most\ dense$ from eq.(3.32) ∈ Amc **do**

3: ExpandClusAmcbyIncremental($ClusId$ of Amc_w, Amc_w)

4: **end**

5: **for** each new non-labelled Amc_w $with\ most\ dense$ from eq.(3.32) ∈ Amc **do**

6: Incremental $ClusId$ by 1

7: Label Amc_w as $ClusId$

8: ExpandClusAmcbyIncremental($ClusId, Amc_w, Amc$)

9: **end**

ภาพที่ 3-34 ขั้นตอนวิธีการในการจัดกลุ่มข้อมูลแบบต่อยอด

Algorithm ExpandClusAmcbyIncremental($ClusId, Amc_w$)

10: **for** each $Amc_i \in Amc$ **do**

11: **if** ($B(Amc_w, Amc_i) \leq 1$) from eq. (3.39) And $\cos(Amc_w, Amc_i) \geq \theta$) from eq. (3.40) **then**

12: merge Amc_w, Amc_i as Amc_w

13: **else if**(($B(Amc_w, Amc_i) \leq \beta$) from eq. (3.39) or ($\delta(Amc_w, Amc_i$) from eq. (3.44) $\leq \emptyset$))

14: **if** ($non\_labelled\ Amc_i$) **then**

15: Label Amc_i as $ClusId$

16: ExpandClusAmc(ClusId, Amc_i)

17: **else if** ($labelled\ Amc_i$) **then**

18: Label All Amc with same class of Amc_i as $ClusId$

19: **end**

20: **end**

21: **end**

ภาพที่ 3-35 ขั้นตอนวิธีการในการแผ่ขยายกลุ่มข้อมูลแบบต่อยอด

บทที่ 4

ผลการดำเนินงาน

ในบทนี้จะกล่าวถึงผลการดำเนินงานโดยการวัดประสิทธิภาพความถูกต้องและเวลาที่ใช้ในการจัดกลุ่มของวิธีการที่นำเสนอ The Dynamic Clustering Algorithm for Streaming Data Using One- Pass- Throw- Away Datum and Hyper- ellipsoidal Micro- Clustering (DyCluStream) และ The Dynamic Clustering Algorithm for Streaming Data Using One-Pass-Throw-Away Datum and Hyper-ellipsoidal Micro-Clustering Plus (DyCluStream+) เปรียบเทียบกับวิธีการ Density-based clustering over an evolving data stream with noise (DenStream) และ Hyper-ellipsoidal clustering technique for evolving data stream (HECES) กับข้อมูลสังเคราะห์และข้อมูลจริง ด้วยวิธีการทดลองบนเครื่องคอมพิวเตอร์ CPU core i5-7500 3.40 GHz และหน่วยความจำ (RAM) 8 GB โดยทำงานบนระบบปฏิบัติการ Windows 10 ซึ่งเขียนโปรแกรมด้วยภาษาโดย R โดยผลลัพธ์ของการจัดกลุ่ม ใช้วิธีการวัดประสิทธิภาพความถูกต้อง Purity, Rand statistic และ Jaccard Index สำหรับพิจารณาความถูกต้องในการจัดกลุ่มของข้อมูล โดยรายละเอียดของผลการดำเนินงานแสดงดังต่อไปนี้

4.1) ข้อมูลที่ใช้ในการทดลอง

4.2) วิธีการวัดประสิทธิภาพความถูกต้อง

4.3) การออกแบบการทดลอง

4.1 ข้อมูลที่ใช้ในการทดลอง

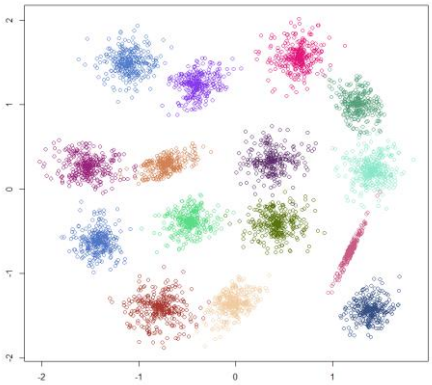
งานวิจัยนี้ได้เปรียบเทียบผลการทดลองทั้งจากข้อมูลจริงและข้อมูลสังเคราะห์ แต่เนื่องจากข้อมูลที่ใช้ไม่ได้มีคุณลักษณะแบบกระแสข้อมูล งานวิจัยนี้จึงได้จำลองข้อมูลที่เข้ามาให้มีคุณลักษณะแบบกระแสข้อมูลโดยวิธีการนำข้อมูลเข้าสู่ขั้นตอนวิธีการส่วนออนไลน์ทีละ 1 ตัวอย่างข้อมูลแบบต่อเนื่อง

4.1.1 ข้อมูลสังเคราะห์

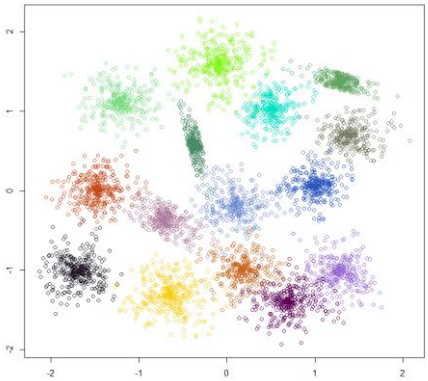
ข้อมูลสังเคราะห์ที่ใช้ในการทดลองมีคุณลักษณะดังนี้ กลุ่มข้อมูลจะมีลักษณะการกระจายตัวที่ไม่เท่ากัน ข้อมูลมีคุณลักษณะของการกระจายตัวแบบ Gaussian กลุ่มข้อมูลมีการซ้อนทับกันกับกลุ่มข้อมูลอื่น กลุ่มข้อมูลภายในถูกล้อมรอบด้วยกลุ่มข้อมูลภายนอก มีสัญญาณรบกวน (noise) อยู่ในตัวอย่างข้อมูล โดยข้อมูลสังเคราะห์มีรายละเอียดดังนี้

- **ชุดข้อมูล S1, S2, S3**

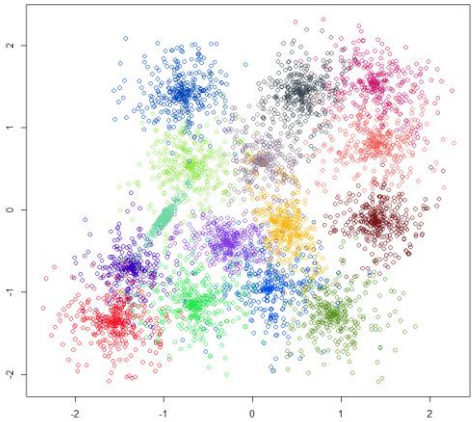
เป็นข้อมูลที่มีลักษณะการกระจายตัวที่ไม่เท่ากันในแต่ละกลุ่มข้อมูล มีคุณลักษณะของการกระจายตัวแบบ Gaussian และมีระดับของการซ้อนทับกันของกลุ่มข้อมูลที่ไม่เท่ากัน โดยข้อมูลมีคุณลักษณะ 2 คุณลักษณะ มีจำนวนข้อมูลทั้งหมด 5,000 ตัวอย่าง ซึ่งมีกลุ่มของข้อมูลทั้งหมด 15 กลุ่ม โดยเป็นข้อมูลที่ได้มาจากแหล่งข้อมูลสาธารณะ https://cs.joensuu.fi/sipu/datasets/ ชุดข้อมูล S1, S2 และ S3 มีคุณลักษณะดังภาพที่ 4-1 ก), 4-1 ข) และ 4-1 ค) ตามลำดับ



ก) คุณลักษณะข้อมูล S1



ข) คุณลักษณะข้อมูล S2



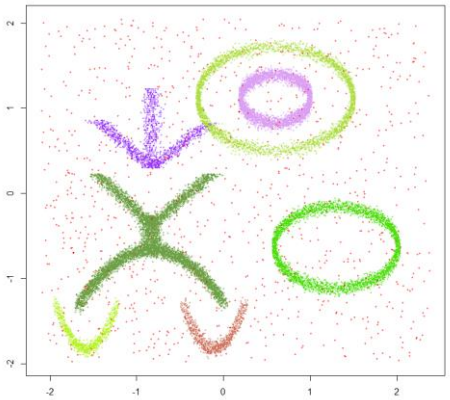
ค) คุณลักษณะข้อมูล S3

ภาพที่ 4-1 คุณลักษณะข้อมูล S3

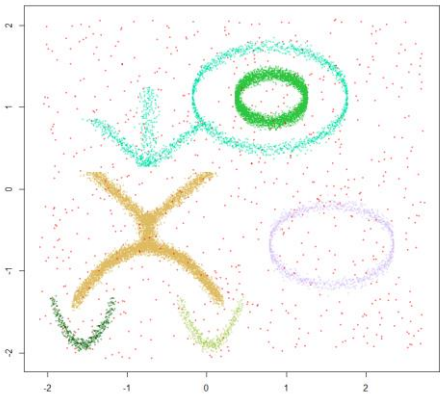
- **ชุดข้อมูล AbitaryHCM1, AbitaryHCM2**

เป็นข้อมูลที่มีลักษณะรูปร่างไม่แน่นอน กลุ่มของข้อมูลบางกลุ่มไม่มีการแบ่งแยกอย่างชัดเจนจากกัน และมีลักษณะของกลุ่มข้อมูลภายในถูกล้อมรอบด้วยกลุ่มข้อมูลภายนอก โดยข้อมูลประกอบไปด้วยคุณลักษณะ 2 คุณลักษณะ โดย **AbitaryHCM1** มีจำนวนตัวอย่างข้อมูลทั้งหมด 20,366

ตัวอย่าง มีกลุ่มของข้อมูลทั้งหมด 7 กลุ่ม และ **AbitaryHCM2** มีจำนวนตัวอย่างข้อมูลทั้งหมด 15,040 ตัวอย่าง มีกลุ่มของข้อมูลทั้งหมด 6 กลุ่ม โดยข้อมูลทั้งสองชุดข้อมูลมี noise 4.7% จากตัวอย่างข้อมูลทั้งหมด ซึ่งเป็นข้อมูลที่ถูกสังเคราะห์ขึ้นเพื่อการทดสอบ ชุดข้อมูล AbitaryHCM1 และ AbitaryHCM2 มีคุณลักษณะดังภาพที่ 4-2 ก) และ 4-2 ข) ตามลำดับ



ก) คุณลักษณะข้อมูล AbitaryHCM1



ข) คุณลักษณะข้อมูล AbitaryHCM2

ภาพที่ 4-2 คุณลักษณะข้อมูล AbitaryHCM1, AbitaryHCM2

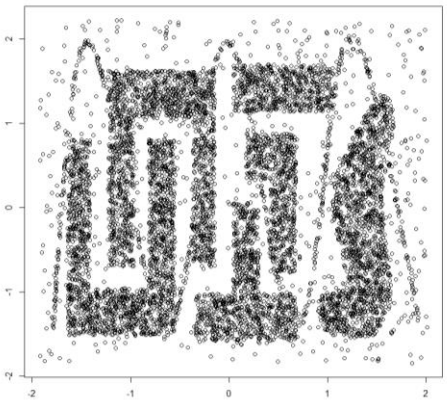
4.1.1.1 ชุดข้อมูล t4.8k, t7.10k, t8.8k

เป็นข้อมูลที่มีลักษณะรูปร่างไม่แน่นอน มี noise รวมอยู่กับกลุ่มข้อมูลซึ่งอาจทำให้เกิดความยากในการแบ่งกลุ่มข้อมูลออกจากกัน โดยข้อมูลประกอบไปด้วยคุณลักษณะ 2 คุณลักษณะ เป็นข้อมูลที่ได้มาจากแหล่งข้อมูลสาธารณะ https://cs.joensuu.fi/sipu/datasets/ โดยแต่ละชุดข้อมูลมีคุณลักษณะที่แตกต่างกันดังนี้

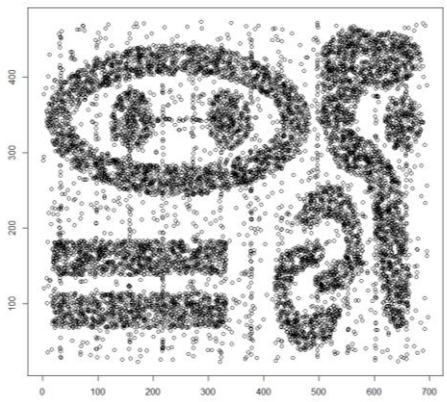
- **ชุดข้อมูล t4.8k** มีจำนวนข้อมูลทั้งหมด 8,000 ตัวอย่าง ซึ่งมีกลุ่มของข้อมูลทั้งหมด 6 กลุ่ม โดยกลุ่มข้อมูลรูปร่างที่ไม่แน่นอน โดยมีคุณลักษณะดังภาพที่ 4-3 ก)

- **ชุดข้อมูล t7.10k** มีจำนวนข้อมูลทั้งหมด 10,000 ตัวอย่าง ซึ่งมีกลุ่มของข้อมูลทั้งหมด 9 กลุ่ม โดยมีกลุ่มข้อมูลที่ถูกล้อมรอบด้วยกลุ่มข้อมูลอื่น โดยมีคุณลักษณะดังภาพที่ 4-3 ข)

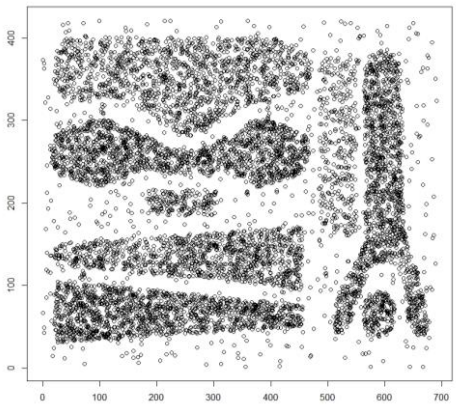
- **ชุดข้อมูล t8.8k** มีจำนวนข้อมูลทั้งหมด 8,000 ตัวอย่าง ซึ่งมีกลุ่มของข้อมูลทั้งหมด 8 กลุ่ม โดย โดยกลุ่มข้อมูลแต่ละกลุ่มอยู่ใกล้ชิดกันและมีความหนาแน่นไม่เท่ากัน โดยมีคุณลักษณะดังภาพที่ 4-3 ค)



ก) คุณลักษณะของข้อมูล t4.8k



ข) คุณลักษณะของข้อมูล t7.10k



ค) คุณลักษณะของข้อมูล t8.8k

ภาพที่ 4-3 คุณลักษณะของข้อมูล t4.8k, t7.10k, t8.8k

4.1.2 ข้อมูลจริง

งานวิจัยนี้ใช้ข้อมูลจริงในการทดลองที่ได้จากฐานข้อมูลที่เผยแพร่สาธารณะ https://archive.ics.uci.edu/ml/ ซึ่งประกอบไปด้วย

- **ชุดข้อมูล Statlog (Shuttle)**

ประกอบไปด้วย 9 คุณลักษณะที่เป็นตัวเลข ข้อมูลแบ่งออกเป็น 7 กลุ่ม โดยข้อมูลจำนวน 80% อยู่ในกลุ่มที่ 1 มีจำนวนข้อมูลทั้งหมด 43,500 ตัวอย่าง

- **ชุดข้อมูล Letter Recognition**

เป็นข้อมูลที่มีวัตถุประสงค์ในการทำนายกลุ่มของข้อมูลที่เป็นตัวอักษรภาษาอังกฤษจากข้อมูลที่อธิบายคุณลักษณะของจุดข้อมูลสีขาวและดำที่อยู่ในภาพจากแบบอักษรทั้งหมด 20 แบบ ซึ่งจะแต่ละแบบและแต่ละตัวอักษรจะถูกสุ่มเพื่อสร้างข้อมูลทั้งหมด 20,000 ตัวอย่าง

- **ชุดข้อมูล KDD Cup 1999**

เป็นข้อมูลที่ใช้ในการแข่งขันทางด้านการขุดเหมืองข้อมูล โดยมีคุณลักษณะของการติดต่อในการเชื่อมต่อข้อมูลทางอินเตอร์เน็ต ข้อมูลแบ่งออกเป็น 23 กลุ่ม แต่ละกลุ่มมี 42 คุณลักษณะ ซึ่งจะ

ใช้เฉพาะคุณลักษณะที่เป็นตัวเลขแค่เพียง 23 คุณลักษณะในการจัดกลุ่มข้อมูลมีข้อมูลทั้งหมด 494,020 ตัวอย่าง

- **ชุดข้อมูล Covertype**

เป็นข้อมูลที่ได้จากการสำรวจชนิดของป่าในพื้นที่ 30x30 ตารางเมตร ซึ่งประกอบไปด้วยข้อมูลที่เกี่ยวกับการทำแผนที่ 54 คุณลักษณะ ซึ่ง 44 คุณลักษณะเป็นข้อมูลเชิงคุณภาพที่เป็นเลขฐานสอง และ 10 คุณลักษณะเป็นข้อมูลเชิงปริมาณ มีกลุ่มข้อมูลทั้งหมด 7 กลุ่ม มีข้อมูลทั้งหมด 581,012 ตัวอย่าง คุณลักษณะของข้อมูล สามารถสรุปได้ดังตารางที่ 4-1

ตารางที่ 4-1 คุณลักษณะของข้อมูล

Datasets	Number of objects	Number of attributes	Number of clusters
Statlog (Shuttle)	43,500	9	7
Letter Recognition	20,000	16	26
KDD Cup 1999	494,020	34	23
Covertype	581,012	54	7

4.2 วิธีการวัดประสิทธิภาพความถูกต้องและหน่วยความจำที่ใช้ของวิธีการจัดกลุ่ม

ขั้นตอนการวัดประสิทธิภาพความถูกต้องของวิธีการจัดกลุ่มที่ใช้ในงานวิจัยนี้ ได้เลือกใช้วิธีการวัดประสิทธิภาพคือ

4.2.1 Purity

เป็นตัววัดประสิทธิภาพในเรื่องของการพิจารณาความถูกต้องของการจัดกลุ่มของข้อมูลโดยการวัดกลุ่มของข้อมูลว่ามีความเป็นหนึ่งเดียวกันหรือไม่ ดังสมการที่ (4.1)

$$Purity = \frac{1}{N}\sum_{i=1}^{k} max_j \left|C_i \cap t_j\right| \quad (4.1)$$

โดยที่ N คือ จำนวนของข้อมูล

k, j คือ จำนวนของกลุ่มข้อมูล

C_i คือ กลุ่มข้อมูลที่ได้จากการจัดกลุ่ม

t_j คือ กลุ่มข้อมูลที่ถูกต้องที่มีข้อมูลสมาชิกของกลุ่ม C_i

ค่าที่ได้จากตัววัดประสิทธิภาพ $Purity$ จะมีค่าอยู่ในช่วง 0 ถึง 1 โดยค่าของตัววัด $Purity$ ที่มีค่าเข้าใกล้ 1 หมายถึง ข้อมูลที่อยู่ในกลุ่มข้อมูลนั้นมีความเป็นหนึ่งเดียวกัน

4.2.2 Rand statistic หรือ Rand Index (RI)

เป็นตัววัดประสิทธิภาพที่พิจารณาความถูกต้องของการจัดกลุ่มของข้อมูล โดยการวัดอัตราส่วนของการจับคู่ข้อมูล 2 ตัวอย่างใด ๆ ที่ถูกจัดกลุ่มได้อย่างถูกต้อง จากการจับคู่ข้อมูล 2

ตัวอย่างที่เป็นไปได้ทั้งหมด หรือเรียกสามารถเรียกได้ว่า เป็นการวัดความแม่นยำของวิธีการจัดกลุ่มข้อมูล ดังสมการที่ (4.2)

$$RI = \frac{TP+TN}{TP+FP+FN+TN} \quad (4.2)$$

โดย TP คือ จำนวนข้อมูล 2 ตัวอย่างที่อยู่ในกลุ่มเดียวกัน ถูกจัดให้อยู่ในกลุ่มเดียวกัน

TN คือ จำนวนข้อมูล 2 ตัวอย่างที่อยู่ต่างกลุ่ม ถูกจัดให้อยู่ต่างกลุ่ม

FP คือ จำนวนข้อมูล 2 ตัวอย่างที่อยู่ต่างกลุ่ม ถูกจัดให้อยู่กลุ่มเดียวกัน

FN คือ จำนวนข้อมูล 2 ตัวอย่างที่อยู่กลุ่มเดียวกัน ถูกจัดให้อยู่ต่างกลุ่ม

ค่าที่ได้จากตัววัดประสิทธิภาพอยู่ในช่วง 0 ถึง 1 โดยค่าของ Rand statistic ที่มีค่าเข้าใกล้ 1 หมายถึง การจัดกลุ่มข้อมูลที่ได้ตรงกับกลุ่มของข้อมูลจริง

4.2.3 Jaccard Index (JI)

เป็นตัววัดประสิทธิภาพในเรื่องของพิจารณาความถูกต้องของการจัดกลุ่มของข้อมูลโดยการวัดผลลัพธ์ในการจัดกลุ่มที่เหมือนกันจากผลลัพธ์ในการจัดกลุ่มทั้งหมด แม้ลักษณะของสมการจะมีความใกล้เคียงกับ Rand Statistic แต่สิ่งที่ Jaccard แตกต่างจาก Rand Statistic คือการสนใจเฉพาะความถูกต้องในการจัดกลุ่มข้อมูลให้อยู่ในกลุ่มเดียวกัน ซึ่งแสดงดังสมการที่ (4.3)

$$JI = \frac{TP}{TP+FP+FN} \quad (4.3)$$

ค่าที่ได้จากตัววัดประสิทธิภาพอยู่ในช่วง 0 ถึง 1 โดยค่าของ Jaccard Index ที่มีค่าเข้าใกล้ 1 หมายถึง การจัดกลุ่มข้อมูลที่ได้ตรงกับกลุ่มของข้อมูลจริง

4.2.4 หน่วยความจำที่ใช้ในการจัดกลุ่ม

ขั้นตอนวิธีการที่ใช้สำหรับเปรียบเทียบผลการทดลองและขั้นตอนวิธีการที่นำเสนอ มีการจัดเก็บตัวแปรในหน่วยความจำหลัก สำหรับเป็นกลุ่มข้อมูลขนาดเล็กเพื่อใช้ในการจัดกลุ่ม ดังตารางที่ 4-2

ตารางที่ 4-2 ตัวแปรที่เก็บในหน่วยความจำหลักของแต่ละขั้นตอนวิธีการ

Variables	Descriptions
DenStream:	
$\overrightarrow{CF_l^1}$	เวกเตอร์ผลรวมของข้อมูล
$\overrightarrow{CF_l^2}$	เวกเตอร์ผลรวมกำลังสองของข้อมูล
n_i	จำนวนสมาชิกที่อยู่ในกลุ่มของข้อมูล
HECES:	
c_i	ค่าเฉลี่ยของข้อมูล
S_i	Covariance matrix
n_i	จำนวนสมาชิกที่อยู่ใน grid-cell
$sample_i$	ข้อมูลที่อยู่ใน grid-cell
DyCluStream:	
c_i	ค่าเฉลี่ยของข้อมูล
S_i	Covariance matrix
n_i	จำนวนสมาชิกที่อยู่ในกลุ่มของข้อมูล
DyCluStream+:	
c_i	ค่าเฉลี่ยของข้อมูล
S_i	Covariance matrix
n_i	จำนวนสมาชิกที่อยู่ในกลุ่มของข้อมูล

หน่วยความจำที่ใช้ต่อหนึ่งกลุ่มข้อมูลของขั้นตอนวิธีการต่างๆในแต่ละชุดข้อมูล สามารถแสดงได้ดังตารางที่ 4-3, 4-4 และ 4-5

ตารางที่ 4-3 หน่วยความจำที่ใช้ต่อหนึ่งกลุ่มข้อมูลขนาดเล็กในขั้นตอนวิธีการ **DyCluStream, DyCluStream+**

Dataset	**DyCluStream, DyCluStream+**			
	c_i	S_i	n_i	**Total Memory usage / bytes**
S1	232	248	56	536
S2	232	248	56	536
S3	232	248	56	536
t4.8k	232	248	56	520
t7.10k	232	248	56	520
t8.8k	232	248	56	520
AbitaryHCM	232	248	56	536
AbitaryHCM2	232	248	56	536
Statlog	334	864	56	1,254
letter	344	2,264	56	2,664
Kdd	480	8,928	56	9,464

ตารางที่ 4-4 หน่วยความจำที่ใช้ต่อหนึ่งกลุ่มข้อมูลขนาดเล็กในขั้นตอนวิธีการ **HECES** โดยจำเป็นต้องเก็บข้อมูลทั้งหมดไว้

Dataset	HECES				Dataset
	c_i	S_i	n_i	Total Memory usage / bytes	Memory usage / bytes
S1	232	248	56	536	40,848
S2	232	248	56	536	40,848
S3	232	248	56	536	40,848
t4.8k	232	248	56	520	128,848
t7.10k	232	248	56	520	160,848
t8.8k	232	248	56	520	128,848
AbitaryHCM	232	248	56	536	326,704
AbitaryHCM2	232	248	56	536	241,488
Statlog	334	864	56	1,254	1,567,800
letter	344	2,264	56	2,664	1,282,528
Kdd	480	8,928	56	9,464	94,856,672

ตารางที่ 4-5 หน่วยความจำที่ใช้ต่อหนึ่งกลุ่มข้อมูลขนาเล็กในขั้นตอนวิธีการ **DenStream**

Dataset	DenStream			
	$\overrightarrow{CF_i^1}$	$\overrightarrow{CF_i^2}$	n_i	Total Memory usage / bytes
S1	232	248	56	520
S2	232	248	56	520
S3	232	248	56	520
t4.8k	232	248	56	520
t7.10k	232	248	56	520
t8.8k	232	248	56	520
AbitaryHCM	232	248	56	520
AbitaryHCM2	232	248	56	520
Statlog	334	864	56	724
letter	344	2,264	56	744
Kdd	480	8,928	56	1,016

4.3 การออกแบบการทดลอง

4.3.1 พารามิเตอร์ของขั้นตอนวิธีการต่าง ๆ

พารามิเตอร์ของแต่ละขั้นตอนวิธีการที่ใช้ในการเปรียบเทียบผลการทดลอง สามารถแสดงได้ดังตารางที่ 4-6

ตารางที่ 4-6 พารามิเตอร์ของขั้นตอนวิธีการต่าง ๆ

Parameters	Descriptions
DenStream:	
$InitPoint$	จำนวนข้อมูลสำหรับสร้าง potential micro-cluster เริ่มต้น
n_ε	จำนวนสมาชิกในกลุ่มข้อมูล
r_ε	รัศมีของ micro-cluster
β	threshold สำหรับจำแนกว่าเป็น outlier micro-clusters
HECES:	
r_ε	ความกว้างของ grid cell
DyCluStream:	
n_ε	จำนวนสมาชิกใน micro-cluster
r_ε	รัศมีของ micro-cluster
$\emptyset$	ค่า threshold สำหรับการเชื่อมต่อ micro-cluster
β	ค่า threshold สำหรับการเชื่อมต่อ micro-cluster โดยใช้ Density ratio
DyCluStream+:	
n_ε	จำนวนสมาชิกใน micro-cluster
r_ε	รัศมีของ micro-cluster
$\emptyset$	ค่า threshold สำหรับการเชื่อมต่อ micro-clusters
β	ค่า threshold สำหรับการเชื่อมต่อ micro-cluster โดยใช้ Bhatthacharyya distance
θ	ค่า threshold สำหรับการรวม micro-clusters โดยใช้ Cosine similarity

4.3.2 การออกแบบการทดลองของข้อมูลสังเคราะห์

การทดลองกับข้อมูลสังเคราะห์ มีวัตถุประสงค์เพื่อแสดงให้เห็นลักษณะการครอบคลุมของกลุ่มข้อมูลขนาดเล็กในแต่ละขั้นตอนวิธีการที่นำมาเปรียบเทียบ ซึ่งการทดลองจะมีการกำหนดค่าพารามิเตอร์ให้อยู่ในช่วงต่าง ๆ และเลือกค่าพารามิเตอร์ที่ให้ผลดีที่สุดของแต่ละขั้นตอนวิธีการ ซึ่งใช้วิธีการวัดประสิทธิภาพโดยเปรียบเทียบในเรื่องของ ความเป็นหนึ่งเดียวกันของกลุ่มข้อมูล ความแม่นยำของวิธีการจัดกลุ่มข้อมูล และความถูกต้องในการจัดกลุ่มข้อมูลให้อยู่ในกลุ่มเดียวกัน โดยมีการกำหนดค่าพารามิเตอร์ดังตารางที่ 4-7

ตารางที่ 4-7 การกำหนดพารามิเตอร์ของขั้นตอนวิธีการต่าง ๆ

Parameters	Range
DenStream:	
$InitPoint$	[500, 1,000] โดยเพิ่มทีละ 500
n_ε	[5, 25] โดยมีค่าเพิ่มทีละ 1
r_ε	[0.05, 0.25] โดยมีค่าเพิ่มทีละ 0.05
β	0.2
HECES:	
r_ε	[5, 25] โดยมีค่าเพิ่มทีละ 5 *ที่ไม่เป็นหลักทศนิยมเท่ากับขั้นตอนวิธีการอื่น เนื่องจาก grid-cell ต้อง normalize ให้อยู่ในช่วง [0, 100]
DyCluStream:	
n_ε	[5, 25] โดยมีค่าเพิ่มทีละ 1
r_ε	[0.05, 0.25] โดยมีค่าเพิ่มทีละ 0.05
$\emptyset$	[2, 20] โดยมีค่าเพิ่มทีละ 0.1
β	[1, 2.5] โดยมีค่าเพิ่มทีละ 0.1
DyCluStream+:	
n_ε	[5, 25] โดยมีค่าเพิ่มทีละ 1
r_ε	[0.05, 0.25] โดยมีค่าเพิ่มทีละ 0.05
$\emptyset$	[2, 20] โดยมีค่าเพิ่มทีละ 0.1
β	[1, 4] โดยมีค่าเพิ่มทีละ 0.1
θ	[0.5, 0.8] โดยมีค่าเพิ่มทีละ 0.1

พารามิเตอร์ที่ให้ผลดีที่สุดหลังการจัดกลุ่มของแต่ละขั้นตอนวิธีการ แสดงได้ดังตาราง 4-8

ตารางที่ 4-8 พารามิเตอร์ที่ให้ผลดีที่สุดหลังการจัดกลุ่ม

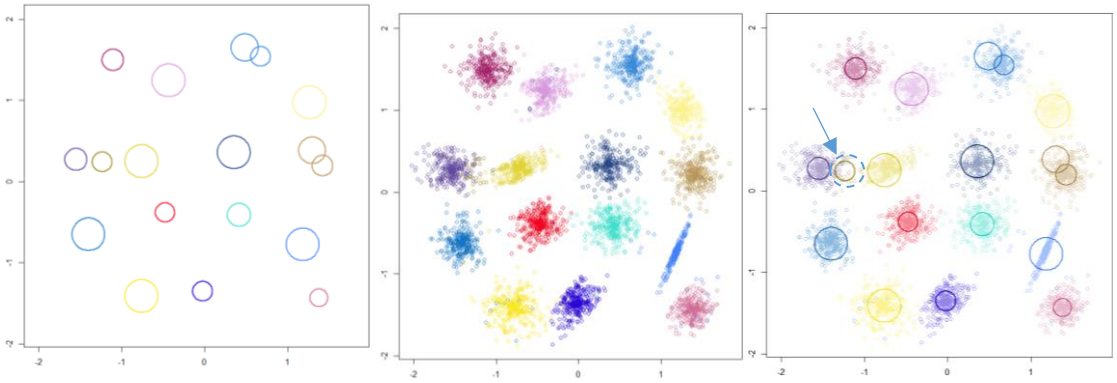
Dataset	DenStream	HECES	DyCluStream	DyCluStream+
S1	n_ε = 5 r_ε= 0.2 InitPoint = 500 β = 0.2	r_ε= 15	n_ε = 9 r_ε= 0.2 $\emptyset$ = 7.5 β = 1.8	n_ε = 20 r_ε= 0.2 $\emptyset$= 4 β = 1 θ = 0.5
S2	n_ε = 10 r_ε= 0.3 InitPoint = 500 β = 0.2	r_ε= 15	n_ε = 12 r_ε= 0.1 $\emptyset$ = 7.5 β = 1.8	n_ε = 14 r_ε= 0.1 $\emptyset$ = 7 β = 1 θ = 0.5
S3	n_ε = 18 r_ε= 0.2 InitPoint = 500 β = 0.2	r_ε= 25	n_ε = 22 r_ε= 0.1 $\emptyset$ = 17 β = 2	n_ε = 22 r_ε= 0.1 $\emptyset$= 17 β = 1 θ = 0.5
t4.8k	n_ε = 7 r_ε= 0.15 InitPoint = 500 β = 0.2	r_ε= 15	n_ε = 7 r_ε= 0.15 $\emptyset$ = 7.8 β = 2.1	n_ε = 14 r_ε= 0.15 $\emptyset$= 3.3 β = 3.1 θ = 0.8
t7.10k	n_ε = 11 r_ε= 0.2 InitPoint = 500 β = 0.2	r_ε= 25		n_ε = 15 r_ε= 0.15 $\emptyset$= 2.7 β = 2.5 θ = 0.7

ตารางที่ 4-8 พารามิเตอร์ที่ให้ผลดีที่สุดหลังการจัดกลุ่ม (ต่อ)

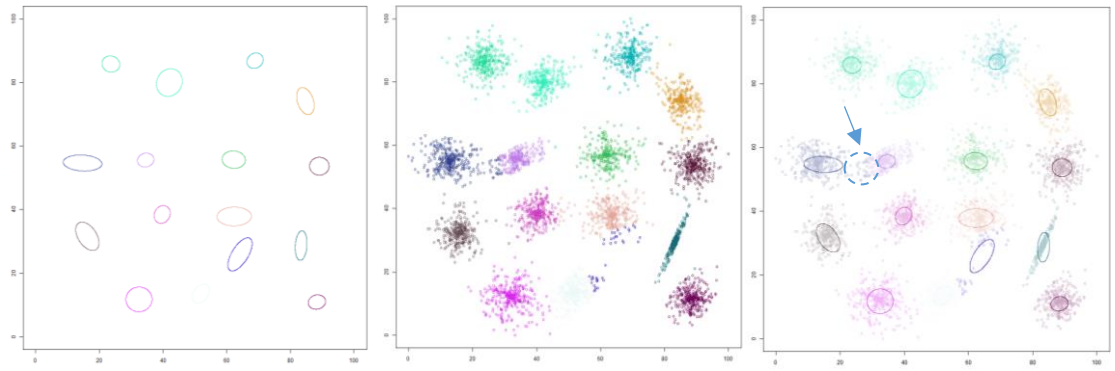
Dataset	DenStream	HECES	DyCluStream	DyCluStream+
t8.8k	n_ε = 9 r_ε= 0.15 InitPoint = 500 β = 0.2	r_ε= 20	n_ε = 9 r_ε= 0.15 $\emptyset$ = 8 β = 1	n_ε = 12 r_ε= 0.15 $\emptyset$= 2.3 β = 2.5 θ = 0.7
AbitaryHCM	n_ε = 20 r_ε= 0.2 InitPoint = 500 β = 0.2	r_ε= 25	n_ε = 20 r_ε= 0.2 $\emptyset$ = 12 β = 2	n_ε = 20 r_ε= 0.2 $\emptyset$= 3.3 β = 1 θ = 0.8
AbitaryHCM2	n_ε = 21 r_ε= 0.1 InitPoint = 500 β = 0.2	r_ε= 25	n_ε = 20 r_ε= 0.2 $\emptyset$ = 15 β = 1	n_ε = 21 r_ε= 0.2 $\emptyset$= 4 β = 2 θ = 0.7

จากการทดสอบการจัดกลุ่มข้อมูล สามารถแสดงผลการทดลองของข้อมูลสังเคราะห์ได้ดังนี้

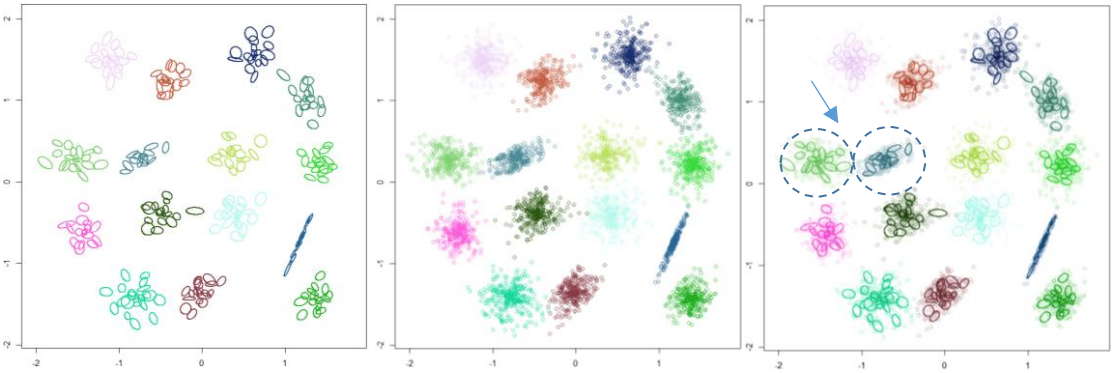
- ผลการทดลองของข้อมูล S1, S2, S3



ก) micro-clusters (DenStream) ข) ผลลัพธ์การจัดกลุ่ม (DenStream) ค) ความครอบคลุม (DenStream)

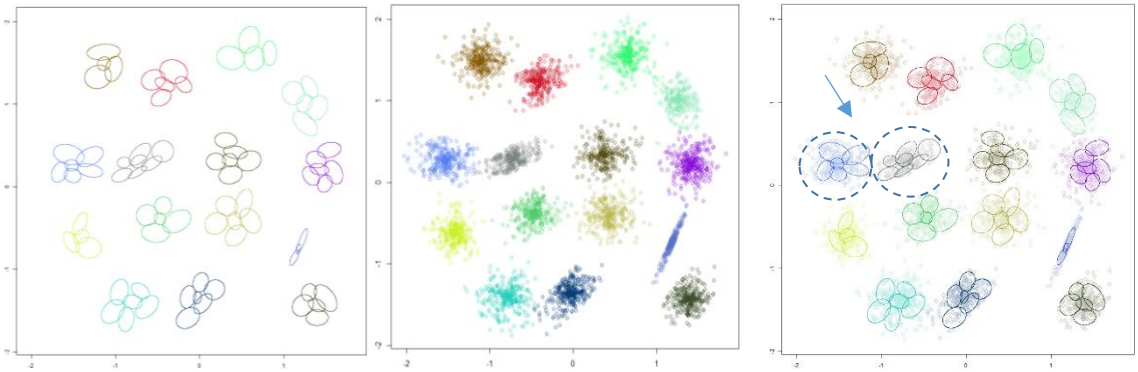


ง) micro-clusters (HECES) จ) ผลลัพธ์การจัดกลุ่ม (HECES) ฉ) ความครอบคลุม (HECES)



ช) micro-clusters (DyCluStream) ซ) ผลลัพธ์การจัดกลุ่ม (DyCluStream) ฌ) ความครอบคลุม (DyCluStream)

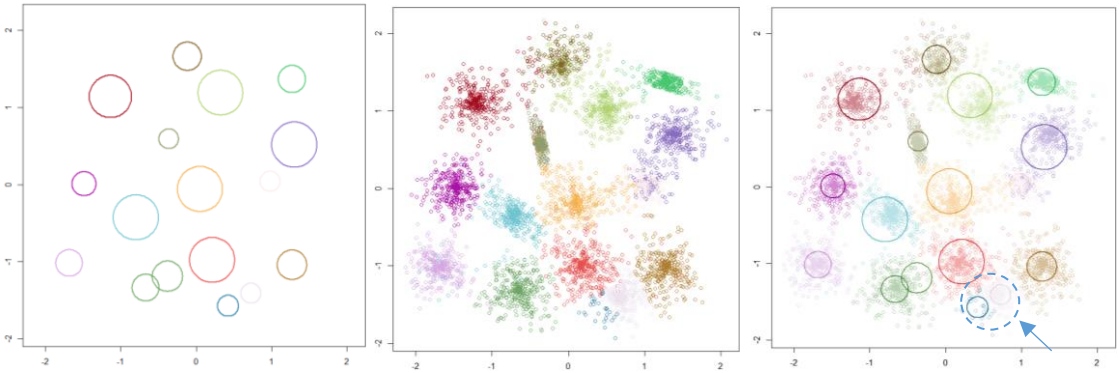
ภาพที่ 4-4 ผลลัพธ์การจัดกลุ่มของข้อมูล S1



ญ) micro-clusters (DyCluStream+)　　ฎ) ผลลัพธ์การจัดกลุ่ม (DyCluStream+)　　ฐ) ความครอบคลุม (DyCluStream+)

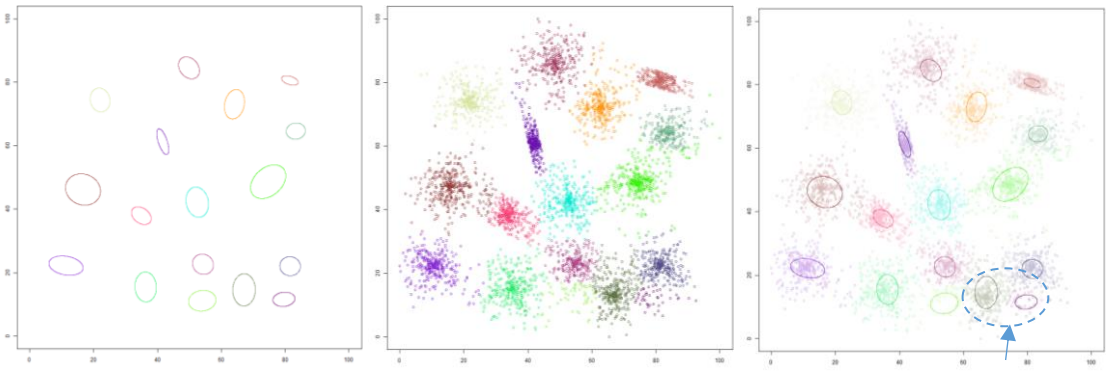
ภาพที่ 4-4 ผลลัพธ์การจัดกลุ่มของข้อมูล S1 (ต่อ)

จากภาพที่ 4-4 แสดงให้เห็นว่าขั้นตอนวิธีการ DyCluStream และ DyCluStream+ ให้ผลการจัดกลุ่มที่ดีกว่า DenStream และ HECES สังเกตได้จากวิธีการแบ่งกลุ่มข้อมูลที่นำเสนอทั้งสองวิธีสามารถแบ่งกลุ่มข้อมูลสองกลุ่มที่อยู่ใกล้กันได้ดีกว่าอย่างชัดเจน ดังภาพที่ 4-4 ฌ) และ 4-4 ฐ) ส่วน DenStream มีการจัดกลุ่มที่ผิดพลาดเพราะแบ่งกลุ่มข้อมูล 2 กลุ่มที่อยู่ใกล้กันเป็น 3 กลุ่ม ดังภาพที่ 4-4 ค) และ HECES มีการจัดกลุ่มที่ผิดพลาดเพราะข้อมูลของกลุ่มข้อมูลหนึ่งถูกจัดให้อยู่ในอีกกลุ่มเป็นส่วนมาก ดังภาพที่ 4-4 ฉ)

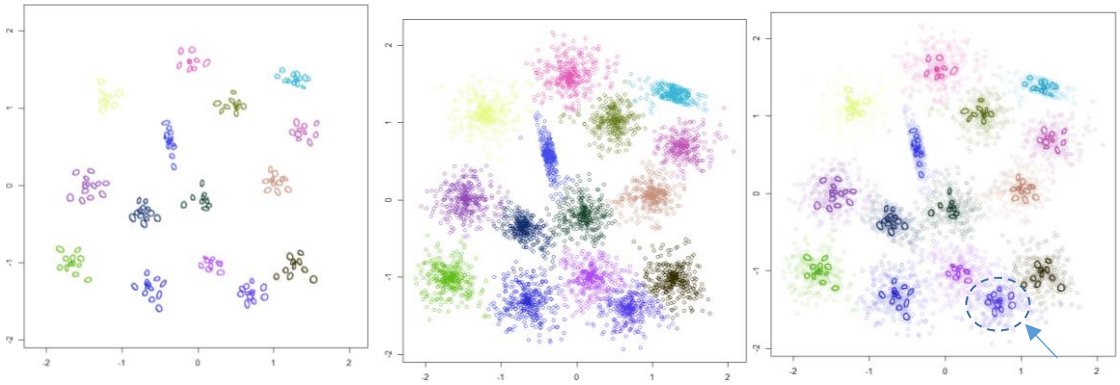


ก) micro-clusters (DenStream)　　ข) ผลลัพธ์การจัดกลุ่ม (DenStream)　　ค) ความครอบคลุม (DenStream)

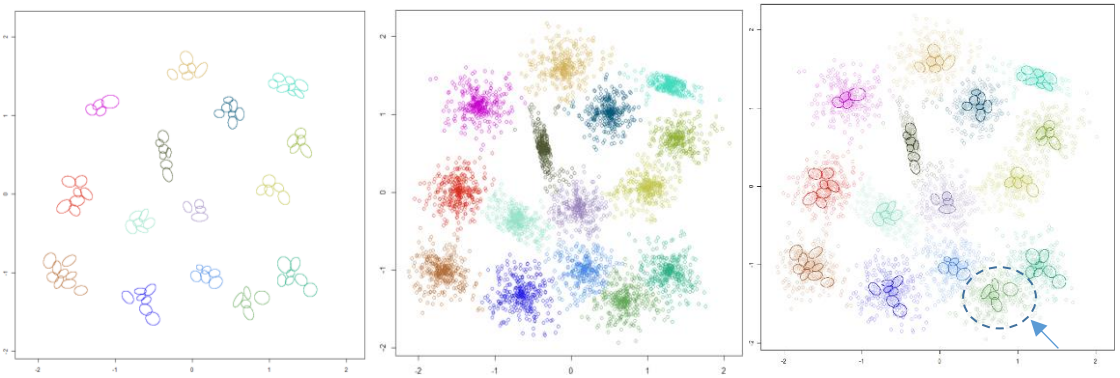
ภาพที่ 4-5 ผลลัพธ์การจัดกลุ่มของข้อมูล S2



ง) micro-clusters (HECES) จ) ผลลัพธ์การจัดกลุ่ม (HECES) ฉ) ความครอบคลุม (HECES)



ช) micro-clusters (DyCluStream) ซ) ผลลัพธ์การจัดกลุ่ม (DyCluStream) ฌ) ความครอบคลุม (DyCluStream)

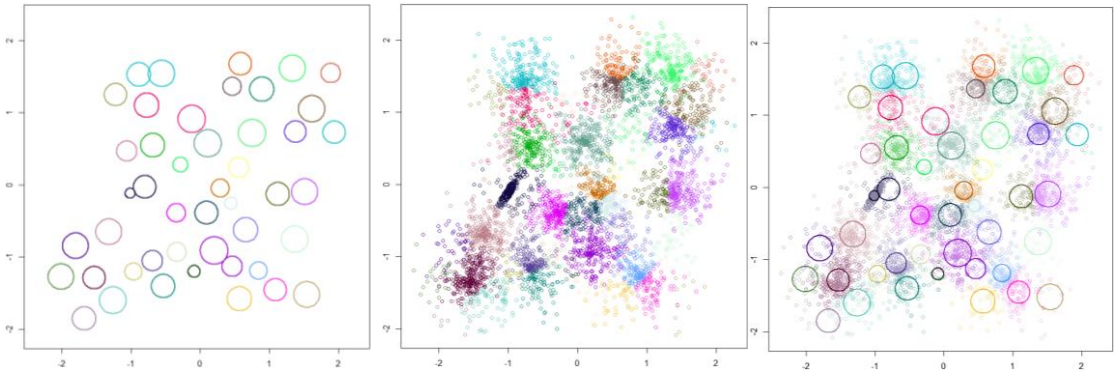


ญ) micro-clusters (DyCluStream+) ฎ) ผลลัพธ์การจัดกลุ่ม (DyCluStream+) ฐ) ความครอบคลุม (DyCluStream+)

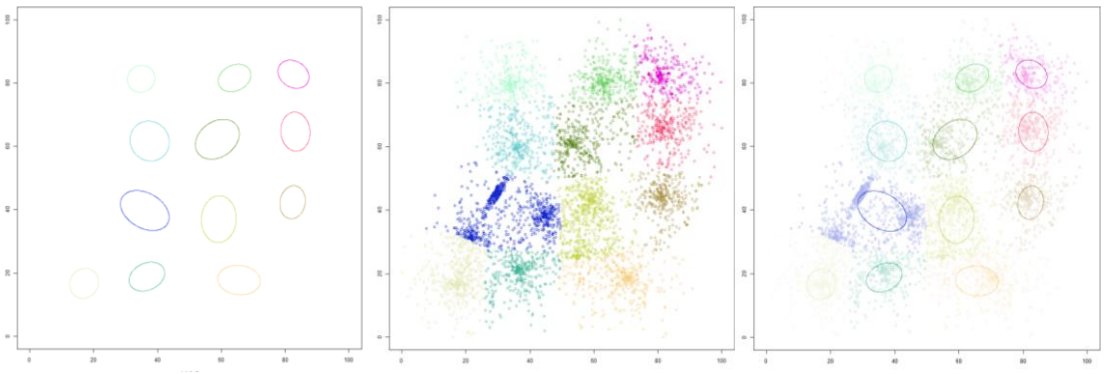
ภาพที่ 4-5 ผลลัพธ์การจัดกลุ่มของข้อมูล S2 (ต่อ)

จากภาพที่ 4-5 แสดงให้เห็นว่าขั้นตอนวิธีการ DyCluStream และ DyCluStream+ ให้ผลการจัดกลุ่มที่ดีกว่า DenStream และ HECES สังเกตได้จากวิธีการแบ่งกลุ่มข้อมูลที่นำเสนอทั้งสองวิธีสามารถแบ่งกลุ่มข้อมูลที่มีการกระจายตัวของข้อมูลมากได้อย่างชัดเจน ดังภาพที่ 4-5 ฌ) และ 4-5 ฐ)

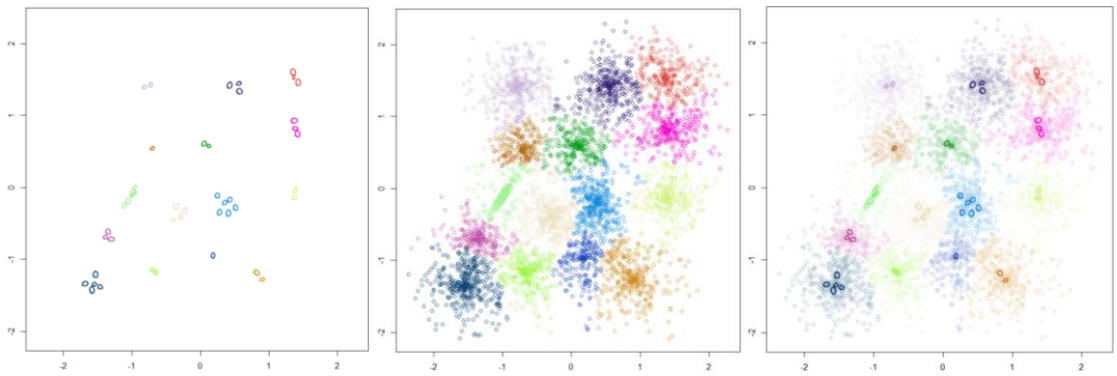
ตามลำดับ ส่วน DenStream และ HECES มีการจัดกลุ่มที่ผิดพลาดเพราะแบ่งออกเป็นสองกลุ่ม ดังภาพที่ 4-5 ค) และ 4-5 ฉ) ตามลำดับ



ก) micro-clusters (DenStream) ข) ผลลัพธ์การจัดกลุ่ม (DenStream) ค) ความครอบคลุม (DenStream)

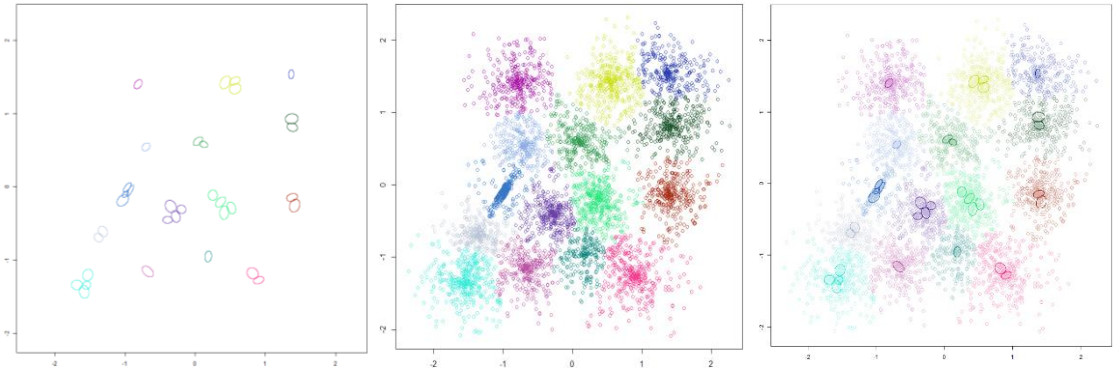


ง) micro-clusters (HECES) จ) ผลลัพธ์การจัดกลุ่ม (HECES) ฉ) ความครอบคลุม (HECES)



ช) micro-clusters (DyCluStream) ซ) ผลลัพธ์การจัดกลุ่ม (DyCluStream) ฌ) ความครอบคลุม (DyCluStream)

ภาพที่ 4-6 ผลลัพธ์การจัดกลุ่มของข้อมูล S3



ญ) micro-clusters (DyCluStream+) ฎ) ผลลัพธ์การจัดกลุ่ม (DyCluStream+) ฐ) ความครอบคลุม (DyCluStream+)

ภาพที่ 4-6 ผลลัพธ์การจัดกลุ่มของข้อมูล S3 (ต่อ)

จากภาพที่ 4-6 แสดงให้เห็นว่าขั้นตอนวิธีการ DyCluStream และ DyCluStream+ ให้ผลการจัดกลุ่มที่ดีกว่า DenStream และ HECES สังเกตได้จากวิธีการแบ่งกลุ่มข้อมูลที่นำเสนอทั้งสองวิธีใช้โครงสร้างกลุ่มข้อมูลขนาดเล็กที่สามารถครอบคลุมส่วนของข้อมูลที่มีความหนาแน่นได้อย่างถูกต้อง ดังภาพที่ 4-6 ฌ) และ 4-6 ฐ) ตามลำดับ แต่กลุ่มข้อมูลขนาดเล็กของ DenStream และ HECES ไม่สามารถครอบคลุมส่วนของข้อมูลที่มีความหนาแน่นได้อย่างถูกต้อง ดังภาพที่ 4-6 ค) และ 4-6 ฉ) ตามลำดับ

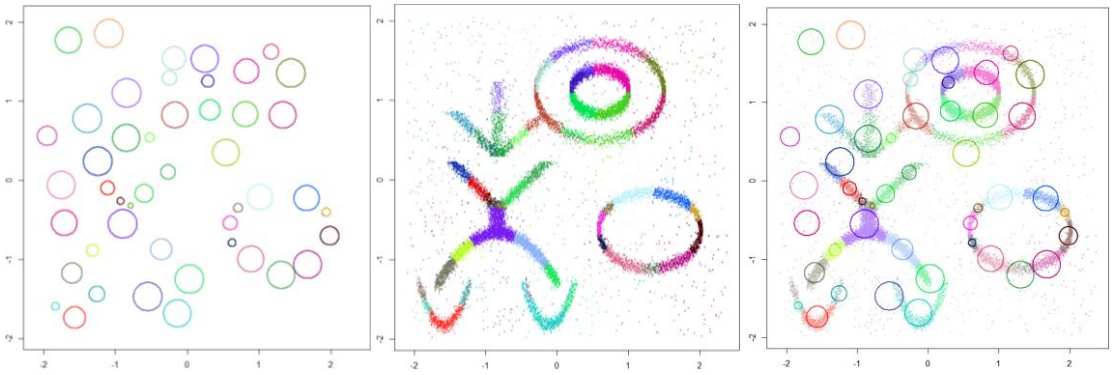
ตารางที่ 4-9 ผลการจัดกลุ่มข้อมูล S1, S2, S3

Dataset	Method	Purity	Jaccard-Index	Rand statistic	Number of micro-clusters	Number of Clusters	Memory Usage/kB
S1	Denstream	0.9764	0.9044	0.9933	18	16	9.36
	HECES	0.9824	0.9311	0.9953	46	16	24.656
	DyCluStream	**0.9934**	**0.974**	**0.9982**	253	15	135.608
	DyCluStream+	0.9932	0.9732	**0.9982**	78	**15**	41.808
S2	Denstream	0.8808	0.6671	0.9727	17	16	8.84
	HECES	0.9440	0.8085	0.9861	44	17	23.584
	DyCluStream	0.9650	0.8733	0.9910	276	15	147.936
	DyCluStream+	**0.9664**	**0.8784**	**0.9914**	93	**15**	49.848
S3	Denstream	**0.8666**	0.4141	0.9567	46	43	23.92
	HECES	0.6904	0.4111	0.9318	18	12	9.648
	DyCluStream	0.8418	0.5682	0.9627	45	15	24.12
	DyCluStream+	0.8484	**0.579**	**0.9641**	34	**15**	18.224

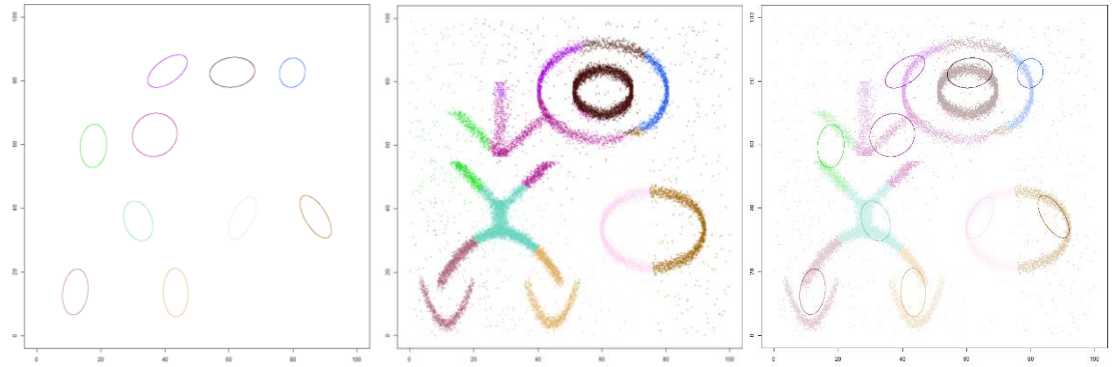
จากตารางที่ 4-9 แสดงให้เห็นว่า DyCluStream และ DyCluStream+ ให้ผลการทดลองที่ดีกว่า Denstream และ HECES เกือบทุกตัววัดประสิทธิภาพและมีจำนวนกลุ่มของข้อมูลที่ถูกต้อง โดย DyCluStream+ จะทำการลดจำนวนของตัวแทนกลุ่มข้อมูลขนาดเล็กของ DyCluStream ลง และยังคงให้ประสิทธิภาพในการจัดกลุ่มข้อมูลที่ใกล้เคียงหรือมากกว่า DyCluStream แต่อย่างไรก็

ตาม DyCluStream และ DyCluStream+ ยังคงใช้หน่วยความจำที่มากกว่า Denstream และ HECES เนื่องมาจากรูปทรง dynamic hyper-ellipsoidal ของ DyCluStream และ DyCluStream+ มีความกระชับกับกลุ่มข้อมูล จึงส่งผลให้กลุ่มข้อมูลขนาดเล็กมีจำนวนมาก

- ผลการทดลองของข้อมูล AbitaryHCM, AbitaryHCM2

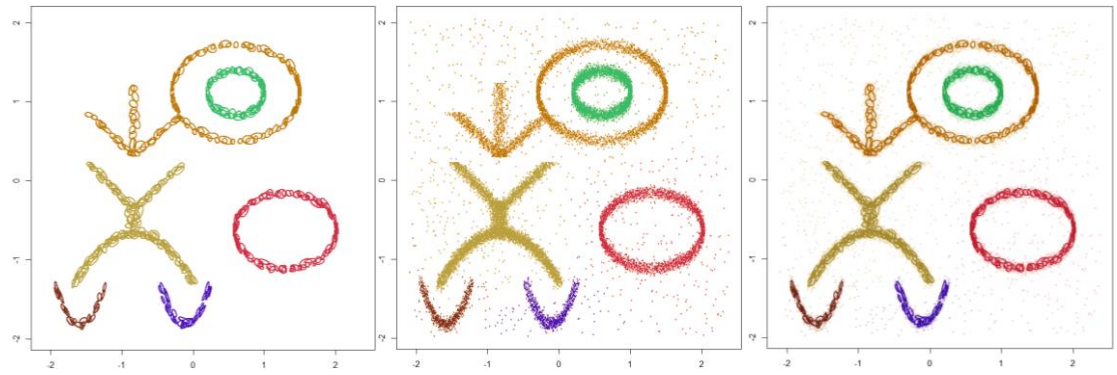


ก) micro-clusters (DenStream) ข) ผลลัพธ์การจัดกลุ่ม (DenStream) ค) ความครอบคลุม (DenStream)



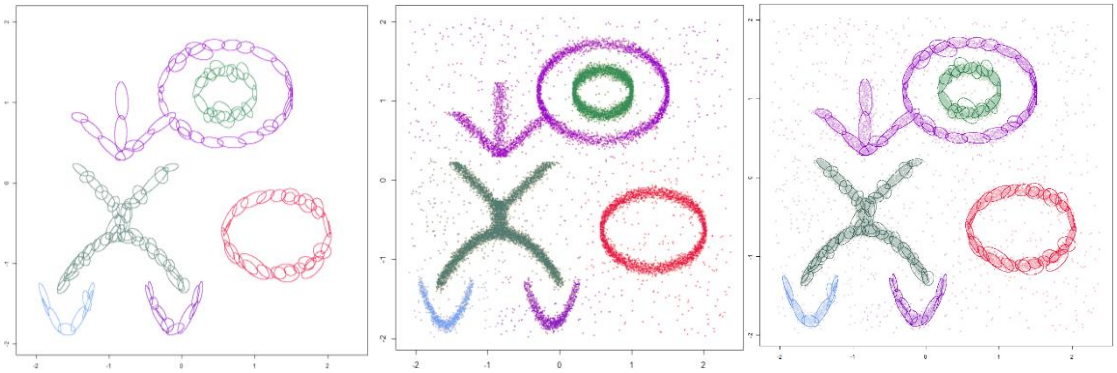
ง) micro-clusters (HECES) จ) ผลลัพธ์การจัดกลุ่ม (HECES) ฉ) ความครอบคลุม (HECES)

ภาพที่ 4-7 ผลลัพธ์การจัดกลุ่มของข้อมูล AbitaryHCM



ช) micro-clusters (DyCluStream) ซ) ผลลัพธ์การจัดกลุ่ม (DyCluStream) ฌ) ความครอบคลุม (DyCluStream)

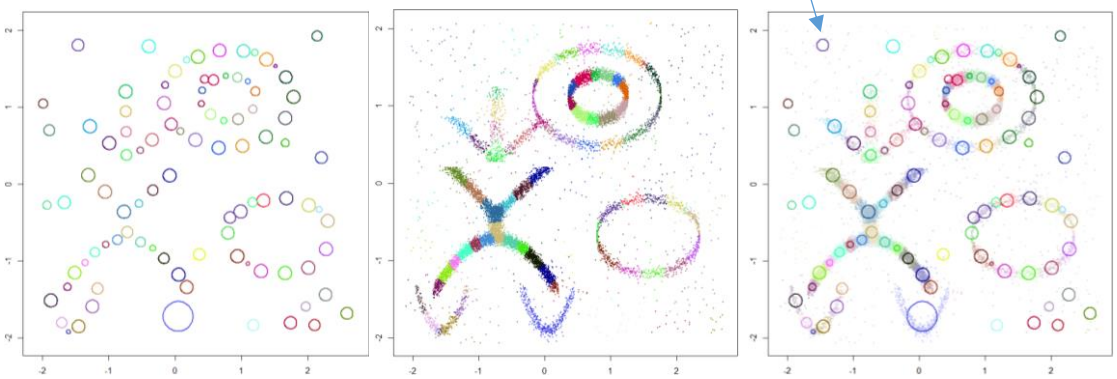
ญ) micro-clusters (DyCluStream+) ฎ) ผลลัพธ์การจัดกลุ่ม (DyCluStream+) ฐ) ความครอบคลุม (DyCluStream+)

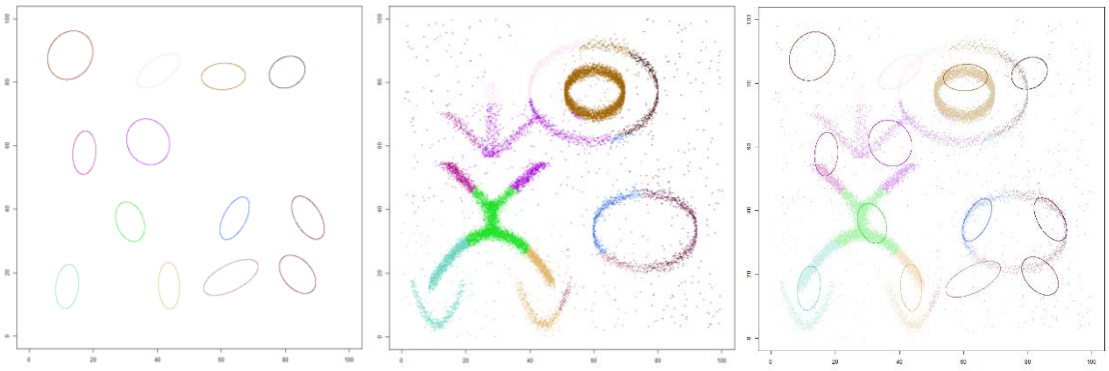


ภาพที่ 4-7 ผลลัพธ์การจัดกลุ่มของข้อมูล AbitaryHCM (ต่อ)

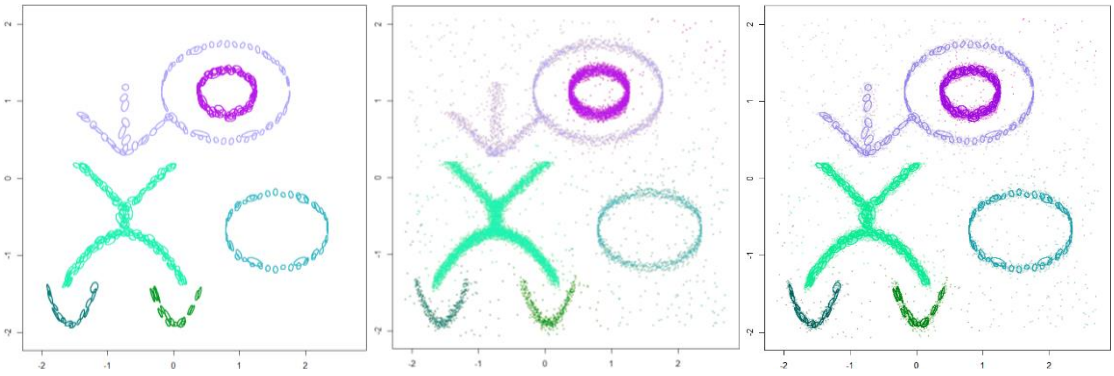
จากภาพที่ 4-7 แสดงให้เห็นว่าขั้นตอนวิธีการ DyCluStream และ DyCluStream+ ให้ผลการจัดกลุ่มที่ดีกว่า DenStream และ HECES สังเกตได้จากกลุ่มข้อมูลขนาดเล็กของวิธีการแบ่งกลุ่มข้อมูลที่นำเสนอทั้งสองวิธี สามารถแทนรูปร่างและทิศทางการกระจายตัวของข้อมูลได้อย่างถูกต้อง และในส่วนของการจัดกลุ่มสำหรับนำไปใช้งาน เกิดจากการเชื่อมต่อกันของกลุ่มข้อมูลขนาดเล็กจึงส่งผลให้ได้กลุ่มข้อมูลสำหรับนำไปใช้งานได้อย่างถูกต้อง ดังภาพที่ 4-7 ฌ) และ 4-7 ฐ) ตามลำดับ แต่กลุ่มข้อมูลขนาดเล็กของ DenStream ไม่สามารถแทนรูปร่างและการกระจายตัวของข้อมูลได้อย่างถูกต้อง ดังภาพที่ 4-7 ค) ส่วนขั้นตอนวิธีการ HECES การจัดกลุ่มข้อมูลขนาดเล็ก โดยใช้วิธีการผสานกลุ่มข้อมูลขนาดเล็กสองกลุ่มที่ใกล้เคียงกันและแทนที่ด้วยกลุ่มข้อมูลใหม่ โดยใช้ค่าเฉลี่ยของข้อมูลในกลุ่มเป็นจุดศูนย์กลาง ส่งผลให้ผลลัพธ์ที่ได้มีความผิดพลาด ดังภาพที่ 4-7 ฉ)

ก) micro-clusters (DenStream) ข) ผลลัพธ์การจัดกลุ่ม (DenStream) ค) ความครอบคลุม (DenStream)



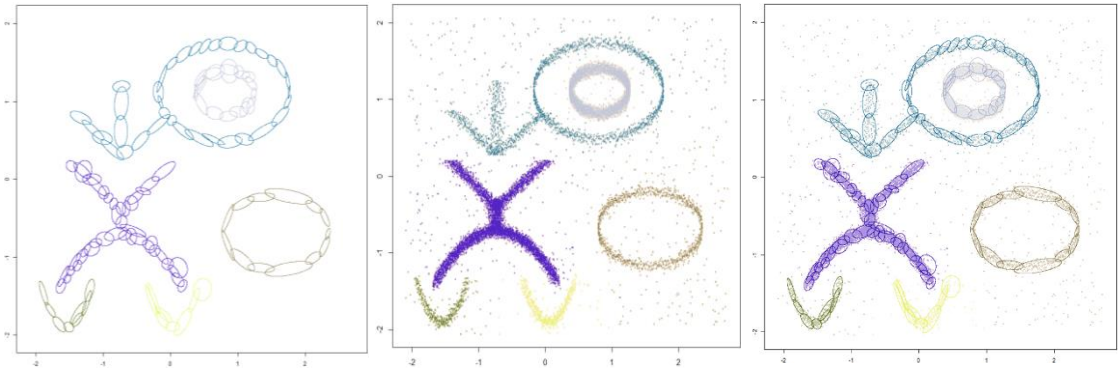


ง) micro-clusters (HECES) จ) ผลลัพธ์การจัดกลุ่ม (HECES) ฉ) ความครอบคลุม (HECES)



ช) micro-clusters (DyCluStream) ซ) ผลลัพธ์การจัดกลุ่ม (DyCluStream) ฌ) ความครอบคลุม (DyCluStream)

ภาพที่ 4-8 ผลลัพธ์การจัดกลุ่มของข้อมูล AbitaryHCM2



ญ) micro-clusters (DyCluStream+) ฎ) ผลลัพธ์การจัดกลุ่ม (DyCluStream+) ฐ) ความครอบคลุม (DyCluStream+)

ภาพที่ 4-8 ผลลัพธ์การจัดกลุ่มของข้อมูล AbitaryHCM2 (ต่อ)

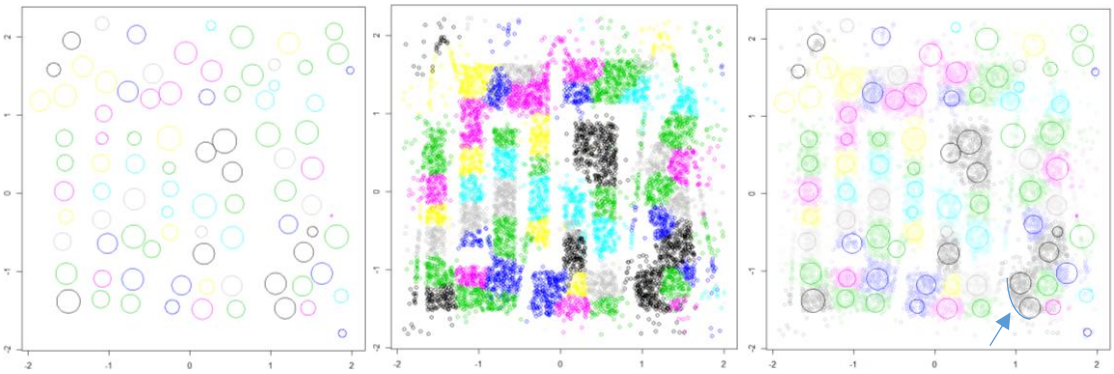
จากภาพที่ 4-8 แสดงให้เห็นว่าขั้นตอนวิธีการ DyCluStream และ DyCluStream+ ให้ผลการจัดกลุ่มที่ดีกว่า DenStream และ HECES สังเกตได้จากวิธีการแบ่งกลุ่มข้อมูลที่นำเสนอทั้งสองวิธีมีโครงสร้างของกลุ่มข้อมูลขนาดเล็ก ที่สามารถแยกกลุ่มข้อมูลกับข้อมูลรบกวนได้อย่างชัดเจน แต่กลุ่มข้อมูลขนาดเล็กของ DenStream และ HECES ไม่สามารถแยกกลุ่มข้อมูลกับข้อมูลรบกวนได้ ดังภาพที่ 4-8 ค) และ 4-8 ฉ) ตามลำดับ

ตารางที่ 4-10 ผลการจัดกลุ่มข้อมูล AbitaryHCM, AbitaryHCM2

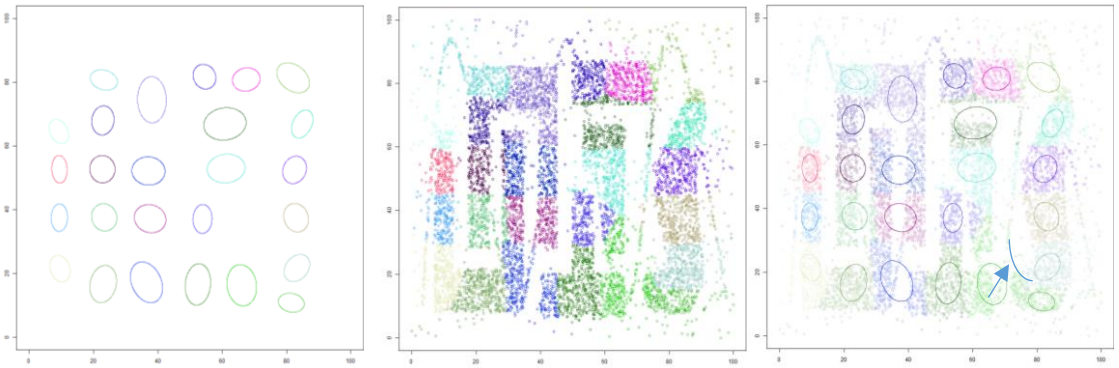
Dataset	Method	Purity	Jaccard-Index	Rand statistic	Number of micro-clusters	Number of Clusters	Memory Usage / kB
AbitaryHCM	Denstream	**0.8888**	0.1925	0.8376	47	46	24.44
	HECES	0.7623	0.4233	0.8684	18	10	9.648
	DyCluStream	0.8846	0.8263	0.9596	506	6	271.216
	DyCluStream+	0.8843	**0.8295**	**0.9604**	161	**6**	86.296
AbitaryHCM2	Denstream	**0.9619**	0.0797	0.7343	95	93	49.4
	HECES	0.8428	0.4648	0.8337	18	13	9.648
	DyCluStream	0.9524	**0.9394**	**0.9816**	379	**6**	203.144
	DyCluStream+	0.9524	0.9385	0.9813	127	6	68.072

จากภาพที่ 4-7 ถึง 4-8 แสดงให้เห็นว่าวิธีการ DyCluStream และ DyCluStream+ สามารถจัดกลุ่มได้ดีกว่า DenStream และ HECES เนื่องมาจากโครงสร้างที่เป็น dynamic hyper-ellipsoidal micro-clusters ทำให้สามารถจัดการกับลักษณะของข้อมูลที่มีลักษณะรูปร่างไม่แน่นอนทำได้ดีกว่า และมาตรวัดระยะทางซึ่งคำนึงถึงลักษณะโครงสร้างของ dynamic hyper-ellipsoidal จึงทำให้ DyCluStream สามารถแผ่ขยายกลุ่มข้อมูลขนาดเล็กที่เชื่อมต่อกันเป็นวงแหวนอย่างถูกต้อง แม้กลุ่มของข้อมูลจะมีลักษณะเป็นวงแหวนล้อมรอบกลุ่มข้อมูลอื่น ๆ อีกทั้งยังสามารถแบ่งกลุ่มข้อมูลขนาดเล็กจากข้อมูลรบกวนได้อย่างชัดเจน และ DyCluStream+ จะลดจำนวนของตัวแทนกลุ่มข้อมูลของ DyCluStream ลง โดยยังคงให้ประสิทธิภาพในการจัดกลุ่มข้อมูลที่ใกล้เคียงหรือมากกว่า DyCluStream ซึ่งสอดคล้องกับตารางที่ 4-10 ซึ่งแสดงให้เห็นว่า DyCluStream และ DyCluStream+ ให้ผลการทดลองที่ดีกว่า DenStream และ HECES เกือบทุกตัววัดประสิทธิภาพ และมีจำนวนกลุ่มของข้อมูลที่ถูกต้อง สำหรับตัววัด Purity ที่ DenStream สามารถทำได้ดีกว่าเนื่องมาจากเป็นตัววัดที่มุ่งเน้นความเป็นหนึ่งเดียวกันของข้อมูล จึงทำให้ DenStream ซึ่งมีกลุ่มข้อมูลจำนวนมากกว่าสามารถทำได้ดีกว่า เพราะแต่ละกลุ่มข้อมูลมีจำนวนของข้อมูลที่น้อยกว่า DyCluStream และ DyCluStream+ แต่อย่างไรก็ตาม DyCluStream+ ยังคงใช้หน่วยความจำที่มากกว่า DenStream และ HECES เนื่องมาจากมีกลุ่มข้อมูลขนาดเล็กจำนวนมาก

- ผลการทดลองของข้อมูล t4.8k, t7.10k, t8.8k

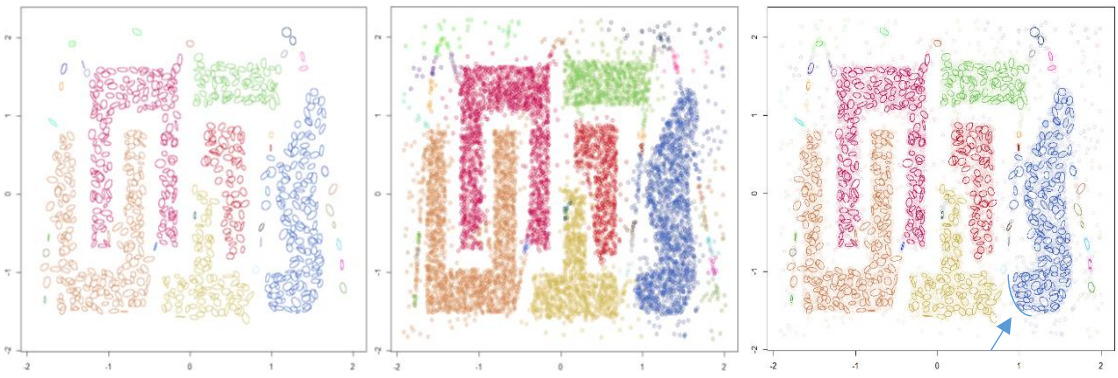


ก) micro-clusters (DenStream) ข) ผลลัพธ์การจัดกลุ่ม (DenStream) ค) ความครอบคลุม (DenStream)

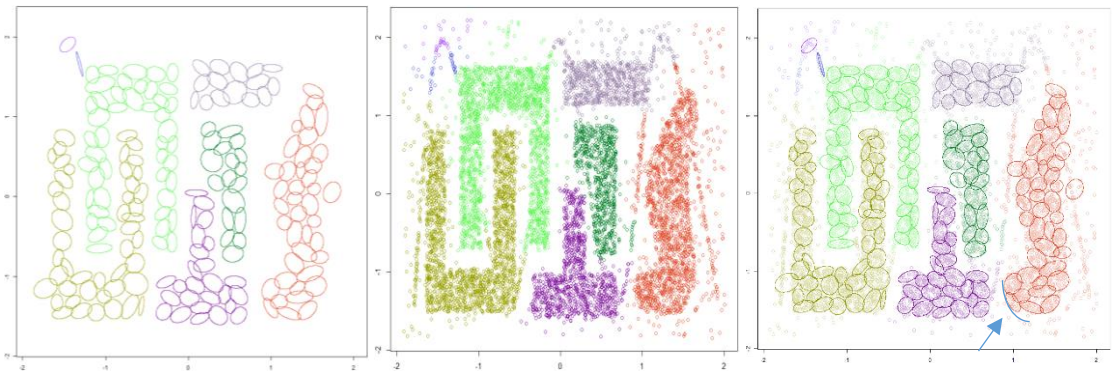


ง) micro-clusters (HECES) จ) ผลลัพธ์การจัดกลุ่ม (HECES) ฉ) ความครอบคลุม (HECES)

ภาพที่ 4-9 ผลลัพธ์การจัดกลุ่มของข้อมูล t4.8k



ช) micro-clusters (DyCluStream) ซ) ผลลัพธ์การจัดกลุ่ม (DyCluStream) ฌ) ความครอบคลุม (DyCluStream)

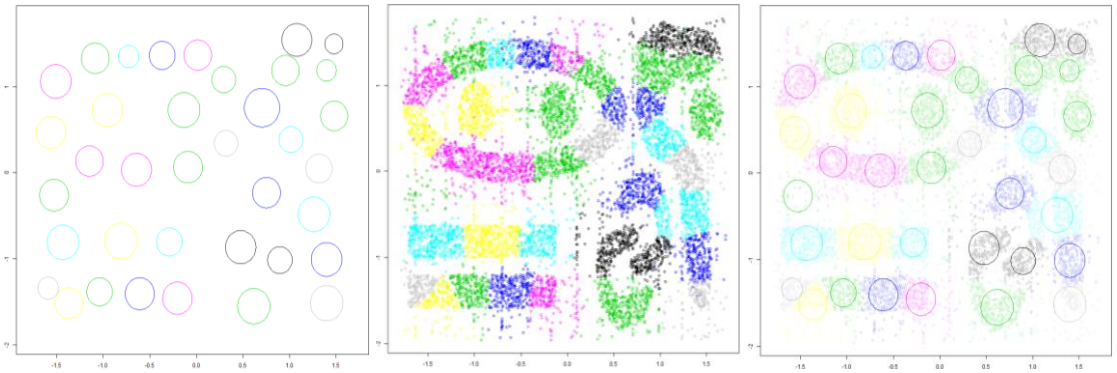


ญ) micro-clusters (DyCluStream+) ฎ) ผลลัพธ์การจัดกลุ่ม (DyCluStream+) ฐ) ความครอบคลุม (DyCluStream+)

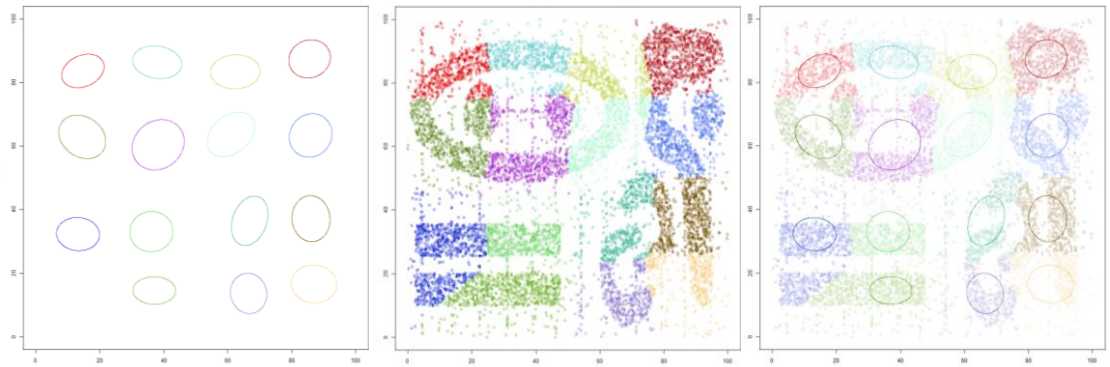
ภาพที่ 4-9 ผลลัพธ์การจัดกลุ่มของข้อมูล t4.8k (ต่อ)

จากภาพที่ 4-9 แสดงให้เห็นว่าขั้นตอนวิธีการ DyCluStream และ DyCluStream+ ให้ผลการจัดกลุ่มที่ดีกว่า DenStream และ HECES สังเกตได้จากกลุ่มข้อมูลขนาดเล็กของวิธีการแบ่งกลุ่มข้อมูลที่นำเสนอทั้งสองวิธี สามารถแทนรูปร่างและทิศทางการกระจายตัวของข้อมูลได้อย่างถูกต้อง จึงสามารถแบ่งขอบการกระจายตัวของกลุ่มข้อมูลได้อย่างชัดเจน และในส่วนของการจัดกลุ่มสำหรับนำไปใช้งาน เกิดจากการเชื่อมต่อกันของกลุ่มข้อมูลขนาดเล็ก ส่งผลให้ได้กลุ่มข้อมูลสำหรับ

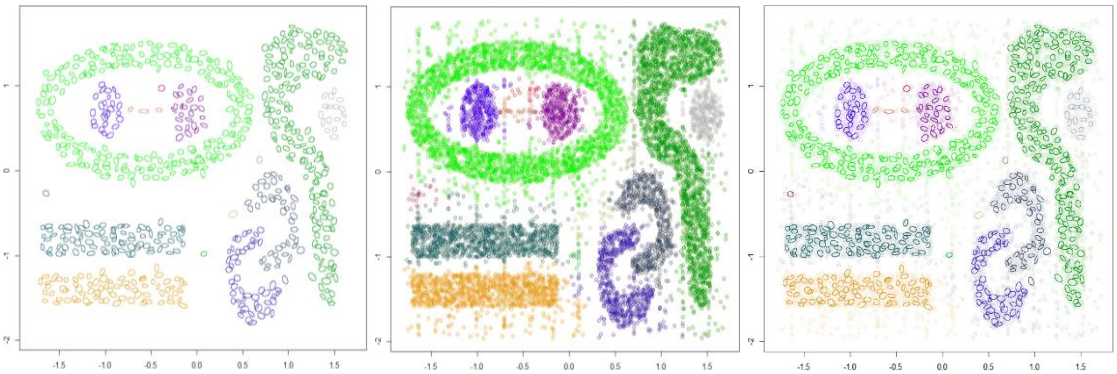
นำไปใช้งานได้อย่างถูกต้อง ดังภาพที่ 4-9 ฌ) และ 4-9 ฐ) ตามลำดับ โดยกลุ่มข้อมูลขนาดเล็กของ DenStream ไม่สามารถแทนรูปร่างและการกระจายตัวของข้อมูลที่เป็นส่วนขอบของกลุ่มข้อมูลได้อย่างถูกต้อง ดังภาพที่ 4-9 ค) ส่วนขั้นตอนวิธีการ HECES การจัดกลุ่มข้อมูลขนาดเล็ก โดยใช้วิธีการผสานกลุ่มข้อมูลขนาดเล็กสองกลุ่มที่ใกล้เคียงกันและแทนที่ด้วยกลุ่มข้อมูลใหม่โดยใช้ค่าเฉลี่ย ของข้อมูลในกลุ่มเป็นจุดศูนย์กลาง ส่งผลให้ผลลัพธ์ที่ได้มีความผิดพลาด ดังภาพที่ 4-9 ฉ)



ก) micro-clusters (DenStream) ข) ผลลัพธ์การจัดกลุ่ม (DenStream) ค) ความครอบคลุม (DenStream)

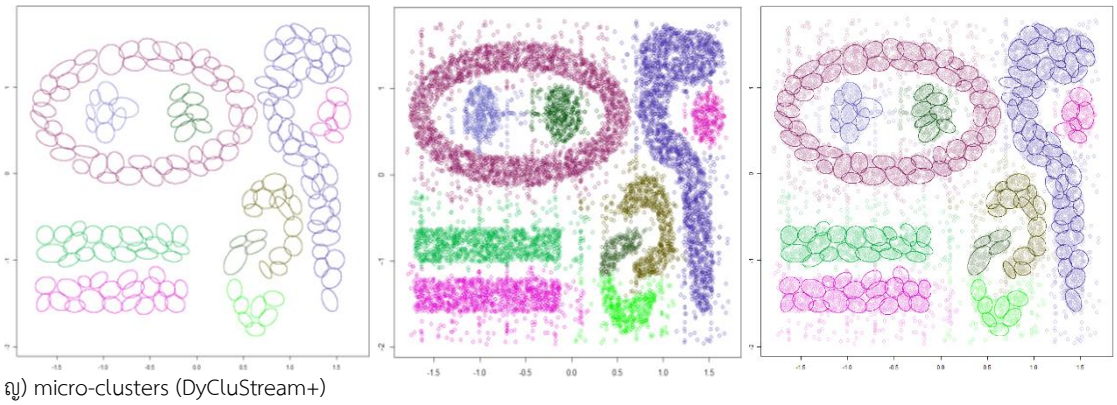


ง) micro-clusters (HECES) จ) ผลลัพธ์การจัดกลุ่ม (HECES) ฉ) ความครอบคลุม (HECES)



ช) micro-clusters (DyCluStream) ซ) ผลลัพธ์การจัดกลุ่ม (DyCluStream) ฌ) ความครอบคลุม (DyCluStream)

ภาพที่ 4-10 ผลลัพธ์การจัดกลุ่มของข้อมูล t7.10k



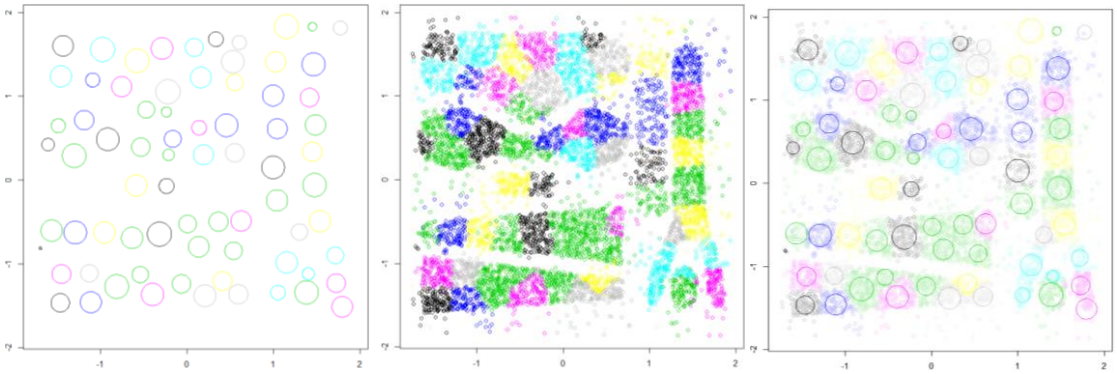
ญ) micro-clusters (DyCluStream+)

ฎ) ผลลัพธ์การจัดกลุ่ม (DyCluStream+)

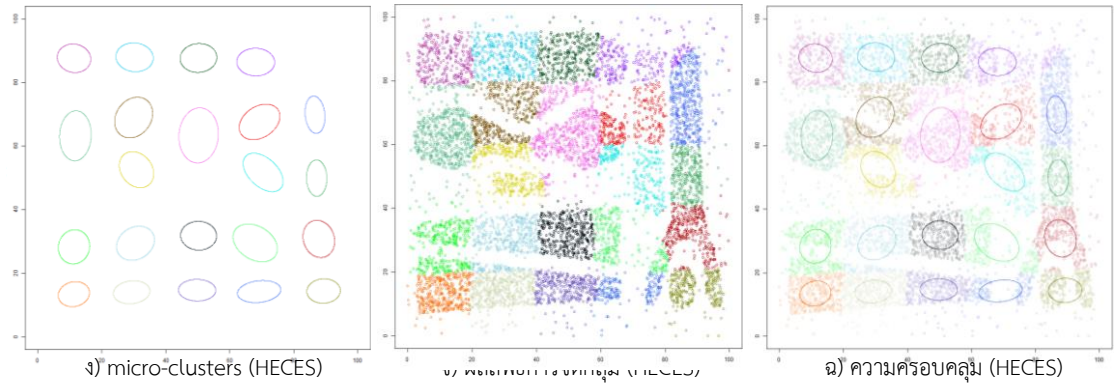
ฐ) ความครอบคลุม (DyCluStream+)

ภาพที่ 4-10 ผลลัพธ์การจัดกลุ่มของข้อมูล t7.10k (ต่อ)

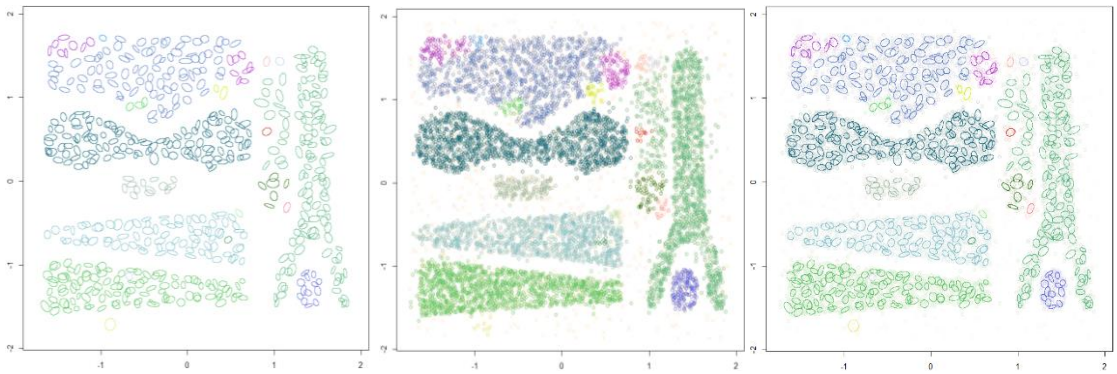
จากภาพที่ 4-10 แสดงให้เห็นว่าขั้นตอนวิธีการ DyCluStream และ DyCluStream+ ให้ผลการจัดกลุ่มที่ดีกว่า DenStream และ HECES สังเกตได้จากจากกลุ่มข้อมูลขนาดเล็กของวิธีการแบ่งกลุ่มข้อมูลที่นำเสนอทั้งสองวิธี สามารถแทนรูปร่างและทิศทางการกระจายตัวของข้อมูลได้อย่างถูกต้องชัดเจน และในส่วนของการจัดกลุ่มสำหรับนำไปใช้งาน เกิดจากการเชื่อมต่อกันของกลุ่มข้อมูลขนาดเล็ก ส่งผลให้ได้กลุ่มข้อมูลสำหรับนำไปใช้งานได้อย่างถูกต้อง ดังภาพที่ 4-10 ฌ) และ 4-10 ฐ) ตามลำดับ โดยกลุ่มข้อมูลขนาดเล็กของ DenStream ไม่สามารถแทนรูปร่างและการกระจายตัวของข้อมูลได้อย่างถูกต้อง ดังภาพที่ 4-10 ค) ส่วนขั้นตอนวิธีการ HECES การจัดกลุ่มข้อมูลขนาดเล็ก โดยใช้วิธีการผสานกลุ่มข้อมูลขนาดเล็กสองกลุ่มที่ใกล้เคียงกันและแทนที่ด้วยกลุ่มข้อมูลใหม่โดยใช้ค่าเฉลี่ยเป็นจุดศูนย์กลางส่งผลให้ได้ผลลัพธ์ที่มีความผิดพลาด ดังภาพที่ 4-10 ฉ)



ก) micro-clusters (DenStream) ข) ผลลัพธ์การจัดกลุ่ม (DenStream) ค) ความครอบคลุม (DenStream)

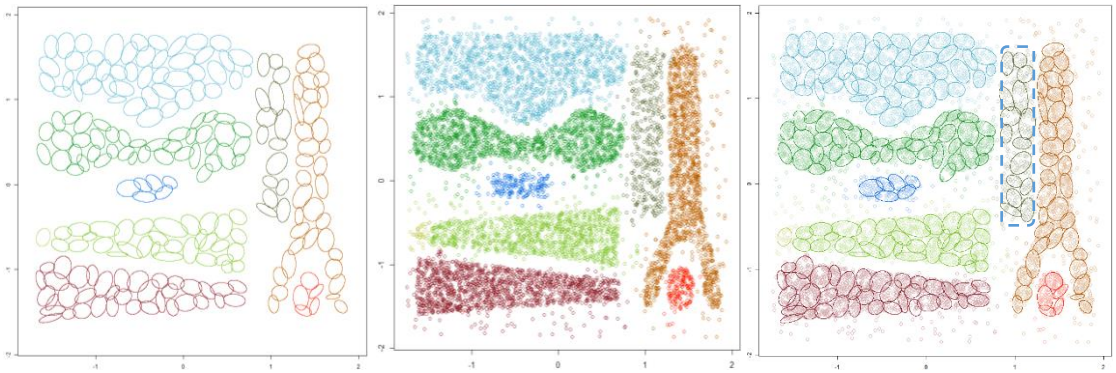


ง) micro-clusters (HECES) ฉ) ความครอบคลุม (HECES)



ช) micro-clusters (DyCluStream) ซ) ผลลัพธ์การจัดกลุ่ม (DyCluStream) ฌ) ความครอบคลุม (DyCluStream)

ภาพที่ 4-11 ผลลัพธ์การจัดกลุ่มของข้อมูล t8.8k



ญ) micro-clusters (DyCluStream+) ฎ) ผลลัพธ์การจัดกลุ่ม (DyCluStream+) ฐ) ความครอบคลุม (DyCluStream+)

ภาพที่ 4-11 ผลลัพธ์การจัดกลุ่มของข้อมูล t8.8k (ต่อ)

จากภาพที่ 4-11 แสดงให้เห็นว่าขั้นตอนวิธีการ DyCluStream และ DyCluStream+ ให้ผลการจัดกลุ่มที่ดีกว่า DenStream และ HECES สังเกตได้จากกลุ่มข้อมูลขนาดเล็กของวิธีการแบ่งกลุ่มข้อมูลที่นำเสนอทั้งสองวิธี สามารถแทนรูปร่างและทิศทางการกระจายตัวของข้อมูลได้อย่างถูกต้องและใช้รูปแบบการเชื่อมต่อกันของกลุ่มข้อมูลขนาดเล็กส่งผลให้ได้กลุ่มข้อมูลสำหรับนำไปใช้งานได้อย่างถูกต้อง ดังภาพที่ 4-11 ฌ) และ 4-11 ฐ) ตามลำดับ และ DyCluStream+ สามารถจัดกลุ่มของข้อมูลที่มีระดับความหนาแน่นแตกต่างกัน ดังเส้นประสีน้ำเงินในภาพที่ 4-11 ฐ) ตามลำดับ แต่กลุ่มข้อมูลขนาดเล็กของ DenStream ไม่สามารถแทนรูปร่างและการกระจายตัวของข้อมูลได้อย่างถูกต้อง ดังภาพที่ 4-11 ค) ส่วนขั้นตอนวิธีการ HECES การจัดกลุ่มข้อมูลขนาดเล็ก โดยใช้วิธีการผสานกลุ่มข้อมูลขนาดเล็กสองกลุ่มที่ใกล้เคียงกันและแทนที่ด้วยกลุ่มข้อมูลใหม่โดยใช้ค่าเฉลี่ยเป็นจุดศูนย์กลาง ส่งผลให้ได้ผลลัพธ์ที่มีความผิดพลาด ดังภาพที่ 4-11 ฉ)

ตารางที่ 4-11 ผลการจัดกลุ่มข้อมูล t4.8k, t4.10k, t8.8k

Dataset	Method	Number of micro-clusters	Number of Clusters	Memory Usage / kB
t4.8k	Denstream	87	85	45.24
	HECES	49	26	26.264
	DyCluStream	687	73	368.232
	DyCluStream+	193	**8**	103.448
t7.10k	Denstream	37	37	19.24
	HECES	18	15	9.648
	DyCluStream	697	35	373.592
	DyCluStream+	181	**10**	97.016
t8.8k	Denstream	74	74	38.48
	HECES	27	22	14.472
	DyCluStream	558	20	299.088
	DyCluStream+	217	**9**	116.312

จากภาพที่ 4-9 ถึง 4-11 แสดงให้เห็นว่าวิธีการ DyCluStream และ DyCluStream+ สามารถจัดกลุ่มได้ดีกว่า DenStream และ HECES แม้กลุ่มของข้อมูลจะมีคุณลักษณะของความหนาแน่นที่แตกต่างกัน และกลุ่มบางกลุ่มมีจำนวนใกล้เคียงกับข้อมูลรบกวน แต่เนื่องจากมีช่องว่างขนาดเล็กมากแบ่งระหว่างกลุ่มข้อมูลและโครงสร้างของ DyCluStream มีลักษณะเป็น dynamic hyper-ellipsoidal ที่มีทิศทาง จึงทำให้กลุ่มข้อมูลขนาดเล็กสามารถแผ่กระจายไปยังกลุ่มข้อมูลใกล้เคียงกันและมีลักษณะของรูปทรงที่แทนขอบของข้อมูลได้ แต่อย่างไรก็ตาม DyCluStream และ DyCluStream+ ยังคงใช้หน่วยความจำที่มากกว่า DenStream และ HECES เนื่องจากมีกลุ่มข้อมูลขนาดเล็กจำนวนมาก ดังตารางที่ 4-11

4.3.3 การออกแบบการทดลองของข้อมูลจริง

การทดลองกับข้อมูลจริง มีวัตถุประสงค์เพื่อแสดงให้เห็นถึงความสามารถในการจัดกลุ่มข้อมูลในแต่ละขั้นตอนวิธีการเมื่อกำหนดพารามิเตอร์ที่เท่ากัน โดยมีการเปรียบเทียบในด้านต่าง ๆ ดังนี้

- **ความสามารถในการจัดการกับจำนวนของข้อมูลที่ไหลเข้ามาในช่วงต่าง ๆ (Streaming speed)**

การทดลองนี้แสดงถึงผลกระทบของจำนวนของข้อมูลที่ไหลเข้ามาในช่วงต่าง ๆ ของแต่ละขั้นตอนวิธีการ โดยการทดลองจะกำหนดช่วงการไหลเข้าของข้อมูลเป็นจำนวนหนึ่ง ซึ่งหลังจากข้อมูลไหลเข้ามาตามจำนวนที่กำหนดแล้วจะดำเนินการจัดกลุ่มข้อมูลตามแต่ละขั้นตอนวิธีการ ซึ่งจะสลับการนำข้อมูลเข้าและการจัดกลุ่มข้อมูลไปเรื่อย ๆ และจะดำเนินการวัดผลเมื่อข้อมูลตัวสุดท้ายไหลเข้ามาและดำเนินการจัดกลุ่มครั้งสุดท้ายแล้ว

พารามิเตอร์ที่กำหนดให้กับ DyCluStream+ เป็นดังนี้

- Streaming Speed (Δ) 1,000, 5,000, 10,000 สำหรับข้อมูล Statlog (Shuttle) และ Letter Recognition
- Streaming Speed (Δ) 5,000, 1,0000, 50,000 สำหรับข้อมูล KDD Cup 1999 และ Covertype
- จำนวนสมาชิกใน n_ε เท่ากับจำนวนคุณลักษณะของข้อมูล
- merge function $\emptyset$ = 0
- Bhatthajariya distance (β) = 1
- Cosine Similairy (θ) = 0.8
- ระยะทาง (r_ε) เป็น 0.3

พารามิเตอร์ที่กำหนดให้กับ DenStream เป็นดังนี้

- Streaming Speed (Δ) 1,000, 5,000, 10,000 สำหรับข้อมูล Statlog (Shuttle) และ Letter Recognition
- Streaming Speed (Δ) 5,000, 10,000, 50,000 สำหรับข้อมูล KDD Cup 1999 และ Covertype
- จำนวนสมาชิกใน n_ε เท่ากับจำนวนคุณลักษณะของข้อมูล
- threshold สำหรับจำแนกว่าเป็น outlier micro-cluster (β) = 0.2

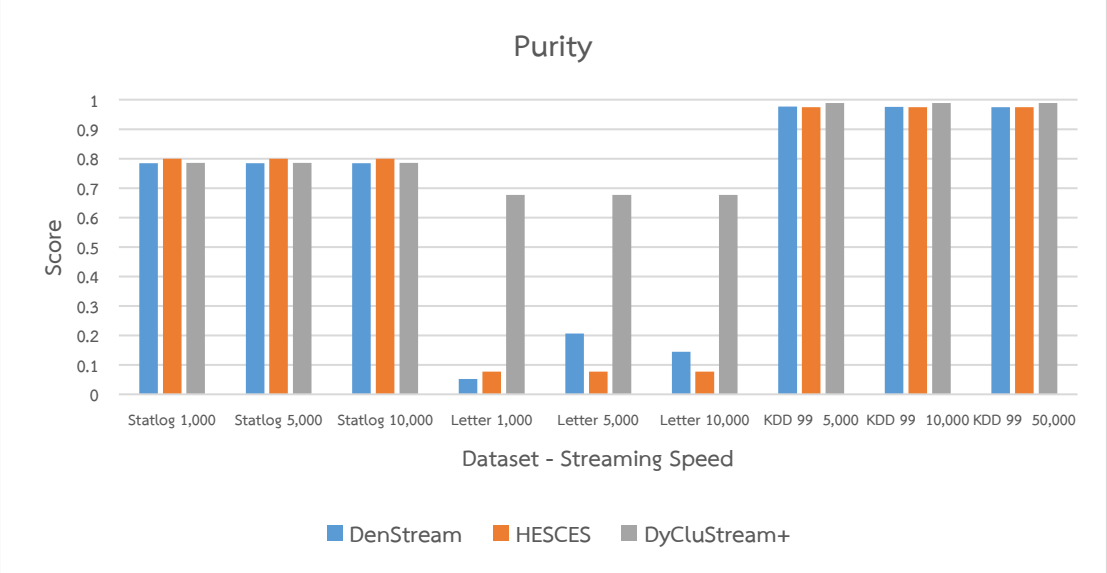
พารามิเตอร์ที่กำหนดให้กับ HECES เป็นดังนี้

- Streaming Speed (Δ) 1,000, 5,000, 10,000 สำหรับข้อมูล Statlog (Shuttle) และ Letter Recognition
- Streaming Speed (Δ) 5,000, 10,000, 50,000 สำหรับข้อมูล KDD Cup 1999 และ Covertype
- ความกว้างของ grid cell (r_ε) เป็น 0.3

สามารถแสดงผลการทดลองได้ดังตารางที่ 4-12 ถึง 4-17 และภาพที่ 4-12 ถึงภาพที่ 4-17

ตารางที่ 4-12 ผลการทดลองของตัววัด Purity สำหรับการปรับเปลี่ยนตัวแปร Stream Speed

Data sets : Stream Speed	Purity		
	DenStream	HECES	DyCluStream+
Statlog (Shuttle) : 1,000	0.784	**0.800**	0.785
Statlog (Shuttle) : 5,000	0.784	**0.800**	0.785
Statlog (Shuttle) : 10,000	0.784	**0.800**	0.785
Letter Recognition : 1,000	0.052	0.077	**0.677**
Letter Recognition : 5,000	0.206	0.077	**0.677**
Letter Recognition : 10,000	0.144	0.077	**0.677**
KDD Cup 1999 : 5,000	0.977	0.974	**0.989**
KDD Cup 1999 : 10,000	0.976	0.974	**0.989**
KDD Cup 1999 : 50,000	0.974	0.974	**0.988**



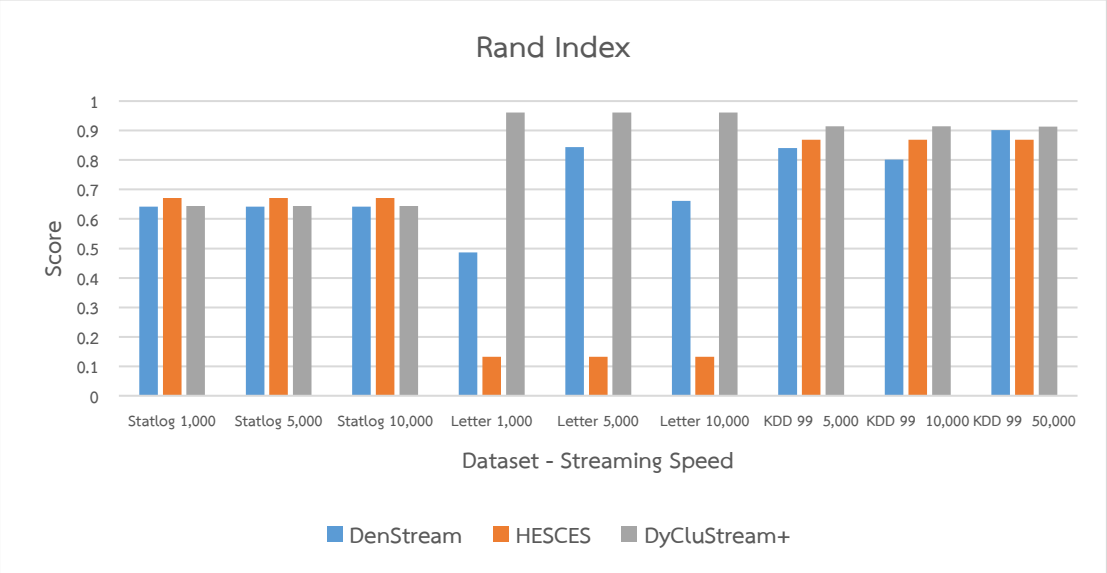
ภาพที่ 4-12 แผนภูมิแท่งเปรียบเทียบผลการทดลองของตัววัด Purity สำหรับการปรับเปลี่ยนตัวแปร Stream Speed

จากตารางที่ 4-12 และภาพที่ 4-12 แสดงให้เห็นว่าวิธีการ DyCluStream+ สามารถจัดกลุ่มในด้านความเป็นหนึ่งเดียวกันของข้อมูลได้ดีกว่า DenStream และ HECES กับชุดข้อมูล Letter Recognition อย่างชัดเจน และค่อนข้างดีกว่าเล็กน้อยในชุดข้อมูล KDD Cup 1999 เนื่องจากกลุ่มข้อมูลขนาดเล็กของ DyCluStream+ สามารถเข้ากับรูปร่างและทิศทางการกระจายตัวของข้อมูลได้ค่อนข้างดีกว่า สำหรับชุดข้อมูล Statlog (Shuttle) แสดงให้เห็นว่า HECES มีผลลัพธ์ที่ดีกว่าเล็กน้อย เนื่องมาจาก HECES ใช้วิธีการเก็บข้อมูลให้อยู่ในรูปแบบ grid-cell ทำให้สามารถกระจายข้อมูลให้อยู่ในระยะของ grid ที่เท่า ๆ กัน จึงสามารถจัดกลุ่มกับข้อมูลที่มีข้อมูลส่วนมากอยู่ในกลุ่มใดกลุ่มหนึ่ง

ค่อนข้างดีกว่าเล็กน้อย สำหรับชุดข้อมูล Letter Recognition แสดงให้เห็นว่า DenStream ค่อนข้างมีผลกระทบในเรื่องความเป็นหนึ่งเดียวกันของข้อมูลเมื่อมีการเปลี่ยนแปลงจำนวนของข้อมูลที่ไหลเข้ามาในช่วงต่าง ๆ เนื่องจาก DenStream ต้องการข้อมูลจำนวนหนึ่ง เพื่อสร้างกลุ่มข้อมูลขนาดเล็กตั้งต้นในระยะแรก

ตารางที่ 4-13 ผลการทดลองของตัววัด Rand statistic สำหรับการปรับเปลี่ยนตัวแปร Streamimg speed

Data sets : Stream Speed	Rand statistic		
	DenStream	HECES	DyCluStream+
Statlog (Shuttle) : 1,000	0.642	**0.671**	0.644
Statlog (Shuttle) : 5,000	0.642	**0.671**	0.644
Statlog (Shuttle) : 10,000	0.642	**0.671**	0.644
Letter Recognition : 1,000	0.486	0.132	**0.961**
Letter Recognition : 5,000	0.844	0.132	**0.961**
Letter Recognition : 10,000	0.661	0.132	**0.961**
KDD Cup 1999 : 5,000	0.841	0.869	**0.914**
KDD Cup 1999 : 10,000	0.801	0.869	**0.914**
KDD Cup 1999 : 50,000	0.901	0.869	**0.913**



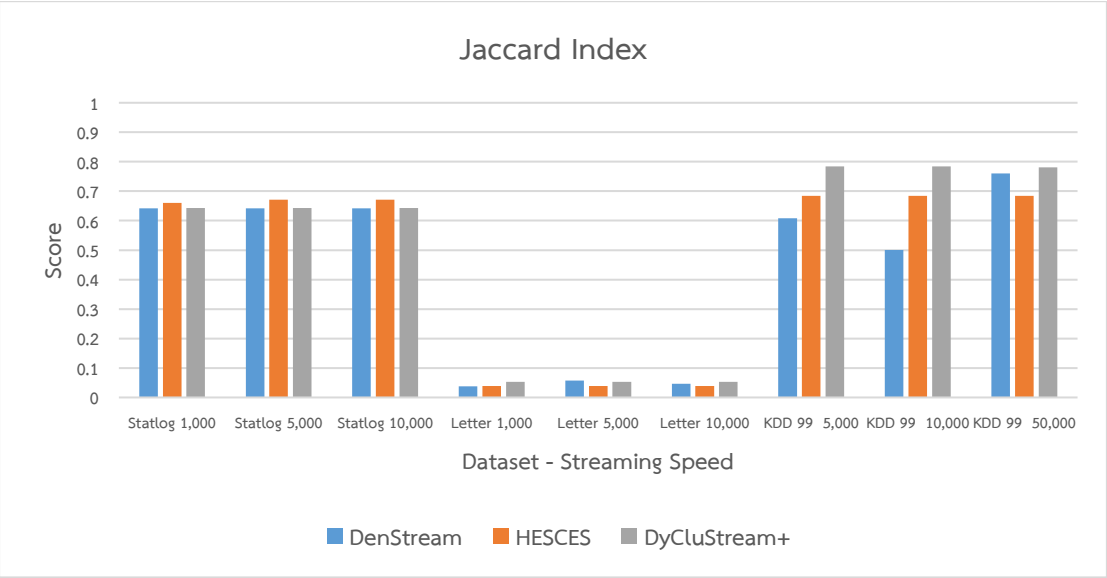
ภาพที่ 4-13 แผนภูมิแท่งเปรียบเทียบผลการทดลองของตัววัด Rand statistic สำหรับการปรับเปลี่ยนตัวแปร Streamimg Speed

จากตารางที่ 4-13 และภาพที่ 4-13 แสดงให้เห็นว่าวิธีการ DyCluStream+ สามารถจัดกลุ่มในด้านความแม่นยำได้ดีกว่า DenStream และ HECES กับชุดข้อมูล Letter Recognition อย่างชัดเจน และค่อนข้างดีกว่าเล็กน้อยในชุดข้อมูล KDD Cup 1999 เนื่องมาจากมาตรวัดในส่วนออฟไลน์

เหมาะสมกับลักษณะของกลุ่มข้อมูลขนาดเล็กที่มีทั้งขนาดและทิศทาง จึงทำให้ได้ผลลัพธ์ของกลุ่มข้อมูลที่มีความถูกต้อง สำหรับชุดข้อมูล Statlog (Shuttle) แสดงให้เห็นว่า HECES มีผลลัพธ์ที่ดีกว่าเล็กน้อย เนื่องมาจากขั้นตอนจัดกลุ่มสำหรับนำไปใช้งานเกิดจากการผสานกลุ่มข้อมูลขนาดเล็ก จึงสามารถจัดกลุ่มกับข้อมูลที่มีข้อมูลส่วนมากอยู่ในกลุ่มใดกลุ่มหนึ่งค่อนข้างดีกว่าเล็กน้อย สำหรับชุดข้อมูล Letter Recognition แสดงให้เห็นว่า DenStream ค่อนข้างมีผลกระทบในเรื่องความแม่นยำของข้อมูลเมื่อมีการเปลี่ยนแปลงจำนวนของข้อมูลที่ไหลเข้ามาในช่วงต่าง ๆ เนื่องจาก DenStream ต้องการข้อมูลจำนวนหนึ่ง เพื่อสร้างกลุ่มข้อมูลขนาดเล็กตั้งต้นในระยะแรก

ตารางที่ 4-14 ผลการทดลองของตัววัด Jaccard Index สำหรับการปรับเปลี่ยนตัวแปร Streaming Speed

Data sets : Stream Speed	Jaccard Index		
	DenStream	HECES	DyCluStream+
Statlog (Shuttle) : 1,000	0.642	**0.660**	0.643
Statlog (Shuttle) : 5,000	0.642	**0.660**	0.643
Statlog (Shuttle) : 10,000	0.642	**0.660**	0.643
Letter Recognition : 1,000	0.038	0.039	**0.053**
Letter Recognition : 5,000	**0.057**	0.039	0.053
Letter Recognition : 10,000	0.046	0.039	**0.053**
KDD Cup 1999 : 5,000	0.608	0.684	**0.784**
KDD Cup 1999 : 10,000	0.501	0.684	**0.784**
KDD Cup 1999 : 50,000	0.760	0.684	**0.781**

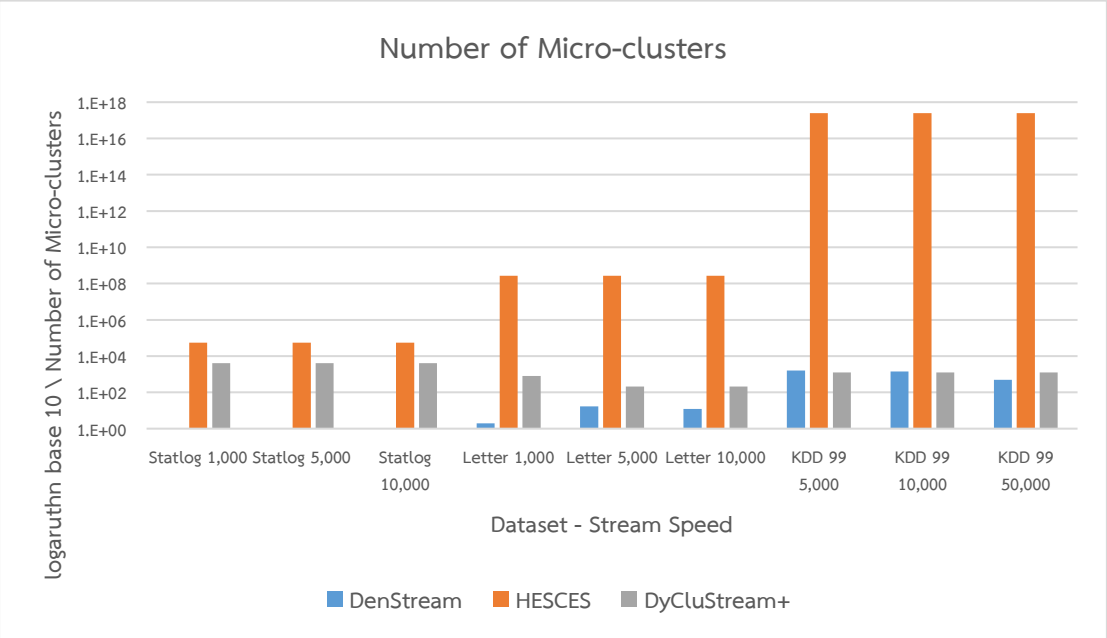


ภาพที่ 4-14 แผนภูมิแท่งเปรียบเทียบผลการทดลองของตัววัด Jaccard Index สำหรับการปรับเปลี่ยนตัวแปร Streaming Speed

จากตารางที่ 4-14 และภาพที่ 4-14 แสดงให้เห็นว่าวิธีการ DyCluStream+ สามารถจัดกลุ่มในด้านความถูกต้องในการจัดกลุ่มข้อมูลให้อยู่ในกลุ่มเดียวกันได้ดีกว่า DenStream และ HECES กับชุดข้อมูล KDD Cup 1999 อย่างชัดเจน เนื่องมาจากโครงสร้างของกลุ่มข้อมูลขนาดเล็กที่สามารถแทนทิศทางและการกระจายตัวของข้อมูลได้อย่างถูกต้องและมีความกระชับ จึงทำให้ชุดข้อมูลที่มีจำนวนกลุ่มมากสามารถถูกจัดกลุ่มได้อย่างถูกต้อง สำหรับชุดข้อมูล Statlog (Shuttle) และ Letter Recognition ที่ให้ผลลัพธ์ในการจัดกลุ่มที่ไม่แตกต่างกับขั้นตอนวิธีการอื่น เนื่องมาจากจำนวนข้อมูลในชุดข้อมูลมีค่อนข้างน้อย จึงทำให้แต่ละกลุ่มข้อมูลยังไม่สามารถแบ่งได้อย่างชัดเจน

ตารางที่ 4-15 ผลลัพธ์ของกลุ่มข้อมูลขนาดเล็กที่ผ่านการจัดกลุ่ม สำหรับการปรับเปลี่ยนตัวแปร Streaming Speed

Data sets : Stream Speed	Number of micro-clusters		
	DenStream	HECES	DyCluStream+
Statlog (Shuttle) : 1,000	1	55,565	4,134
Statlog (Shuttle) : 5,000	1	55,565	4,134
Statlog (Shuttle) : 10,000	1	55,565	4,134
Letter Recognition : 1,000	2	272,396,740	801
Letter Recognition : 5,000	17	272,396,740	802
Letter Recognition : 10,000	12	272,396,740	802
KDD Cup 1999 : 5,000	1,626	249,806,611,870,909,000	1,276
KDD Cup 1999 : 10,000	1,438	249,806,611,870,909,000	1,276
KDD Cup 1999 : 50,000	490	249,806,611,870,909,000	1,273

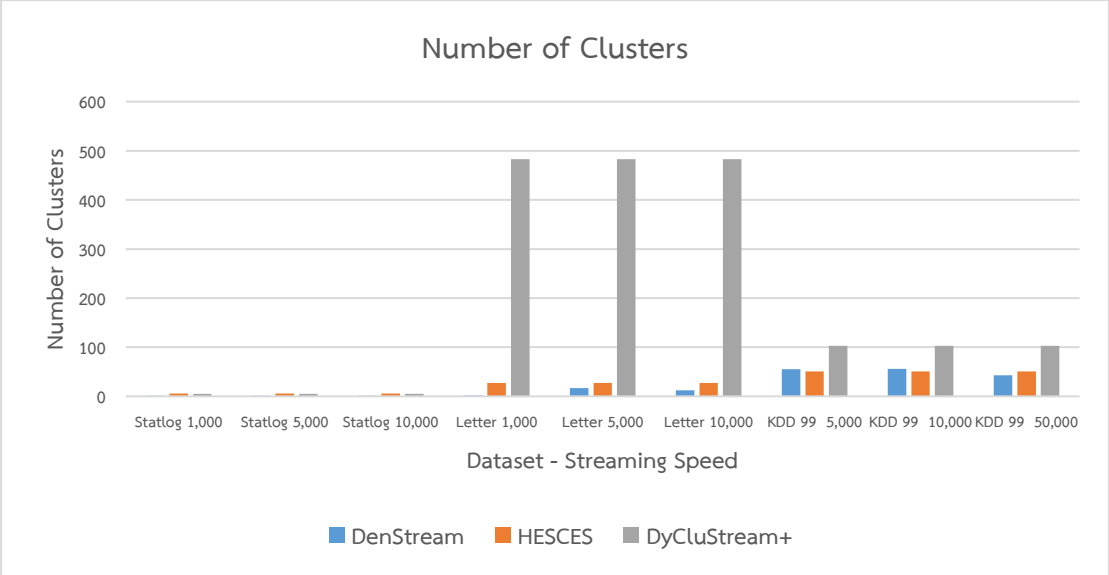


ภาพที่ 4-15 แผนภูมิแท่งเปรียบเทียบผลลัพธ์ของกลุ่มข้อมูลขนาดเล็กที่ผ่านการจัดกลุ่ม สำหรับการปรับเปลี่ยนตัวแปร Streaming Speed

จากตารางที่ 4-15 และภาพที่ 4-15 แสดงถึงจำนวน micro-cluters ของ DyCluStream+ ที่มีจำนวนมากกว่า DenStream เป็นส่วนมากเนื่องจาก DyCluStream+ จะมีการปรับรูปร่างของ micro-cluster ที่เป็น dynamic hyper-ellipsoidal ให้มีลักษณะที่สามารถครอบคลุมกับข้อมูลที่เป็นสมาชิก จึงทำให้มีจำนวน micro-cluster มากกว่า DenStream ซึ่งใช้ micro-clusters ที่มีโครงสร้างเป็น spherical จึงกินพื้นที่ขนาดใหญ่ ส่งผลให้จำนวน micro-clusters น้อย ส่วน HECES นั้นใช้ grid-cells ที่มีระยะความกว้างมากจึงกินพื้นที่ขนาดใหญ่เช่นเดียวกับ DenStream แต่อย่างไรก็ตาม HECES มีการกำหนดจำนวนของ grid-cells ที่มีจำนวนมากตามมิติของข้อมูลเพื่อใช้ในการเก็บข้อมูลสถิติ จึงทำให้มีจำนวน micro-cluster เยอะตามจำนวน grid-cells และสำหรับชุดข้อมูล Letter แสดงให้เห็นว่า DenStream ค่อนข้างมีผลกระทบในเรื่องความถูกต้องของการจัดกลุ่มเมื่อมีการเปลี่ยนแปลงจำนวนของข้อมูลที่ไหลเข้ามาในช่วงต่าง ๆ

ตารางที่ 4-16 ผลลัพธ์ของกลุ่มข้อมูลที่ผ่านการจัดกลุ่ม สำหรับการปรับเปลี่ยนตัวแปร Streaming Speed

Data sets : Stream Speed	Number of Clusters		
	DenStream	HECES	DyCluStream+
Statlog (Shuttle) : 1,000	1	6	5
Statlog (Shuttle) : 5,000	1	6	5
Statlog (Shuttle) : 10,000	1	6	5
Letter Recognition : 1,000	2	27	483
Letter Recognition : 5,000	17	27	483
Letter Recognition : 10,000	12	27	483
KDD Cup 1999 : 5,000	55	51	103
KDD Cup 1999 : 10,000	56	51	103
KDD Cup 1999 : 50,000	43	51	103

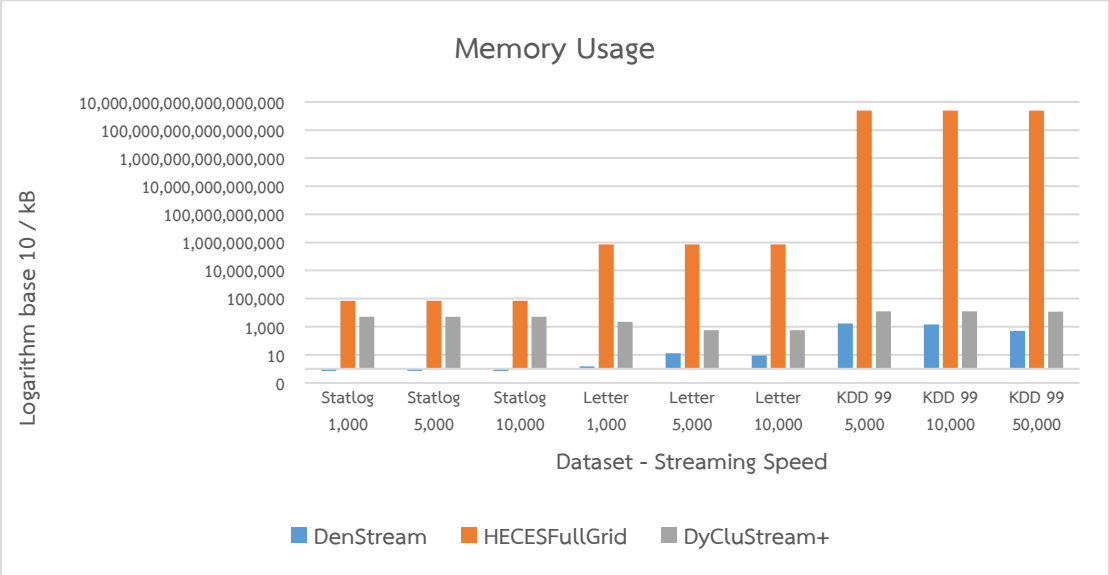


ภาพที่ 4-16 แผนภูมิแท่งเปรียบเทียบผลลัพธ์ของกลุ่มข้อมูลที่ผ่านการจัดกลุ่ม สำหรับการปรับเปลี่ยนตัวแปร Streaming Speed

จากตารางที่ 4-16 และภาพที่ 4-16 แสดงถึงจำนวนกลุ่มข้อมูลของ DyCluStream+ ที่มากกว่าเนื่องมาจาก DyCluStream+ ใช้ลักษณะของ micro-clusters ที่เป็น dynamic hyper-ellipsoidal ที่เข้ากับลักษณะของข้อมูล ทำให้มีจำนวน micro-clusters มาก ซึ่งหลังจากดำเนินการในส่วนออฟไลน์ เพื่อให้ได้กลุ่มข้อมูลขนาดใหญ่สำหรับนำไปใช้งานจริงของ DyCluStream+ ซึ่งใช้มาตรวัดระยะทางที่เน้นเฉพาะ micro-clusters ที่มีการซ้อนทับกันเท่านั้น จึงมี micro-cluster บางกลุ่มที่ไม่สามารถถูกแผ่กระจายได้เนื่องจากอยู่ไกลกับกลุ่มอื่นหรือเป็นข้อมูลที่เป็น จึงทำให้ได้จำนวนกลุ่มข้อมูลสำหรับนำไปใช้งานที่มีจำนวนมาก

ตารางที่ 4-17 หน่วยความจำที่ใช้งาน สำหรับการปรับเปลี่ยนตัวแปร Streaming Speed

Data sets : Stream Speed	Memory Usage/kB		
	DenStream	HECES	DyCluStream+
Statlog (Shuttle) : 1,000	1	71,246	5,184
Statlog (Shuttle) : 5,000	1	71,246	5,184
Statlog (Shuttle) : 10,000	1	71,246	5,184
Letter Recognition : 1,000	1.488	725,666,198	2,137
Letter Recognition : 5,000	13	725,666,198	554
Letter Recognition : 10,000	9	725,666,198	554
KDD Cup 1999 : 5,000	1,652	2,364,169,774,746,380,000	12,076
KDD Cup 1999 : 10,000	1,461	2,364,169,774,746,380,000	12,076
KDD Cup 1999 : 50,000	497.84	2,364,169,774,746,380,000	12,048



ภาพที่ 4-17 แผนภูมิแท่งเปรียบเทียบหน่วยความจำที่ใช้งาน สำหรับการปรับเปลี่ยนตัวแปร Streaming Speed

จากตารางที่ 4-17 และภาพที่ 4-17 แสดงถึงปริมาณหน่วยความจำที่ใช้ของขั้นตอนวิธีการต่าง ๆ โดย HECES จะใช้หน่วยความจำมากที่สุด เนื่องจากจะมีการจองพื้นที่ในหน่วยความจำเพื่อสร้าง grid-cells รวมไปถึงการเก็บข้อมูลในแต่ละ grid-cell เพื่อใช้ในการคำนวณหา Covariance matrix ในขั้นตอนการรวมกลุ่มข้อมูล ส่วน DenStream และ DyCluStream+ มีปริมาณการใช้หน่วยความจำที่ค่อนข้างน้อยกว่า HECES เพราะจะมีการจองหน่วยความจำก็ต่อเมื่อมี micro-clusters กลุ่มใหม่เท่านั้น และเนื่องจาก DyCluStream+ ใช้กลุ่มข้อมูลขนาดเล็กที่กระชับกับ

ลักษณะของข้อมูลในกลุ่ม จึงทำให้มีกลุ่มข้อมูลขนาดเล็กมีจำนวนมาก ส่งผลทำให้ใช้หน่วยความจำมากตามจำนวน micro-clusters

- **การมีผลของระยะทางที่ใช้ในการจัดกลุ่มข้อมูล**

เนื่องจากแต่ละขั้นตอนวิธีการใช้มาตรวัดระยะทางที่แตกต่างกัน ดังนั้นระยะทางที่กำหนดให้เท่ากันในแต่ละขั้นตอนวิธีการจึงมีผลต่อประสิทธิภาพในการจัดกลุ่มของข้อมูล

พารามิเตอร์ที่กำหนดให้กับ DyCluStream+ เป็นดังนี้

- ระยะทาง (r_ε) เป็น 0.2, 0.25, 0.3
- จำนวนสมาชิกใน (n_ε) เท่ากับจำนวนคุณลักษณะของข้อมูล
- merge function $(\emptyset)$ = 0
- Bhatthajariya distance (β) = 1
- Cosine Similairy (θ) = 0.8
- Streaming Speed (Δ) 5,000 สำหรับข้อมูล Statlog (Shuttle) และ Letter Recognition
- Streaming Speed (Δ) 10,000 สำหรับข้อมูล KDD Cup 1999 และ Covertype

พารามิเตอร์ที่กำหนดให้กับ DenStream เป็นดังนี้

- ระยะทาง (r_ε) เป็น 0.2, 0.25, 0.3
- จำนวนสมาชิกใน (n_ε) เท่ากับจำนวนคุณลักษณะของข้อมูล
- threshold สำหรับจำแนกว่าเป็น outlier micro-cluster (β) = 0.2
- Streaming Speed (Δ) 5,000 สำหรับข้อมูล Statlog (Shuttle) และ Letter Recognition
- Streaming Speed (Δ) 10,000 สำหรับข้อมูล KDD Cup 1999 และ Covertype

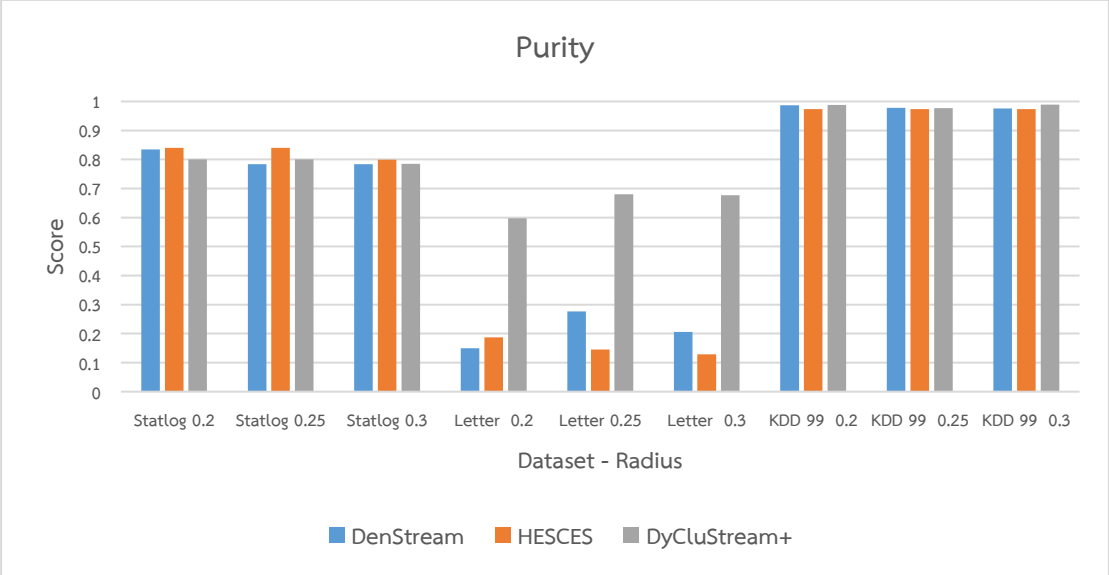
พารามิเตอร์ที่กำหนดให้กับ HECSE เป็นดังนี้

- ระยะทาง (r_ε) เป็น 0.2, 0.25, 0.3

สามารถแสดงผลการทดลองได้ดังภาพที่ 18 ถึงภาพที่ 23

ตารางที่ 4-18 ผลการทดลองของตัววัด Purity สำหรับการปรับเปลี่ยนตัวแปร Radius

Data sets : Radius	Purity		
	DenStream	HECES	DyCluStream+
Statlog (Shuttle) : 0.2	0.835	**0.840**	0.800
Statlog (Shuttle) : 0.25	0.784	0.784	**0.800**
Statlog (Shuttle) : 0.3	0.784	**0.800**	0.785
Letter Recognition : 0.2	0.150	0.203	**0.597**
Letter Recognition : 0.25	0.277	0.107	**0.680**
Letter Recognition : 0.3	0.206	0.077	**0.677**
KDD Cup 1999 : 0.2	0.987	0.778	**0.988**
KDD Cup 1999 : 0.25	**0.978**	0.956	0.977
KDD Cup 1999 : 0.3	0.976	0.974	**0.989**



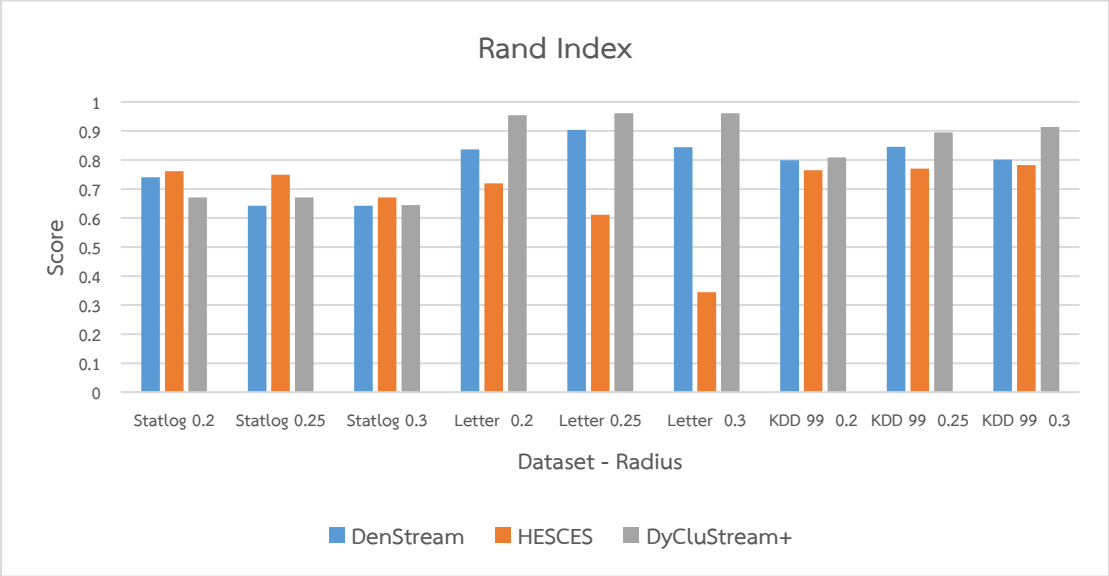
ภาพที่ 4-18 แผนภูมิแท่งเปรียบเทียบผลการทดลองของตัววัด Purity สำหรับการปรับเปลี่ยนตัวแปร Radius

จากตารางที่ 18 และภาพที่ 4-18 แสดงให้เห็นว่าวิธีการ DyCluStream+ สามารถจัดกลุ่มในด้านความเป็นหนึ่งเดียวกันของข้อมูลได้ดีกว่า DenStream และ HECES กับชุดข้อมูล Letter Recognition อย่างชัดเจน และค่อนข้างดีกว่าเล็กน้อยในชุดข้อมูล KDD Cup 1999 ในแต่ละค่าระยะทางเนื่องจากใช้ micro-cluster และมาตรวัดระยะทางที่เป็น dynamic hyper-ellipsoidal ที่สามารถแทนลักษณะและทิศทางการกระจายตัวของข้อมูลจึงสามารถจัดกลุ่มได้อย่างถูกต้อง สำหรับชุดข้อมูล Statlog (Shuttle) ของ HECES ที่ให้ผลลัพธ์ในการจัดกลุ่มที่ดีกว่า DyCluStream+

เล็กน้อย เนื่องมากจาก HECES ใช้โครงสร้างของข้อมูลที่มีลักษณะเป็น grid-cells ที่สามารถแบ่งช่วงของข้อมูลที่เป็นสมาชิกได้จำนวนที่เท่า ๆ กัน ส่งผลให้มาตรวัดระยะทางที่ประยุกต์ใช้ในจัดกลุ่มข้อมูลในส่วนออฟไลน์มีค่าค่อนข้างใกล้กันจึงทำให้ผลลัพธ์การจัดกลุ่มข้อมูล ที่ข้อมูลส่วนมากอยู่ในกลุ่มใดกลุ่มหนึ่งเป็นจำนวนมากสามารถทำได้ดี

ตารางที่ 4-19 ผลการทดลองของตัววัด Rand statistic สำหรับการปรับเปลี่ยนตัวแปร Radius

Data sets : Radius	Rand statistic		
	DenStream	HECES	DyCluStream+
Statlog (Shuttle) : 0.2	0.740	**0.747**	0.671
Statlog (Shuttle) : 0.25	0.642	0.643	**0.671**
Statlog (Shuttle) : 0.3	0.642	**0.671**	0.644
Letter Recognition : 0.2	0.836	0.764	**0.954**
Letter Recognition : 0.25	0.904	0.176	**0.961**
Letter Recognition : 0.3	0.844	0.132	**0.961**
KDD Cup 1999 : 0.2	0.799	0.675	**0.809**
KDD Cup 1999 : 0.25	0.845	0.829	**0.895**
KDD Cup 1999 : 0.3	0.801	0.869	**0.914**

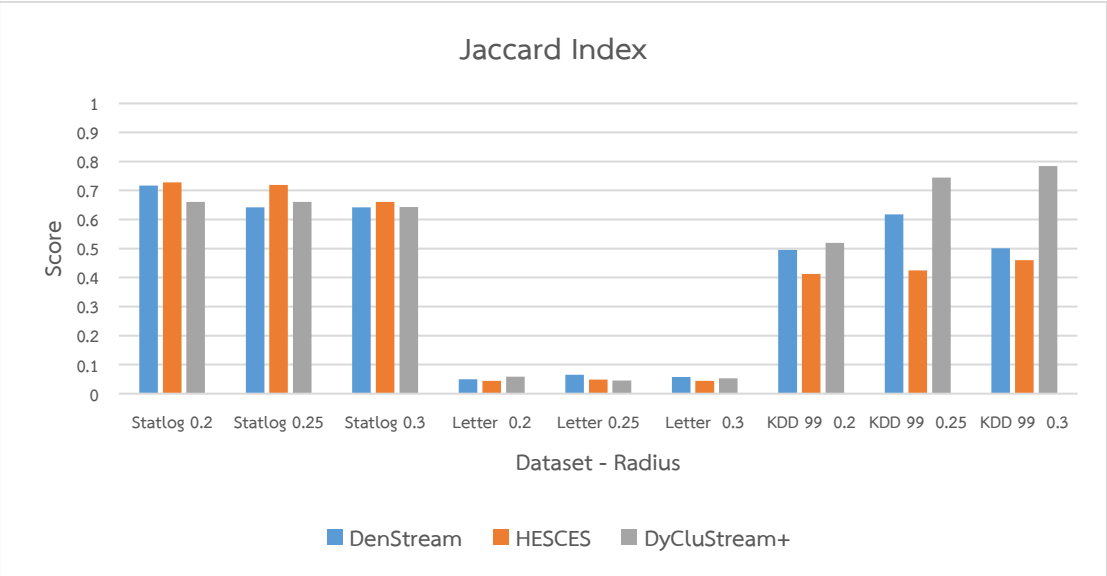


ภาพที่ 4-19 แผนภูมิแท่งเปรียบเทียบผลการทดลองของตัววัด Rand statistic สำหรับการปรับเปลี่ยนตัวแปร Radius

จากตารางที่ 4-19 และภาพที่ 4-19 แสดงให้เห็นว่าวิธีการ DyCluStream+ สามารถจัดกลุ่มในด้านความแม่นยำได้ดีกว่า DenStream และ HECES กับชุดข้อมูล Letter Recognition อย่างชัดเจน และค่อนข้างดีกว่าเล็กน้อยในชุดข้อมูล KDD Cup 1999 ซึ่งการเปลี่ยนแปลงค่าของตัวแปร Radius ส่งผลกระทบต่อผลลัพธ์ของขั้นตอนวิธีการ DyCluStream+ ค่อนข้างน้อยกว่า DenStream และ HECES ส่วนชุดข้อมูล Letter Recognition แสดงให้เห็นถึงผลกระทบของการเปลี่ยนแปลงค่าของตัวแปร Radius ของ HECES ที่สังเกตเห็นได้ว่า เมื่อค่าตัวแปร Radius มีค่าเพิ่มมากขึ้น ความแม่นยำก็จะน้อยลง เนื่องจาก grid-cell ของ HECES ไม่สามารถที่จะแทนที่การกระจายตัวของข้อมูลได้อย่างถูกต้อง

ตารางที่ 4-20 ผลการทดลองของตัววัด Jaccard Index สำหรับการปรับเปลี่ยนตัวแปร Radius

Data sets : Radius	Jaccard Index		
	DenStream	HECES	DyCluStream+
Statlog (Shuttle) : 0.2	**0.717**	0.717	0.661
Statlog (Shuttle) : 0.25	0.642	0.642	**0.661**
Statlog (Shuttle) : 0.3	0.642	**0.660**	0.643
Letter Recognition : 0.2	0.050	0.045	**0.058**
Letter Recognition : 0.25	**0.065**	0.040	0.045
Letter Recognition : 0.3	**0.057**	0.039	0.053
KDD Cup 1999 : 0.2	0.495	0.353	**0.519**
KDD Cup 1999 : 0.25	0.618	0.603	**0.745**
KDD Cup 1999 : 0.3	0.501	0.684	**0.784**

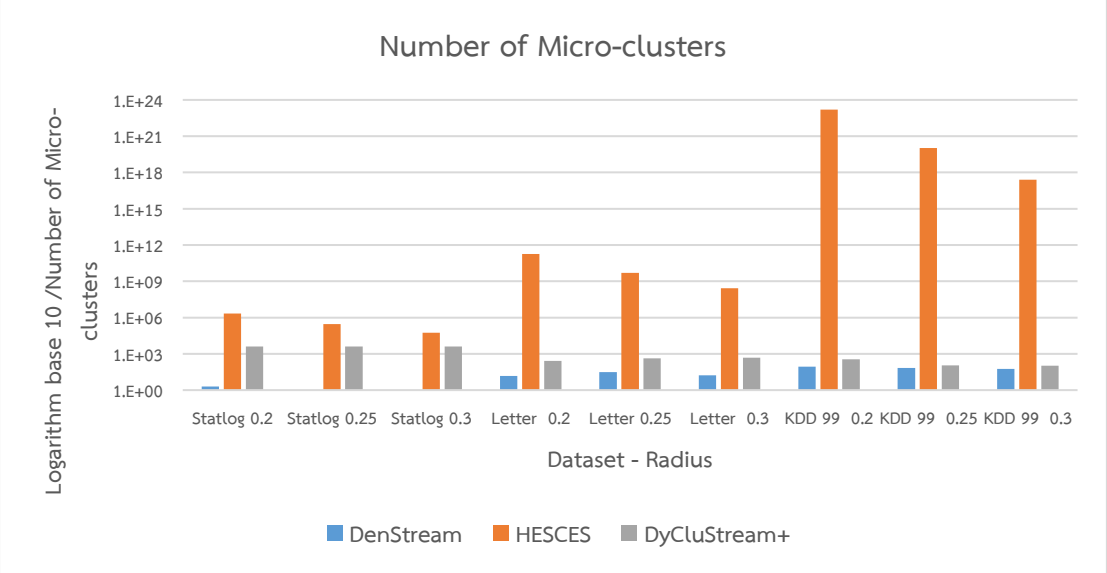


ภาพที่ 4-20 แผนภูมิแท่งเปรียบเทียบผลการทดลองของตัววัด Jaccard Index สำหรับการปรับเปลี่ยนตัวแปร Radius

จากตารางที่ 4-20 และภาพที่ 4-20 แสดงให้เห็นว่าวิธีการ DyCluStream+ สามารถจัดกลุ่มในด้านความถูกต้องในการจัดกลุ่มข้อมูลให้อยู่ในกลุ่มเดียวกันได้ดีกว่า DenStream และ HECES กับชุดข้อมูล KDD Cup 1999 อย่างชัดเจน โดยเมื่อทำการปรับเปลี่ยนตัวแปร Radius ให้มากขึ้น ความสามารถในการจัดกลุ่มให้อยู่ในกลุ่มเดียวกันของ DyCluStream+ ก็เพิ่มมากขึ้นด้วย ส่วนชุดข้อมูล Letter Recognition ผลลัพธ์การจัดกลุ่มของทั้งสามวิธีค่อนข้างใกล้เคียงกัน แต่สำหรับชุดข้อมูล Statlog (Shuttle) ขั้นตอนวิธีการ HECES ค่อนข้างทำได้ดีกว่าเล็กน้อย เนื่องจากลักษณะโครงสร้าง grid-cell ที่ช่วยในการกระจายข้อมูลให้อยู่ในแต่ละกลุ่มเท่าๆกัน จึงทำให้ข้อมูลส่วนมากที่อยู่ในกลุ่มใดกลุ่มหนึ่ง สามารถผสานกันได้เนื่องจากมีระยะทางระหว่างกลุ่มที่น้อย

ตารางที่ 4-21 ผลลัพธ์ของกลุ่มข้อมูลขนาดเล็กที่ผ่านการจัดกลุ่ม สำหรับการปรับเปลี่ยนตัวแปร Radius

Data sets : Radius	Number of Micro-clusters		
	DenStream	HECES	DyCluStream+
Statlog (Shuttle) : 0.2	2	2,136,105	4,086
Statlog (Shuttle) : 0.25	1	286,704	4,097
Statlog (Shuttle) : 0.3	1	55,565	4,134
Letter Recognition : 0.2	15	178,921,302,021	285
Letter Recognition : 0.25	31	5,036,186,932	562
Letter Recognition : 0.3	17	272,396,740	801
KDD Cup 1999 : 0.2	1,665	161,664,803,199,983,000,000,000	1,234
KDD Cup 1999 : 0.25	1,561	102,467,242,406,119,000,000	1,440
KDD Cup 1999 : 0.3	1,438	249,806,611,870,909,000	1,276

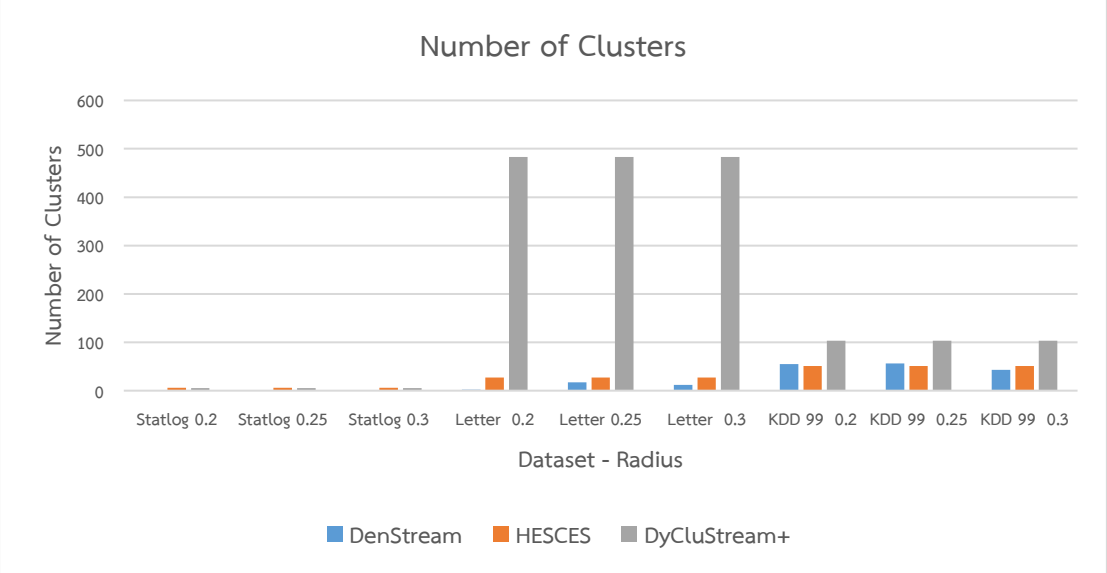


ภาพที่ 4-21 แผนภูมิแท่งเปรียบเทียบผลลัพธ์ของกลุ่มข้อมูลขนาดเล็กที่ผ่านการจัดกลุ่ม สำหรับการปรับเปลี่ยนตัวแปร Radius

จากตารางที่ 4-21 และภาพที่ 4-21 แสดงถึงจำนวน micro-cluters ของ HECES ที่มีจำนวนมากที่สุดเนื่องจาก grid-cells จะมีจำนวนมากตามมิติของข้อมูล และจะมีขนาดลงลงตามตัวแปร Radius ส่วน DyCluStream+ ที่มีจำนวน micro-clusters มากกว่า DenStream เนื่องจากมีการใช้ micro-cluster ที่เป็น dynamic hyper-ellipsoidal ที่สามารถปรับให้เข้ากับการกระจายตัวของข้อมูลได้ ทำให้ใช้พื้นที่น้อยกว่า spherical ของ DenStream ซึ่งส่งผลให้ micro-cluster มีจำนวนมาก ส่วน micro-cluster ของ DyCluStream+ ที่จำนวนไม่แปรผันตามตัวแปร Radius เนื่องจากจำนวนครั้งในการรวมกันของกลุ่มข้อมูลขนาดเล็กไม่เท่ากันในแต่ละตัวแปร Radius

ตารางที่ 4-22 ผลลัพธ์ของกลุ่มข้อมูลที่ผ่านการจัดกลุ่ม สำหรับการปรับเปลี่ยนตัวแปร Radius

Data sets : Radius	Number of Clusters		
	DenStream	HECES	DyCluStream+
Statlog (Shuttle) : 0.2	2	8	2
Statlog (Shuttle) : 0.25	1	9	5
Statlog (Shuttle) : 0.3	1	6	5
Letter Recognition : 0.2	15	121	255
Letter Recognition : 0.25	31	65	428
Letter Recognition : 0.3	17	27	483
KDD Cup 1999 : 0.2	87	30	360
KDD Cup 1999 : 0.25	67	66	108
KDD Cup 1999 : 0.3	56	51	103

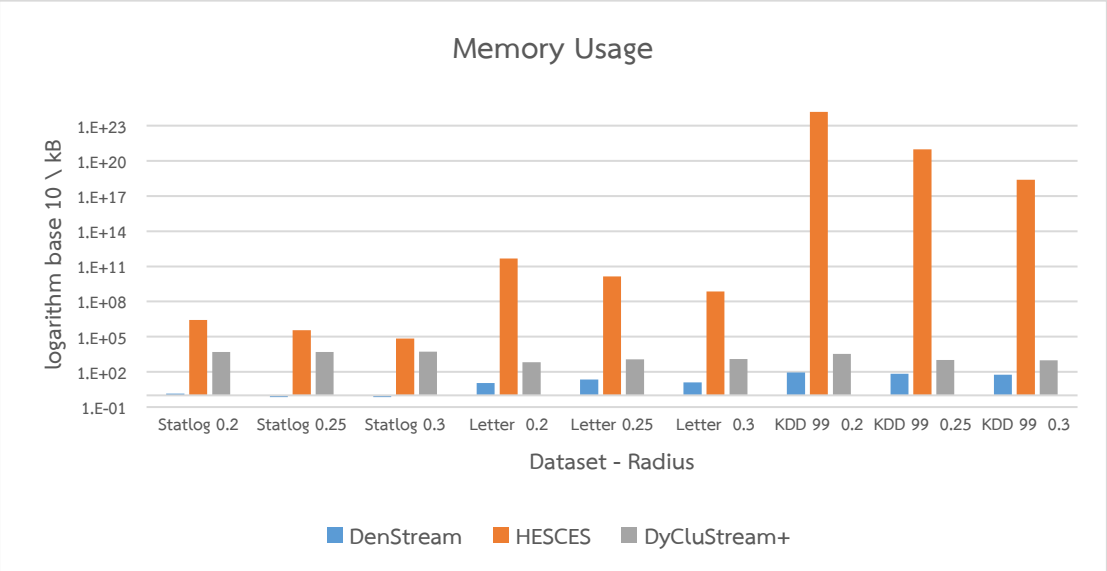


ภาพที่ 4-22 แผนภูมิแท่งเปรียบเทียบผลลัพธ์ของกลุ่มข้อมูลที่ผ่านการจัดกลุ่ม สำหรับการปรับเปลี่ยนตัวแปร Radius

จากตารางที่ 4-22 และภาพที่ 4-22 แสดงถึงจำนวนกลุ่มข้อมูลของ DyCluStream+ ที่มีจำนวนมากกว่า DenStream และ HECES ซึ่งจำนวนกลุ่มข้อมูลที่มากกว่าเกิดจาก micro-cluster ที่จะถูกจัดกลุ่มโดยการแผ่ขยายการเชื่อมต่อกันเพื่อเป็นกลุ่มข้อมูลขนาดใหญ่มีจำนวนที่มาก ซึ่งส่งผลให้หลังการจัดกลุ่มอาจมี micro-cluster ที่อยู่ไกลไม่ได้รับการจัดกลุ่มจึงถูกนับว่าเป็นกลุ่มใหม่เพราะ DyCluStream+ จะเชื่อมต่อ micro-cluster ที่มีการซ้อนทับกันเท่านั้น แต่อย่างไรก็ตามจากตัวมาตรวัดที่งานวิจัยนี้ใช้เพื่อบอกประสิทธิภาพของการจัดกลุ่มข้อมูล แสดงให้เห็นว่าวิธีการ DyCluStream+ ให้ประสิทธิภาพที่ดีกว่า DenStream และ HECES เป็นส่วนมาก

ตารางที่ 4-23 หน่วยความจำที่ใช้งาน สำหรับการปรับเปลี่ยนตัวแปร Radius

Data sets : Radius	Memory Usage/kB		
	DenStream	HECES	DyCluStream+
Statlog (Shuttle) : 0.2	1	2,680,243	5,124
Statlog (Shuttle) : 0.25	1	361,095	5,138
Statlog (Shuttle) : 0.3	1	71,246	5,184
Letter Recognition : 0.2	11	476,646,349,866	679
Letter Recognition : 0.25	23	13,416,403,269	1,140
Letter Recognition : 0.3	13	725,666,198	1,287
KDD Cup 1999 : 0.2	88	1,529,995,697,484,640,000,000,000	3,407
KDD Cup 1999 : 0.25	68	969,749,982,131,510,000,000	1,022
KDD Cup 1999 : 0.3	57	2,364,169,774,746,380,000	975



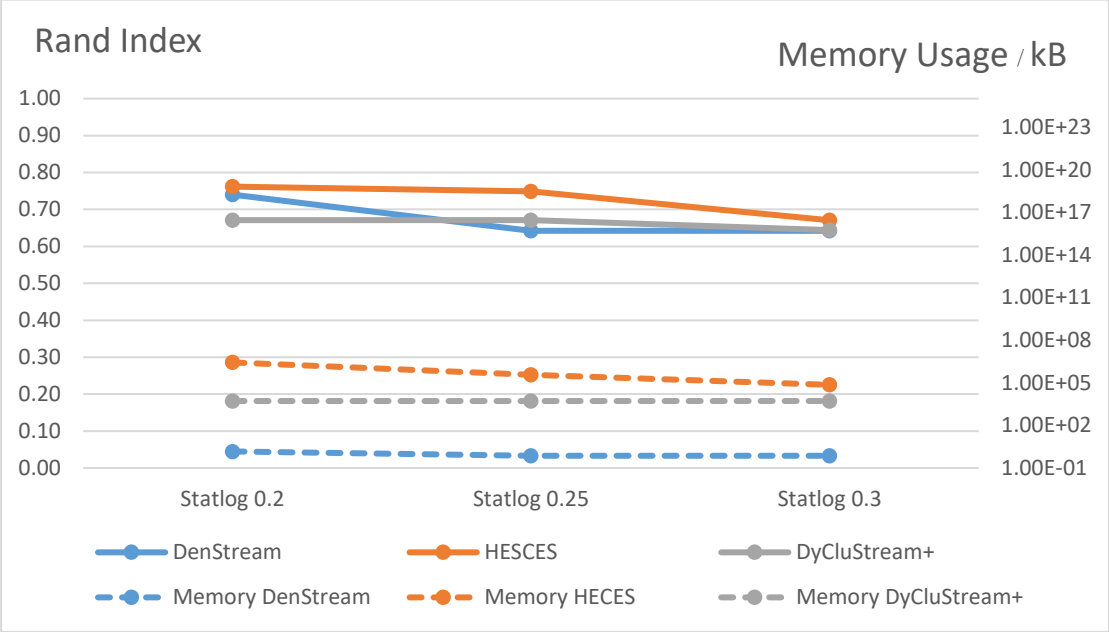
ภาพที่ 4-23 แผนภูมิแท่งเปรียบเทียบหน่วยความจำที่ใช้งาน สำหรับการปรับเปลี่ยนตัวแปร Radius

จากตารางที่ 4-23 และภาพที่ 4-23 แสดงถึงปริมาณหน่วยความจำที่ใช้ในแต่ละขั้นตอนวิธีการ โดย HECES จะใช้หน่วยความจำมากที่สุด เนื่องจากมีจำนวน grid-cell ที่มาก และขนาดของ

หน่วยความจำจะลดลงเรื่อยๆตามตัวแปร Radius เนื่องจาก grid-cell มีลักษณะของการคลอบคลุมที่กว้าง ส่วน DyCluStream+ จะใช้หน่วยความจำที่มากกว่า DenStream เนื่องจากมีจำนวน micro-clusters ที่มากกว่า DenStream

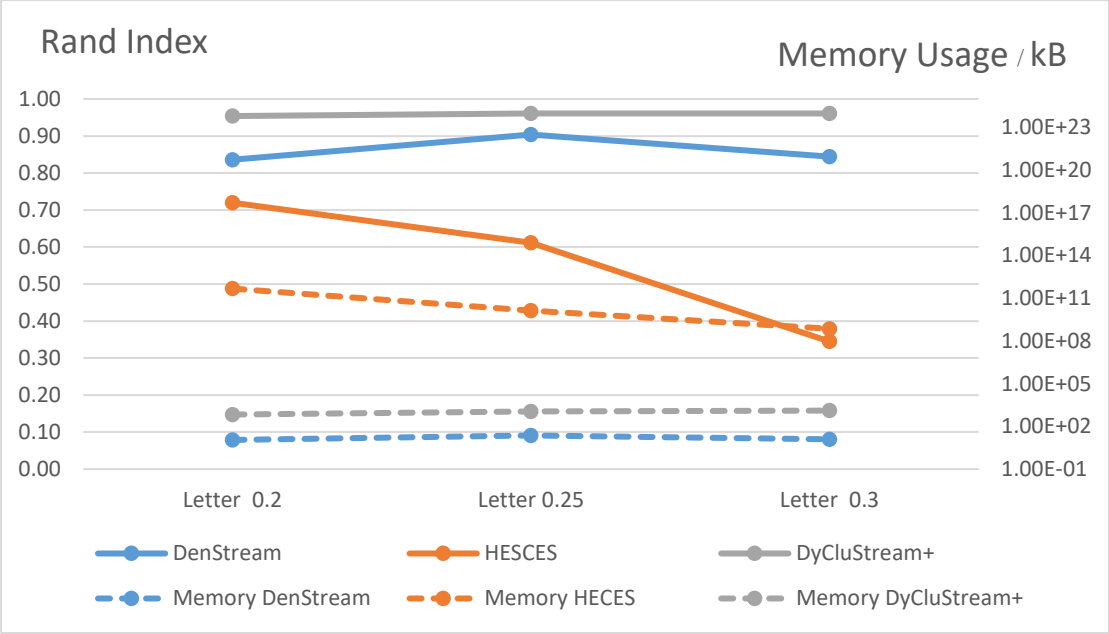
- **ผลกระทบระหว่างหน่วยความจำที่ใช้งานและความแม่นยำ**

การทดลองนี้แสดงถึงผลกระทบระหว่างหน่วยความจำที่ใช้งานและความแม่นยำในแต่ละขั้นตอนวิธีการ โดยหน่วยความจำที่ใช้งานจะมีการเปลี่ยนแปลงจากตัวแปร Radius สามารถแสดงได้ดังภาพที่ 4-24 ถึง 4-26



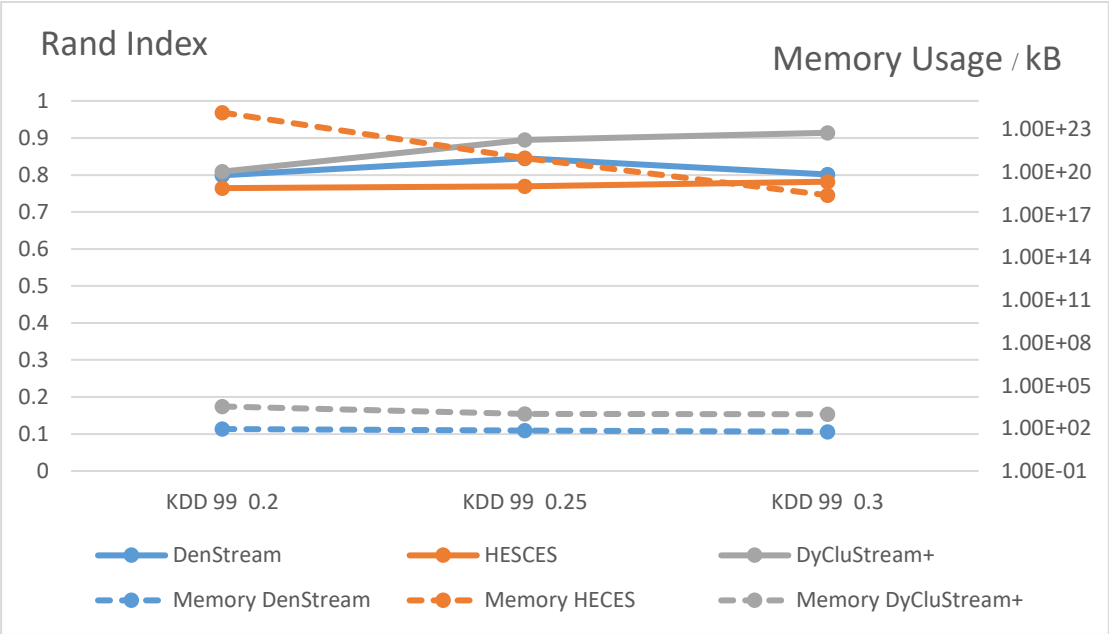
ภาพที่ 4-24 แผนภูมิกราฟเปรียบเทียบผลกระทบระหว่างหน่วยความจำที่ใช้งานกับความแม่นยำในแต่ละขั้นตอนวิธีการของชุดข้อมูล Statlog Shutter

สำหรับชุดข้อมูล Statlog (Shuttle) แสดงให้เห็นว่า ขั้นตอนวิธีการ DyCluStream+ และ DenStream มีผลกระทบระหว่างหน่วยความจำที่ใช้งานและความแม่นยำที่ค่อนข้างน้อย ส่วนขั้นตอนวิธีการ HECES นั้น เมื่อหน่วยความจำที่ใช้ลดลง ส่งผลให้ความแม่นยำลดลงตามไปด้วย



ภาพที่ 4-25 แผนภูมิกราฟเปรียบเทียบผลกระทบระหว่างหน่วยความจำที่ใช้งานกับความแม่นยำในแต่ละขั้นตอนวิธีการของชุดข้อมูล Letter Recognition

สำหรับชุดข้อมูล Letter Recognition แสดงให้เห็นว่า ขั้นตอนวิธีการ HECES มีผลลัพธ์ความแม่นยำลดลงอย่างมากเมื่อหน่วยความจำที่ใช้ลดลง ส่วน DenStream และ DyCluStream+ มีผลกระทบระหว่างหน่วยความจำที่ใช้งานและความแม่นยำค่อนข้างน้อย โดย DyCluStream+ มีการใช้หน่วยความจำที่มากกว่า DenStream แต่ก็ส่งผลให้มีความแม่นยำมากขึ้น



ภาพที่ 4-26 แผนภูมิกราฟเปรียบเทียบผลกระทบระหว่างหน่วยความจำที่ใช้งานกับความแม่นยำในแต่ละขั้นตอนวิธีการของชุดข้อมูล KDD Cup 1999

สำหรับชุดข้อมูล KDD Cup 1999 แสดงให้เห็นว่า ขั้นตอนวิธีการ HECES ใช้หน่วยความจำที่มากที่สุด แต่ยังคงให้ผลลัพธ์ที่น้อยกว่า DenStream และ DyCluStream ส่วน DyCluStream ใช้หน่วยความจำที่มากกว่า DenStream เล็กน้อย แต่ก็ให้ค่าความแม่นยำที่มากขึ้น

- **การวัดประสิทธิภาพเชิงเวลา**

สำหรับการวัดประสิทธิภาพของวิธีการที่นำเสนอ The Dynamic Clustering Algorithm for Streaming Data Using One-Pass-Throw-Away Datum and Hyper-ellipsoidal Micro-Clustering (DyCluStream) และ The Dynamic Clustering Algorithm for Streaming Data Using One-Pass-Throw-Away Datum and Hyper-ellipsoidal Micro-Clustering Plus (DyCluStream+) เปรียบเทียบกับวิธีการ Density-based clustering over an evolving data stream with noise (DenStream) และ Hyper-ellipsoidal clustering technique for evolving data stream (HECES) ในเชิงเวลา สามารถอธิบายได้ดังนี้

วิธีการ DenStream ต้องการข้อมูลนำเข้าจำนวนหนึ่งเพื่อใช้ในการสร้างกลุ่มข้อมูลขนาดเล็กในตอนเริ่มต้น ด้วยวิธีการ DBSCAN ดังภาพที่ 4-24 ซึ่งมีการทำซ้ำในบรรทัดที่ 2 และ 11 จึงมีความซับซ้อนเชิงเวลาคือ $O(initPoint^2)$ โดยที่ $initPoint$ คือ จำนวนข้อมูลที่ใช้ในการสร้างกลุ่มข้อมูลขนาดเล็กเริ่มต้น (เท่ากับ stream speed) เมื่อได้กลุ่มข้อมูลขนาดเล็กแล้วจึงทำในส่วนออนไลน์คือ การจัดกลุ่มข้อมูลที่เข้ามาให้อยู่ในกลุ่มข้อมูลขนาดเล็ก จึงมีความซับซ้อนเชิงเวลาคือ $O(mn)$ โดย m คือ จำนวนกลุ่มข้อมูลขนาดเล็กที่เกิดจากการตรวจสอบกลุ่มข้อมูลในภาพที่ 4-28 บรรทัดที่ 2 และ 6 และ n คือ จำนวนข้อมูลนำเข้า เมื่อมีความต้องการกลุ่มข้อมูลสำหรับไปใช้งานจริง จะเชื่อมต่อกลุ่มข้อมูลขนาดเล็กในส่วนออฟไลน์ ด้วยวิธีแบบ DBSCAN อีกครั้ง โดยเปลี่ยนจากการเชื่อมต่อข้อมูลเป็นการเชื่อมต่อกลุ่มข้อมูล ดังภาพที่ 4-27 ซึ่งมีการทำซ้ำในบรรทัดที่ 2 และ 11 ดังนั้นจึงมีความซับซ้อนเชิงเวลาคือ $O(m^2)$

Algorithm DBSCAN ($initPoint, r_\varepsilon, n_\varepsilon, \beta$)

ClusId = 0
for each unvisited point x_i in $initPoint$
    mark x_i as visited
    sphere_points = regionQuery(x_i, r)
    if sizeof(sphere_points) $\geq n$
        ClusId = ClusId +1
        expandCluster(x_i, sphere_points, ClusId, r_ε , n_ε)
    end
end

Algorithm expandCluster(x_i, sphere_points, ClusId, r_ε, n_ε):

add $x_i$to cluster ClusId
for each point x_i' in sphere_points
    if x_i' is not visited
        mark x_i' as visited
        sphere_points' = regionQuery(x_i', r_ε)
        if sizeof(sphere_points') $\geq n_\varepsilon$
            sphere_points = sphere_points joined with sphere_points'
            If x_i' is not yet member of any cluster
                add x_i' to cluster ClusId
            end
        end
    end
end

Algorithm regionQuery(x_i, r_ε):

return all points within the n-dimensional sphere centered at x_i with radius r_ε (including P)

ภาพที่ 4-27 ขั้นตอนวิธีการ DBSCAN

Algorithm DenStreamOnline ($x_i, r_\varepsilon, n_\varepsilon, \beta$)

1: Get the arriving data point x_i
2: Try to merge x_i into its nearest p-micro-cluster cp
3: **If** ro (the new radius of cp) $\leq r_\varepsilon$ then
4: Merge x_i into cp;
5: **else**
6: Try to merge *p* into its nearest o-micro-cluster CO;
7: **If** ro (the new radius of co) $\leq r_\varepsilon$ then
8: Merge x_i into co;
9: **If** *w* (the new weight of co) > $\beta \times n_\varepsilon$ then
10: Remove co from outlier-buffer and create a new p-micro-cluster by co;
11: **end**
12: **else**
13: Create a new o-micro-cluster by x_i and insert it into the outlier-buffer;
14: **end**
15: **end**

ภาพที่ 4-28 ขั้นตอนวิธีการ DenStream ส่วนออนไลน์

วิธีการ HECES จะเริ่มจากส่วนออนไลน์คือการจัดกลุ่มข้อมูลที่เข้ามาให้อยู่ในรูปแบบ grid-cells ดังนั้นจึงมีความซับซ้อนเชิงเวลาคือ $O(n)$ โดย n คือ จำนวนข้อมูลนำเข้า ดังภาพที่ 4-26 บรรทัดที่ 1 ถึง 6 หลังจากนั้น เมื่อมีความต้องการกลุ่มข้อมูลสำหรับไปใช้งานจริงจะเริ่มกระบวนการส่วนออฟไลน์ โดยทำการลบ grid-cells ที่มีนัยสำคัญต่ำออกจากหน่วยความจำ ดังภาพที่ 4-26 บรรทัดที่ 8 จึงมีความซับซ้อนเชิงเวลาคือ $O(m)$ แล้วจึงแทน grid-cells ด้วยกลุ่มข้อมูลขนาดเล็กรูปทรง hyper-ellipsoid และทำการผสานกลุ่มข้อมูลรูปทรง hyper-ellipsoid ที่อยู่ใกล้กันแล้วแทนที่ด้วยกลุ่มข้อมูลใหม่โดยใช้ค่าเฉลี่ยเป็นจุดศูนย์กลาง ดังภาพที่ 4-29 ซึ่งมีการทำซ้ำในบรรทัดที่ 13 ถึง 16 จึงมีความซับซ้อนเชิงเวลาเป็น $O(m^2)$ หลังจากนั้น จึงลบกลุ่มข้อมูลรูปทรง ellipsoid ที่มีความซ้ำซ้อนกัน ดังภาพที่ 4-29 ซึ่งมีการทำซ้ำในบรรทัดที่ 17 ถึง 21 จึงมีความซับซ้อนเชิงเวลาเป็น $O(m^2)$

Algorithm HECESOnline(Amc)
1: Initialize $gridList(G)$ with an empty set
2: **while** (Stream S has more instances) **do**
3: Read an incoming d dimensional point $x_l = (x_{l,1}, x_{l,2}, \dots, x_{l,d})$ in the stream S
4: Assigning the object x_l to the corresponding grid-Cell g_i
5: Calculate the statistical summary of g_i
6: Update G by adding g_i with the updated statistics
7: **end**
Algorithm HECESOffline(Amc)
8: Remove the empty grid-Cells and grid-Cells having cardinality less than threshold $\pi = (\mu_G - \sigma_G)$
9: Update the $gridList\ G$
\\ μ_G, σ_G are the mean and standard deviation of data in all $g_i \epsilon G$, respectively.
10: Calculate the covariance of data in g_i
11: **Call $CovEstt$ (covariance matrix in g_i)**
12: Replace cells with ellipsoids (initial clusters) by computing the Mahalanobis distance covering 95% of data in the cells
13: Calculate the Mahalanobis distance M_{ab} between hyper-ellipsoids e_a and e_b
14: **if** $M_{ab} < M_{threshold}$ **then**
15: Merge the Ellipsoids
16: **end**
17: **for all** elliptical clusters e_l, e_m **do**
18: **if** $|e_l - e_m| < \frac{\sqrt{d}}{2} \times w$ **then**
19: Remove the ellipsoid having smaller number of data points among e_l or e_m
20: **end**
21: **end**
22: **end**
25: $t_c \leftarrow t_c + 1$
26: **end**

ภาพที่ 4-29 ขั้นตอนวิธีการ HECES

Algorithm Online Phase($x_i, r_\varepsilon, n_\varepsilon$)

1: Get the arriving data point x_i
2: **If** $Amc \neq \emptyset$ **then**
3: Find the winning Actual micro-cluster Amc_w using (3.25)
4: **If** x_i is satisfies the conditions in eq. (3.26) with respect Amc_w **then**
5: Update the parameters of Amc_w from eq. (3.27), (3.28), (3.29)
6: **return**
7: **end**
8: **end**
9: **If** $Smc \neq \emptyset$ **then**
10: Find the winning Small micro-cluster Smc_w using eq. (3.14)
11: **If** $mahaDist(x_i, Smc_w) \leq r_\varepsilon$ then
12: Update the parameters of Smc_w using eq. (3.15), (3.16), (3.17)
13: **If** $n_w > n_\varepsilon$ then
14: Create a new $Amc_{|Amc|+1}$ by Smc_w using eq. (3.18), (3.19), (3.20)
15: **return**
16: **end**
17: **return**
18: **end**
19: **end**
20: Create new $Smc_{|Smc|+1}$ by the arriving data point x_i using eq. (3.20), (3.21), (3.22)
21: **return**

ภาพที่ 4-30 ขั้นตอนวิธีการ DyCluStream ส่วนออนไลน์

วิธีการ DyCluStream จะเริ่มจากส่วนออนไลน์คือการจัดกลุ่มข้อมูลที่เข้ามาให้อยู่ในกลุ่มข้อมูลขนาดเล็ก ดังภาพที่ 4-30 บรรทัดที่ 3 และ 10 จึงมีความซับซ้อนเชิงเวลาคือ $O(mn)$ โดย m คือ จำนวนกลุ่มข้อมูลขนาดเล็ก และ n คือ จำนวนข้อมูลนำเข้า เมื่อมีความต้องการกลุ่มข้อมูลสำหรับไปใช้งานจริง จะเชื่อมต่อกลุ่มข้อมูลขนาดเล็กในส่วนออฟไลน์ ดังภาพที่ 4-31 ซึ่งมีการทำซ้ำในบรรทัดที่ 3 และ 8 ดังนั้นจึงมีความซับซ้อนเชิงเวลาคือ $O(m^2)$

ภาพที่ 4-31 ขั้นตอนวิธีการ DyCluStream ส่วนออฟไลน์

Algorithm ClusteringAmc(Amc)

1: Label Amc_i in Amc as non-label
2: Initial $ClusId = 0$
3: **for** each non-labelled Amc_w $with\ most\ dense$ from eq.(3.32) ∈ Amc do
4: Incremental ClusId by 1
5: Label Amc_w as $ClusId$
6: ExpandClusAmc(ClusId, Amc_w, Amc)
7: **end**

Algorithm ExpandClusAmc($ClusId, Amc_w, Amc$)

8: **for** each non-labelled $Amc_w \in Amc$ **do**
9: **if** ($min\delta(Amc_w, Amc_i)$ from eq. (3.34) ≤ ∅)) **then**
10: Label Amc_i as $ClusId$
11: ExpandClusAmc($ClusId, Amc_i$)
12: **else if** ($densityRatio(Amc_w, Amc_i)$ from eq. (3.35) ≤ β)
13: Label Amc_i as $ClusId$
14: ExpandClusAmc($ClusId, Amc_i$)
15: **end**
16: **end**

อย่างไรก็ตาม วิธีการ DenStream, HECES และ DyCluStream จำเป็นจะต้องเป็นต้อง

ดำเนินการจัดกลุ่มกับตัวแทนกลุ่มข้อมูลขนาดเล็กใหม่ทั้งหมดทุกครั้งที่มีความต้องการกลุ่มข้อมูลสำหรับไปใช้งานจริง ดังนั้นจึงมีความซับซ้อนเชิงเวลาคือ $O(m^2)$

สำหรับวิธีการ DyCluStream+ ซึ่งมีความซับซ้อนเชิงเวลาในขั้นตอนออนไลน์ เช่นเดียวกับ DyCluStream คือ จะเริ่มจากส่วนออนไลน์คือการจัดกลุ่มข้อมูลที่เข้ามาให้อยู่ในกลุ่มข้อมูลขนาดเล็ก ดังภาพที่ 4-32 บรรทัดที่ 3 และ 10 จึงมีความซับซ้อนเชิงเวลาคือ $O(mn)$ โดย m คือ จำนวนกลุ่มข้อมูลขนาดเล็ก และ n คือ จำนวนข้อมูลนำเข้า เมื่อมีความต้องการกลุ่มข้อมูลสำหรับไปใช้งานจริงจะเชื่อมต่อกลุ่มข้อมูลขนาดเล็กในส่วนออฟไลน์ ดังภาพที่ 4-33 ซึ่งมีการทำซ้ำในบรรทัดที่ 5, 10 เมื่อมีการทำจัดกลุ่มสำหรับนำไปใช้งานในครั้งแรก ดังนั้นจึงมีความซับซ้อนเชิงเวลาคือ $O(m^2)$ และเมื่อมีวามต้องการจัดกลุ่มสำหรับนำไปใช้งานในครั้งต่อไป จะมีการเชื่อมต่อกลุ่มข้อมูลขนาดเล็กในภาพที่ 4-33 บรรทัดที่ 2, 5 และ 11 โดยบรรทัดที่ 2 และ 5 คือการเชื่อมต่อกลุ่มข้อมูลที่ได้รับการปรับปรุงในส่วนออนไลน์เท่านั้น จึงทำให้มีความซับซ้อนเชิงเวลาคือ $O(mu)$ โดย u คือกลุ่มของข้อมูลขนาดเล็กที่ถูกปรับปรุง

Algorithm Online Phase($x_i, r_\varepsilon, n_\varepsilon$)

1: Get the arriving data point x_i
2: If $Amc \neq \emptyset$ **then**
3: Find the winning Actual micro-cluster Amc_w using (3.25)
4: If x_i is satisfies the conditions in eq. (3.37) with respect Amc_w **then**
5: Update the parameters of Amc_w from eq. (3.27), (3.28), (3.29)
6: **return**
7: end
8: end
9: If $Smc \neq \emptyset$ **then**
10: Find the winning Small micro-cluster Smc_w using eq. (3.14)
11: **If** $mahaDist(x_i, Smc_w) \leq r_\varepsilon$ then
12: Update the parameters of Smc_w using eq. (3.15), (3.16), (3.17)
13: **If** $n_w > n_\varepsilon$ then
14: Create a new $Amc_{|Amc|+1}$ by Smc_w using eq. (3.18), (3.19), (3.20)
15: **return**
16: **end**
17: **return**
18: **end**
19: end
20: Create new $Smc_{|Smc|+1}$ by the arriving data point x_i using eq. (3.20), (3.21), (3.22)
21: return

ภาพที่ 4-32 ขั้นตอนวิธีการ DyCluStream+ ส่วนออนไลน์

Algorithm ClusteringAmcbyIncremental(Amc)

1: Label all new Amc as non-labelled

2: for each updated Amc_w $with\ most\ dense$ from eq.(3.32) ∈ Amc **do**

3: ExpandClusAmcbyIncremental($ClusId$ of Amc_w , Amc_w)

4: end

5: for each new non-labelled Amc_w $with\ most\ dense$ from eq.(3.32) ∈ Amc **do**

6: Incremental $ClusId$ by 1

7: Label Amc_w as $ClusId$

8: ExpandClusAmcbyIncremental($ClusId$, Amc_w)

9: end

Algorithm ExpandClusAmcbyIncremental($ClusId$, Amc_w)

10: for each $Amc_i \in Amc$ **do**

11: **if** ($B(Amc_w, Amc_i) \leq 1$) from eq. (3.39) And $\cos(Amc_w, Amc_i) \geq \theta$) from eq. (3.40) **then**

12: merge Amc_w, Amc_i as Amc_w

13: **else if**(($B(Amc_w, Amc_i) \leq \beta$) from eq. (3.39) or ($\delta(Amc_w, Amc_i$) from eq. (3.44) $\leq \emptyset$))

14: **if** ($non\_labelled\ Amc_i$) **then**

15: Label Amc_i as $ClusId$

16: ExpandClusAmc(ClusId, Amc_i)

17: **else if** ($labelled\ Amc_i$) **then**

18: Label All Amc with same class of Amc_i as $ClusId$

19: end

20: **end**

21: end

ภาพที่ 4-33 ขั้นตอนวิธีการ DyCluStream+ ส่วนออฟไลน์

การเปรียบเทียบประสิทธิภาพเชิงเวลาของ DenStream, HECES, DyCluStream และ DyCluStream+ สามารถแสดงได้ดังตารางที่ 4-23

ตารางที่ 4-24 การเปรียบเทียบประสิทธิภาพเชิงเวลาของ DenStream, HECES, DyCluStream และDyCluStream+

Algorithm	Initial	Online Phase	Offline Phase			Incremental Offline Phase
			Delete empty grid-cell	Offline Phase	Remove the redundant cluster	
DenStream	$O(initPoint^2)$	$O(mn)$	-	$O(m^2)$	-	$O(m^2)$
HECES	-	$O(n)$	$O(m)$	$O(m^2)$	$O(m^2)$	$O(m) + O(m^2) + O(m^2)$
DyCluStream	-	$O(mn)$	-	$O(m^2)$	-	$O(m^2)$
DyCluStream+	-	$O(mn)$	-	$O(m^2)$	-	$O(mu)$

บทที่ 5
สรุปผลการดำเนินงานและข้อเสนอแนะ

สำหรับงานวิจัยนี้ ผู้จัดทำงานวิจัยจะกล่าวถึงการสรุปและวิจารณ์ผลการดำเนินงาน ได้แก่ สรุปผลการทดลอง และการวิจารณ์เกี่ยวกับข้อดี ข้อบกพร่อง และข้อเสนอแนะของงานวิจัย

5.1 สรุปผลการดำเนินงาน

งานวิจัยนี้ได้นำเสนอขั้นตอนวิธีการจัดกลุ่มข้อมูล **The Dynamic Clustering Algorithm for Streaming Data Using One-Pass-Throw-Away Datum and Hyper-ellipsoidal Micro-Clustering** หรือเรียกว่า **DyCluStream** ซึ่งเป็นขั้นตอนวิธีการจัดกลุ่มสำหรับข้อมูลสตรีมมิ่งที่ไม่สามารถทราบลักษณะการกระจายตัวและจำนวนกลุ่มได้ล่วงหน้า และมีโครงสร้างสำหรับเก็บข้อสรุปแต่ละกลุ่มเพื่อเป็นตัวแทนของกลุ่มข้อมูล โดยใช้เพียงข้อมูลที่เข้ามาใหม่ ซึ่งแบ่งออกเป็นสองขั้นตอน ดังนี้

ขั้นตอนออนไลน์ ข้อมูลที่เข้ามาจะถูกเก็บให้อยู่ในรูปแบบฟังก์ชันทางคณิตศาสตร์และสถิติ ซึ่งอยู่ในรูปแบบกลุ่มของข้อมูลขนาดเล็กที่มีรูปร่างเป็น spherical ซึ่งประกอบไปด้วย ค่าเฉลี่ย จำนวนสมาชิกที่อยู่ในกลุ่มข้อมูล และ Covariance matrix โดยเมื่อมีความหนาแน่นในระดับที่ผู้ใช้กำหนด ก็จะเปลี่ยนเป็นกลุ่มข้อมูลขนาดเล็กรูปทรง dynamic hyper-ellipsoidal ซึ่งเข้ากับรูปร่างของกลุ่มข้อมูลดีกว่ารูปทรงแบบ spherical

ขั้นตอนออฟไลน์ กลุ่มข้อมูลสำหรับนำไปใช้งานจะเกิดจากการเชื่อมต่อกันไปเรื่อย ๆ ของกลุ่มข้อมูลขนาดเล็ก โดยเริ่มจากกลุ่มข้อมูลขนาดเล็กที่มีความหนาแน่นมากที่สุดเป็นตัวแผ่ขยาย

และวิธีการ **The Dynamic Clustering Algorithm for Streaming Data Using One-Pass-Throw-Away Datum and Hyper-ellipsoidal Micro-Clustering Plus** หรือเรียกว่า **DyCluStream+** ที่ปรับปรุงขั้นตอนวิธีการจัดกลุ่มจากวิธีการ **DyCluStream** ให้มีประสิทธิภาพมากยิ่งขึ้น และสามารถนำไปใช้กับข้อมูลจริงได้ โดยวิธีการ **DyCluStream+** มีการปรับปรุงหลัก ๆ ดังนี้

ขั้นตอนออนไลน์ มีการปรับเปลี่ยนมาตรวัดระยะทางส่วนออนไลน์ให้มีความแม่นยำมากขึ้น

ขั้นตอนออฟไลน์ มีการปรับเปลี่ยนเงื่อนไขในการเชื่อมต่อกันของกลุ่มข้อมูลขนาดเล็ก และในระหว่างการเชื่อมต่อ ถ้ากลุ่มข้อมูลขนาดเล็กสามารถรวมกลุ่มกันแล้วแทนด้วยกลุ่มข้อมูลใหม่ได้ก็จะถูกรวมกลุ่ม และขั้นตอนในส่วนออฟไลน์นี้สามารถดำเนินการต่อจากรอบก่อนหน้า โดยพิจารณาถึงกลุ่มข้อมูลขนาดเล็กที่ได้รับผลกระทบจากขั้นตอนในส่วนออนไลน์ แล้วดำเนินการแผ่กระจายเฉพาะตัวแทนกลุ่มข้อมูลขนาดเล็กที่ได้รับผลกระทบเท่านั้น

การประยุกต์ใช้กับข้อมูลจริง เพิ่มการประมาณค่า Covariance Matrix ที่มีปัญหาในเรื่อง Not Positive Definite Matrix ที่เกิดขึ้นกับข้อมูลจริง ให้สามารถใช้ขั้นตอนวิธีการนี้ได้อย่างมีประสิทธิภาพ

จากผลการทดลองการกับข้อมูลสังเคราะห์ โดยวัดประสิทธิภาพความถูกต้องด้วยตัววัดประสิทธิภาพ Purity เพื่อวัดความเป็นหนึ่งเดียวกันของข้อมูลที่อยู่ในกลุ่ม ตัววัดประสิทธิภาพ Rand Index เพื่อวัดความสามารถในการจัดกลุ่มข้อมูล 2 ตัวใด ๆ ที่อยู่ในกลุ่มเดียวกันได้อย่างถูกต้อง และ Jaccard Index ที่วัดความถูกต้องในการจัดกลุ่มข้อมูลให้อยู่ในกลุ่มเดียวกัน แสดงให้เห็นว่า **DyCluStream** และ **DyCluStream+** ให้ผลลัพธ์ในการจัดกลุ่มที่ดีกว่า **DenStream** และ **HECES** เนื่องจากรูปร่างของตัวแทนกลุ่มข้อมูลที่สามารถแทนลักษณะการกระจายตัวของข้อมูลได้ดีกว่า และมาตรวัดระยะทางที่ใช้มีความเหมาะสมกับลักษณะของตัวแทนกลุ่มข้อมูล และ **DyCluStream+** ซึ่งถูกพัฒนาต่อจาก **DyCluStream** สามารถให้ผลลัพธ์ที่ดีกว่า **DyCluStream** และใช้เวลาในการจัดกลุ่มได้รวดเร็วกว่าเนื่องจากมีความสามารถจัดกลุ่มต่อจากรอบการจัดกลุ่มในรอบก่อนหน้า และมีความสามารถในการลดจำนวนกลุ่มของข้อมูลขนาดเล็กลงด้วยวิธีการรวมกลุ่มข้อมูลขนาดเล็กที่คุณลักษณะใกล้เคียงกันเข้าด้วยกันแล้วแทนที่ด้วยกลุ่มข้อมูลขนาดเล็กใหม่ที่เกิดจากคุณลักษณะของกลุ่มข้อมูลขนาดเล็กที่รวมกัน

5.2 วิจารณ์ผลการดำเนินงาน

5.2.1 ข้อดีของงานวิจัย

- ขั้นตอนวิธีการ **DyCluStream+** สามารถจัดกลุ่มข้อมูลได้โดยที่ไม่จำเป็นต้องรู้จำนวนกลุ่มล่วงหน้า และสามารถจัดกลุ่มข้อมูลแบบสตรีมมิ่งได้

- ขั้นตอนวิธีการ **DyCluStream+** สามารถจัดกลุ่มได้อย่างรวดเร็วและถูกต้องเมื่อเปรียบเทียบกับขั้นตอนวิธีการที่นำมาเปรียบเทียบ เนื่องจากสามารถจัดกลุ่มตัวแทนกลุ่มของข้อมูลที่มีผลกระทบเท่านั้น

- ขั้นตอนวิธีการ **DyCluStream+** สามารถลดจำนวนตัวแทนกลุ่มของข้อมูลลงเพื่อประหยัดหน่วยความจำ โดยยังให้ประสิทธิภาพที่ดีกว่าหรือเท่ากันกับก่อนการลดจำนวนตัวแทนกลุ่ม

5.2.2 ข้อจำกัดของงานวิจัย

- เนื่องจากขั้นตอนวิธีการ **DyCluStream+** ถูกจัดอยู่ในขั้นตอนวิธีการแบบ Density-based จึงยังคงต้องกำหนดจำนวน parameter มากกว่าขั้นตอนวิธีการแบบอื่น

5.2.3 ข้อเสนอแนะของงานวิจัย

- ความสามารถในการจัดกลุ่มได้ต่อจากในครั้งก่อนหน้า หรือความสามารถในการจัดกลุ่มตัวแทนกลุ่มของข้อมูลที่มีผลกะทบ ควรพิจารณาเพิ่มเติมว่าเมื่อปรับปรุงกลุ่มตัวแทนกลุ่มของข้อมูลแล้ว กลุ่มข้อมูลจะมีลักษณะของการแบ่งออกเป็นสองกลุ่มหรือไม่

บรรณานุกรม

Ackermann, M.R., Märtens, M., Raupach, C., Swierkot, K., Lammersen, C., & Sohler, C. (2010). StreamKM++: A clustering algorithm for data streams. ACM Journal of Experimental Algorithmics, 17.

Aggarwal, C.C., Han, J., Wang, J., & Yu, P.S. (2003). A Framework for Clustering Evolving Data Streams. VLDB.

Amini, A., Teh, Y.W., & Saboohi, H. (2014). On Density-Based Data Streams Clustering Algorithms: A Survey. Journal of Computer Science and Technology, 29, 116-141.

Bhattacharyya, A. (1946). On a Measure of Divergence between Two Multinomial Populations. Sankhya: The Indian Journal of Statistics (1933-1960), 7(4), 401-406. Retrieved from http://www.jstor.org/stable/25047882

Cao, F., Ester, M., Qian, W., & Zhou, A. (2006). Density-Based Clustering over an Evolving Data Stream with Noise. SDM.

Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. KDD.

Guha, S., Meyerson, A., Mishra, N., Motwani, R., & O'Callaghan, L. (2003). Clustering Data Streams: Theory and Practice. IEEE Trans. Knowl. Data Eng., 15, 515-528.

Hahsler, M., & Bolaños, M. (2016). Clustering Data Streams Based on Shared Density between Micro-Clusters. IEEE Transactions on Knowledge and Data Engineering, 28, 1449-1461.

Han, J., Kamber, M. (2006). Data Mining: Concepts and Techniques. Morgan Kaufmann.

Higham, N.J. (2002). Computing the Nearest Correlation Matrix---A Problem from Finance.

Ho, Shen-Shyang & Wechsler, H. (2010). A Martingale Framework for Detecting Changes in Data Streams by Testing Exchangeability. IEEE transactions on pattern analysis and machine intelligence. 32. 2113-27. 10.1109/TPAMI.2010.48.

Jaiyen, S., Lursinsap, C., & Phimoltares, S. (2010). A Very Fast Neural Learning for Classification Using Only New Incoming Datum. IEEE Transactions on Neural Networks, 21, 381-392.

Lin, C., & Chen, M. (2005). Combining partitional and hierarchical algorithms for robust and efficient data clustering with cohesion self-merging. IEEE Transactions on Knowledge and Data Engineering, 17, 145-159.

Li, Y., Li, D., Wang, S. & Zhai, Y. (2014). Incremental Entropy-Based Clustering on Categorical Data Streams with Concept Drift. Knowledge-Based Systems. 59. 10.1016/j.knosys.2014.02.004.

Lughofer, E., & Mouchaweh, M.S. (2015). Autonomous data stream clustering implementing split-and-merge concepts - Towards a plug-and-play approach. Inf. Sci., 304, 54-79.

Rehman, M.Z., Li, T., Yang, Y., & Wang, H. (2014). Hyper-ellipsoidal clustering technique for evolving data stream. Knowl.-Based Syst., 70, 3-14.

Wattanakitrungroj, N., & Lursinsap, C. (2011). Memory-less unsupervised clustering for data streaming by versatile ellipsoidal function. CIKM.

Wattanakitrungroj, N., Maneeroj, S., & Lursinsap, C. (2017). Versatile Hyper-Elliptic Clustering Approach for Streaming Data Based on One-Pass-Thrown-Away Learning. Journal of Classification, 34, 108-147.

ภาคผนวก

ภาคผนวก ก

เอกสารเผยแพร่ผลงานวิจัย