การนำฟังก์ชันความหนาแน่นร่วมกับอัลกอริทึมการค้นหาคำตอบด้วยแรงโน้มถ่วงเพื่อแก้ปัญหาการจัดแบ่งกลุ่มข้อมูล

APPLYING THE DENSITY FUNCTION TO THE GRAVITATIONAL SEARCH ALGORITHM FOR CLUSTERING PROBLEMS

พงษ์เลิศ สังกะพัฐ

PONGLERT SANGKAPAS

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2554

KMITL-2011-IT-M-001-005

การนำฟังก์ชันความหนาแน่นร่วมกับอัลกอริทึมการค้นหาคำตอบด้วยแรงโน้มถ่วงเพื่อแก้ปัญหาการจัดแบ่งกลุ่มข้อมูล

APPLYING THE DENSITY FUNCTION TO THE GRAVITATIONAL SEARCH ALGORITHM FOR CLUSTERING PROBLEMS

พงศ์เลิศ สังกะเพศ

PONGLERT SANGKAPAS

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2554

KMITL-2011-IT-M-001-005

APPLYING THE DENSITY FUNCTION TO THE GRAVITATIONAL SEARCH ALGORITHM FOR CLUSTERING PROBLEMS

PONGLERT SANGKAPAS

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

FACULTY OF INFORMATION TECHNOLOGY

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2011

KMITL-2011-IT-M-001-005

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ใบรับรองวิทยานิพนธ์

หัวข้อวิทยานิพนธ์ การนำฟังก์ชันความหนาแน่นร่วมกับอัลกอริทึมการค้นหาคำตอบด้วย แรงโน้มถ่วงเพื่อแก้ปัญหาการจัดแบ่งกลุ่มข้อมูล

Applying the Density Function to The Gravitational Search Algorithm for Clustering Problems

นักศึกษา นายพงศ์เลิศ สังกะเพศ

รหัสประจำตัว 51066414

ปริญญา วิทยาศาสตรมหาบัณฑิต

สาขาวิชา เทคโนโลยีสารสนเทศ

อาจารย์ที่ปรึกษาวิทยานิพนธ์ รองศาสตราจารย์ ดร.อาริต ธรรมโน

คณะกรรมการสอบวิทยานิพนธ์	ลายมือชื่อ
รองศาสตราจารย์ ดร.วรพจน์ กรีสุระเดช	
รองศาสตราจารย์ ดร.กฤษณะ ไวยมัย	
รองศาสตราจารย์ ดร.อาริต ธรรมโน	
ผู้ช่วยศาสตราจารย์ ดร.พรฤดี เนติโสภากุล	
ดร.กิติ์สุชาต พสุภา	

วัน / เดือน/ ปี ที่สอบ วันพุธที่ 21 ธันวาคม 2554 เวลา 10.00 น.

สถานที่สอบ ณ ห้อง 335 (ชั้น 3) คณะเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศรับรองแล้ว

(รองศาสตราจารย์ ดร.จันทร์บูรณ์ สถิตวิริยวงศ์)

คณบดีคณะเทคโนโลยีสารสนเทศ

วันที่ 19 **เดือน** มกราคม **พ.ศ.** 2555

หัวข้อวิทยานิพนธ์	การนำฟังก์ชันความหนาแน่นร่วมกับอัลกอริทึมการค้นหาคำตอบด้วยแรงโน้มถ่วงเพื่อแก้ปัญหาการจัดแบ่งกลุ่มข้อมูล
นักศึกษา	นายพงศ์เลิศ สังกะเพศ
รหัสนักศึกษา	51066414
ปริญญา	วิทยาศาสตร์มหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
พ.ศ.	2554
อาจารย์ที่ปรึกษา	รศ.ดร.อาริต ธรรมโน

บทคัดย่อ

วิทยานิพนธ์ฉบับนี้ นำเสนออัลกอริทึมการจัดกลุ่มที่ใช้แนวคิดของอัลกอริทึมการค้นหาคำตอบด้วยแรงโน้มถ่วงร่วมกับฟังก์ชันหนาแน่นเพื่อทำให้เกิดการเคลื่อนที่ของจุดศูนย์กลางเคลื่อนที่ไปยังบริเวณที่มีข้อมูลหนาแน่นสูงและจำนวนกลุ่มจะแบ่งโดยอัตโนมัติ ในงานวิจัยนี้ได้มีการนำขั้นตอนต่างๆ มาใช้ร่วมกันอย่างเป็นระบบไม่ว่าจะ อัลกอริทึมการค้นหาคำตอบด้วยแรงโน้มถ่วงร่วมกับฟังก์ชันหนาแน่นใช้ในการปรับค่าจุดศูนย์กลางของคลัสเตอร์ การรวมจุดศูนย์กลางของคลัสเตอร์ที่ใกล้ชิดกัน และการลบคลัสเตอร์ โดยผลลัพธ์ที่ได้จากอัลกอริทึมคือจำนวนคลัสเตอร์และจุดศูนย์กลางของแต่ละคลัสเตอร์ ซึ่งจากการทดลองพบว่าอัลกอริทึมที่นำเสนอได้จำนวนคลัสเตอร์ที่เหมาะสมและมีประสิทธิภาพดีกว่าอัลกอริทึมการจัดกลุ่มข้อมูลแบบอื่นๆ ในบางปัญหา

Thesis	Applying the density function to the Gravitational Search Algorithm for clustering problems
Student	Mr.Ponglert Sangkapas
Student ID.	51066414
Degree	Master of Science
Program	Information Technology
Major	Information Science
Year	2011
Thesis Advisor	Assoc. Prof. Dr. Arit Thammano

ABSTRACT

This thesis proposes a clustering algorithm that applies the density function to the gravitational search algorithm. The proposed algorithm moves the cluster centers to the high density areas and, at the same time, automatically determines the numbers of clusters. In addition, the proposed algorithm also combines closely located clusters into a single cluster and deletes the unnecessary clusters. The result from this algorithm shows that its performance is better than other clustering algorithms on most of the problems.

กิตติกรรมประกาศ

วิทยานิพนธ์เล่มนี้สำเร็จ ได้ด้วยความกรุณาจากอาจารย์ที่ปรึกษา รศ.ดร.อาริต ธรรมโน ที่ให้ความช่วยเหลือ ให้คำชี้แนะช่วยแก้ปัญหาตลอดจนให้ความรู้และประสบการณ์ที่ดีแก่ข้าพเจ้า

ขอกราบพระคุณคณาจารย์คณะเทคโนโลยีสารสนเทศสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบังทุก ๆ ท่านที่ได้ประสิทธิ์ประสาทวิชาให้กับข้าพเจ้า

ขอขอบคุณเพื่อน ๆ ทั้งในคณะเทคโนโลยีสารสนเทศสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบังและเพื่อนสนิทมิตรสหายที่ผ่านเข้ามาในชีวิตทุกคนที่ให้คำแนะนำและคอยให้กำลังใจเสมอมา

สุดท้ายนี้ข้าพเจ้าขอกราบขอบพระคุณบิดามารดาและพี่ของข้าพเจ้าที่คอยให้กำลังใจ ให้คำปรึกษาและคอยสนับสนุนในทุก ๆ เรื่องที่ทำให้ข้าพเจ้าสามารถทำวิทยานิพนธ์เล่มนี้สำเร็จลุล่วงไปได้ด้วยดี

สำหรับคุณงามความดีอันใดที่เกิดจากวิทยานิพนธ์ฉบับนี้ ข้าพเจ้าขอมอบให้กับบิดามารดาซึ่งเป็นที่รักและเคารพยิ่ง ตลอดจนครูอาจารย์ที่เคารพทุกท่านที่ได้ประสิทธิ์ประสาทวิชาความรู้และถ่ายทอดประสบการณ์ที่ดีให้แก่ข้าพเจ้า

พงศ์เลิศ สังกะเพศ

สารบัญ

หน้า

สารบัญ(ต่อ)

หน้า

สารบัญ(ต่อ)

หน้า

สารบัญ(ต่อ)

หน้า

สารบัญตาราง

สารบัญตาราง(ต่อ)

สารบัญตาราง(ต่อ)

สารบัญตาราง(ต่อ)

สารบัญรูป

สารบัญรูป(ต่อ)

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การจัดแบ่งกลุ่มข้อมูล (Clustering) เป็นขั้นตอนในการค้นหาเซตของโมเดลที่ใช้อธิบายหรือจำแนกความแตกต่างของข้อมูล โดยที่เราไม่รู้ว่าข้อมูลที่พิจารณาอยู่นั้นเป็นประเภทไหน ตั้งแต่อดีตจนถึงปัจจุบันมีผู้คิดค้นอัลกอริทึม (Algorithm) ที่ใช้ในการจัดแบ่งกลุ่มข้อมูลมากมายหลายอัลกอริทึม ซึ่งอัลกอริทึมที่ค่อนข้างได้รับความนิยมคือ อัลกอริทึมเคมีน (K-means Algorithm) เป็นอัลกอริทึมที่หาค่าเฉลี่ยของข้อมูลใช้ในการจัดกลุ่ม อัลกอริทึมฟัชชี่ซีมีน (C-means Algorithm) ใช้แนวคิดของฟัชชี่ (Fuzzy) ในการจัดแบ่งกลุ่มข้อมูล วิทยานิพนธ์ฉบับนี้พยายามสร้างอัลกอริทึมที่มีการทำงานในลักษณะลอกเลียนปรากฏการทางฟิสิกส์ด้วยกฎแรงแนวโน้มถ่วงและกฎการเคลื่อนที่ของนิวตัน โดยแนวคิดการนำกฎแรงแนวโน้มถ่วงและกฎการเคลื่อนที่ของนิวตันได้มีงานวิจัยเกี่ยวกับการค้นหาคำตอบที่เหมาะสมได้นำไปใช้อยู่แล้ว ซึ่งงานวิจัยนั้นมีชื่อว่า อัลกอริทึมการค้นหาคำตอบด้วยแรงโน้มถ่วง (The Gravitational Search Algorithm) ในวิทยานิพนธ์ฉบับนี้ก็ได้ทำการศึกษา และได้นำมาใช้ร่วมกับฟังก์ชันความหนาแน่น เพื่อทำให้เกิดแรงและการเคลื่อนที่เกิดจากแรงไปสู่บริเวณที่มีข้อมูลรวมกันอยู่หนาแน่น ซึ่งบริเวณที่มีข้อมูลอยู่รวมกันหนาแน่นมากจะมีแรงแนวโน้มถ่วงมากจะมีแรงดึงดูดข้อมูลที่อยู่บริเวณที่มีความหนาแน่นข้อมูลน้อยกว่าเข้าไปบริเวณที่มีข้อมูลอยู่รวมกันหนาแน่นจะถือว่าจุดนั้นเป็นตัวแทนกลุ่ม

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

1. เพื่อพัฒนาอัลกอริทึมที่ใช้ในการจัดแบ่งกลุ่มข้อมูลที่มีแนวคิดลอกเลียนแบบปรากฏการทางฟิสิกส์ โดยลอกเลียนกฎแรงโน้มถ่วงและกฎการเคลื่อนที่ของนิวตัน
2. อัลกอริทึมนี้ใช้ได้กับข้อมูลที่เป็นตัวเลขเท่านั้น
3. เพื่อพัฒนาอัลกอริทึมที่มีการสร้างจำนวนคลัสเตอร์ที่เหมาะสมกับข้อมูลแต่ละแบบ

1.3 ขอบเขตการวิจัย

1. เป็นการประมวลผลแบบ off-line

2. ในการวัดประสิทธิภาพจะใช้ตัววัดประสิทธิภาพดังนี้ ความบริสุทธิ์ (Purity), เอนโทรปี (Entropy), NMI (Normalized Mutual Information) และ F measure ในการจัดแบ่งกลุ่มข้อมูลเป็นตัววัดประสิทธิภาพของอัลกอริทึม

1.4 ขั้นตอนของการศึกษา

1. ศึกษาอัลกอริทึมพื้นฐานที่ใช้ในการจัดแบ่งกลุ่มข้อมูล (Clustering)
2. ศึกษาทฤษฎีที่เกี่ยวกับกฎแรงโน้มถ่วงและกฎการเคลื่อนที่ของนิวตัน
3. ศึกษาอัลกอริทึมการค้นหาคำตอบด้วยแรงโน้มถ่วง (The Gravitational Search Algorithm)
4. ทดลองนำแนวคิดกฎแรงโน้มถ่วงและกฎการเคลื่อนที่ของนิวตันของอัลกอริทึมการค้นหาคำตอบด้วยแรงโน้มถ่วงมาประยุกต์ใช้ร่วมกับการจัดกลุ่มข้อมูล
5. วิเคราะห์และสรุปผลเพื่อเลือกหาแนวทางที่เหมาะจะนำมาสร้างอัลกอริทึมใหม่
6. สร้างอัลกอริทึมที่ใช้ในการจัดแบ่งกลุ่มข้อมูลแบบใหม่
8. นำอัลกอริทึมใหม่ไปทดลองกับข้อมูลลักษณะต่างๆ
9. แก้ไขปรับปรุงโมเดลให้มีความเหมาะสมมากขึ้น
10. สรุปผลการทดลอง

1.5 โครงสร้างของวิทยานิพนธ์

วิทยานิพนธ์ฉบับนี้ได้แบ่งเนื้อออกเป็น 5 บทด้วยกันคือ

บทที่ 1 กล่าวถึงความเป็นมาของงานวิจัย ความมุ่งหมาย วัตถุประสงค์ ขอบเขตของการวิจัย และขั้นตอนการศึกษา

บทที่ 2 กล่าวถึงทฤษฎีพื้นฐานที่ใช้ในการวิจัย

บทที่ 3 กล่าวถึงทฤษฎีการนำฟังก์ชันความหนาแน่นร่วมกับอัลกอริทึมการค้นหาคำตอบด้วยแรงโน้มถ่วงเพื่อแก้ปัญหาการจัดแบ่งกลุ่มข้อมูล

บทที่ 4 กล่าวถึงการทดลองและผลการทดลอง

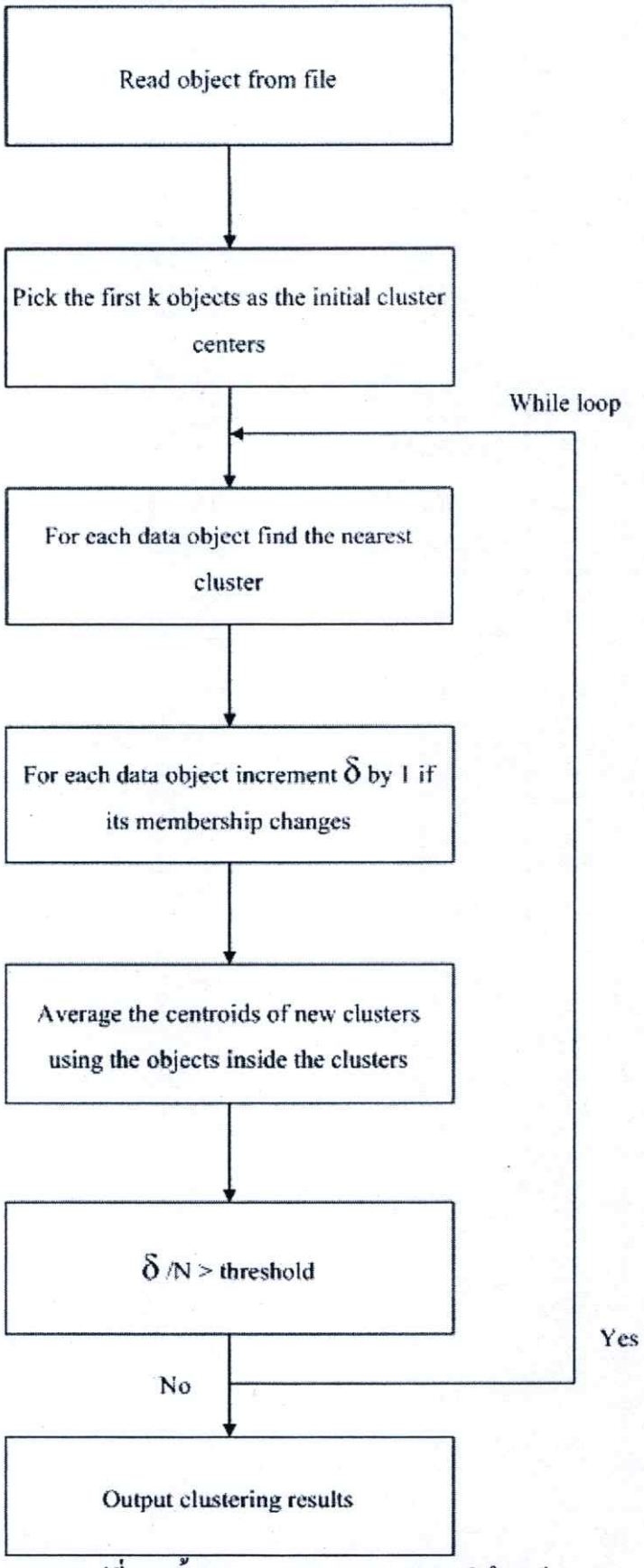
บทที่ 5 เป็นบทสรุปผลการวิจัยและข้อเสนอแนะ

บทที่ 2
ทฤษฎีพื้นฐานที่เกี่ยวข้อง

การจัดแบ่งกลุ่มข้อมูล (Clustering) เป็นขั้นตอนในการค้นหาเซตของโมเดลที่อธิบายหรือจำแนกความแตกต่างของข้อมูล โดยที่เราไม่รู้ว่าข้อมูลที่พิจารณาอยู่นั้นเป็นประเภทไหน [1] การจัดแบ่งกลุ่มข้อมูลเป็นการเรียนรู้แบบไม่มีเป้าหมาย (Unsupervised Learning) คือ การนำข้อมูลที่ไม่รู้เป้าหมายมาเรียนรู้ โดยดูจากลักษณะความเหมือนกันและความต่างกันของข้อมูลใช้ในการสร้างโมเดล การจัดแบ่งกลุ่มข้อมูลจะแบ่งขั้นตอนการทำงานออกเป็น 3 ขั้นตอนดังนี้ (1) Feature Selection/Extraction เป็นขั้นตอนการเลือกคุณสมบัติข้อมูลที่จะนำมาจัดแบ่งกลุ่มข้อมูล (2) Inter Pattern Similarity เป็นขั้นตอนการวัดรูปแบบความเหมือน (3) Grouping เป็นขั้นตอนจัดกลุ่มข้อมูลโดยดูจากความเหมือนของข้อมูลที่วัดได้จากขั้นตอนที่ (2) เป้าหมายของการจัดแบ่งกลุ่มคือต้องการกำหนดกลุ่มที่แท้จริงให้กับข้อมูลที่ไม่รู้เป้าหมาย แต่จะรู้ได้อย่างไรว่าจัดแบ่งกลุ่มดีหรือไม่ ซึ่งการทำการจัดแบ่งกลุ่ม (Clustering) มันจะดูจากลักษณะความเหมือนกันและความต่างกันของข้อมูลอย่างเดียว มันไม่มีเกณฑ์ที่จะบอกว่าการจัดแบ่งกลุ่มได้ดีที่สุด เกณฑ์ที่ว่าจะต้องให้ผู้ใช้เป็นจัดทำเองในแนวทางที่ทำให้เหมาะสมกับผลลัพธ์ที่ได้ ตัวอย่างเช่น ผู้ใช้สนใจในการค้นหาตัวแทนของกลุ่มที่มีข้อมูลเหมือนกัน (Data Reduction), สนใจในการค้นหา "ธรรมชาติของคลัสเตอร์" และอธิบายคุณสมบัติที่ไม่รู้ของข้อมูล (Data Type) , สนใจในการค้นหาการจัดแบ่งกลุ่มที่มีประโยชน์และเหมาะสม (Data Classes) หรือสนใจในการค้นหาข้อมูลที่ผิดปกติ (Outlier Detection)

2.1 อัลกอริทึมเคมีนคลัสเตอร์ริง (K-means Clustering Algorithm)

อัลกอริทึมเคมีนคลัสเตอร์ริง (K-means Clustering Algorithm) เป็นวิธีการแบ่งกลุ่มข้อมูลทั้งหมดออกตามจำนวนกลุ่มที่ต้องการ K กลุ่ม และค่า K ต้องน้อยกว่าจำนวนของข้อมูลทั้งหมด (N) ซึ่งจำนวนกลุ่มต้องเป็นเลขจำนวนเต็มบวก และการจัดกลุ่มวัตถุต้องอาศัยความเหมือนของข้อมูลโดยวัดจากระยะห่างที่น้อยที่สุดระหว่างข้อมูลกับจุดศูนย์กลางของกลุ่มทั้งหมด เพื่อจัดข้อมูลเข้าสู่กลุ่มต่างๆ ตามจำนวนกลุ่มที่ต้องการ นิยมใช้การวัดระยะแบบยูคลิด (Euclidean distance) และจุดศูนย์กลางเหล่านั้นเป็นตัวแทนของกลุ่มต่างๆ สามารถคำนวณได้จากค่าเฉลี่ยของข้อมูลที่อยู่ภายในกลุ่มเดียวกัน ซึ่งข้อมูลเหล่านั้นมีคุณลักษณะเหมือนกัน ส่วนวัตถุที่อยู่ต่างกลุ่มก็มีคุณลักษณะที่แตกต่างกัน และอัลกอริทึมเคมีนแบบลำดับ เมื่อใช้หนึ่งหน่วยประมวลผลต้องใช้เวลาสำหรับการประมวลผลเท่ากับ O(RKN) และมีขั้นตอนย่อยในการทำงานดังนี้ [2]

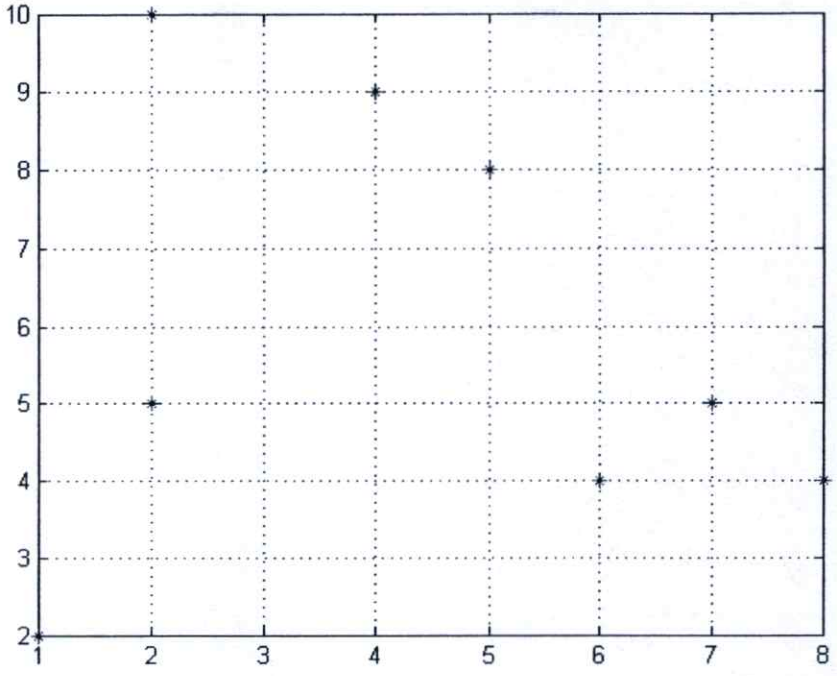


รูปที่ 2.1 ขั้นตอนการทำงานของอัลกอริทึมเคมีน

ตัวอย่าง ถ้ามีข้อมูลอยู่ 8 ตัวในตารางที่ 2.1 [3] ต้องการจัดการจัดแบ่งข้อมูล k = 3 กลุ่ม และใช้การวัดระยะแบบยูคลิดเดียน (Euclidean distance)

ตารางที่ 2.1 แสดงตัวอย่างจำนวนข้อมูลที่ต้องการจัดแบ่งกลุ่ม

ID	X	Y
A1	2	10
A2	2	5
A3	8	4
A4	5	8
A5	7	5
A6	6	4
A7	1	2
A8	4	9



รูปที่ 2.2 แสดงรูปข้อมูลที่ต้องการจัดแบ่งกลุ่มโดย * แทนข้อมูลแต่ละจุด

2.1.1 สุ่มค่าจุดศูนย์กลางเริ่มต้น จำนวน k ค่า เรียกว่า Cluster Centers (Centroid) สมมติ k =3 แสดงว่า c1, c2 และ c3 เป็นจุดศูนย์กลางเริ่มต้นที่เราสุ่มขึ้นมา c1(2, 10), c2(5, 8) และ c3(1, 2)

2.1.2 หาสมาชิกของจุดศูนย์กลาง ทำการหาค่าระยะห่างระหว่างข้อมูลกับจุดศูนย์กลาง หากข้อมูลไหนใกล้ค่าจุดศูนย์กลางตัวไหนมากที่สุดอยู่กลุ่มนั้น หาความห่างกันระหว่างข้อมูล 2 ข้อมูล คือ หาความห่างจากข้อมูล A = (x1, y1) และ centroid = (x2, y2) ตามสมการที่ 2.1

$$\text{distance}(a, b) = |x2 - x1| + |y2 - y1| \qquad (2.1)$$

distance(A1, c1) = |x2 – x1| + |y2 – y1| = |2 – 2| + |10 – 10| = 0 + 0 = 0

distance(A1, c2) = |x2 – x1| + |y2 – y1| = |5 – 2| + |8 – 10| = 3 + 2 = 5

distance(A1, c3) = |x2 – x1| + |y2 – y1| = |1 – 2| + |2 – 10| = 1 + 8 = 9

เมื่อหาค่าระยะห่างระหว่างข้อมูลกับจุดศูนย์กลาง กับทุกจุดข้อมูลแล้วจะได้ระยะห่างตามตารางที่ 2.2

ตารางที่ 2.2 แสดงตัวอย่างระยะห่างระหว่างข้อมูลกับจุดศูนย์กลาง

		c1(2, 10)	c2 (5, 8)	c3(1, 2)	
	Point	Dist Mean 1	Dist Mean 2	Dist Mean 3	Cluster
A1	(2, 10)	**0**	5	9	1
A2	(2, 5)	5	6	**4**	3
A3	(8, 4)	12	**7**	9	2
A4	(5, 8)	5	**0**	10	2
A5	(7, 5)	10	**5**	9	2
A6	(6, 4)	10	**5**	7	2
A7	(1, 2)	9	10	**0**	3
A8	(4, 9)	3	**2**	10	2

Cluster 1 มีสมาชิก A1 Cluster 2 มีสมาชิก A3, A4, A5, A6 และ A8 Cluster 3 มีสมาชิก A2 และ A7

2.1.3 ปรับตำแหน่งจุดศูนย์กลางใหม่ หาค่าเฉลี่ย (Mean) แต่ละกลุ่ม ให้เป็นค่าจุดกลาง (Centroid) ใหม่

- สำหรับ Cluster 1 มีจุดเดียวคือ A1(2, 10) แสดงว่า C1(2,10) ยังคงเดิม
- สำหรับ Cluster 2 มี 5 จุดอยู่กลุ่มเดียวกัน เพราะฉะนั้นหา C2 ใหม่ ((8+5+7+6+4)/5, (4+8+5+4+9)/5) = C2(6, 6)
- สำหรับ Cluster 3 มี 2 จุดอยู่กลุ่มเดียวกัน ((2+1)/2, (5+2)/2) = C3(1.5, 3.5)

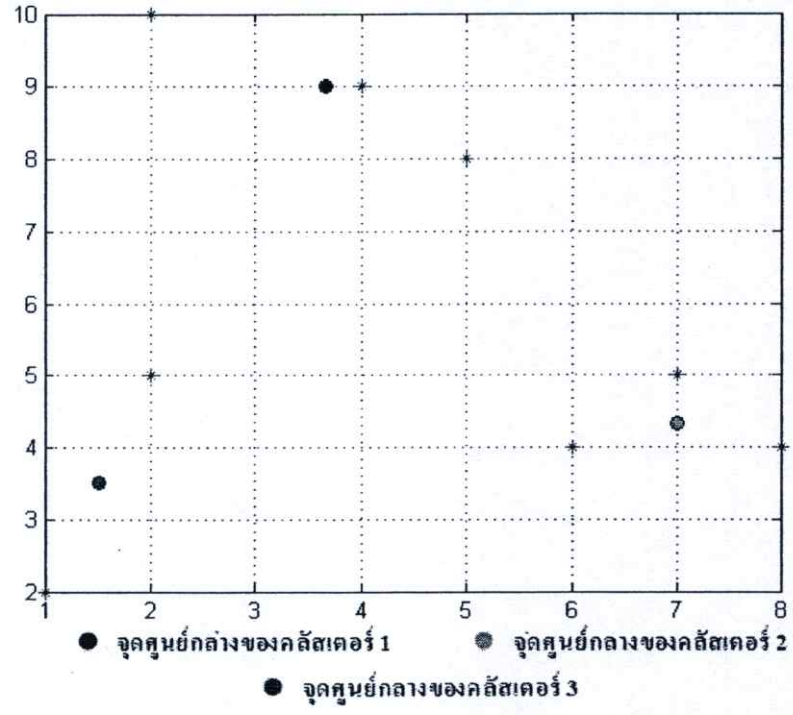
2.1.4 การวนซ้ำ กลับไปยังข้อ 2.1.2 จนกระทั่ง ค่าเฉลี่ยในแต่ละกลุ่มจะไม่เปลี่ยนแปลง

ตารางที่ 2.3 แสดงค่าจุดศูนย์กลางของคลัสเตอร์ที่ได้จากอัลกอริทึมเคมีนคลัสเตอร์ริง

คลัสเตอร์	ค่าจุดศูนย์กลาง	
	X	Y
1	3.6667	9

ตารางที่ 2.3 (ต่อ) แสดงค่าจุดศูนย์กลางของคลัสเตอร์ที่ได้จากอัลกอริทึมเคมีนคลัสเตอร์ริง

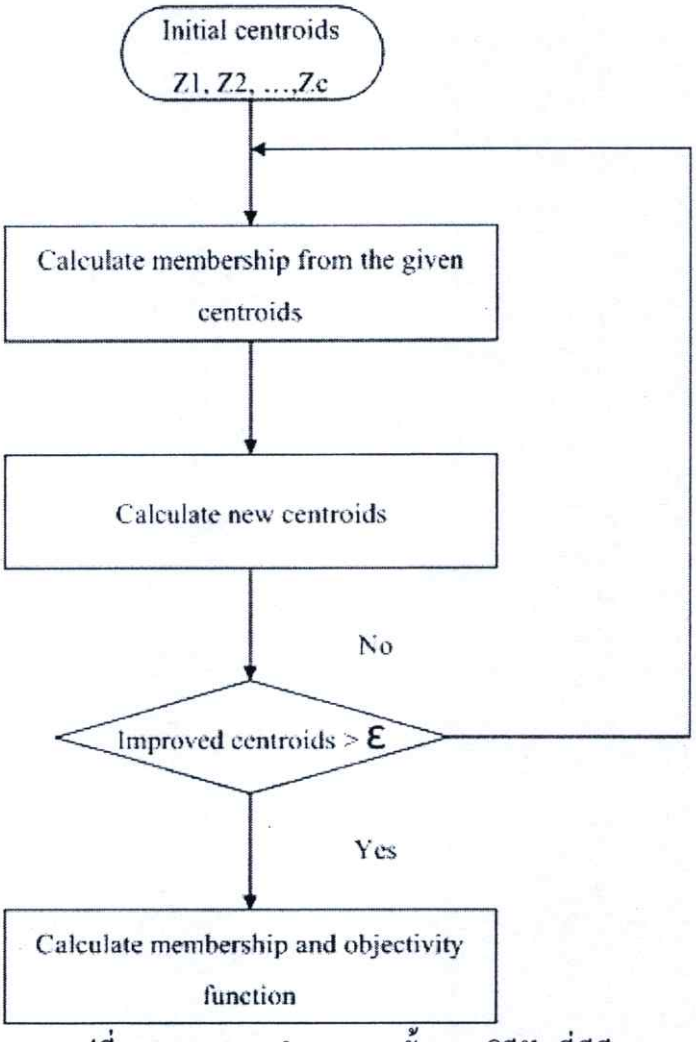
2	7	4.033
3	1.5	3.5



รูปที่ 2.3 แสดงจุดศูนย์กลางของคลัสเตอร์ 1, 2 และ 3 หลังการทำอัลกอริทึมเคมีนคลัสเตอร์ริง

2.2 อัลกอริทึมฟัชซี่ซีมีนคลัสเตอร์ริง (Fuzzy C-Means Clustering Algorithm)

การแบ่งกลุ่มลักษณะนี้ต่างจาก การแบ่งกลุ่มข้อมูลที่ข้อมูลสามารถเป็นสมาชิกได้เพียงกลุ่มใดกลุ่มหนึ่งเรียกการจัดกลุ่มนั้นว่า crisp clustering โดยการแบ่งกลุ่มแบบฟัชซี่จะเป็นการให้ค่าหรือโอกาสการเป็นสมาชิก (degree of membership) ของข้อมูลต่อกลุ่มข้อมูลต่างๆ การทำงานตามขั้นตอนการทำงานของฟัชซี่ซีมีน (Fuzzy C-Means) มีดังนี้ [4]



รูปที่ 2.4 แสดงการทำงานของขั้นตอนวิธีฟัซซี่ซีมีน

ตัวอย่าง จากข้อมูลอยู่ 8 ตัวในตารางที่ 2.1 ต้องการจัดการจัดแบ่งข้อมูล C = 3 กลุ่ม

2.2.1 สุ่มค่าจุดศูนย์กลางเริ่มต้น กำหนดกลุ่มข้อมูลที่ต้องการจัดกลุ่ม C = 3 แสดงว่า z1, z2 และ z3 เป็น จุดศูนย์กลางเริ่มต้นที่เราสุ่มขึ้นมา z1(2,8) , z2(3,5) และ z3(4,4) กำหนดค่าเพื่อเป็นเงื่อนไขในการให้ข้อมูลหยุดการจัดกลุ่ม ε = 0.00001 กำหนดค่าฟัซซี่พารามิเตอร์ m = 2 ซึ่งต้องมากกว่าหนึ่ง

2.2.2 การหาค่าความเป็นสมาชิก คำนวณค่าการเป็นสมาชิกของข้อมูลต่อ กลุ่มข้อมูลต่างๆ การหาค่าการเป็นสมาชิก μ_{ij} แสดงได้จากสมการที่ 2.2

$$\mu_{ij} = \frac{[1/d^2(x_j - z_i)]^{1/(m-1)}}{\sum_{i=1}^{c}[1/d^2(x_j - z_i)]^{1/(m-1)}} \qquad (2.2)$$

c แทนจำนวนกลุ่มข้อมูล μ_{ij} คือค่าการเป็นสมาชิก (membership) ของข้อมูลที่ j ในกลุ่มที่ i $d^2(X_j - Z_i)$ แทนระยะทางยกกำลังสองระหว่างข้อมูล x ที่ j และจุดศูนย์กลางของข้อมูล z กลุ่มที่ i

เมื่อหาค่าการเป็นสมาชิกของข้อมูลทุกกับทุกจุดศูนย์กลางตามตารางที่ 2.4

ตารางที่ 2.4 แสดงตัวอย่างค่าเป็นสมาชิกของข้อมูลกับจุดศูนย์กลาง

	Point	μ_{z1} z1(2,8)	μ_{z2} z2(3,5)	μ_{z2} z3(4,4)
A1	(2, 10)	0.8475	0.0678	0.0847
A2	(2, 5)	0.3165	0.1139	0.5696
A3	(8, 4)	0.0331	0.8595	0.1074
A4	(5, 8)	0.4501	0.3116	0.2383
A5	(7, 5)	0	1	0
A6	(6, 4)	0.0400	0.6400	0.3200
A7	(1, 2)	0.2142	0.1761	0.6097
A8	(4, 9)	0.7143	0.1429	0.1429

2.2.3 ปรับตำแหน่งจุดศูนย์กลางใหม่ คำนวณจุดศูนย์กลางกลุ่มข้อมูลใหม่และตรวจสอบเงื่อนไขโดยตรวจสอบค่าการเป็นสมาชิกใหม่ ลบค่าการเป็นสมาชิกก่อนหน้า การคำนวณค่าจุดศูนย์กลางกลุ่มข้อมูลใหม่ Z_i ตามสมการที่ 2.3

$$z_i = \frac{\sum_{j=1}^{n}(\mu_{ij})^m x_j}{\sum_{j=1}^{n}(\mu_{ij})^m} \qquad (2.3)$$

n แทนจำนวนข้อมูล

- สำหรับ Z1 = $[0.8475^2(2,10) + 0.3165^2(2,5) + 0.0331^2(8,4) + 0.4501^2(5,8) + 0(7,5) + 0.0400^2(6,4) + 0.2142^2(1,2) + 0.7143^2(4,9)] / [0.8475^2 + 0.3165^2 + 0.0331^2 + 0.4501^2 + 0 + 0.0400^2 + 0.2142^2 + 0.7143^2] = (3.0098, 8.8610)$
- สำหรับ Z2 = $[0.0678^2(2,10) + 0.1139^2(2,5) + 0.8595^2(8,4) + 0.3116^2(5,8) + 1^2(7,5) + 0.6400^2(6,4) + 0.1761^2(1,2) + 0.1429^2(4,9)] / [0.0678^2 + 0.1139^2 + 0.8595^2 + 0.3116^2 + 1 + 0.6400^2 + 0.1761^2 + 0.1429^2] = (6.9135, 4.6347)$

- สำหรับ Z3 = $[0.0847^2(2,10) + 0.5696^2(2,5) + 0.1074^2(8, 4) + 0.2383^2(5, 8) + 0^2(7, 5) + 0.3200^2(6, 4) + 0.6097^2(1, 2) + 0.1429^2(4, 9)] / [0.0847^2 + 0.5696^2 + 0.1074^2 + 0.2383^2 + 0 + 0.3200^2 + 0.6097^2 + 0.1429^2] = (2.3559, 3.9478)$

2.2.4 การวนซ้ำ ถ้าค่าจุดศูนย์กลางใหม่กับค่าจุดศูนย์กลางก่อนหน้าต่างกันน้อยกว่าค่าเกณฑ์ที่กำหนดไว้ คือ 0.00001 เป็นจริงคำนวณค่าการเป็นสมาชิกของแต่ละจุดข้อมูลกับจุดศูนย์กลางใหม่และ Objective Function ถ้าเงื่อนไขเป็นเท็จ กลับไปยังข้อ 2.2.2 จนกระทั้ง ค่าจุดศูนย์กลางกลุ่มจะไม่เปลี่ยนแปลง การคำนวณ Objective Function สามารถแสดงดังสมการที่ 2.4

$$J_{FCM} = \sum_{i=1}^{c}\sum_{j=1}^{n}(\mu_{ij})^{m} d^2(x_j - z_i) \qquad (2.4)$$

J_{FCM} = แทน Objective Function ของขั้นตอนวิธีฟัซซี่ซีมีน

ตารางที่ 2.5 แสดงตัวอย่างค่าเป็นสมาชิกของข้อมูลกับจุดศูนย์กลางใหม่

	Point	μ_{z1} z1(3.6066,9.0186)	μ_{z2} z2(6.9762,4.3866)	μ_{z2} z3(1.4107,3.2189)
A1	(2, 10)	0.8776	0.0553	0.0671
A2	(2, 5)	0.1413	0.1053	0.7534
A3	(8, 4)	0.0255	0.9486	0.0259
A4	(5, 8)	0.7942	0.1395	0.0663
A5	(7, 5)	0.0133	0.9760	0.0107
A6	(6, 4)	0.0328	0.9202	0.0469
A7	(1, 2)	0.0277	0.0375	0.9349
A8	(4, 9)	0.9911	0.0051	0.0038

สามารถหาค่าของ J_{FCM} ได้ดังนี้

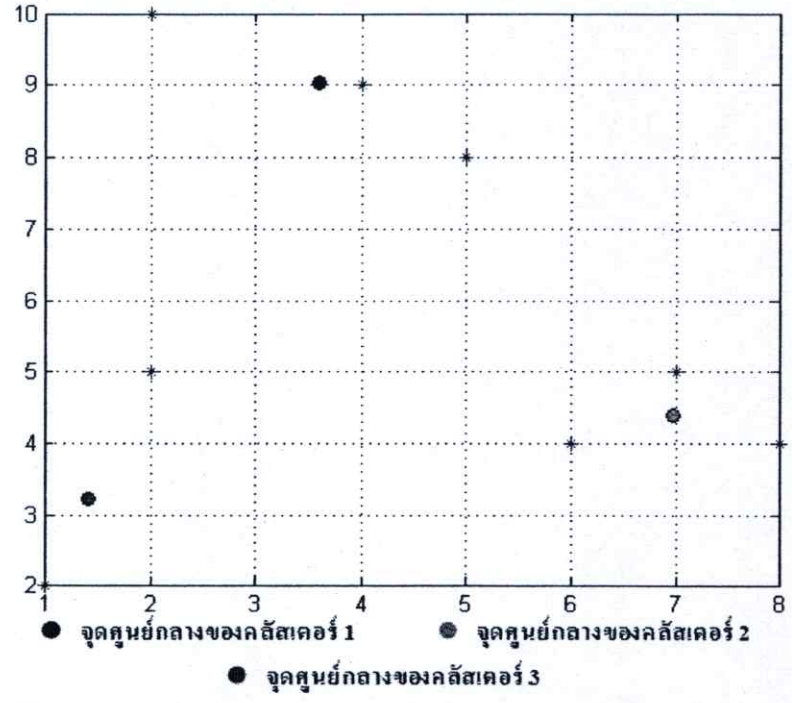
- สำหรับ $J_{F1} = (0.8776^2 \times 1.8826) + (0.1413^2 \times 4.3279) + (0.0255^2 \times 6.67) + (0.7942^2 \times 1.726) + (0.0133^2 \times 5.2597) + (0.0328^2 \times 5.5601) + (0.0277^2 \times 7.487) + (0.9911^2 \times 0.3938) = 3.0289$

- สำหรับ $J_{F2} = (0.0553^2 \times 7.5015) + (0.1053^2 \times 5.0139) + (0.9486^2 \times 1.0944) + (0.1395^2 \times 4.1185) + (0.976^2 \times 0.6138) + (0.9202^2 \times 1.05) + (0.0375^2 \times 6.4351) + (0.0051^2 \times 5.4401) = 2.6265$
- สำหรับ $J_{F3} = (0.0671^2 \times 6.8062) + (0.7534^2 \times 1.8741) + (0.0259^2 \times 6.0291) + (0.0663^2 \times 5.9747) + (0.0107^2 \times 5.8602) + (0.0461^2 \times 4.6490) + (0.9349^2 \times 1.2882) + (0.0638^2 \times 6.3319) = 2.261$

$J_{FCM} = 3.0289 + 2.6265 + 2.2621 = 7.9175$

ตารางที่ 2.6 แสดงค่าจุดศูนย์กลางของคลัสเตอร์ที่ได้จากอัลกอริทึมฟัซซี่ซีมีนคลัสเตอร์ริง

คลัสเตอร์	ค่าจุดศูนย์กลาง	
	X	Y
1	3.6066	9.0186
2	6.9762	4.3866
3	1.4107	3.2189



รูปที่ 2.5 แสดงจุดศูนย์กลางของคลัสเตอร์ 1, 2 และ 3 หลังการทำอัลกอริทึมฟัซซี่ซีมีนคลัสเตอร์ริง

2.3 อัลกอริทึมจัดแบ่งกลุ่มข้อมูลย้ายตามค่าเฉลี่ย (Mean-Shift Clustering Algorithm)

ขั้นตอนการย้ายตามค่าเฉลี่ย (Mean-Shift Algorithm) เดิมได้ถูกนำเสนอในชื่อว่า "valley – seeking procedure" [5] ขั้นตอนวิธีการย้ายตามค่าเฉลี่ยเป็นวิธีหาจุดที่มีความหนาแน่นของข้อมูลที่มากที่สุดโดยเริ่มจากการคำนวณค่าความหนาแน่นของทุกจุดเทียบกับจุดพิจารณา แล้วคำนวณค่าเฉลี่ยเพื่อหาตำแหน่งการย้ายใหม่ การเปลี่ยนแปลงนี้จะเป็นลักษณะการเดินทางตามเส้นทางไปยังตำแหน่งที่มีความหนาแน่นของข้อมูลมากขึ้นจนกระทั่งตำแหน่งการย้ายที่เกิดจากการคำนวณไม่มีการเปลี่ยนตำแหน่งหรือเปลี่ยนแปลงน้อยกว่าค่าที่ยอมรับได้ ก็จะถือว่าตำแหน่งที่ทำการย้ายมานั้นเป็นจุดที่มีความหนาแน่นของข้อมูลสูงที่สุดและจุดที่ลู่เข้าไปที่จุดเดียวกันจะถือว่าอยู่ในกลุ่มเดียวกัน ขั้นตอนวิธีการย้ายตามค่าเฉลี่ย (Mean-Shift Algorithm) กำหนดให้ข้อมูลนำเข้าเป็น $X \in R^m$ โดยที่ $X = (X_1, X_2, X_3, \dots, X_n)$ และ $X_i = [X_{1,i}, X_{2,i}, X_{3,i}, \dots, X_{m,i}]^T$ การประมาณความหนาแน่นของข้อมูล ณ ตำแหน่ง X ใดๆสามารถคำนวณได้สมการ 2.5

$$p(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{\|X-X_i\|^2}{\sigma}\right) \tag{2.5}$$

โดยที่ σ เป็นค่าคงที่และ K(t) เป็นฟังก์ชันเคอร์เนลตำแหน่งที่มีความหนาแน่นสูงสุดของบริเวณที่พิจารณานี้คือตำแหน่งที่ $\nabla p(x) = 0$ ซึ่งสามารถหาได้โดยการคำนวณเพื่อย้ายตำแหน่งหลายๆรอบ โดยการย้ายในแต่ละรอบคำนวณได้ดังสมการ 2.6

$$X^{(t+1)} = f(X^t) \tag{2.6}$$

โดยที่

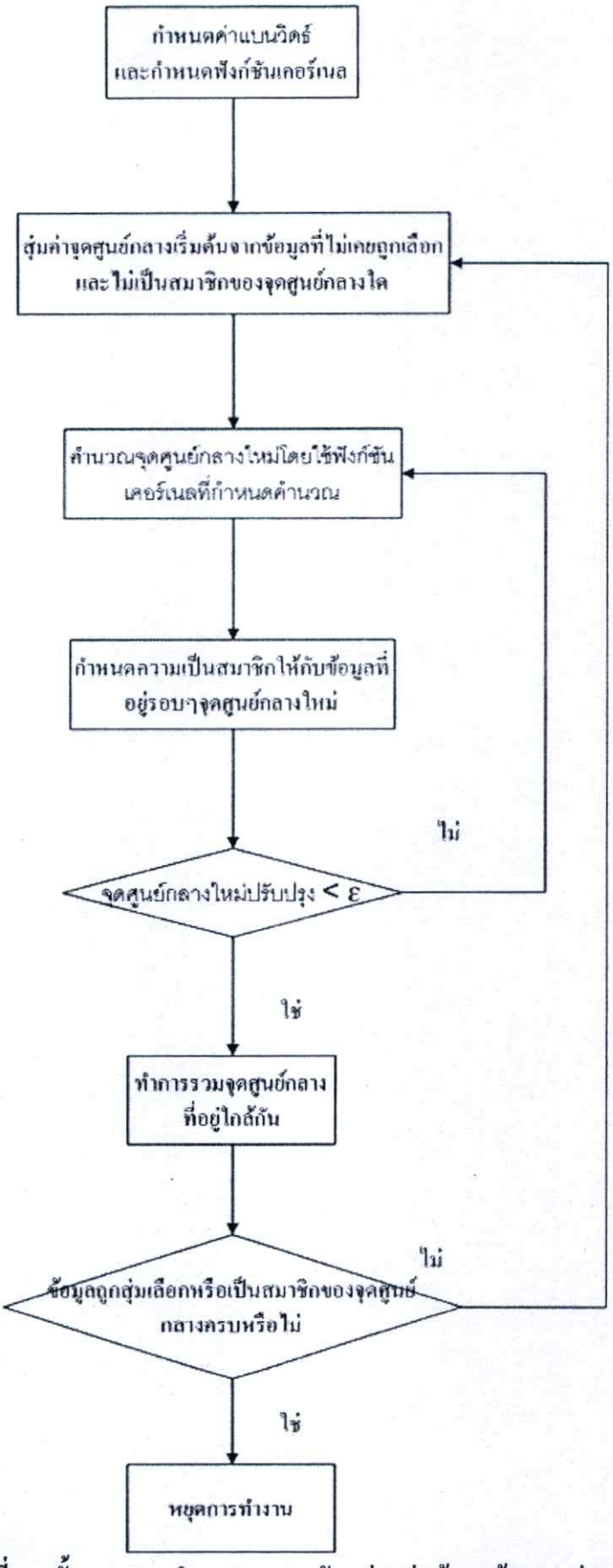
$$f(x) = \frac{\sum_i^n K'\left(\frac{\|x-x_i\|^2}{\sigma}\right)x_i}{\sum_{j=1}^n K'\left(\frac{\|x-x_j\|^2}{\sigma}\right)} \tag{2.7}$$

เมื่อ $K'(t) = \frac{\partial K(i)}{\partial t}$ ขั้นตอนวิธีนี้เรียกว่า ขั้นตอนวิธีการย้ายตามค่าเฉลี่ยจาก [7] ถ้าฟังก์ชันเคอร์เนลเป็นฟังก์ชันเกาส์เซียน ($K(t) = e^{-t/2}$) สมการที่ 2.6, 2.7 จะเป็นดังสมการ 2.8, 2.9

$$X^{(t+1)} = \sum_{i=1}^{n} p(i|X^t)X_i \tag{2.8}$$

$$p(i|X^t) = \frac{\exp\left(-\frac{1}{2}\frac{\|X^t - X_i\|^2}{\sigma}\right)}{\sum_{j=1}^{n} \exp\left(-\frac{1}{2}\frac{\|X^t - X_j\|^2}{\sigma}\right)} \quad (2.9)$$

เมื่อสิ้นสุดขั้นตอนวิธีการย้ายตามค่าเฉลี่ย ทุกตำแหน่งจะถูกย้ายไปที่ตำแหน่งที่มีความหนาแน่นสูงสุดในบริเวณของตำแหน่งนั้นหลังจากนั้นก็จะสามารถดำเนินการจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีในการจัดกลุ่มต่อได้ การทำงานตามขั้นตอนจัดแบ่งกลุ่มข้อมูลย้ายตามค่าเฉลี่ย (Mean-Shift Clustering) มีดังนี้



รูปที่ 2.6 ขั้นตอนการทำงานของการจัดแบ่งกลุ่มข้อมูลย้ายตามค่าเฉลี่ย

ตัวอย่าง จากข้อมูลอยู่ 8 ตัวในตารางที่ 2.1 ต้องการจัดการจัดแบ่งข้อมูลโดยย้ายตามค่าเฉลี่ย

2.3.1 การกำหนดพารามิเตอร์ กำหนดค่าแบนวิดธ์ (bandwidth) = 1.77 กำหนดค่าเพื่อเป็นเงื่อนไขในการหยุดการปรับปรุงจุดศูนย์ $\varepsilon = 0.0018$

2.3.2 สุ่มค่าจุดศูนย์กลางเริ่มต้น C(1) = (8,4)

รอบการทำงานที่ 1

2.3.3 คำนวณค่าจุดศูนย์กลางใหม่ พิจารณาข้อมูลรอบๆจุดศูนย์จะได้ในตารางที่ 2.7

ตารางที่ 2.7 แสดงข้อมูลที่อยู่รอบตัวจุดศูนย์กลาง C(1) = (8,4)

		C(1) = (8,4)	
	Point	Dist C(1)	Dist C(1) < 0.177
A1	(2, 10)	8.4853	False
A2	(2, 5)	6.0828	False
A3	(8, 4)	0	True
A4	(5, 8)	5	False
A5	(7, 5)	1.4142	True
A6	(6, 4)	2	False
A7	(1, 2)	7.2801	False
A8	(4, 9)	6.4031	False

หาค่าเฉลี่ยของจุดศูนย์ใหม่จากข้อมูลที่ผ่านเงื่อนไขในตารางที่ 2.7

$$C(2) = \left(\frac{8+7}{2}, \frac{4+5}{2}\right) = (7.5,4.5)$$

2.3.4 กำหนดความเป็นสมาชิก กำหนดให้ A3, A5 เป็นสมาชิกของ C(2)

2.3.5 การบันทึกตำแหน่งจุดศูนย์กลางใหม่ พิจารณาเงื่อนไขจุดศูนย์กลางใหม่ปรับปรุงหรือไม่ โดยใช้สมการที่ (2.10)

$$\|C(t+1) - C(t)\| < \varepsilon \quad (2.10)$$

จะได้ว่า

$$\|C(2) - C(1)\| < 0.0018$$

$$\|(7.5,4.5) - (8,4)\| < 0.0018$$

$0.7071 < 0.0018$ ดังนั้น กลับไปทำขั้นตอนที่ (2.3.3) โดยนำค่าจุดกลางใหม่ที่ได้คือ C(2) = (7.5,4.5) ไปคำนวณการเคลื่อนที่ของจุดศูนย์กลางอีกรอบ จนกระทั่งเมื่อถึงรอบที่ 3 ค่า C(4) = (7,4.3333) และ ค่า C(3) = (7,4.3333) จะได้เงื่อนตามนี้ $0 < 0.0018$ ซึ่งค่าของจุดศูนย์กลางรอบที่ 4 กับ รอบที่ 3 มีค่าต่างกันน้อยว่า 0.0018 ดังนั้นจะได้ค่าของ C(4) เป็นจุดศูนย์กลางของคลัสเตอร์ที่ 1 และทำการบันทึกเก็บไว้แล้วไปทำขั้นตอนที่ 2.3.6

2.3.6 การรวมตำแหน่งจุดศูนย์กลางของคลัสเตอร์ ถ้าจำนวนของคลัสเตอร์ทั้งหมดเท่ากับ 1 ก็ไม่ต้องทำการรวมค่าตำแหน่งจุดศูนย์กลาง ถ้าจำนวนของคลัสเตอร์ทั้งหมดมากกว่า 1 ก็จะทำการรวมตำแหน่งจุดศูนย์กลาง เมื่อโปรแกรมทำงานจนมาถึงจำนวนของคลัสเตอร์ทั้งหมดเท่ากับ 2 ค่าจุดศูนย์ของคลัสเตอร์ที่ 1 C_1 = (7,4.3333) และค่าจุดศูนย์ของคลัสเตอร์ที่ 2 C_2 = (4.5,8.5) ซึ่งเป็นคลัสเตอร์ที่บันทึกใหม่

พิจารณาการรวมตำแหน่งจุดศูนย์กลางโดยใช้สมการที่ 2.11

$$\|C_i - C_K\| < \frac{\text{bandwidth}}{2} \quad (2.11)$$

โดยที่ $i = 1, 2, 3, \ldots K$

จะได้ว่า

$$\|(7,4.3333) - (4.5,8.5)\| < \frac{1.77}{2}$$

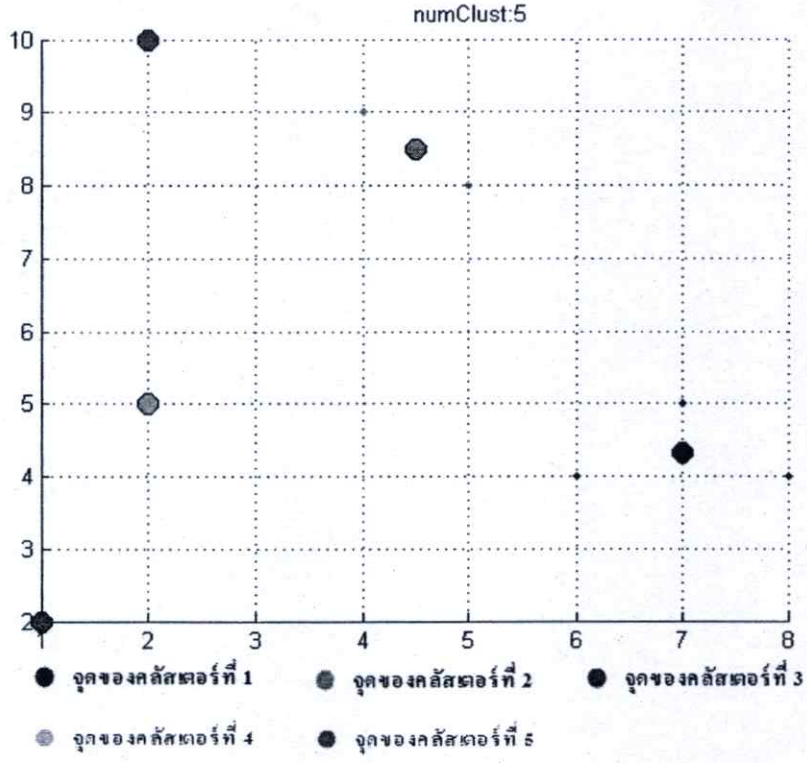
$0.0441 < 0.1175$ ดังนั้นไม่ต้องทำการรวมคลัสเตอร์ที่ 1 กับ ของคลัสเตอร์ที่ 2 ในกรณีที่เงื่อนไขเป็นจริงจะทำการรวมตำแหน่งจุดศูนย์กลางใช้สมการที่ 2.12

$$C_i = \frac{C_i + C_K}{2} \quad (2.12)$$

2.3.7 ขั้นตอนการวนซ้ำ ถ้าข้อมูลถูกสุ่มเลือกหรือเป็นสมาชิกของจุดศูนย์กลางยังไม่ครบทั้งหมดในเซตข้อมูลก็ให้กลับไปทำขั้นตอนที่ 2.3.2 ถ้าไม่ก็หยุดทำงาน

ตารางที่ 2.8 แสดงค่าจุดศูนย์กลางของคลัสเตอร์ที่ได้จากอัลกอริทึมจัดแบ่งกลุ่มข้อมูลย้ายตามค่าเฉลี่ย

คลัสเตอร์	ค่าจุดศูนย์กลาง	
	X	Y
1	7	4.333
2	4.5	8.5
3	1	2
4	2	5
5	2	10



รูปที่ 2.7 แสดงค่าจุดศูนย์กลางของคลัสเตอร์ที่ได้จากอัลกอริทึมจัดแบ่งกลุ่มข้อมูลย้ายตามค่าเฉลี่ย

2.4 อัลกอริทึมจัดแบ่งกลุ่มเชิงพื้นที่ความหนาแน่น (Density Based Spatial Clustering Of Application With Noise Algorithm: DBSCAN Algorithm)

ขั้นตอน DBSCAN [6] จัดอยู่ในประเภทของการแบ่งกลุ่มแบบ Center-Based-Approach ซึ่งอาศัยความหนาแน่นของจุดมาใช้ในการจัดกลุ่ม กลุ่มที่ได้จากการแบ่งกลุ่มแบบ DBSCAN จะมีรูปร่างที่ไม่แน่นอน รวมทั้งสามารถแยกจุดที่ไม่สามารถเข้ากับกลุ่มใดๆได้ (Noise) ออกจากกลุ่มได้ด้วย ในการประมาณความหนาแน่นของจุดทำได้โดยการนับจำนวนจุดที่อยู่ภายใต้รัศมีของจุดกึ่งกลาง และนับจุดกึ่งกลางด้วย ความหนาแน่นของจุดขึ้นอยู่กับรัศมีที่กำหนด DBSCAN ไม่ต้องกำหนดจำนวนกลุ่มที่ต้องการ เมื่อทำตามวิธีการของ DBSCAN แล้ว จะสามารถทราบได้ว่าจำนวนกลุ่มที่สามารถจัดได้มีทั้งหมดกี่กลุ่ม ซึ่งสามารถค้นพบกลุ่มที่ K-means ไม่สามารถค้นพบได้ แต่ใน DBSCAN ต้องทำการกำหนดจำนวนขั้นต่ำของจุดภายในกลุ่ม และรัศมีจากจุดศูนย์กลางของกลุ่มที่อยู่ของจุด และความหนาแน่นของทุกๆจุด P นั้นสามารถนิยามได้จาก 2 พารามิเตอร์คือ

(1) Eps (ε) หรือ Radius คือบริเวณที่อยู่ใกล้เคียงของจุด P ซึ่งทำให้สามารถบอกได้ว่าจุดไหนอยู่ภายกลุ่มหรืออยู่ภายนอกกลุ่ม

(2) MinPts คือ จำนวนขั้นต่ำของจุดภายในกลุ่ม

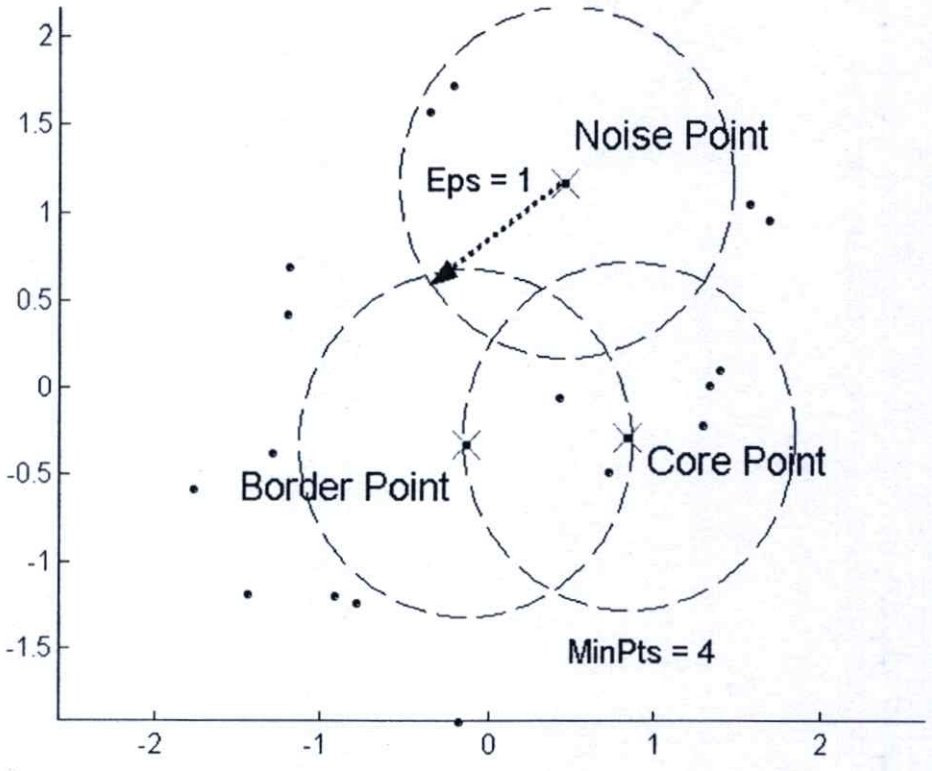
ในการเลือก พารามิเตอร์สำหรับ DBSCAN คือ การเลือกค่าของ Eps และ MinPts นั้นจะดูจากพฤติกรรมของระยะห่างของจุด เรียกว่า ค่า k-dist ถ้าจุดอยู่ในกลุ่มเดียวกันจะมีค่า k-dist ที่ใกล้เคียงกัน คือมีค่า k-dist น้อยๆ แต่ถ้าเป็น Noise ค่า k-dist จะแตกต่างจากจุดอื่น คือมีค่ามาก ทำให้เราสามารถกำหนดค่า Eps และ MinPts ได้อย่างเหมาะสม

ในการแบ่งกลุ่มของ DBSCAN ตามแบบ Center-Based-Approach สามารถแบ่งจุดได้ 3 แบบ คือ

(1) Core point คือ จุดที่อยู่ภายในรัศมีของกลุ่ม และจะเป็นจุดกึ่งกลางเมื่อมีจุดอื่นที่อยู่ภายในกลุ่มอยู่รอบๆจุดนั้น

(2) Border point คือ จุดที่อยู่ขอบของกลุ่ม แต่ไม่ใช่จุด Core point คือจะเป็นจุดที่อยู่ใกล้ๆกับ Core point

(3) Noise point คือ จุดที่ไม่สามารถเข้าร่วมกลุ่มกับกลุ่มใดๆได้เลย



รูปที่ 2.8 แสดงประเภทของจุดใน DBSCAN

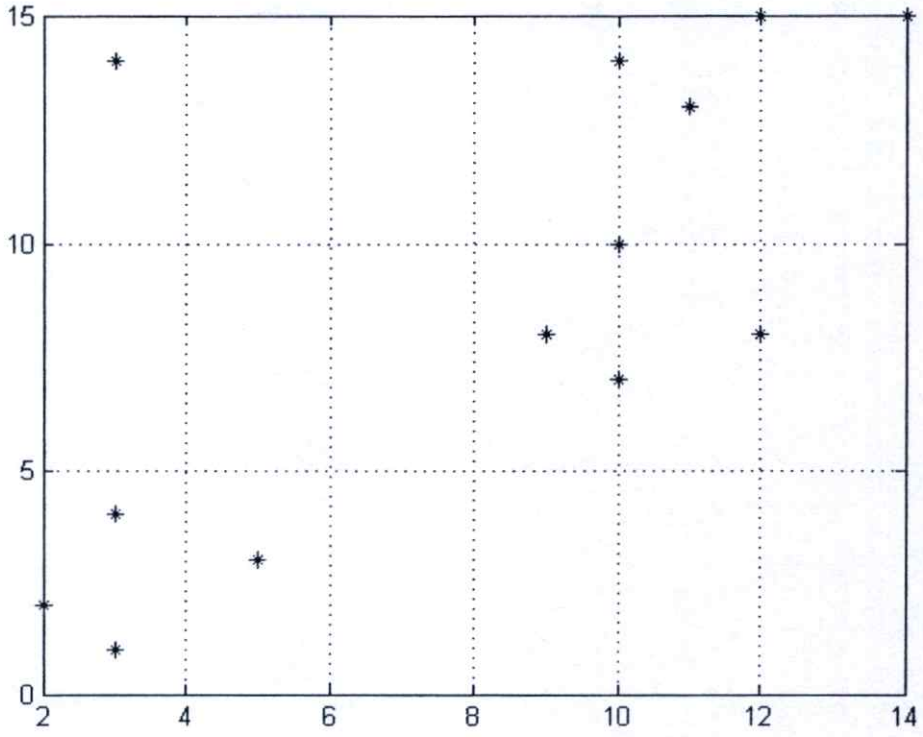
ตัวอย่าง ถ้ามีข้อมูลอยู่ 13 ตัวในตารางที่ 2.9 ต้องการจัดแบ่งกลุ่มโดยใช้อัลกอริทึม DBSCAN

ตารางที่ 2.9 แสดงตัวอย่างเซตข้อมูลที่ต้องการจัดแบ่งกลุ่มโดยใช้อัลกอริทึม DBSCAN

ID	X	Y
A1	2	2
A2	3	1
A3	3	4
A4	5	3

ตารางที่ 2.9 (ต่อ) แสดงตัวอย่างเซตข้อมูลที่ต้องการจัดแบ่งกลุ่มโดยใช้อัลกอริทึม DBSCAN

ID	X	Y
A5	9	8
A6	10	7
A7	10	10
A8	12	8
A9	3	14
A10	10	14
A11	11	13
A12	12	15
A13	14	15



รูปที่ 2.9 แสดงรูปข้อมูลที่ต้องการจัดแบ่งกลุ่มโดย * แทนข้อมูลแต่ละจุด

2.4.1 กำหนดพารามิเตอร์ ค่าจำนวนขั้นต่ำของจุดภายในกลุ่ม (MinPts) = 4 และคำนวณค่าของ Eps โดยการประมาณรัศมีพื้นที่ใกล้เคียง

$$\text{Eps} = \left(\frac{\prod_{d=1}^{D}(\max_{j\in\{1,2,3,\ldots N\}} A_j^d - \min_{j\in\{1,2,3,\ldots N\}} A_j^d)\times(0.5\times D)!\times \text{MinPts}}{N\times\sqrt{\pi^D}}\right)^{\frac{1}{D}} \qquad (2.13)$$

โดยที่

D = จำนวนมิติของเซตข้อมูล

N = จำนวนข้อมูลของเซตข้อมูลทั้งหมด

จากตัวอย่างสามารถคำนวณค่า Eps ได้ดังนี้

$$\text{Eps} = \left(\frac{((14-2)\times(15-1))\times(0.5\times 2)!\times 4}{13\times\sqrt{3.1416^2}}\right)^{\frac{1}{2}} = 4.0564$$

2.4.2 ทำการสุ่มเลือกจุดเริ่มต้น ในที่นี้เลือก P = (2, 2) C1 = {A1(2, 2)}

2.4.3 ทำการหาจุดที่อยู่ใกล้จุด P คือจุด A1(2,2) และจุดนั้นอยู่ภายใต้ข้อกำหนดของ Eps หรือไม่

ตารางที่ 2.10 แสดงจุดที่ใกล้ P ที่สุดที่ค้นพบภายใต้รัศมี Eps

		P = (2,2)	
	Point	Dist P	Dist P < 4.0564
A1	(2, 2)	0	True
A2	(3, 1)	1.4142	True
A3	(3, 4)	2.2361	True
A4	(5, 3)	3.1623	True
A5	(9, 8)	9.2195	False
A6	(10, 7)	9.4340	False
A7	(10, 10)	11.3137	False
A8	(12, 8)	11.6619	False
A9	(3, 14)	12.0416	False
A10	(10,14)	14.4222	False
A11	(11, 13)	14.2127	False
A12	(12, 15)	16.4012	False
A13	(14,15)	17.6918	False

จากนั้นจะทำการเพิ่มเป็นกลุ่มเป็นกลุ่มใหม่ที่ค้นพบโดยเพิ่มเข้าไปใน C1 ตอนนี้ C1 = {A1, A2, A3, A4}

2.4.4 ทำการค้นหาทุกจุดที่ใกล้จุดอื่นที่อยู่ในกลุ่ม จะทำการเลือกจุดอื่นที่อยู่ในกลุ่มเป็นจุด P ทำการค้นหาทุกจุดที่อยู่ใกล้ วนทำตามขั้นตอนที่ 2.4.3 – 2.4.4 ทำไปเรื่อยๆ จนหมด จะได้ว่า C1 = {A1(2, 2), A2(3, 1), A3(3,4), A4(5,3)}

2.4.5 ทำการตรวจสอบจำนวนขั้นต่ำของกลุ่ม ถ้าจำนวนจุดภายในเซต C1 น้อยกว่าจำนวนขั้นต่ำของจุดภายในกลุ่ม (MinPts) หรือไม่ ถ้าน้อยกว่าก็กำหนดจุดในเซต C1 เป็น Noise point ในที่นี้จะได้ว่า C1 ไม่เป็น Noise point เพราะมีจำนวนจุดเท่ากับ 4 พอดี จะได้ว่าเซตของ C1 เป็นจุดที่อยู่ในคลัสเตอร์ที่ 1

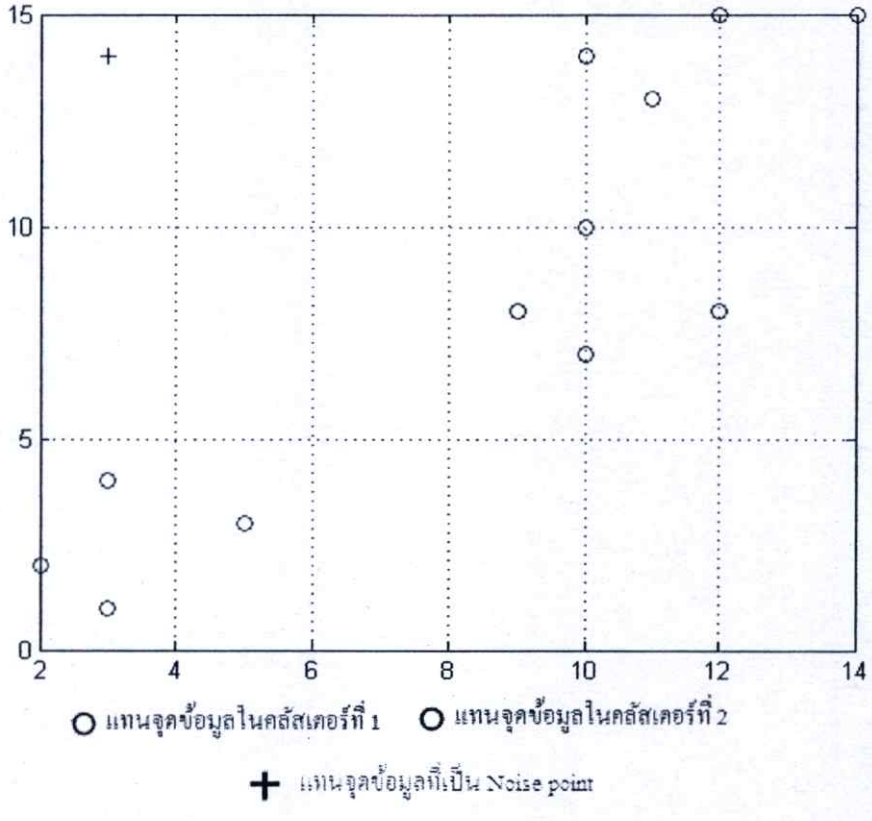
2.4.6 ทำการวนซ้ำ ตรวจสอบว่าทุกจุดในเซตของข้อมูลทั้งหมดถูกระบุบเป็นจุดในคลัสเตอร์ต่างครบหรือไม่ ถ้ายังไม่ครบให้กลับไปทำขั้นตอนที่ 2.4.2 เพื่อหาเซตของจุดที่อยู่คลัสเตอร์ลำดับถัดไป โดยจะสุ่มเลือกตัวที่ไม่เคยถูกสุ่มและไม่ได้เป็นจุดที่อยู่เซตของคลัสเตอร์ใดๆมาก่อน แต่ถ้าครบแล้วก็หยุดการทำงานอัลกอริทึมนี้

ตารางที่ 2.11 แสดงข้อมูลแต่ละจุดอยู่คลัสเตอร์ใดที่ได้จากอัลกอริทึม DBSCAN

ค่าจุดข้อมูล		คลัสเตอร์
X	Y	
2	2	1
3	1	1
3	4	1
5	3	1
9	8	2
10	7	2
10	10	2
12	8	2
3	14	Noise point
10	14	2
11	13	2

ตารางที่ 2.11 (ต่อ) แสดงข้อมูลแต่ละจุดอยู่คลัสเตอร์ใดที่ได้จากอัลกอริทึม DBSCAN

ค่าจุดข้อมูล		คลัสเตอร์
X	Y	
12	15	2
14	15	2



รูปที่ 2.10 แสดงข้อมูลแต่ละจุดอยู่คลัสเตอร์ใดที่ได้จากอัลกอริทึม DBSCAN

2.5 อัลกอริทึมการค้นหาคำตอบด้วยแรงโน้มถ่วง (Gravitational Search Algorithm: GSA)

นิวตันได้ค้นพบธรรมชาติพื้นฐานของแรงดึงดูดโน้มถ่วงระหว่างวัตถุใดๆ สองวัตถุโดยนิวตันได้ตีพิมพ์กฎแรงโน้มถ่วง เอาไว้ดังนี้ "ทุกอนุภาคสสารในเอกภพดึงดูดทุกอนุภาคอื่นด้วยแรงซึ่งแปรผันตรงกับผลคูณของมวลของอนุภาคและแปรผกผันกับกำลังสองของระยะห่างระหว่างอนุภาคทั้งสองนั้น" เมื่อแปลข้อความนี้เป็นสมการจะได้ตาม 2.14 [7]

$$F = G\frac{M_1 M_2}{R^2} \tag{2.14}$$

F คือขนาดของแรงโน้มถ่วงซึ่งทำต่ออนุภาคใดอนุภาคหนึ่ง, M_1 และ M_2 เป็นมวลของอนุภาคทั้งสอง, R คือระยะห่างระหว่างอนุภาคและ G คือค่าคงตัวฟิสิกส์พื้นฐานที่เรียกว่าค่าคงตัวโน้มถ่วง

อัลกอริทึมการค้นหาคำตอบด้วยแรงโน้มถ่วงนี้ อนุภาคจะถูกมองว่าคือวัตถุและมวลของวัตถุคือตัววัดประสิทธิภาพ วัตถุจะดึงดูดวัตถุอื่นแต่ละวัตถุโดยแรงโน้มถ่วง และด้วยแรงโน้มถ่วงนี้เองทำให้เกิดการเคลื่อนที่ของวัตถุไปสู่วัตถุที่หนักกว่า โดยมวลจะใช้รูปแบบการคำนวณเพื่อปรับตำแหน่งด้วยแรงโน้มถ่วง มวลที่หนักก็เป็นจุดที่มีคำตอบที่ดี แล้วเคลื่อนที่ช้ากว่ามวลที่มีน้ำหนักเบากว่า เพื่อความเข้าใจมากขึ้นขออธิบายมวลในทางฟิสิกส์ในย่อหน้าถัดไป

มวลในทางฟิสิกส์แบ่งออกเป็น 3 ชนิดด้วยกันได้แก่

1) Active Gravitational Mass เป็นตัวบอกถึงอิทธิพลของแรงโน้มถ่วงเนื่องจากวัตถุนั้น วัตถุที่มีมวลมากกว่าก็ย่อมที่จะมีสนามโน้มถ่วงที่แรงกว่าวัตถุมวลน้อย

2) Passive Gravitational Mass เป็นปริมาณที่บ่งบอกความแรงของอันตรกิริยาระหว่างวัตถุกับสนามโน้มถ่วง สำหรับสนามโน้มถ่วงที่มีความเข้มเท่ากันวัตถุที่มีมวลมากก็จะรู้สึกถึงแรงโน้มถ่วงมากกว่าวัตถุที่มีมวลน้อย แรงโน้มถ่วงที่มาทำกับวัตถุนี้เราเรียกกันว่า น้ำหนักนั่นเอง

3) Inertial Mass เป็นปริมาณบอกถึงสภาพการต่อต้านการเคลื่อนที่ของวัตถุเมื่อมีแรงมากระทำวัตถุขนาดใหญ่ก็จะมีสภาพการต่อต้านการเคลื่อนที่ของวัตถุมากและวัตถุขนาดเล็กจะมีสภาพการต่อต้านการเคลื่อนที่ของวัตถุน้อย

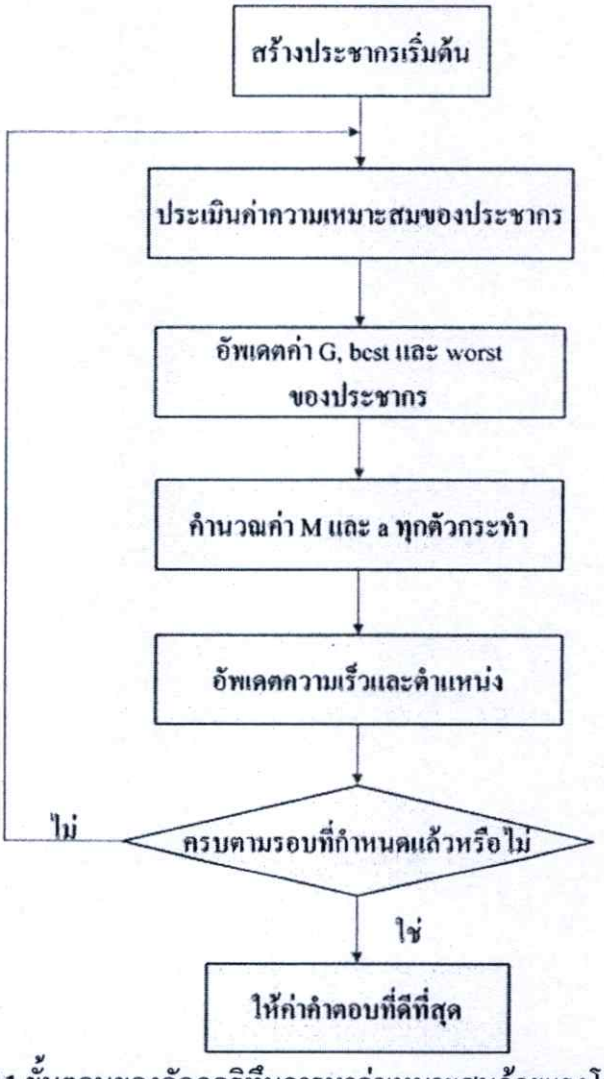
ใน GSA แต่ละมวล (อนุภาค) มีข้อมูลที่ต้องพิจารณาดังนี้ ตำแหน่งของมวล, Inertial Mass, Active Gravitational Mass และ Passive Gravitational Mass โดยตำแหน่งของมวลจะคือคำตอบของปัญหา ซึ่งมวลถูกแทนด้วยฟังก์ชันค่าเหมาะสม โดยอัลกอริทึมจะพยายามหาเส้นทางเพื่อการปรับตำแหน่งของมวลอย่างเหมาะสม แล้วเมื่ออัลกอริทึมทำงานจนครบจำนวนรอบที่กำหนดไว้ ในบทความนี้คาดหวังว่ามวลจะถูกดึงดูดโดยมวลที่หนักที่สุด มวลนี้ก็จะถูกแทนเป็นคำตอบที่เหมาะสมที่สุดในพื้นที่ค้นหา

GSA สามารถแยกการพิจารณามวลออกจากระบบได้ เหมือน Artificial world เล็กของมวลที่ทำตามกฎแรงโน้มถ่วงและกฎการเคลื่อนที่ของนิวตัน เพื่อความเข้าใจมากขึ้นของอธิบายกฎที่มวลถูกกระทำในย่อหน้าถัดไป

1) กฎของแรงโน้มถ่วง แต่ละอนุภาคดึงดูดทุกอนุภาค และแรงโน้มถ่วงระหว่าง 2 อนุภาคคือสัดส่วนของการคูณของมวลและหารด้วยระยะทางระหว่างมวลคือ R ในบทความนี้จะใช้ R แทน R^2 เพราะเป็นที่ยอมรับกันจากผลลัพธ์การทดลองของเรา R ให้ผลลัพธ์ที่ดีกว่า R^2 ในการทดลองทุกกรณี

2) กฎของการเคลื่อนที่ ความเร็วปัจจุบันของมวลใดๆ เท่ากับผลรวมความเร็วก่อนหน้าบวกด้วยอัตราการเปลี่ยนแปลงของความเร็ว อัตราการเปลี่ยนแปลงของความเร็วหรือความเร่งของมวลใดๆ เท่ากับแรงที่ถูกกระทำบนระบบหารด้วย Inertial Mass

ขั้นตอนต่างๆ ของอัลกอริทึมการหาค่าเหมาะสมด้วยแรงโน้มถ่วงดังรูปที่ 2.11 [9]



รูปที่ 2.11 ขั้นตอนของอัลกอริทึมการหาค่าเหมาะสมด้วยแรงโน้มถ่วง

2.5.1 วิธีการคำนวณค่ามวล

ประสิทธิภาพจะถูกวัดเหมือนกับมวล มวลแรงโน้มถ่วงและมวลเฉื่อย (inertia mass) ถูกคำนวณอย่างง่ายๆ โดยใช้การประเมินค่าเหมาะสม มวลที่หนักกว่าหมายถึงอนุภาคที่มีผลกระทบมากกว่า ความหมายนี้หมายถึงอนุภาคที่ดีกว่ามีแรงดึงดูดสูงกว่าและเคลื่อนที่ช้ากว่า การตั้งสมมุติฐานให้มวลแรงโน้มถ่วงและมวลเฉื่อยเท่ากัน ค่าของมวลถูกคำนวณโดยการใช้การ map ของค่าเหมาะสม สามารถอัพเดตมวลแรงโน้มถ่วงและมวลเฉื่อยโดยตามสมการนี้

$$M_{ai} = M_{pi} = M_{ii} = M_i \quad i = 1, 2, \ldots, N \tag{2.14}$$

$$m_i(t) = \frac{fit_i(t) - worst(t)}{best(t) - worst(t)} \tag{2.15}$$

$$M_i(t) = \frac{m_i(t)}{\sum_{j=1}^{N} m_j(j)} \tag{2.16}$$

โดย $fit_i(t)$ แทนค่าเหมาะสมของอนุภาค i ที่เวลา t

worst(t) และ best(t) ถูกกำหนดตามสมการ 2.17 และ 2.18 นี้ (สำหรับปัญหาค่าน้อยสุด)

$$best(t) = \min_{j \in \{1,\ldots,N\}} fit_j(t) \tag{2.17}$$

$$worst(t) = \max_{j \in \{1,\ldots,N\}} fit_j(t) \tag{2.18}$$

สำหรับปัญหาค่ามากสุดสมการ 2.17 และ 2.18 ถูกเปลี่ยนเป็นสมการ 2.19 และ 2.20 ตามลำดับ

$$best(t) = \max_{j \in \{1,\ldots,N\}} fit_j(t) \tag{2.19}$$

$$worst(t) = \min_{j \in \{1,\ldots,N\}} fit_j(t) \tag{2.20}$$

2.5.2 การปรับความเร็วและตำแหน่ง

กำหนดตำแหน่งของอนุภาคลำดับที่ i โดย

$$X_i = \left(X_i^1, \ldots, X_i^d, \ldots, X_i^n\right) \tag{2.21}$$

โดย $i = 1, 2, \ldots, N$

X_i^d แทนตำแหน่งของอนุภาคลำดับที่ i ในมิติลำดับที่ d ณ เวลาที่ t กำหนดให้แรงกระทำบนมวล i จากมวล j เป็นตามสมการที่ 2.22

$$F_{ij}^d = G(t)\frac{M_{pi}(t)\times M_{aj}(t)}{R_{ij}(t)+\varepsilon}\left(X_j^d(t) - X_i^d(t)\right) \tag{2.22}$$

โดย M_{aj}คือมวลแบบ Active gravitational ที่เกี่ยวข้องกับอนุภาค j

M_{pi} คือมวลแบบ Passive gravitational ที่เกี่ยวข้องกับอนุภาค i

G(t) คือ ค่าคงที่แรงโน้มถ่วงที่เวลา t

คือค่าคงที่น้อยๆ

$R_{ij}(t)$ คือ ระยะทางแบบ Euclidian ระหว่าง 2 อนุภาค i และ j

เพื่อให้อัลกอริทึมมีลักษณะแบบ stochastic สมมุติให้แรงทั้งหมดกระทำบนอนุภาคที่ i ในมิติที่ d และมีการสุ่มผลรวมน้ำหนักของมิติที่ d เป็นส่วนประกอบของแรงกระทำจากอนุภาคอื่น

$$F_i^d(t) = \sum_{j=1,j\neq i}^{N} \mathrm{rand}_j F_{ij}^d(t) \tag{2.23}$$

โดย rand_j คือ ตัวเลขสุ่มในช่วง [0,1]

นอกจากนี้ด้วยกฎการเคลื่อนที่ ความเร่งของอนุภาคที่ i ที่เวลา t และในทิศทาง d เป็นตามสมการที่ 2.24

$$a_i^d(t) = \frac{F_i^d(t)}{M_{ii}(t)} \tag{2.24}$$

โดย M_{ii} คือ มวยเฉื่อยของอนุภาคที่ i

นอกจากนี้ ความเร็วถัดไปของอนุภาคถูกพิจารณาเป็นเลขเศษส่วนของความเร็วปัจจุบันของอนุภาคบวกความเร่งอนุภาค เพราะฉะนั้นตำแหน่งอนุภาคและความเร็วของอนุภาคสามารถคำนวณได้ตามสมการที่ 2.25 และ 2.26

$$v_i^d(t+1) = rand_i \times v_i^d(t) + a_i^d(t) \quad (2.25)$$

$$x_i^d(t+1) = x_i^d(t) + v_i^d(t+1) \quad (2.26)$$

โดย $rand_i$ คือรูปแบบค่าตัวแปรสุ่มในช่วง [0,1] เราใช้เลขสุ่มนี้ให้มีลักษณะการสุ่มในการค้นหา

ค่าคงที่แรงโน้มถ่วง G ถูกเริ่มต้นที่จุดเริ่มต้นและจะลดลงด้วยเวลาในควบคุมความถูกต้อง อีกในหนึ่ง G คือฟังก์ชันของค่าเริ่มต้น (G_0) และเวลา (t)

$$G(t) = G(G_0, t) \quad (2.27)$$

2.6 อัลกอริทึมการจัดแบ่งกลุ่มข้อมูลแบบใหม่ด้วยแรงโน้มถ่วง (A New Gravitational Clustering Algorithm)

อัลกอริทึมการจัดแบ่งกลุ่มข้อมูลแบบใหม่ด้วยแรงโน้มถ่วง [12] เป็นเทคนิคการจัดแบ่งที่อาศัยแนวคิดกฎแรงโน้มถ่วงและกฎการเคลื่อนข้อสองของนิวตัน โดยพื้นฐานแล้วทุกจุดข้อมูลจะถูกมองว่าเป็นวัตถุที่สามารถเคลื่อนโดยใช้แรงโน้มถ่วงและกฎการเคลื่อนข้อมูลสองของนิวตัน โดยแนวคิดการจัดแบ่งกลุ่มด้วยแรงโน้มถ่วงนี้ได้มีการนำไปพัฒนาเป็นงานวิจัยบ้างแล้ว งานวิจัยดังกล่าวถูกนำเสนอโดย Wright [13] ในงานวิจัยของเขาเป็นการทำคลัสเตอร์แบบ An hierarchical agglomerative algorithm โดยเขาใช้แนวคิดแรงโน้มถ่วงในกลไกการรวมกันของอนุภาค (จุดข้อมูล) จนกระทั่งเหลือเพียงหนึ่งอนุภาคเท่านั้นในระบบ

ต่อไปนี้เป็นหัวข้อที่แสดงถึงคุณสมบัติที่สำคัญของการเรียนรู้โมเดลแบบไดนามิกที่นำเสนอโดย Wright

- การกำหนดตำแหน่งใหม่ของอนุภาคมาหนึ่งตัวจากอนุภาคที่เหลืออยู่ทั้งหมดในเซตข้อมูลที่ถูกใช้งาน
- เมื่อสองอนุภาค (จุดข้อมูลเริ่มต้น) ใกล้ชิดกันเพียงพอจะถูกรวมกันแล้วจะมีหนึ่งอนุภาคที่ถูกลบไปจากการเรียนรู้โมเดลนี้และมวลของอนุภาคอื่นๆ จะถูกเพิ่มโดยมวลของอนุภาคที่ถูกลบไป
- ระยะทางสูงสุดที่แต่ละจุดข้อมูลสามารถเคลื่อนที่ได้ในแต่ละรอบของอัลกอริทึม ระยะทางสูงสุดเป็นพารามิเตอร์ซึ่งถูกกำหนดโดยผู้ใช้
- วิธีที่กล่าวไปก่อนหน้านี้ จะทำการเรียนรู้โมเดลจนจบก็ต่อเมื่อเหลือหนึ่งอนุภาคในระบบ

ในงานวิจัยนี้ได้นำเสนอการเรียนรู้โมเดลไดนามิกแบบใหม่ซึ่งสามารถใช้ในอัลกอริทึมการทำคลัสเตอร์ได้ ข้อมูลดีหลักในงานวิจัยของ Wright คือความเร็ว, ความทนทานต่อความผิดพลาดและเป็นการเรียนแบบ unsupervised (ไม่จำเป็นต้องรู้จำนวนคลัสเตอร์ที่เหมาะสม)

2.6.1 กฎการเคลื่อนที่ของนิวตัน

ไอแซก นิวตัน ได้พัฒนาแบบจำลองการเคลื่อนที่เรียกว่า "กฎการเคลื่อนที่ของนิวตัน" ซึ่งอธิบายการเคลื่อนที่ของวัตถุขนาดใหญ่ในเอกภพแต่ไม่สามารถอธิบายการเคลื่อนที่ของวัตถุขนาดเล็กในระดับอะตอมได้ ในส่วนนี้จะอธิบายกฎการเคลื่อนที่ของนิวตันกับวัตถุที่เคลื่อนที่ในวิถีเส้นตรงและมีมิติการเคลื่อนที่ n มิติ

2.6.1.1 กฎการเคลื่อนที่ในหนึ่งมิติ

x คือวัตถุในพื้นที่ Euclidean และ t เป็นจำนวนจริงแทนเวลา x(t) เป็นตำแหน่งของวัตถุที่เวลา t s(t) คือความเร็วของวัตถุที่เวลา t a(t) คือความเร่งของวัตถุที่เวลา t และ F(t) คือแรงที่กระทำกับวัตถุที่เวลา t แสดงสมการการเคลื่อนที่ตามสมการที่ 2.28 – 2.29

$$s(t) = s(0) + at \tag{2.28}$$

$$x(t) = x(0) + s(0)t + \frac{at^2}{2} \tag{2.29}$$

เมื่อ x(0) คือตำแหน่งของวัตถุเริ่มต้น s(0) คือ ความเร็วของวัตถุเริ่มต้นและ a คือค่าคงที่ของความเร่งของวัตถุ

สมการที่ (2.28) และ (2.29) สามารถใช้จำลองการเคลื่อนที่ของวัตถุเมื่อตำแหน่งและความเร่งไม่สามารถวิเคราะห์ได้ ในวิธีนี้ความเร่งเป็นค่าคงระหว่างช่วงเวลา $\Delta(t)$ และความเร็วกับตำแหน่งของวัตถุที่เวลา $t + \Delta(t)$ ตามสมการที่ (2.30) – (2.31)

$$s\big(t + \Delta(t)\big) = s(t) + a(t)\Delta(t) \tag{2.30}$$

$$x\big(t + \Delta(t)\big) = x(t) + s(t)\Delta(t) + \frac{a(t)\Delta(t)^2}{2} \tag{2.31}$$

สุดท้ายถ้า m_x คือมวลของวัตถุ x ดังนั้นแรงที่กระทำบนวัตถุที่ถูกนำมาใช้ในกฎการเคลื่อนข้อที่สองของนิวตันตามสมการที่ (2.32)

$$F(t) = m_x a(t) \tag{2.32}$$

2.6.1.2 กฎการเคลื่อนที่วิถีเส้นตรงในหลายมิติ

กฎการเคลื่อนที่หลายมิติสำหรับการเคลื่อนที่วิถีเส้นตรงเป็นการเคลื่อนที่ที่มีเวกเตอร์มากกว่าหนึ่งมิติ x เป็นวัตถุในพื้นที่ Euclidean แบบ n มิติ ซึ่งการเคลื่อนที่มีทิศทางเป็นเวกเตอร์ $\vec{d}$ และ t เป็นจำนวนจริงที่แทนเวลา x(t) เป็นตำแหน่งของวัตถุที่เวลา t v(t) เป็นความเร็วของวัตถุที่เวลา t $\overrightarrow{a(t)}$ เป็นความเร่งของวัตถุที่เวลา t แสดงสมการการเคลื่อนที่ในหลายมิติตามสมการที่ 2.33 – 2.34

$$s(t) = s(0) + \vec{a}t \tag{2.33}$$

$$x(t) = x(0) + s(0)t + \frac{\vec{a}t^2}{2} \tag{2.34}$$

ถ้าความเร่งไม่คงที่ ดังนั้นการเคลื่อนที่ของวัตถุจะได้ตามสมการที่ 2.35 – 2.36

$$v\big(t + \Delta(t)\big) = v(t) + \overrightarrow{a(t)}\Delta(t) \tag{2.35}$$

$$x\big(t + \Delta(t)\big) = x(t) + v(t)\Delta(t) + \frac{\overrightarrow{a(t)}\Delta(t)^2}{2} \tag{2.36}$$

2.6.2 กฎแรงโน้มถ่วง

นิวตันได้ค้นพบธรรมชาติพื้นฐานของแรงดึงดูดโน้มถ่วงระหว่างวัตถุใดๆ สองวัตถุโดยนิวตันได้ตีพิมพ์กฎแรงโน้มถ่วง เอาไว้ดังนี้ "ทุกอนุภาคสสารในเอกภพดึงดูดทุกอนุภาคอื่นด้วยแรงซึ่งแปรผันตรงกับผลคูณของมวลของอนุภาคและแปรผกผันกับกำลังสองของระยะห่างระหว่างอนุภาคทั้งสองนั้น"

แรงที่จากวัตถุ x กระทำบนวัตถุ y แสดงได้ตามสมการที่ (2.37)

$$F(t) = \frac{Gm_x m_y}{d(x(t),y(t))^2} \tag{2.37}$$

เมื่อ m_x และm_y คือมวลของสองวัตถุ $d(x(t), y(t))$ คือระยะทางแบบ Euclidean ระหว่างสองวัตถุและ G คือค่าคงที่แรงโน้มถ่วงสากล $6.67 * 10^{-11}$

เมื่อสมการที่ 2.37 มาใหม่เป็นเวกเตอร์ที่มีทิศทางจะได้ตามสมการที่ 2.38

$$F(t) = \frac{Gm_x m_y}{\overrightarrow{d(t)}^2} \tag{2.38}$$

เมื่อ $\vec{d} = x(t) - y(t)$ จากสมการแรงโน้มถ่วงกล่าวมาข้างต้นเมื่อพิจารณาการเคลื่อนที่ของวัตถุ x ไปยังวัตถุ y สามารถนำไปแทนในสมการการเคลื่อนที่ได้ทำให้สามารถเขียนสมการที่ 2.35 กับ 2.36 ได้ใหม่ตามสมการที่ 2.39 กับ 2.40

$$v(t + \Delta(t)) = v(t) + \vec{d}\frac{Gm_x}{\|\vec{d}\|^3}\Delta(t) \tag{2.39}$$

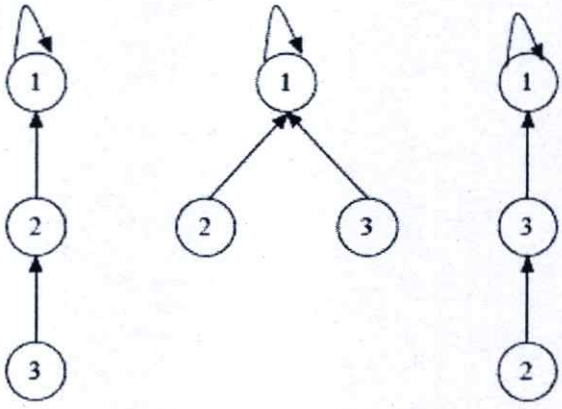
$$x(t + \Delta(t)) = x(t) + v(t)\Delta(t) + \overrightarrow{d(t)}\frac{Gm_x\Delta(t)^2}{2\|\overrightarrow{d(t)}\|^3} \tag{2.40}$$

2.6.3 โครงสร้าง Optimal Disjoint Set Union-Find

โครงสร้าง disjoint set Union-Find เป็นโครงสร้างซึ่งมีตัวปฏิบัติการ 3 ตัวด้วยกัน [14]

- MAKESET(x): สร้างเซตใหม่ประกอบไปด้วยเอลิเมนต์ x เดียว
- Union(x,y): แทนที่ของสองเซตที่ประกอบไปด้วย x และ y โดยใช้การยูเนียน
- Find(x): คืนค่าชื่อของเซตที่ประกอบด้วยเอลิเมนต์ x

ในโครงสร้าง optimal disjoint set Union-Find แต่ละเซตจะถูกแทนด้วยต้นไม้ที่มีรูทโหนดเป็น canonical element ของเซตและแต่ละโหนดลูกจะมีตัวชี้ไปชี้ที่โหนดพ่อ (รูทโหนดตัวชี้ชี้ตัวมันเอง) รูปที่ 2.12 แสดงรูปแบบต้นไม้ที่เป็นไปได้ทั้งหมดของเซต {1, 2, 3} โดยมี canonical element คือ 1

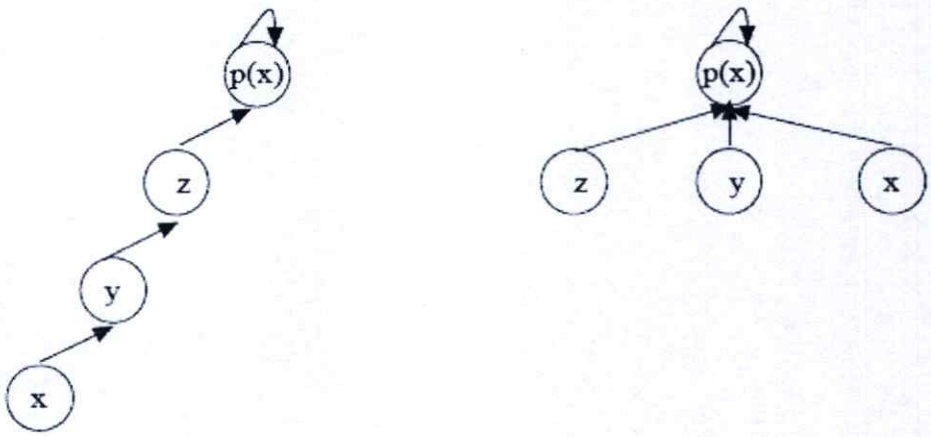


รูปที่ 2.12 แสดงการแทนค่าเซตของ canonical element และโครงสร้างต้นไม้

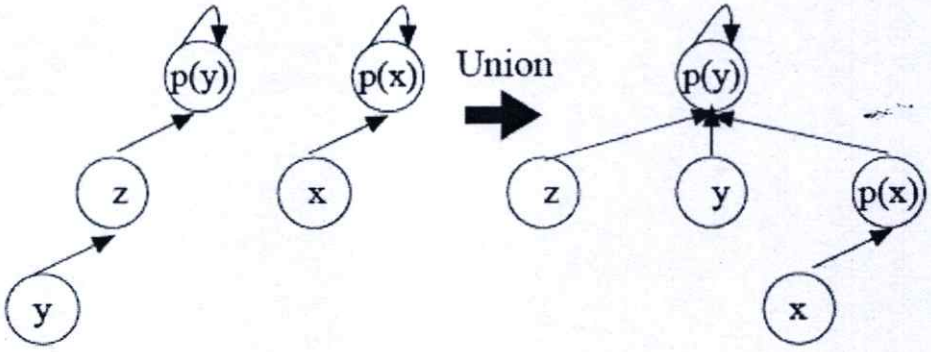
- The MAKESET(x) สร้างเซตของจุด x
- The FIND(x) คืนค่า canonical element ของเซตที่ประกอบไปด้วย x โดยท่องลงไปในต้นไม้โดยตัวชี้ที่ชี้ไปยังโหนดพ่อและทำการบีบอัดเส้นทางระหว่างโหนด x และ canonical element ของเซตโดยการกำหนดให้ canonical element เหมือนเป็นโหนดพ่อของแต่ละเอลิ

เมนต์ในเส้นทางจากไป canonical element รูปที่ 2.13 แสดงการบีบอัดเส้นทางโดยตัวปฏิบัติการ FIND

- The UNION(x,y) ตัวปฏิบัติการทำการบีบเส้นทางบน 2 ต้นไม้และจะทำการสร้าง canonical element ของเซตจาก y ไป canonical element ของ x ข้อมูลเกี่ยวกับเซตถูกเก็บใน canonical element ของเซตใหม่ รูปที่ 2.14 แสดงการทำงานตัวปฏิบัติการ UNION



รูปที่ 2.13 แสดงการบีบอัดเส้นทางของตัวปฏิบัติการ FINE



รูปที่ 2.14 แสดงตัวปฏิบัติการ Union

2.6.4 อัลกอริทึมการจัดแบ่งกลุ่มข้อมูลแบบใหม่ด้วยแรงโน้มถ่วง

การพัฒนาเทคนิคการทำคลัสเตอร์ที่ใช้แนวคิดกฎแรงโน้มถ่วงและกฎการเคลื่อนที่ ในวิธีนี้สามารถใช้กับเซตข้อมูลที่มี n มิติและมีจำนวน N จุดข้อมูลแต่ละจุดจะถูกพิจารณาเป็นวัตถุใน n มิติ และมีมวลเท่ากับ 1 แต่ละจุดในเซตข้อมูลจะเคลื่อนที่โดยใช้สมการที่ 2.40 แนวคิดการนำกฎแรงโน้มถ่วงมีดังนี้

1) จุดข้อมูลในบางคลัสเตอร์ที่ออกแรงกระทำกับจุดข้อมูลที่อยู่คลัสเตอร์เดียวกันจะมีแรงโน้มถ่วงกระทำสูงกว่าจุดข้อมูลที่ไม่อยู่คลัสเตอร์เดียวกัน ดังนั้นจุดในคลัสเตอร์จะเคลื่อนที่ไปในทิศทางของจุดศูนย์กลางของคลัสเตอร์ ในเทคนิคที่นำเสนอนี้จะระบุจำนวนคลัสเตอร์ในเซตข้อมูลเองอัตโนมัติ
2) ถ้าบางจุดเป็น noise จุดจะไม่อยู่ในคลัสเตอร์ได้ดังนั้นแรงโน้มถ่วงที่ทำกระจุดนั้นจะมีค่าน้อยทำให้จุดนั้นเกือบจะไม่เคลื่อนที่ ดังนั้นจุดที่เป็น noise จะไม่ถูกกำหนดให้เป็นคลัสเตอร์

สมการที่ 2.40 ใช้สำหรับการเคลื่อนที่จุดข้อมูลตามสนามแรงโน้มถ่วงที่จุดอื่นๆ สร้างขึ้น (y)

$$x(t+1) = x(t) + \vec{d}\frac{G}{\|\vec{d}\|^3} \tag{2.41}$$

เมื่อ $\vec{d} = \vec{y} - \vec{x}$ และ G เป็นค่าคงที่หารด้วย 2 ตามที่กำหนดไว้ในสมการที่ 2.40

ในงานวิจัยนี้พิจารณาความเร็วที่เวลาใด v(t) เป็นเวกเตอร์เท่ากับ 0 และ $\Delta(t) = 1$ด้วยวิธีนี้อัลกอริทึมไม่จำเป็นต้องใช้หน่วยจำในการเก็บความเร็วของแต่ละจุด ดังนั้นอัลกอริทึมจึงทำงานได้เร็วขึ้นเพราะจำนวนของโอเปอร์เรชั่นน้อยกว่าการใช้เวกเตอร์ความเร็ว

เนื่องจากค่าคงที่แรงโน้มถ่วงจะมีค่ามากในการคำนวณในช่วงรอบหลัง ทำให้ทุกจุดเคลื่อนที่เกือบเหมือนกันทั้งหมด ดังนั้นจึงให้แต่ละคลัสเตอร์มีค่าคงที่แรงโน้มถ่วงลดลงเรื่อยๆ เมื่อจำนวนรอบเพิ่มขึ้น

```
GRAVITATIONAL( x, G, Δ(G), M, ε)
1    for i=1 to N do
2      MAKE(i)
3    for i=1 to M do
4      for j=1 to N do
5        k = random point index such that k ≠ j
6        MOVE( x_j , x_k )  (see Eq (2.22))
7        if dist^2( x_j , x_k ) ≤ε  then UNION( j, k)
8     G = (1-Δ(G))*G
9    for i=1 to N do
10     FIND(i)
11   return disjoint-sets
```

รูปที่ 2.15 แสดงอัลกอริทึมการจัดแบ่งกลุ่มข้อมูลด้วยแรงโน้มถ่วง

แต่ละรอบอัลกอริทึมจะสร้างคลัสเตอร์โดยใช้โครงสร้าง Optimal disjoint set union-find และระยะทางระหว่างวัตถุ (หลังจากการเรียนรู้โมเดลที่ใช้การเคลื่อนของวัตถุโดยประยุกต์ใช้แรงโน้มถ่วง) เมื่อสองจุดรวมกัน ทั้งสองจะถูกเก็บไว้ในระบบขณะที่ความสัมพันธ์โครงสร้างของเซตถูกปรับเปลี่ยน ในการกำหนดตำแหน่งใหม่ของแต่ละจุดอัลกอริทึมในงานวิจัยนี้จะเลือกจุดข้อมูลใหม่โดยใช้วิธีการสุ่มและเคลื่อนที่โดยใช้สมการที่ 2.41 (ฟังก์ชัน MOVE) อัลกอริทึมจะคืนค่าของเซตที่ถูกเก็บไว้ในโครงสร้าง disjoint set union-find ซึ่งสามารถที่จะเขียนโปรแกรมสร้างโครงสร้าง disjoint set union-find โดยใช้อาเรย์ชนิดข้อมูลจำนวนเต็มขนาด N เมื่อตำแหน่ง i เก็บดัชนีของ canonical object ของเซตที่บรรจุจุดข้อมูล i

เพราะอัลกอริทึมที่กล่าวมาข้างต้นกำหนดให้ทุกจุดในเซตข้อมูลเป็นสมาชิกได้เพียงหนึ่งคลัสเตอร์ มันจำเป็นจะต้องพิจารณาคลัสเตอร์นั้นมาความถูกต้องหรือไม่ โดยพิจารณาจากจำนวนสมาชิกในคลัสเตอร์ว่ามีจำนวนมากว่าเกณฑ์หรือไม่ (ในผลลัพธ์ที่ได้มีคลัสเตอร์ที่เป็น noisy รวมกับคลัสเตอร์ที่ปกติ) ในงานวิจัยนี้เลยใช้พารามิเตอร์การพิจารณาคือ α กำหนดจำนวนต่ำสุดของสมาชิกในแต่ละคลัสเตอร์ (เปอร์เซ็นของชุดข้อมูลที่ใช้เรียนรู้) ในงานวิจัยนี้ได้เพิ่มฟังก์ชัน GETCLUSTERS ซึ่งจะดึง disjoint set ที่ได้จากอัลกอริทึมการจัดแบ่งกลุ่มด้วยแรงโน้มถ่วงและคืนค่าชุดของคลัสเตอร์ที่มีจำนวนสมาชิกที่ผ่านเกณฑ์

```
GetClusters( clusters, alpha, data )
1  newClusters = ∅
2  MIN_POINTS = αN
3  for i=0 to number of clusters do
4    if size( cluster_i ) ≥MIN_POINTS then
5      newClusters = newClusters ∪ { cluster_i }
6  return newClusters
```

รูปที่ 2.16 แสดงอัลกอริทึมลบคลัสเตอร์ที่เป็น noisy ออก

บทที่ 3

การนำฟังก์ชันความหนาแน่นร่วมกับอัลกอริทึมการค้นหาคำตอบด้วยแรงโน้มถ่วงเพื่อแก้ปัญหาการจัดแบ่งกลุ่มข้อมูล

3.1 บทนำ

แรงโน้มถ่วงเป็นหนึ่งในอันตรกิริยาพื้นฐานสี่อย่างที่พบในธรรมชาติและเป็นอันตรกิริยาแรกสุดที่ได้มีการศึกษาอย่างกว้างขวาง นิวตันค้นหาพบในคริสต์สตวรรษที่สิบเจ็ดว่าอันตรกิริยาที่ทำให้ลูกแอปเปิลตกจากต้นไม้เป็นอันตรกิริยาเดียวกันกับที่ทำให้ดาวเคราะห์โคจรรอบดวงอาทิตย์ด้วย ซึ่งเป็นการศึกษาพลศาสตร์ของวัตถุในอวกาศ แต่ตัวอย่างแรงดึงดูดโน้มถ่วงที่คุณน่าจะคุ้นเคยที่สุด คือ น้ำหนักของคุณซึ่งดึงดูดคุณเข้าหาโลก ในระหว่างการศึกษาการเคลื่อนที่ของดาวเคราะห์และดวงจันทร์ นิวตันได้ค้นพบธรรมชาติพื้นฐานของแรงดึงดูดโน้มถ่วงระหว่างวัตถุใดๆ สองวัตถุโดย นิวตันตีพิมพ์กฎแรงโน้มถ่วงพร้อมกับการเคลื่อนที่สามข้อของเขาในปี ค.ศ.1687 เอาไว้ดังนี้ “ทุกอนุภาคสสารในเอกภพดึงดูดทุกอนุภาคอื่นด้วยแรงซึ่งแปรผันตรงกับผลคูณของมวลของอนุภาคและแปรผกผันกับกำลังสองของระยะห่างระหว่างอนุภาคทั้งสองนั้น” เมื่อแปลข้อความนี้เป็นสมการจะได้ตาม 3.1 [7]

$$F = G\frac{M_1 M_2}{R^2} \tag{3.1}$$

F คือขนาดของแรงโน้มถ่วงซึ่งทำต่ออนุภาคใดอนุภาคหนึ่ง M_1 และ M_2 เป็นมวลของอนุภาคทั้งสอง R คือระยะห่างระหว่างอนุภาคและ G คือค่าคงตัวฟิสิกส์พื้นฐานที่เรียกว่าค่าคงตัวโน้มถ่วง ในทางฟิสิกส์แล้วค่าคงตัวแรงโน้มถ่วงนั้นขึ้นอยู่กับอายุขัยที่แท้จริงของจักรวาลตามสมการที่ 3.2 [7] การลดลงของค่าคงที่แรงโน้มถ่วง G กับอายุ

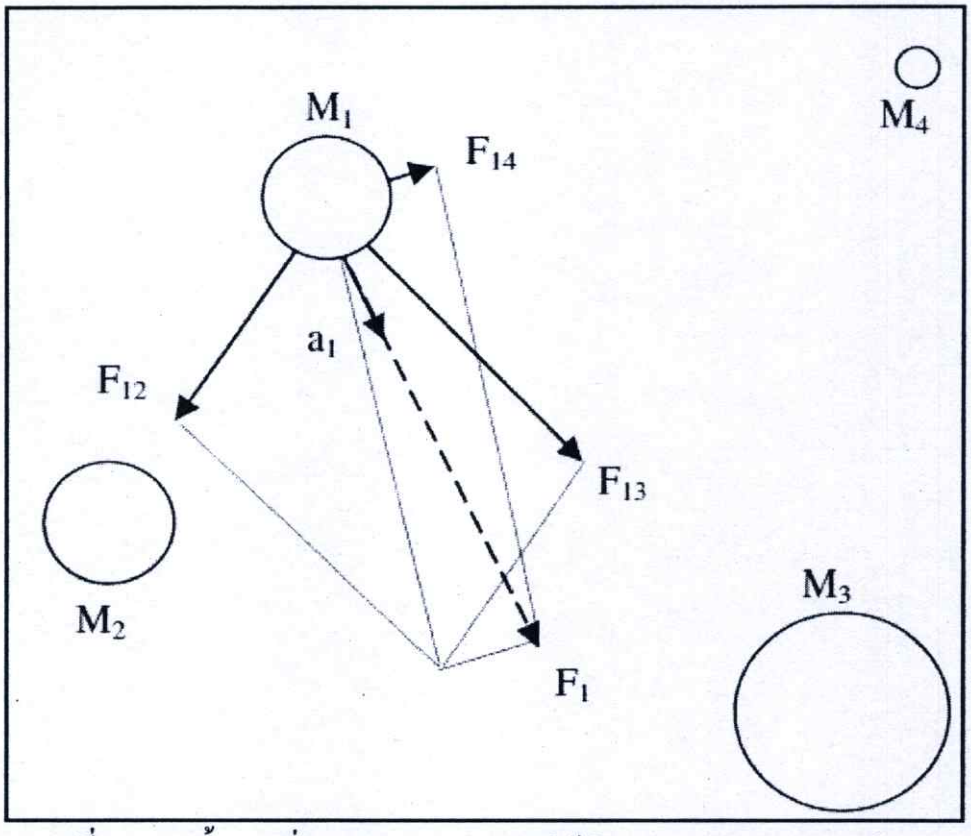
$$G(t) = G(t_0) \times \left(\frac{t_0}{t}\right)^{\beta},\ \beta < 1 \tag{3.2}$$

โดย $G(t_0)$ คือ ค่าคงที่แรงโน้มถ่วง cosmic quantum-interval ช่วงแรกๆ ของเวลาที่ t_0 นอกจากนี้นิวตันยังค้นพบกฎการเคลื่อนที่สามข้อ ซึ่งในกฎการเคลื่อนที่ข้อที่สองของนิวตัน

สามารถนำมาใช้รวมกับกฎแรงโน้มถ่วงได้ กฎการเคลื่อนที่ข้อที่สองได้กล่าวไว้ว่า "เมื่อมีแรงลัพธ์ที่มีค่าไม่เท่ากับศูนย์มากระทำต่อวัตถุ จะทำให้วัตถุเกิดความเร่งในทิศเดียวกับแรงลัพธ์ที่มากระทำ และขนาดของความเร่งนี้จะแปรผันตรงกับขนาดของแรงลัพธ์ และแปรผกผันกับมวลของวัตถุ" ข้อความนี้เป็นสมการตาม 3.3 [7]

$$a = \frac{F}{M} \tag{3.3}$$

จากสมการ 3.1 และ 3.3 แรงโน้มถ่วงมากขึ้นเมื่ออนุภาคมีขนาดใหญ่ขึ้นและมีระยะทางที่ใกล้กันมากขึ้น การเพิ่มระยะทางระหว่าง 2 อนุภาคหมายถึงการลดแรงโน้มถ่วง เมื่อในเอกภพมีอนุภาคหลายอนุภาคเช่นตัวอย่างในรูปที่ 3.1 ในรูปที่ 3.1 นี้ F_{1j} เป็นแรงซึ่งกระทำบน M_1 และ M_j และ F_1



รูปที่ 3.1 แรงทั้งหมดที่กระทำบนมวล M_1 ทำให้เกิดแรงรวม F_1 และความเร่ง a_1

มวลแบ่งได้ออกเป็น 3 ชนิดด้วยกันในทางฟิสิกส์:

Active Gravitational Mass M_a เป็นตัวบอกถึงอิทธิพลของแรงโน้มถ่วงเนื่องจากวัตถุนั้น วัตถุที่มีมวลมากกว่าก็ย่อมที่จะมีสนามโน้มถ่วงที่แรงกว่าวัตถุมวลน้อย

Passive Gravitational Mass M_p เป็นปริมาณที่บ่งบอกความแรงของอันตรกิริยาระหว่างวัตถุกับสนามโน้มถ่วงสำหรับสนามโน้มถ่วงที่มีความเข้มเท่ากันวัตถุที่มีมวลมากก็จะรู้สึกถึงแรงโน้มถ่วงมากกว่าวัตถุที่มีมวลน้อย แรงโน้มถ่วงที่มาทำกับวัตถุนี้เราเรียกกันว่า น้ำหนักนั่นเอง

Inertial Mass M_i เป็นปริมาณบอกถึงสภาพการต่อต้านการเคลื่อนที่ของวัตถุเมื่อมีแรงมากระทำ วัตถุขนาดใหญ่ก็จะมีสภาพการต่อต้านการเคลื่อนที่ของวัตถุมากและวัตถุขนาดเล็กจะมีสภาพการต่อต้านการเคลื่อนที่ของวัตถุน้อย

พิจารณาสิ่งที่กล่าวมาข้างต้น เราสามารถเขียน กฎของนิวตันใหม่ได้ แรงโน้มถ่วง F_{ij} ซึ่งกระทำบนมวล i โดย มวล j เป็นผลคูณ ของ Active Gravitational Mass ของมวล j และ Passive Gravitational Mass ของมวล i และหารด้วยระยะทางของพวกมันยกกำลัง a_i เป็นสัดส่วนของ F_{ij} และหารด้วย Inertial Mass ของ i จะสามารถเขียนสมการ 3.1 และ 3.3 ใหม่ได้ดังนี้

$$F_{ij} = G \frac{M_{aj} M_{pi}}{R^2} \tag{3.4}$$

$$a_i = \frac{F_{ij}}{M_{ii}} \tag{3.5}$$

ขั้นตอนที่นำเสนอเป็นการจัดกลุ่มของข้อมูลที่ใช้หลักการของกฎแรงโน้มถ่วงและกฎการเคลื่อนที่ของนิวตัน โดยแนวคิดคือว่า ทุกอนุภาคในจักรวาลดึงดูดอนุภาคอื่นๆทุกอนุภาค เข้าหากัน บริเวณที่มีอนุภาครวมกันอยู่สูงจะมีแรงดึงดูดมากกว่าบริเวณที่มีอนุภาครวมกันอยู่น้อย อนุภาคที่กระจายอยู่ในจักรวาลจะถูกดึงดูดไปบริเวณที่มีอนุภาครวมกันอยู่หนาแน่น ทำให้เกิดการเคลื่อนของอนุภาคที่กระจายอยู่ เราจะใช้การเคลื่อนนี้ในการหาจุดที่มีความหนาแน่นสูงของข้อมูล จุดที่มีข้อมูลหนาแน่นสูงเราจะถือว่าเป็น จุดศูนย์กลางของคลัสเตอร์

3.2 โครงสร้างการทำงาน

3.2.1 คำนวณค่าแบนวิดธ์ ดังรายละเอียดที่แสดงในหัวข้อ 3.3.1

3.2.2 สุ่มตำแหน่งจุดศูนย์กลางเริ่มต้นจากเซตข้อมูล ดังรายละเอียดที่แสดงในหัวข้อ 3.3.2

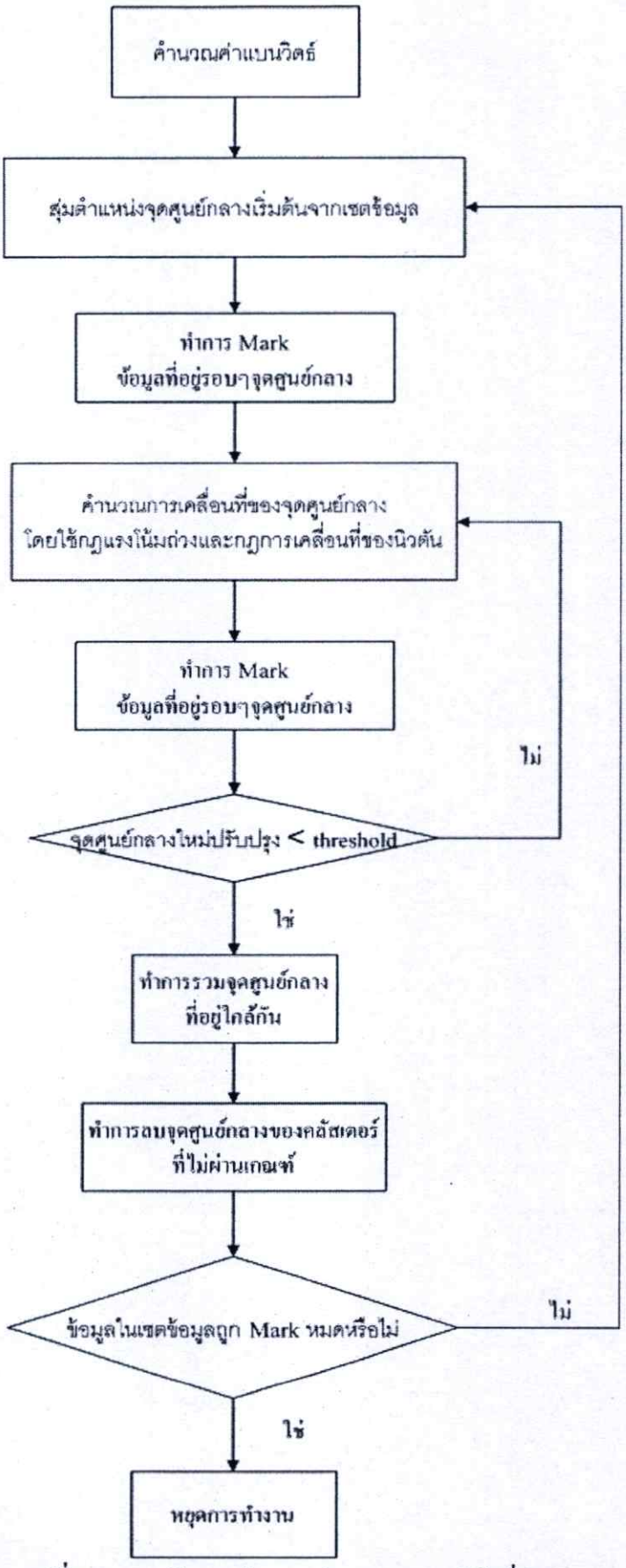
3.2.3 คำนวณการเคลื่อนที่ของจุดศูนย์กลางที่ถูกเลือกโดยใช้กฎแรงโน้มถ่วงและกฎการเคลื่อนที่ของนิวตัน ดังรายละเอียดที่แสดงในหัวข้อ 3.3.3

3.2.4 พิจารณาการหยุดการเคลื่อนของจุดศูนย์กลางและทำการบันทึกตำแหน่งจุดศูนย์กลาง ดังรายละเอียดที่แสดงในหัวข้อ 3.3.4

3.2.5 ทำการรวมตำแหน่งจุดศูนย์กลางที่อยู่ใกล้กัน ดังรายละเอียดที่แสดงในหัวข้อ 3.3.5

3.2.6 ทำการลบจุดศูนย์กลางของคลัสเตอร์ โดยดูว่าคลัสเตอร์ที่มีสมาชิกน้อยก็จะถูกลบออกไป ดังรายละเอียดที่แสดงในหัวข้อ 3.3.6

3.2.7 ถ้าข้อมูลในเซตข้อมูลยังถูก mark ไม่หมดให้กลับไปทำขั้นตอนที่ 3.2.2 ถ้าไม่ก็หยุดการทำงาน ดังรายละเอียดที่แสดงในหัวข้อ 3.3.7



รูปที่ 3.2 แสดงโครงสร้างการทำงานของงานวิจัยที่นำเสนอ

3.3 ขั้นตอนการเรียนรู้โมเดล

ทำการกำหนดค่าเริ่มต้นของตัวแปรต่างๆ ดังนี้

- ค่าที่กำหนดให้จุดศูนย์กลางหยุดการเคลื่อนที่ (threshold)
- ค่าเกณฑ์ในการลบคลัสเตอร์ (γ)

ขั้นตอนการเรียนรู้ของโมเดลมีขั้นตอนการทำงานดังนี้

3.3.1 ขั้นตอนคำนวณค่าแบนวิดธ์

ค่าแบนวิดธ์จะเป็นตัวกำหนดจำนวนคลัสเตอร์ ถ้าค่าแบนวิดธ์ มีความเหมาะสมก็จะทำให้จำนวนคลัสเตอร์มีความเหมาะสมกับชุดข้อมูลได้เช่นกัน จากการศึกษาของ Silverman (1986) [8] พบว่าค่าแบนวิดธ์ ที่ดีสำหรับการประมาณฟังก์ชันความหนาแน่นของความน่าจะเป็นแบบเคอร์เนลที่คือแบนวิดธ์ที่มีค่าเท่ากับตามสมการที่ 3.6

$$S = \frac{0.9A}{\sqrt[5]{N}} \tag{3.6}$$

$$A = \min\left[\sigma, \frac{IQR}{1.34}\right] \tag{3.7}$$

โดย S คือ ค่าแบนวิดธ์

N คือ จำนวนของเซตข้อมูลทั้งหมด

σ คือ ส่วนเบี่ยงเบนมาตรฐานในเซตข้อมูล สามารถคำนวณตามสมการที่ 3.8

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} \|X_i - \bar{X}\|}{N}} \tag{3.8}$$

IQR คือ ค่าพิสัยระหว่างควอไทล์ของเซตข้อมูล (Interquartile range = $Q_3 - Q_1$)

$\bar{X}$ คือ ค่าเฉลี่ยของเซตข้อมูลทั้งหมด

3.3.2 ขั้นตอนการเลือกจุดศูนย์กลางเริ่มต้นจากเซตข้อมูล

3.3.2.1 ขั้นตอนการเลือกจุดศูนย์กลางเริ่มต้น เราจะสร้างตัวแปรที่ชื่อว่า mark มีขนาดเท่ากับจำนวนของเซตข้อมูลทั้งหมด โดยตอนเริ่มต้น mark ทุกตัวจะมีค่าเท่ากับ 0

3.3.2.2 เราสุ่มข้อมูลจากเซตข้อมูลเพื่อนำมาเป็นจุดศูนย์กลางเริ่มต้น และเซตค่า mark ของข้อมูลที่สุ่มมานั้นให้เท่ากับ 1 เรากำหนดตำแหน่งของจุดศูนย์กลางตามสมการ 3.9

$$C = (C^1, \ldots, C^d, \ldots, C^D) \tag{3.9}$$

โดย D คือ จำนวนมิติทั้งหมดของข้อมูล

C^d แทนตำแหน่งของจุดศูนย์กลางในมิติที่ d

3.3.2.3 เราก็จะกำหนดให้ตำแหน่งของ mark ที่ดัชนีเดียวกับเซตข้อมูลที่ถูกเลือกมีค่าเท่ากับ 1

3.3.3 ขั้นตอนคำนวณการเคลื่อนที่ของจุดศูนย์กลาง

การเคลื่อนที่ของจุดกลางศูนย์กลางจะถูกคำนวณด้วยกฎแรงโน้มถ่วงและกฎการเคลื่อนที่สองของนิวตัน จุดศูนย์กลางจะคำนวณแรงโดยใช้กฎแรงโน้มถ่วงกับข้อมูลที่อยู่รอบตัวในระยะรัศมีค่าแบนวิดธ์ จากนั้นนำผลรวมของแรงที่คำนวณได้ไปคำนวณหาตำแหน่งถัดไปที่จุดศูนย์กลางนั้นเคลื่อนที่โดยใช้กฎการเคลื่อนที่ของนิวตัน

3.3.3.1 เริ่มต้นให้ทำการเซตค่า mark ของข้อมูลที่อยู่รอบตัวจุดศูนย์กลางในระยะรัศมีค่าแบนวิดธ์ให้เท่ากับ 1 ตามเงื่อนไขในสมการ 3.10 โดยให้กำหนดตำแหน่งของข้อมูลที่อยู่รอบตัวจุดศูนย์กลางในระยะรัศมีแบนวิดธ์ตามสมการ 3.11

$$\|X_j - C\| < S \tag{3.10}$$

$$X_j = (X_j^1, \ldots, X_j^d, \ldots, X_j^D) \tag{3.11}$$

โดย $j = 1, 2, \ldots, P$

P คือ จำนวนของข้อมูลที่อยู่รอบๆจุดศูนย์กลางในระยะรัศมีแบนวิดธ์

X_j^d คือ แทนตำแหน่งของข้อมูลที่อยู่รอบจุดศูนย์กลางในระยะรัศมีแบนวิดธ์ลำดับที่ j ในมิติที่ d

3.3.3.2 ทำการคำนวณแรงทั้งหมดกระทำกับจุดศูนย์กลางโดยใช้กฎแรงโน้มถ่วงของนิวตันตามสมการ 3.12 [9]

$$F_C^d(t) = \sum_{j=1}^{P} F_{cj}^d(t) \tag{3.12}$$

โดย t คือ จำนวนรอบที่จุดศูนย์กลางเคลื่อนที่

F_{Cj}^d คือ แรงที่ข้อมูลลำดับที่ j กระทำกับจุดศูนย์กลาง C สามารถคำนวณได้ตามสมการ 3.13 [9]

$$F_{Cj}^d(t) = G(t)\frac{M_{pc}(t) \times M_{aj}(t)}{R_{Cj}(t)+\varepsilon}(X_j^d(t) - C^d(t)) \tag{3.13}$$

โดย $M_{aj}(t)$ คือ Active Gravitational Mass ที่สัมพันธ์กับข้อมูลลำดับที่ j

$M_{pc}(t)$ คือ Passive Gravitational Mass ที่สัมพันธ์กับจุดศูนย์กลาง C

ในการคำนวณมวลสามารถคำนวณได้โดยค่าความหนาแน่นของข้อมูล จุดที่มีมวลมากหมายถึงจุดที่มีความหนาแน่นของข้อมูลมากและเป็นจุดที่มีแรงดึงดูดมาก วัตถุจะเคลื่อนที่ออกจากจุดนั้นได้ยาก เราเลยกำหนดให้จุดที่มีความหนาแน่นสูงหรือมวลมากเป็นจุดที่จะเป็นจุดศูนย์กลางของคลัสเตอร์ เรากำหนดให้มวลสามารถคำนวณได้ตามสมการ 3.14 – 3.19

$$M_{a\bullet} = M_{p\bullet} = M_{c\bullet} = M_\bullet \tag{3.14}$$

$$M_\bullet(t) = \frac{m_\bullet(t)}{\sum_{q=1}^{P} m_q(t)} \tag{3.15}$$

$$m_\bullet(t) = \frac{density_\bullet(t) - min\_density}{max\_density - min\_density} \tag{3.16}$$

$$min\_density = \min_{q \in \{1,2,\dots,P\}} density_q(t) \tag{3.17}$$

$$max\_density = \max_{q \in \{1,2,\dots,P\}} density_q(t) \tag{3.18}$$

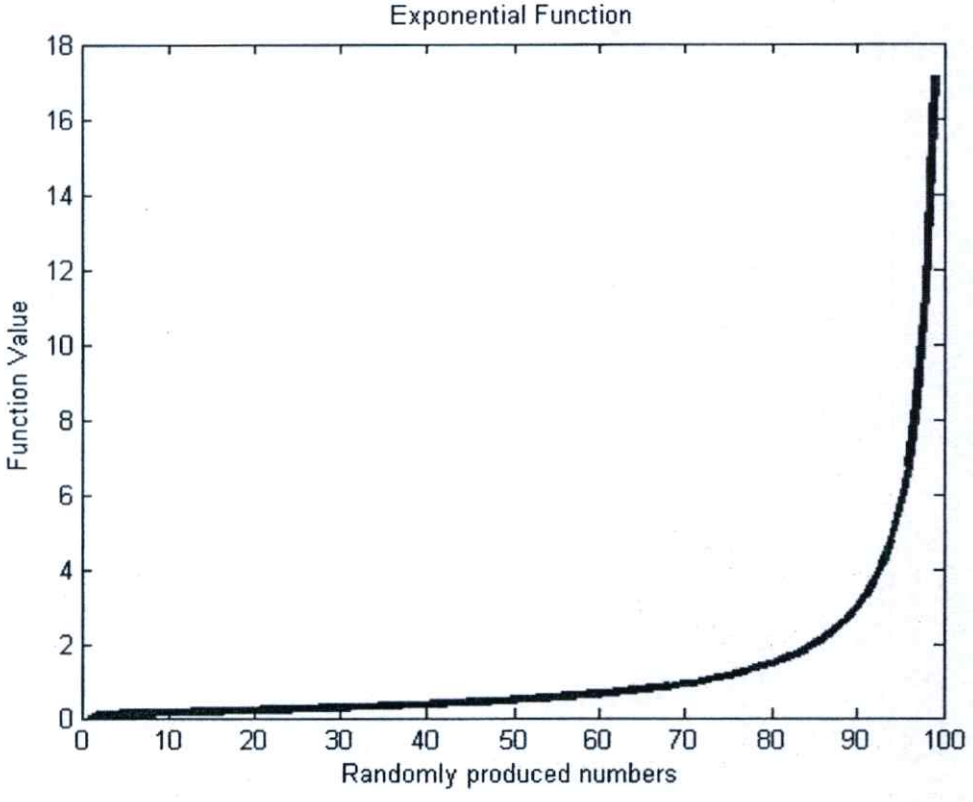
โดย $density_c(t)$ คือ ค่าความหนาแน่นรอบๆจุดศูนย์กลาง C สามารถคำนวณได้ตามสมการ 3.19

$$density_c(t) = \left[-\frac{1}{2 \times \log_2\left(\frac{P}{N}\right) - \varepsilon}\right] \times \frac{1}{P}\left(\sum_{j=1}^{P} e^{-\frac{\|x_j(t) - C(t)\|^2}{S^2}}\right) \tag{3.19}$$

จากสมการ 3.14 – 3.18 เป็นการหาค่าความน่าจะเป็นความหนาแน่นของจุดใดที่อยู่รอบๆ จุดศูนย์กลาง ในช่วงของค่าความหนาแน่ระหว่างค่าความหนาแน่ที่น้อยที่สุดกับค่าความหนาแน่มากที่สุดของจุดที่อยู่รอบจุดศูนย์กลาง

จากสมการที่ 3.19 สามารถอธิบายของแต่ละเทอมได้ดังนี้

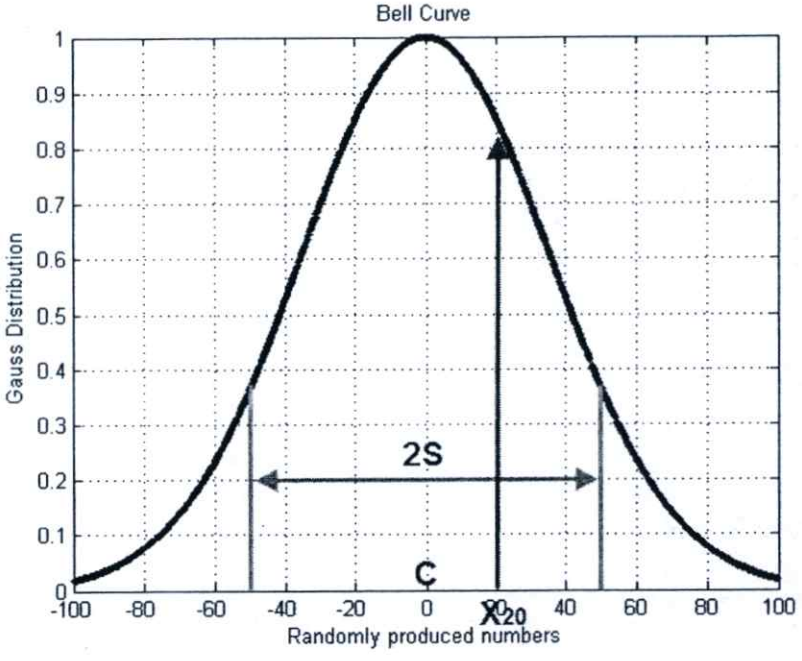
ก. เทอม $\left[-\frac{1}{2\times\log_2\left(\frac{P}{N}\right)-\varepsilon}\right]$ คือค่าอัตราส่วนจำนวนข้อมูลที่อยู่รอบ C ในรัศมีค่าแบนวิดธ์เมื่อเปรียบเทียบกับจำนวนข้อมูลของเซตข้อมูลทั้งหมด ค่าในเทอมนี้บ่งบอกถึงปริมาณต่อพื้นที่โดยถ้า P มีค่าเพิ่มขึ้นจะทำให้ค่าในเทอมนี้มีค่าเพิ่มขึ้นในลักษณะเอ็กโปแนนเชียลและP มีค่ามากสุดได้เท่ากับ N นั้น ถ้าในเทอมนี้มากก็แสดงว่าบริเวณรอบจุดศูนย์นั้นมีความหนาแน่นสูง



รูปที่ 3.3 แสดงกราฟของสมการ 3.19 ในส่วนของเทอมแรกเมื่อ P เพิ่มขึ้นอย่างต่อเนื่อง

จากรูปที่ 3.3 เมื่อให้ N มีค่าเท่ากับ 100 และ P เริ่มต้นเท่ากับ 0 แล้วเพิ่มขึ้นอย่างต่อเนื่องไปถึง 99 ถ้า P เท่ากับ 100 จะทำให้ค่าในลอการิทึมเท่ากับ 0 ส่งผลให้ค่าส่วนมีค่าน้อยมากตามค่า ε ฟังก์ชันจึงมีค่าสูงมาก

ข. เทอม $\frac{1}{P}\left(\sum_{j=1}^{P} e^{-\frac{\|X_j(t)-C(t)\|^2}{S^2}}\right)$ คือ ค่าเฉลี่ยของฟังก์ชันเกาส์เซียนที่ใช้วัดความห่างของข้อมูลที่อยู่รอบๆ C ถ้ามีค่ามากแสดงว่าข้อมูลที่อยู่รอบ C อยู่ใกล้จุด C และกระจายตัวน้อยรวมกันอยู่หนาแน่น ถ้ามีค่าน้อยแสดงว่าข้อมูลที่อยู่รอบ C อยู่ห่างจากจุด C และกระจายตัวมาก

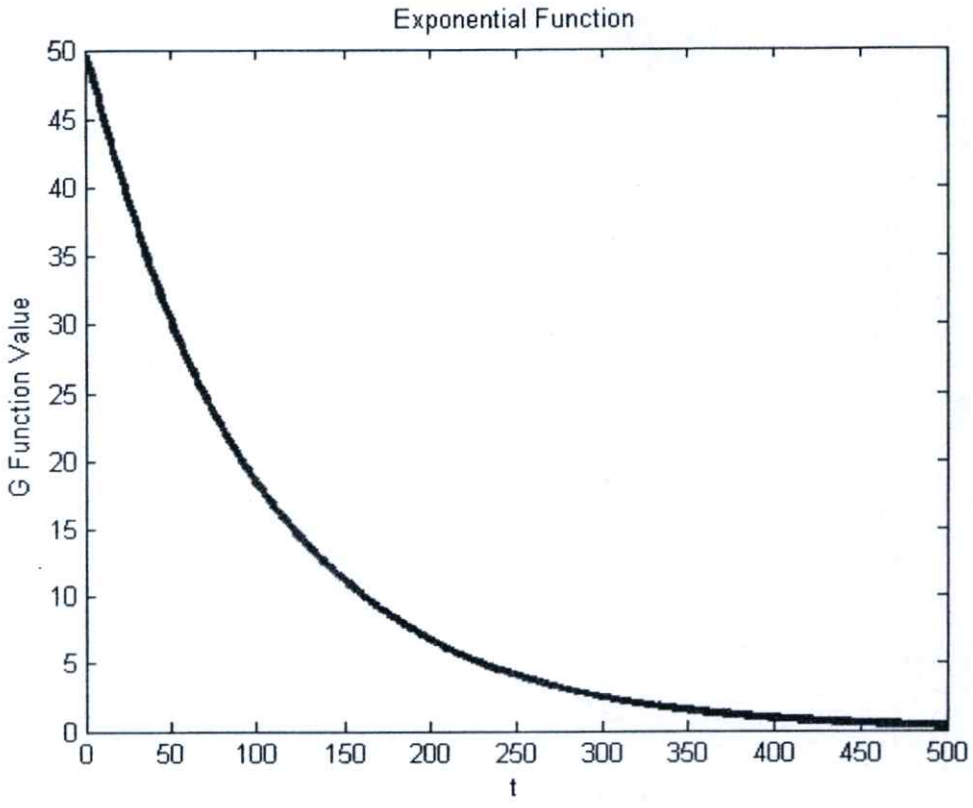


รูปที่ 3.4 แสดงกราฟของสมการ 3.19 ในส่วนของเทอมสอง

จากรูปที่ 3.4 S มีค่าเท่ากับ 50 และมีจุดศูนย์กลาง C อยู่ที่ตำแหน่ง 0 ซึ่งมาค่าใกล้ชิดสูงสุดคือ 1 และต่ำสุดประมาณ 0.36 ซึ่งทำให้ค่า X ตำแหน่งที่ 20 มีค่าประมาณ 0.8 เมื่อผ่านฟังก์ชันนี้

G(t) คือ ค่าคงที่แรงโน้มถ่วง ในทางฟิสิกส์แล้วค่าคงที่แรงโน้มถ่วงนั้นขึ้นอยู่กับอายุขัยที่แท้จริงของจักรวาลตามสมการที่ 3.2 ใน GCA สามารถคำนวณค่าคงที่แรงโน้มถ่วงได้ในสมการ 3.20 ซึ่งเราลอกเลียนพฤติกรรมค่าคงที่ในทางฟิสิกส์เหมือนกัน จะใช้ฟังก์ชันลดแบบเอ็กโปเนนเชียลโดยจะมีค่าลดลงเรื่อยๆเมื่อรอบเพิ่มขึ้น G(t) จะควบคุมให้จุดศูนย์กลางเคลื่อนที่ลู่สู่บริเวณที่หนาแน่นภายท้องถิ่น (Local) อย่างเหมาะสม และยังควบคุมให้จุดศนย์กลางเคลื่อนที่ในระยะไม่เกินขอบเขตของแบนวิดธ์

$$G(t) = S \times e^{-\frac{t}{100}} \qquad (3.20)$$



รูปที่ 3.5 แสดงกราฟของสมการ 3.20 เมื่อ t มีค่าเพิ่มขึ้นอย่างต่อเนื่อง

จากรูปที่ 3.5 S มีค่าเท่ากับ 50 และ t มีค่าเริ่มต้นเท่ากับ 1 แล้วเพิ่มขึ้นอย่างต่อเนื่องไปถึงจำนวน 500 ส่งทำให้ค่า G มีค่าลดลงเรื่อยๆเมื่อ t เพิ่มขึ้น

$R_{Cj}(t)$ คือ ค่าระยะทางแบบยูคลิดเดียน (Euclidean Distance) ระหว่างจุดศูนย์กลางกับข้อมูลลำดับที่ j ตามสมการ 3.21

$$R_{Cj}(t) = \|X_j(t) - C(t)\| \tag{3.21}$$

3.3.3.3 คำนวณอัตราเร่งของจุดศูนย์กลาง C ในมิติที่ d จากผลรวมของแรงโดยใช้กฎการเคลื่อนที่ข้อที่ 2 ของนิวตัน ตามสมการที่ 3.22

$$a_c^d(t) = \frac{F_c^d(t)}{M_{cc}} \tag{3.22}$$

โดย M_{cc} คือ มวลเฉื่อยของจุดศูนย์กลาง C สามารถคำนวณได้ตามสมการ 3.14 – 3.19

3.3.3.4 คำนวณความเร็วของจุดศูนย์กลาง C ในมิติที่ d ตามสมการที่ 3.23

$$V_C^d(t+1) = rand_C \times V_C^d(t) + \beta_C a_c^d(t) \quad (3.23)$$

โดย $rand_C$ คือ จำนวนสุ่มที่อยู่ในช่วง [0,1]

β_C คือ ค่าตัวแปรควบคุมอัตราเร่ง จะใช้สมการเป็นฟังก์ชันลดแบบเอ็กโปเนนเชียลตามสมการ 3.25 เมื่อตอนเริ่มต้นที่มีค่าเท่ากับ 1 จะลดลงเรื่อยเมื่อตำแหน่งจุดศูนย์กลาง C เข้าใกล้ตำแหน่งจุดศูนย์กลางของคลัสเตอร์ที่ถูกบันทึกไว้ โดยใช้เงื่อนไขตามสมการ 3.24

$$\|C_i - C\| < (2 \times S) \quad (3.24)$$

โดย $i = 1, 2, \ldots, K$

K คือ จำนวนของคลัสเตอร์ทั้งหมด

$$\beta_C = e^{-\frac{increa}{10}} \quad (3.25)$$

โดย increa คือ จำนวนที่ตอนเริ่มต้นเท่ากับ 0 แล้วจะเพิ่มขึ้นเรื่อยเมื่อเงื่อนไขในสมการที่ 3.24 เป็นจริง

β_C จะควบคุมไม่ให้จุดศูนย์กลางวิ่งไปยังตำแหน่งจุดศูนย์กลางของคลัสเตอร์ที่ถูกบันทึกไว้ เพื่อไม่ให้เกิดการซับซ้อนและรวมตัวกันของจุดศูนย์กลาง ซึ่งจะทำให้จำนวนของคลัสเตอร์มีน้อยเกินไปจนทำให้เกิดความผิดพลาดในการทำนาย ดังนั้นจึงปรับอัตราเร่งเมื่อจุดศูนย์กลางที่กำลังเคลื่อนที่เดินเข้าหาจุดศูนย์กลางของคลัสเตอร์ที่ถูกบันทึกไว้

3.3.3.5 ปรับปรุงตำแหน่งของจุดศูนย์กลาง C ในมิติที่ d ตามสมการที่ 3.26

$$C^d(t+1) = C^d(t) + V_C^d(t+1) \quad (3.26)$$

3.3.4 ขั้นตอนการบันทึกตำแหน่งจุดศูนย์กลางใหม่

พิจารณาการหยุดการเคลื่อนของจุดศูนย์กลางตามสมการ 3.27

$$\|C(t+1) - C(t)\| < threshold \quad (3.27)$$

ถ้าค่าตำแหน่งของจุดศูนย์กลางรอบปัจจุบันกับค่าตำแหน่งของจุดศูนย์กลางรอบก่อนหน้าวัดระยะแบบยูคลิดเดียน (Euclidean Distance) แล้วมากกว่าค่า threshold ให้กลับไปทำในขั้นตอนที่

3.3.3 แต่ถ้าค่าตำแหน่งของจุดศูนย์กลางรอบปัจจุบันกับค่าตำแหน่งของจุดศูนย์กลางรอบก่อนหน้าวัดระยะแบบยูคลิดเดียน (Euclidean Distance) แล้วน้อยกว่าค่า threshold ให้ทำการบันทึกค่าตำแหน่งของจุดศูนย์กลางรอบปัจจุบันและทำขั้นตอนที่ 3.3.5

3.3.5 ขั้นตอนการรวมตำแหน่งจุดศูนย์กลาง

เรากำหนดตำแหน่งจุดศูนย์กลางของคลัสเตอร์ที่ถูกบันทึกไว้ก่อนหน้าตามสมการที่ 3.28

$$C_i = (C_i^d, \dots, C_i^d, \dots, C_i^d) \tag{3.28}$$

โดย i = 1, 2, …, K-1

K คือ จำนวนของคลัสเตอร์ทั้งหมด

เรากำหนดตำแหน่งจุดศูนย์กลางของคลัสเตอร์ที่ถูกบันทึกล่าสุดจากขั้นตอนที่ 3.3.4 ตามสมการที่ 3.29

$$C_K = (C_K^d, \dots, C_K^d, \dots, C_K^d) \tag{3.29}$$

3.3.5.1 รับค่าตำแหน่งจุดศูนย์กลางที่ถูกบันทึกไว้ก่อนหน้ามาทีละ 1 เพื่อนำมาพิจารณาตามเงื่อนไขของจุดศูนย์กลางของคลัสเตอร์ที่ i สมการที่ 3.31

$$\|C_i - C_K\| < \frac{S}{2} \tag{3.31}$$

ถ้าค่าตำแหน่งของจุดศูนย์กลางดัชนีที่ i กับค่าตำแหน่งของจุดศูนย์กลางถูกบันทึกล่าสุด วัดระยะแบบยูคลิดเดียน (Euclidean Distance) แล้วน้อยกว่าค่าแบนวิดธ์หารสอง แสดงว่าค่าตำแหน่งจุดศูนย์กลางทั้งสองใกล้ชิดก็จะทำการรวมกันในขั้นตอนที่ 3.3.5.2 แต่ถ้ามากกว่าค่าแบนวิดธ์หารสองให้ไปทำขั้นตอนที่ 3.3.5.5

3.3.5.2 ทำการรวมตำแหน่งจุดศูนย์กลางตามสมการที่ 3.32

$$C_i = \frac{\text{density}_{C_i}}{\text{density}_{C_i} + \text{density}_{C_K}} C_i + \frac{\text{density}_{C_K}}{\text{density}_{C_i} + \text{density}_{C_K}} C_K \tag{3.32}$$

โดย density_{Ci} คือ ค่าความหนาแน่นของจุดศูนย์กลาง C_i สามารถคำนวณได้ตามสมการที่ 3.19

การรวมตำแหน่งจุดศูนย์เป็นการหาค่าเฉลี่ยของสองจุดโดยเฉลี่ยตามค่าความหนาแน่นของทั้งสองจุด แล้วบันทึกค่าตำแหน่งที่คำนวณได้ แทนค่าในตำแหน่งดัชนีที่ i

3.3.5.3 ทำการลบตำแหน่งจุดศูนย์กลางที่ดัชนีที่ K และทำการลดจำนวน K ลงอีกหนึ่ง

3.3.5.4 ทำการสลับค่าตำแหน่งจุดศูนย์กลางของคลัสเตอร์ในดัชนีที่ i กับ ดัชนีที่ K แล้วทำเริ่มวนทำซ้ำจุดศูนย์กลางของคลัสเตอร์จากดัชนีที่ 1 ใหม่ เพื่อดูว่าตำแหน่งที่ถูกรวมกันใหม่นั้นเปลี่ยนตำแหน่งไปใกล้ตำแหน่งจุดศูนย์กลางของคลัสเตอร์ไหนหรือไม่โดยพิจารณาตามสมการที่ 3.31

3.3.5.5 วนทำซ้ำข้อ 3.3.5.1 จนกระทั่งครบทุกตำแหน่งของจุดศูนย์กลางของคลัสเตอร์ที่เหลืออยู่

3.3.6 ขั้นตอนการลบจุดศูนย์กลางของคลัสเตอร์

3.3.6.1 รับค่าตำแหน่งจุดศูนย์กลางที่ถูกบันทึกไว้มาทีละ 1 เพื่อมาคำนวณหาค่าปริมาณสมาชิกของคลัสเตอร์ที่ i ($rate_i$) ตามสมการที่ 3.33

$$rate_i = -\frac{1}{2\log_2\frac{h_i}{N} - \varepsilon} \tag{3.33}$$

โดย h_i คือ จำนวนสมาชิกของคลัสเตอร์ที่ i

N คือ จำนวนของเซตข้อมูลทั้งหมด

$rate_i$ คือ อัตราส่วนจำนวนของสมาชิกที่อยู่ในคลัสเตอร์ที่ i เมื่อเปรียบเทียบกับจำนวนข้อมูลของเซตข้อมูลทั้งหมด

3.3.6.2 เปรียบเทียบค่า $rate_i$ กับค่า γ ที่กำหนดไว้ โดยถ้า $rate_i < \gamma$ แล้วแสดงว่าคลัสเตอร์ที่ i มีค่าจำนวนสมาชิกน้อยกว่าเกณฑ์ที่ตั้งไว้ ดังนั้น คลัสเตอร์ที่ i จะถูกลบทิ้ง

3.3.6.3 วนทำซ้ำข้อ 3.3.6.1 จนกระทั่งครบทุกคลัสเตอร์

3.3.7 ขั้นตอนการวนซ้ำ

ถ้าข้อมูลใน mark ยังมีค่า 0 อยู่ให้สุ่มข้อมูลจากเซตข้อมูลเพื่อนำมาเป็นจุดศูนย์กลางเริ่มต้น ต้องตรวจสอบก่อนว่าดัชนีที่เลือกมามีค่าของ mark ตรงดัชนีนั้นเป็นค่าเท่าไร ถ้ามีค่าเท่ากับ 0 ก็จะสามารถถูกเลือกได้เลย แต่ถ้ามีค่าเท่ากับ 1 ก็ให้สุ่มตำแหน่งมาใหม่ จนกว่าจะเจอตำแหน่งที่ mark

เท่ากับ 0 และเซตค่า mark ของข้อมูลที่สุ่มมานั้นให้เท่ากับ 1 แล้วเซตค่า t เท่ากับ 1 และกลับไปทำขั้นตอนที่ 3.3.3 แต่ถ้า mark ทุกค่ามีค่าเท่ากับ 1 ก็หยุดการทำงาน

3.4 กระบวนการทดสอบโมเดล

เมื่อผ่านขั้นตอนการเรียนรู้โมเดลแล้วสิ่งที่ได้จากการเรียนรู้ และนำมาใช้ประโยชน์ต่อได้แก่ จำนวนคลัสเตอร์และจุดศูนย์กลางของคลัสเตอร์ โดยในขั้นตอนนี้เราจะทำการทดสอบว่าเมื่อนำข้อมูลทดลองมาผ่านขั้นตอนการเรียนรู้แล้วเราได้จำนวนคลัสเตอร์ และจุดศูนย์กลางของคลัสเตอร์ไปทำนาย โดยนำจำนวนคลัสเตอร์ และจุดศูนย์กลางของแต่ละคลัสเตอร์ที่ได้มาใช้กับข้อมูลทดสอบ (Testing Data)

เลือกข้อมูลมาจากชุดข้อมูลทดสอบมาหนึ่งตัวแล้วใช้วัดระยะแบบยูคลิดเดียน (Euclidean Distance)ในการคำนวณหาค่าระยะห่างระหว่างข้อมูลที่เลือกจากชุดทดสอบกับแต่ละคลัสเตอร์แล้วเลือกคลัสเตอร์ที่มีระยะห่างน้อยสุดตามสมการที่ 3.34

$$J = \arg\min_{j\in\{1,\ldots,K\}} \|O - C_i\| \tag{3.34}$$

โดย K คือ จำนวนคลัสเตอร์

O คือ ค่าข้อมูลที่เลือกมาจากชุดข้อมูลทดสอบ

C_i คือ ค่าจุดศูนย์กลางของคลัสเตอร์ที่ i

J คือ ดัชนีของคลัสเตอร์ที่ข้อมูล O เป็นสมาชิก

3.5 ตัวอย่างการทำงาน

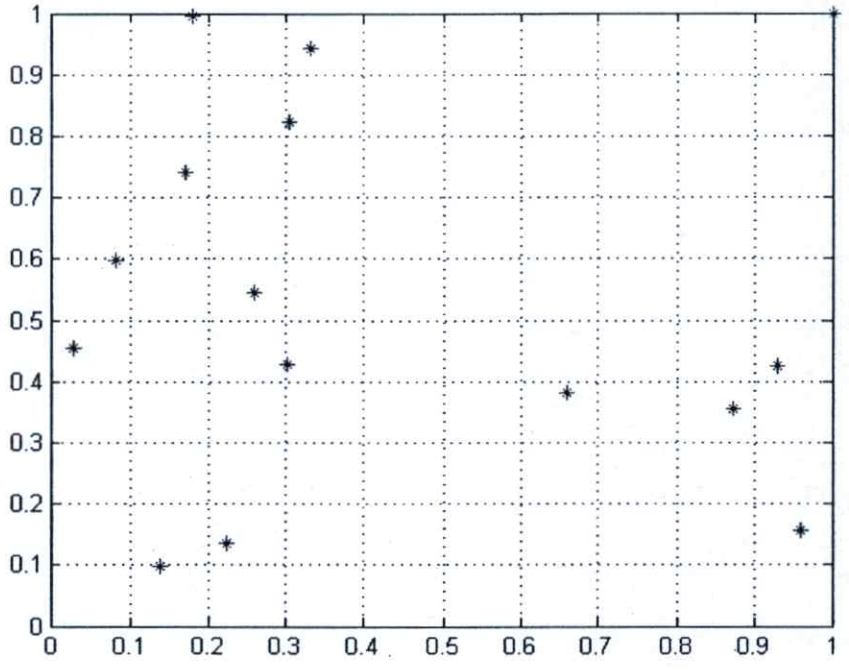
- ข้อมูลที่ใช้มีจำนวน 15 ตัว
- จำนวนมิติเท่ากับ 2 คือ X กับ Y

ตารางที่ 3.1 แสดงตัวอย่างข้อมูลที่ต้องการจัดแบ่งกลุ่มโดยใช้งานวิจัยที่นำเสนอ

ID	X	Y
A1	0.1684	0.7401
A2	0.8728	0.3555
A3	0.1788	0.9964
A4	0.9591	0.1560
A5	0.9289	0.4246
A6	0.2231	0.1364
A7	0.2602	0.5445
A8	0.0822	0.5981
A9	0.0292	0.4549
A10	0.3325	0.9437
A11	0.3031	0.4282
A12	0.6610	0.3811

ตารางที่ 3.1 (ต่อ) แสดงตัวอย่างข้อมูลที่ต้องการจัดแบ่งกลุ่มโดยงานวิจัยที่นำเสนอ

ID	X	Y
A13	0.1362	0.0966
A14	0.3052	0.8239
A15	1	1



รูปที่ 3.6 แสดงรูปข้อมูลที่ต้องการจัดแบ่งกลุ่มโดย * แทนข้อมูลแต่ละจุด

กำหนดค่าตัวแปรต่างๆ

- ค่าที่กำหนดให้จุดศูนย์กลางหยุดการเคลื่อนที่ (threshold) = 0.02
- ค่าเกณฑ์ในการลบคลัสเตอร์ (γ) = 0.08

3.5.1 คำนวณค่าแบนวิดธ์

คำนวณค่าเฉลี่ยในแต่ละมิติได้ $\bar{X} = (0.4294, 0.5387)$ คำนวณค่า IQR ได้เท่ากับ 0.7846 คำนวณส่วนเบี่ยงเบนมาตรฐานโดยใช้สมการที่ 3.8

$$\sigma = \sqrt{\frac{\|A1 - \bar{X}\| + \|A2 - \bar{X}\| + \|A3 - \bar{X}\| + \ldots + \|A15 - \bar{X}\|}{N}}$$

$$\sigma = \sqrt{\frac{0.3297 + 0.4798 + 0.5218 + 0.6535 + 0.5124 + 0.4521 + 0.1693 + 0.3522 + 0.4082 + 0.4165 + 0.1678 + 0.2801 + 0.5305 + 0.3111 + 0.7338}{15}}$$

$$\sigma = 0.4488$$

คำนวณหา A สมการที่ 3.7

$$A = \min\left[0.4488, \frac{0.7846}{1.34}\right] = 0.4488$$

คำนวณค่าแบนวิดธ์โดยใช้สมการที่ 3.6

$$S = \frac{0.9(0.4488)}{\sqrt[5]{15}} = 0.2350$$

3.5.2 การเลือกจุดศูนย์กลางเริ่มต้นจากเซตข้อมูล

3.5.2.1 สร้างตัวแปร mark ขนาดเท่ากับ 15 และเซตค่าเริ่มต้นเท่ากับ 0 ทุกตัวตามตารางที่ 3.2

ตารางที่ 3.2 แสดงตัวอย่างข้อมูลที่เซตค่า mark เริ่มต้นเท่ากับ 0

ID	X	Y	Mark
A1	0.1684	0.7401	0

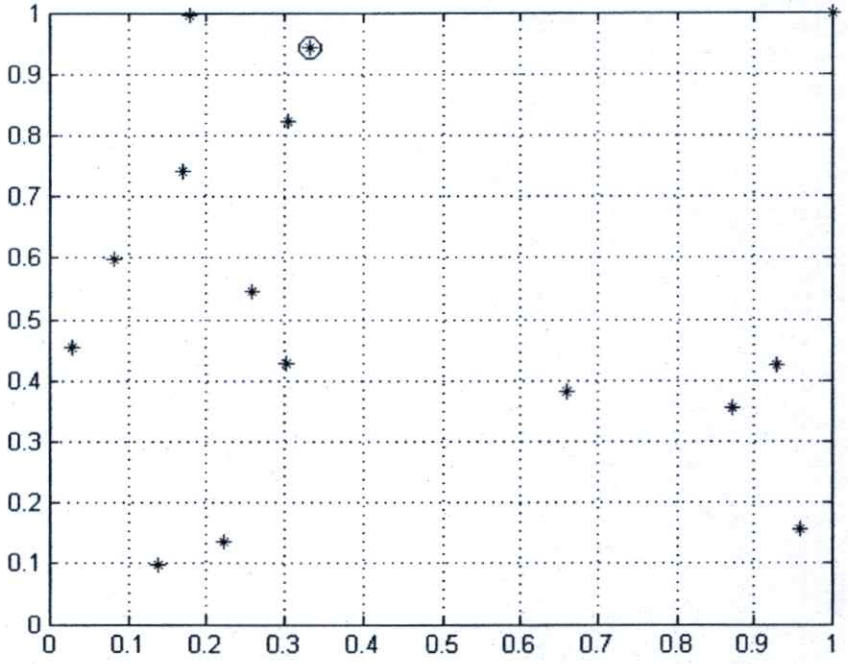
ตารางที่ 3.2 (ต่อ) แสดงตัวอย่างข้อมูลที่เซตค่า mark เริ่มต้นเท่ากับ 0

ID	X	Y	Mark
A2	0.8728	0.3555	0
A3	0.1788	0.9964	0
A4	0.9591	0.1560	0
A5	0.9289	0.4246	0
A6	0.2231	0.1364	0
A7	0.2602	0.5445	0
A8	0.0822	0.5981	0
A9	0.0292	0.4549	0
A10	0.3325	0.9437	0
A11	0.3031	0.4282	0
A12	0.6610	0.3811	0
A13	0.1362	0.0966	0

ตารางที่ 3.2 (ต่อ) แสดงตัวอย่างข้อมูลที่เซตค่า mark เริ่มต้นเท่ากับ 0

ID	X	Y	Mark
A14	0.3052	0.8239	0
A15	1	1	0

3.5.2.2 ทำการสุ่มจุดศูนย์กลางเริ่มต้นมาหนึ่งตำแหน่งได้ จุดศูนย์กลางคือ C(1) = (0.3325, 0.9437) แล้วเซตค่า mark ของ A10 เท่ากับ 1



รูปที่ 3.7 แสดงจุดศูนย์กลางเริ่มต้นของคลัสเตอร์โดย O แทนจุดศูนย์กลางของคลัสเตอร์ที่ 1

รอบการทำงานที่ 1

3.5.3 คำนวณการเคลื่อนที่ของจุดศูนย์กลาง

3.5.3.1 ทำการเซตค่า mark ของข้อมูลที่อยู่รอบตัวจุดศูนย์กลาง C(1) ในระยะรัศมีค่าแบนวิดธ์ให้เท่ากับ 1 ตามเงื่อนไขในสมการ 3.10

ตารางที่ 3.3 แสดงข้อมูลที่อยู่รอบตัวจุดศูนย์กลาง C(1)

		C(1) = (0.3325, 0.9437)		
	Point	Dist C(1)	Dist C(1) < 0.235	mark
A1	(0.1686,0.7401)	0.2615	True	0
A2	(0.8728,0.3555)	0.7987	False	0
A3	(0.1788,0.9964)	0.1625	True	1
A4	(0.9591,0.9289)	1.0065	False	0
A5	(0.9289,0.4246)	0.7907	False	0
A6	(0.2231,0.1364)	0.8147	False	0
A7	(0.2602,0.5445)	0.4057	True	0
A8	(0.0822,0.5981)	0.4267	True	0
A9	(0.0292,0.4549)	0.5753	False	0
A10	(0.3325,0.9437)	0	True	1
A11	(0.3031,0.4282)	0.5163	True	0
A12	(0.6610,0.3811)	0.6515	False	0
A13	(0.1362,0.0966)	0.8695	False	0
A14	(0.3052,0.8239)	0.1229	True	1

ตารางที่ 3.3 (ต่อ) แสดงข้อมูลที่อยู่รอบตัวจุดศูนย์กลาง C(1)

		C(1) = (0.3325, 0.9437)		
	Point	Dist C(1)	Dist C(1) < 0.235	mark
A15	(1,1)	0.6699	False	0

3.5.3.2 ทำการเลือกแต่เฉพาะข้อมูลที่ผ่านเงื่อนไขมี A3, A10 และ A14 ในตารางที่ 3.3 คำนวณแรงทั้งหมดกระทำกับจุดศูนย์กลาง C(1) โดยใช้สมการ 3.12 – 3.21

3.5.3.2.1 คำนวณค่าความหนาแน่นของจุดข้อมูลคำนวณความหนาแน่นที่ A3 เลือกแต่ข้อมูลมีข้อมูลที่อยู่ระยะรัศมีค่าแบนวิดธ์ของ A3 มาคำนวณ โดยใช้สมการที่ 3.19

ตารางที่ 3.4 แสดงข้อมูลที่อยู่รอบตัวจุดศูนย์กลาง A3(0.1788,0.9964)

		A3(0.1788,0.9964)	
	Point	Dist A3	Dist A3 < 0.235
A1	(0.1686,0.7401)	0.2565	False
A2	(0.8728,0.3555)	0.9447	False
A3	(0.1788,0.9964)	0	True
A4	(0.9591,0.9289)	1.1468	False
A5	(0.9289,0.4246)	0.9432	False
A6	(0.2231,0.1364)	0.8611	False
A7	(0.2602,0.5445)	0.4592	False
A8	(0.0822,0.5981)	0.4098	False
A9	(0.0292,0.4549)	0.5618	False
A10	(0.3325,0.9437)	0.1625	True

ตารางที่ 3.4 (ต่อ) แสดงข้อมูลที่อยู่รอบตัวจุดศูนย์กลาง A3(0.1788,0.9964)

		A3(0.1788,0.9964)	
	Point	Dist A3	Dist A3 < 0.235
A11	(0.3031,0.4282)	0.5816	False
A12	(0.6610,0.3811)	0.7817	False
A13	(0.1362,0.0966)	0.9008	False
A14	(0.3052,0.8239)	0.2139	True
A15	(1,1)	0.8212	False

จะได้ข้อมูล A3, A10 และ A14 มาคำนวณ

$$\text{density}_{A3} = \left[-\frac{1}{2\times\log_2\left(\frac{3}{15}\right)-(2.2204\times10^{-16})}\right] \times \frac{1}{3}\left(e^{-\frac{\|A3-A3\|^2}{0.235^2}} + e^{-\frac{\|A10-A3\|^2}{0.235^2}} + e^{-\frac{\|A14-A3\|^2}{0.235^2}}\right)$$

$$= \left[-\frac{1}{2\times\log_2\left(\frac{3}{15}\right)-(2.2204\times10^{-16})}\right] \times \frac{1}{3}\left(e^{-\frac{0^2}{0.235^2}} + e^{-\frac{0.1625^2}{0.235^2}} + e^{-\frac{0.2139^2}{0.235^2}}\right)$$

$$\text{density}_{A3} = 0.2153 \times 0.6856 = 0.1476$$

หาความหนาแน่นในแต่ละจุดดังนี้ $\text{density}_{A3} = 0.1476$, $\text{density}_{A10} = 0.1709$, $\text{density}_{A14} = 0.1852$ และ $\text{density}_{C} = 0.1709$

หาค่า min_density และ max_density ได้ดังนี้ min_density = 0.1476 และ max_density = 0.1852

3.5.3.2.2 คำนวณค่ามวลโดยใช้สมการที่ 3.14 – 3.16

$$m_{A3} = \frac{\text{density}_{A3} - \text{min\_density}}{\text{max\_density} - \text{min\_density}}$$

$$m_{A3} = \frac{0.1476 - 0.1476}{0.1852 - 0.1476} = 0$$

หาค่า m_i แต่ละจุดได้ดังนี้ $m_{A3} = 0$, $m_{A10} = 0.6191$, $m_{A14} = 1$ และ $m_C = 0.6191$

$$M_{A3} = \frac{m_{A3}}{m_{A3} + m_{A10} + m_{A14} + m_C}$$

$$M_{A3} = \frac{0}{0 + 0.6191 + 1 + 0.6191} = 0$$

หาค่า $M_{\bullet}$ แต่ละมวลได้ดังนี้ $M_{A3} = 0$, $M_{A10} = 0.2766$, $M_{A14} = 0.4468$, และ $M_C = 0.2766$

3.5.3.2.3 คำนวณค่าคงที่แรงโน้มถ่วงโดยใช้สมการที่ 3.20

$$G = 0.235 \times e^{-\frac{1}{100}} = 0.2327$$

3.5.3.2.4 นำค่ามวลและค่าคงที่แรงโน้มถ่วงที่คำนวณข้างต้นมาคำนวณค่าแรงที่กระทำกับจุดศูนย์กลางโดยใช้สมการที่ 3.13

$$F^x_{CA3} = 0.2327 \frac{M_c \text{ x } M_{A3}}{R_{CA3} + \varepsilon}(A3^x - C(1)^x)$$

$$= 0.2327 \frac{0.2766 \text{ x } 0}{0.1625 + (2.2204 \times 10^{-16})}(0.1788 - 0.3325) = 0$$

$$F^y_{CA3} = 0.2327 \frac{M_c \text{ x } M_{A3}}{R_{CA3} + \varepsilon}(A3^y - C(1)^y)$$

$$= 0.2327 \frac{0.2766 \text{ x } 0}{0.1625 + (2.2204 \times 10^{-16})}(0.9964 - 0.9437) = 0$$

$$F_{CA3} = (0,0)$$

หาค่า $F_{C\bullet}$ แต่ละแรงได้ดังนี้ $F_{CA3} = (0,0)$. $F_{CA10} = (0,0)$ และ $F_{CA14} = (-0.0064, -0.028)$

3.5.3.2.5 คำนวณผลรวมของที่กระทำกับจุดศูนย์กลางสมการที่ 3.12

$$F^X_C = (0 + 0 - 0.0064) = -0.0064$$

$$F^Y_C = (0 + 0 - 0.028) = -0.028$$

$$F_C = (-0.0064, -0.028)$$

3.5.3.3 คำนวณค่าอัตราเร่งของจุดศูนย์กลาง โดยใช้สมการที่ 3.22

$$a_C^X = \frac{F_C^X}{M_C} = \frac{-0.0064}{0.2766} = -0.0231$$

$$a_C^Y = \frac{F_C^Y}{M_C} = \frac{-0.028}{0.2766} = -0.1014$$

$$a_C = (-0.0231, -0.1014)$$

3.5.3.4 คำนวณค่าความเร็วของจุดศูนย์กลาง

3.5.3.4.1 คำนวณค่า β_C โดยใช้สมการที่ 3.25

$$\beta_C = e^{-\frac{0}{10}} = 1$$

3.5.3.4.2 คำนวณค่าความเร็วของจุดศูนย์กลางโดยใช้สมการที่ 3.23 โดยที่ $rand_C = 0.9$

$$V_C^X = 0.9 \times 0 + (1 \times -0.0231) = -0.0231$$

$$V_C^Y = 0.9 \times 0 + (1 \times -0.1014) = -0.1014$$

$$V_C = (-0.0231\ , -0.1014)$$

3.5.3.5 คำนวณค่าตำแหน่งของจุดศูนย์กลางใหม่โดยใช้สมการที่ 3.26

$$C(2)^X = 0.3325 - 0.0231 = 0.3094$$

$$C(2)^Y = 0.9437 - 0.1014 = 0.8423$$

$$C(2) = (0.3094, 0.8423)$$

3.5.4 ขั้นตอนการบันทึกตำแหน่งจุดศูนย์กลางใหม่

พิจารณาการหยุดการเคลื่อนของจุดศูนย์กลางโดยใช้สมการที่ 3.27

$$\|C(2) - C(1)\| < 0.02$$

$$\|(0.3094,0.8423) - (0.3325,0.9437)\| < 0.02$$

0.1040 < 0.02 ดังนั้น กลับไปทำขั้นตอนที่ (3.5.3) โดยนำค่าจุดกลางใหม่ที่ได้คือ C(2) = (0.3094,0.8423) ไปคำนวณการเคลื่อนที่ของจุดศูนย์กลางอีกรอบ จนกระทั่งเมื่อถึงรอบที่ 4 ค่า C(5) = (0.2293,0.7076) และ ค่า C(4) = (0.2342,0.7206) จะได้เงื่อนตามนี้ 0.014 < 0.02 ซึ่งค่าของจุดศูนย์กลางรอบที่ 5 กับ รอบที่ 4 มีค่าต่างกันน้อยว่า 0.02 ดังนั้นจะได้ค่าของ C(5) เป็นจุดศูนย์กลางของคลัสเตอร์ที่ 1 และทำการบันทึกเก็บไว้แล้วไปทำขั้นตอนที่ 3.5.5

3.5.5 ทำการรวมตำแหน่งจุดศูนย์กลางของคลัสเตอร์

ถ้าจำนวนของคลัสเตอร์ทั้งหมดเท่ากับ 1 ก็ไม่ต้องทำการรวมค่าตำแหน่งจุดศูนย์กลาง แต่ถ้าจำนวนของคลัสเตอร์ทั้งหมดมากกว่า 1 ก็จะทำการรวมตำแหน่งจุดศูนย์กลาง เมื่อโปรแกรมทำงานจนมาถึงจำนวนของคลัสเตอร์ทั้งหมดเท่ากับ 5 ค่าจุดศูนย์ของคลัสเตอร์ที่ 1 C_1 = (0.2293,0.7076), ค่าจุดศูนย์ของคลัสเตอร์ที่ 2 C_2 = (0.1342,0.1038), ค่าจุดศูนย์ของคลัสเตอร์ที่ 3 C_3 = (1,1), ค่าจุดศูนย์ของคลัสเตอร์ที่ 4 C_4 = (0.9062,0.3758) และค่าจุดศูนย์ของคลัสเตอร์ที่ 5 C_5 = (0.2324,0.6636)

3.5.5.1 พิจารณาการรวมตำแหน่งจุดศูนย์กลางโดยใช้สมการที่ 3.31

$$\|(0.2293,0.7076) - (0.2324,0.6636)\| < \frac{0.235}{2}$$

0.0441 < 0.1175 ดังนั้นจะทำการรวมคลัสเตอร์ที่ 1 กับ ของคลัสเตอร์ที่ 5

3.5.5.2 ทำการรวมตำแหน่งจุดศูนย์กลางโดยใช้สมการที่ 3.32 และใช้สมการที่ 3.19 หาค่าความหนาแน่นที่จุดศูนย์กลาง C_1 และ C_5 จะได้ $density_{C1}$ = 0.1818 และ $density_{C5}$ = 0.1825

$$C_1 = \frac{0.1818}{0.1818 + 0.1825}(0.2293,0.7076) + \frac{0.1825}{0.1818 + 0.1825}(0.2324,0.6636)$$

$$C_1 = (0.2309,0.6856)$$

3.5.5.3 ทำการลบตำแหน่งจุดศูนย์กลางของคลัสเตอร์ที่ 5 ออก

3.5.6 ทำลบจุดศูนย์กลางของคลัสเตอร์

ถ้าจำนวนของคลัสเตอร์ทั้งหมดเท่ากับ 1 ก็ไม่ต้องทำลบจุดศูนย์กลางของคลัสเตอร์ ถ้าจำนวนของคลัสเตอร์ทั้งหมดมากกว่า 1 ก็จะทำลบจุดศูนย์กลางของคลัสเตอร์ เมื่อโปรแกรมทำงานจนมาถึงจำนวนของคลัสเตอร์ทั้งหมดเท่ากับ 4 ค่าจุดศูนย์ของคลัสเตอร์ที่ 1 C_1 = (0.2309,0.6856), ค่าจุดศูนย์ของคลัสเตอร์ที่ 2 C_2 = (0.1342,0.1038), ค่าจุดศูนย์ของคลัสเตอร์ที่ 3 C_3 = (1,1) และค่าจุดศูนย์ของคลัสเตอร์ที่ 4 C_4 = (0.9062,0.3758)

3.5.6.1 คลัสเตอร์ที่ 1

$$\text{rate}_{C_1} = -\frac{1}{2\log_2\frac{8}{15} - (2.2204 \times 10^{-16})} = 0.5513$$

0.5513 < 0.08 ดังนั้น จุดศูนย์กลางของคลัสเตอร์ที่ 1 ไม่ถูกลบทิ้ง

3.5.6.2 คลัสเตอร์ที่ 2

$$\text{rate}_{C_2} = -\frac{1}{2\log_2\frac{2}{15} - (2.2204 \times 10^{-16})} = 0.1720$$

0.1720 < 0.08 ดังนั้น จุดศูนย์กลางของคลัสเตอร์ที่ 2 ไม่ถูกลบทิ้ง

3.5.6.3 คลัสเตอร์ที่ 3

$$\text{rate}_{C_3} = -\frac{1}{2\log_2\frac{1}{15} - (2.2204 \times 10^{-16})} = 0.128$$

0.128 < 0.08 ดังนั้น จุดศูนย์กลางของคลัสเตอร์ที่ 3 ไม่ถูกลบทิ้ง

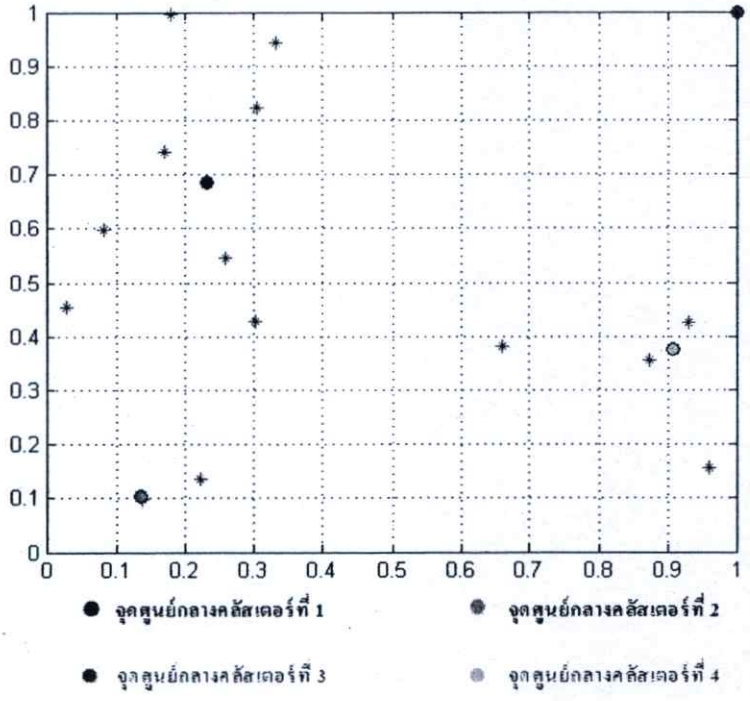
3.5.6.4 คลัสเตอร์ที่ 4

$$\text{rate}_{C_4} = -\frac{1}{2\log_2\frac{4}{15} - (2.2204 \times 10^{-16})} = 0.2622$$

0.2622 < 0.08 ดังนั้น จุดศูนย์กลางของคลัสเตอร์ที่ 4 ไม่ถูกลบทิ้ง

ตารางที่ 3.5 แสดงค่าจุดศูนย์กลางของคลัสเตอร์ที่ได้จากขั้นตอนของงานวิจัยที่นำเสนอ

คลัสเตอร์	ค่าจุดศูนย์กลาง	
	X	Y
1	0.2309	0.6856
2	0.1342	0.1038
3	1	1
4	0.9062	0.3758



รูปที่ 3.8 แสดงจุดศูนย์กลางของคลัสเตอร์ 1, 2, 3 และ 4 หลังการทำตามขั้นตอนของงานวิจัยที่นำเสนอ

3.6 ความแตกต่างระหว่างงานวิจัยที่นำเสนอกับ Gravitational Search Algorithm: GSA

1) **แนวคิดการนำกฎแรงโน้มถ่วงมาใช้ต่างกัน** โดยที่ในงานวิจัยที่นำเสนอใช้แนวแรงโน้มถ่วงแก้ปัญหาเกี่ยวกับการจัดแบ่งกลุ่มข้อมูลแบบไม่รู้เป้าหมาย (Clustering) เพื่อทำให้จุดศูนย์กลางนั้นเคลื่อนที่ไปยังบริเวณที่มีข้อมูลรวมตัวกันอยู่หนาแน่น ขณะที่ GSA ใช้แนวแรงโน้มถ่วงแก้ปัญหาเกี่ยวกับการค้นหาคำตอบที่ดีที่สุด เพื่อทำให้อนุภาคเคลื่อนที่ไปยังบริเวณจุดที่มีคำตอบที่ดีที่สุด

2) **ขั้นตอนของอัลกอริทึมที่ต่างกัน** งานวิจัยที่นำเสนอมีขั้นตอนดังนี้ การคำนวณค่าแบนวิดธ์, ขั้นตอนการปรับตำแหน่งจุดศูนย์กลาง, ขั้นตอนการรวมตำแหน่งจุดศูนย์กลางและขั้นตอนการลบตำแหน่งจุดศูนย์กลาง ซึ่งในส่วน GSA จะมีเฉพาะขั้นตอนการปรับตำแหน่งของอนุภาคที่ใช้แรงโน้มถ่วงเช่นเดียวกับขั้นตอนการปรับตำแหน่งจุดศูนย์กลางในงานวิจัยที่นำเสนอ

3) **วิธีการคำนวณมวลแตกต่างกัน** GSA มวลสามารถคำนวณได้จากฟังก์ชันความเหมาะสม ซึ่งมันสามารถถูกแทนได้ในฟังก์ชันใดก็ได้ขึ้นอยู่กับปัญหาที่ต้องการแก้แต่ในงานวิจัยที่นำเสนอมวลคำนวณได้จากฟังก์ชันความหนาแน่นในสมการที่ 3.19

3.7 ความแตกต่างระหว่างงานวิจัยที่นำเสนอกับ A New Gravitational Clustering Algorithm

1) แนวคิดการประยุกต์ใช้กฎแรงโน้มถ่วงที่ต่างกัน

งานวิจัยที่นำเสนอใช้แนวคิดของกฎแรงโน้มถ่วงเพื่อทำให้จุดศูนย์กลางนั้นเคลื่อนที่ไปยังบริเวณที่มีข้อมูลรวมตัวกันอยู่หนาแน่น อัลกอริทึมการจัดแบ่งกลุ่มข้อมูลแบบใหม่ด้วยแรงโน้มถ่วงใช้แนวคิดของกฎแรงโน้มถ่วงเพื่อทำให้แต่ละในเซตข้อมูลปรับการเคลื่อนที่แล้วก็ตรวจสอบจุดไหนเคลื่อนที่มาใกล้ก็ทำการระบุว่าเป็นคลัสเตอร์เดียวกัน

2) ขั้นตอนของอัลกอริทึมที่ต่างกัน

งานวิจัยที่นำเสนอมีขั้นตอน เริ่มจากทำการสุ่มรูปแบบอินพุตหนึ่งตัวมาเป็นจุดศูนย์กลางเริ่มต้นทำการปรับค่าจุดศูนย์กลางของคลัสเตอร์ โดยดูจากจุดที่อยู่รอบข้างซึ่งออกแรงกระทำต่อจุดศูนย์กลาง ซึ่งในการคำนวณแรงเราจะใช้สมการแรงโน้มถ่วงคำนวณแรงที่กระทำบนจุดศูนย์กลางนั้น จากนั้นก็หาแรงรวมที่เกิดขึ้นบนจุดศูนย์กลางแล้วไปคำนวณหาการเคลื่อนที่โดยใช้สมการการเคลื่อนของนิวตัน เมื่อทำการปรับค่าจุดศูนย์กลางไปแต่ละรอบแล้วเราจะทำการตรวจสอบว่าตำแหน่งจุดศูนย์ที่ปรับไปนั้นมีค่าเปลี่ยนแปลงไปจากรอบที่แล้วหรือไม่ ถ้าเริ่มไม่เปลี่ยนแปลงแล้วก็บันทึกตำแหน่งจุดศูนย์กลางใหม่ที่ได้ หลังจากนั้นก็นำตำแหน่งจุดศูนย์กลางใหม่ที่ได้มาทำการ

ตรวจสอบว่าอยู่ใกล้กันหรือไม่ ถ้าอยู่ใกล้ก็จะทำการรวมจุดกลางศูนย์ที่อยู่ใกล้ชิดกัน จากนั้นก็ทำการตรวจสอบการลบคลัสเตอร์โดยจะนับจำนวนสมาชิกของแต่ละคลัสเตอร์แล้วดูว่าจำนวนสมาชิกในแต่ละคลัสเตอร์มีค่าผ่านเกณฑ์ที่กำหนดไว้หรือไม่ ถ้าไม่ผ่านก็จะทำการลบคลัสเตอร์นั้น ขั้นตอนทั้งหมดจะถูกทำซ้ำจนกระทั่งข้อมูลทุกจุดในเซตข้อมูลถูกกระทำทั้งหมด

อัลกอริทึมการจัดแบ่งกลุ่มข้อมูลแบบใหม่ด้วยแรงโน้มถ่วง เริ่มทำการสร้างเซตของคลัสเตอร์ โดยทุกจุดข้อมูลในเซตข้อมูลเป็นคลัสเตอร์ ทำการสุ่มจุดมาหนึ่งจุดแล้วนำไปทำการปรับการเคลื่อนที่กับจุดในเซตข้อมูล แล้วตรวจสอบว่าจุดที่ถูกปรับการเคลื่อนที่กับจุดข้อมูลในเซตข้อมูลนั้นมีระยะทางใกล้หรือไม่ ถ้าใกล้กันก็จะทำการยูเนียนสองจุดนั้นเป็นคลัสเตอร์เดียวกันโดยใช้ปฏิบัติการ UNION จะทำซ้ำแบบนี้จนทุกจุดข้อมูลเป็นสมาชิกในเซตของคลัสเตอร์ หลังจากนั้นก็ทำการบีบอัดเส้นทางในทุกเซตคลัสเตอร์โดยตัวปฏิบัติการ FIND จากนั้นก็ทำการตรวจสอบการลบคลัสเตอร์ โดยจะนับจำนวนสมาชิกในเซตของคลัสเตอร์แล้วดูว่าจำนวนสมาชิกในแต่ละคลัสเตอร์มีค่าผ่านเกณฑ์ที่กำหนดไว้หรือไม่ ถ้าไม่ผ่านก็จะทำการลบคลัสเตอร์นั้น

3) การคำนวณการเคลื่อนที่ต่างกัน

งานวิจัยที่นำเสนอใช้สมการแรงโน้มถ่วงตามสมการที่ 3.12 – 3.19 โดยแรงโน้มถ่วงดังกล่าวมีการคำนวณ M มาจากสมการความแน่นหนาของจุดข้อมูลที่อยู่รอบข้างจุดกำลังพิจารณาอยู่ หาผลรวมของแรงของจุดที่กำลังจะเคลื่อนที่ หลังจากนั้นก็นำแรงโน้มถ่วงที่คำนวณไปความเร่งตามสมการที่ 3.22 นำความเร่งที่ได้ไปคำนวณความเร็วตามสมการที่ 3.23 แล้วสุดจึงจะนำความเร็วที่ได้ไปปรับตำแหน่งการเคลื่อนของจุดตามสมการที่ 3.26 แต่ในอัลกอริทึมการจัดแบ่งกลุ่มข้อมูลแบบใหม่ด้วยแรงโน้มถ่วง ใช้สมการการเคลื่อนที่ตามสมการที่ 2.41 ในการปรับการเคลื่อนที่ของจุด โดยมอง v(t) มีค่าเท่ากับ 0 และ M ของทุกจุดมีค่าเท่ากับ 1 และการปรับตำแหน่งจะพิจารณาเป็นคู่จุด ไม่เหมือนงานวิจัยที่นำเสนอจะพิจารณาจากทุกจุดข้อมูลที่อยู่รอบข้างในระยะรัศมีที่คำนวณได้

4) โครงสร้างของผลลัพธ์ที่ได้

งานวิจัยที่นำเสนอผลลัพธ์ที่ได้นั้นเป็นค่าตำแหน่งของจุดศูนย์กลางของแต่ละคลัสเตอร์ แต่ในอัลกอริทึมการจัดแบ่งกลุ่มข้อมูลแบบใหม่ด้วยแรงโน้มถ่วง ผลลัพธ์ออกมาเป็นเซตของโครงสร้าง disjoint set union-find ของแต่ละคลัสเตอร์ซึ่งโครงสร้างที่คล้ายกับต้นไม้

บทที่ 4

การทดสอบการนำฟังก์ชันความหนาแน่นร่วมกับอัลกอริทึมการค้นหาคำตอบด้วยแรงโน้มถ่วงเพื่อแก้ปัญหาการจัดแบ่งกลุ่มข้อมูล

ในบทนี้จะกล่าวถึงค่าพารามิเตอร์ที่ใช้ในการทดลองและผลที่ได้จากการทดลอง โดยข้อมูลที่นำมาใช้ในการทดลองมี 2 ชุด คือ 1) ข้อมูลมาตรฐาน [10] ซึ่งประกอบด้วย 6 ชุดข้อมูลคือ Dermatology, Libras Movement, Large Soybean, Wine Recognition, Iris Plant และ Bupa Liver Disorders 2) ข้อมูลที่ผู้วิจัยสร้างขึ้นเองซึ่งประกอบด้วย 4 ชุดข้อมูล คือ Normal Distribution, Fan, Ring และ Shape ในการทดลองนี้จะทำการเปรียบเทียบประสิทธิภาพ 4 ตัววัดด้วยกัน คือ ความบริสุทธิ์ (Purity), เอนโทรปี (Entropy), NMI (Normalized Mutual Information) และ F measure โดยทำการเปรียบเทียบกับ K-means Clustering, Fuzzy C-means Clustering และ DBSCAN Algorithm ซึ่งเป็นอัลกอริทึมทางด้านการจัดแบ่งกลุ่มข้อมูลที่ได้รับการยอมรับและนำไปประยุกต์ใช้ในหลายๆ ด้าน

4.1 เกณฑ์ที่ใช้วัดประสิทธิภาพในการทดลอง

4.1.1 ความบริสุทธิ์ (Purity) การวัดประสิทธิภาพในเทอมของความบริสุทธิ์ (Purity) ในการจัดแบ่งกลุ่มข้อมูล ในการคำนวณความถูกต้องของความบริสุทธิ์ ความถูกต้องจะวัดโดยผลรวมของการนับจำนวนของคลาสที่พบมากที่สุดของแต่ละคลัสเตอร์แล้วหารด้วยจำนวนสมาชิกทั้งหมด ตามสมการที่ 4.1 [11]

$$\text{Purity}(\Omega, \phi) = \frac{1}{N}\sum_k \max_j |\omega_k \cap C_j| \qquad (4.1)$$

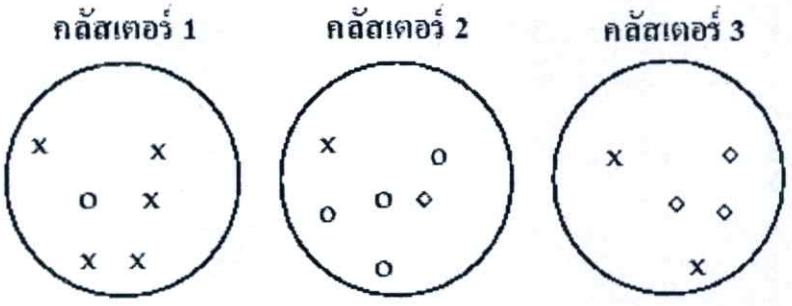
โดย $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$ คือ เซตของคลัสเตอร์ทั้งหมด

$\phi = \{C_1, C_2, \ldots, C_J\}$ คือ เซตของคลาสทั้งหมด

K คือ จำนวนคลัสเตอร์ทั้งหมด

J คือ จำนวนคลาสทั้งหมด

N คือ จำนวนเซตข้อมูลทั้งหมด



รูปที่ 4.1 แสดงสมาชิกและคลาสส่วนมากของ 3 คลัสเตอร์ X เป็นคลาสส่วนมากของคลัสเตอร์ 1,O เป็นคลาสส่วนมากของคลัสเตอร์ 2 และ ◊ เป็นคลาสส่วนมากของคลัสเตอร์ 3

จากรูปที่ 4.1 สามารถคำนวณ Purity ได้ดังนี้ X, 5 (คลัสเตอร์ 1); O, 4 (คลัสเตอร์ 2); และ ◊, 3 (คลัสเตอร์ 3)

$$\text{Purity} = \frac{1}{17} \times (5 + 4 + 3) \approx 0.71$$

ถ้าค่า Purity มีค่า 0 แสดงว่าการทำคลัสเตอร์มีคุณภาพต่ำสุดแต่ถ้าค่า Purity มีค่า 1 แสดงว่าการทำคลัสเตอร์มีคุณภาพสูงสุด

4.1.2 เอนโทรปี (Entropy) เป็นการวัดความบริสุทธิ์ของคลาสว่าเป็นคลาสเดียวกันมากน้อยเท่าไร คล้ายกับ Purity แต่มีวิธีการคำนวณที่ต่างกัน ในการคำนวณเอนโทรปีจะหาความน่าจะเป็นที่เจอคลาสในแต่ละคลัสเตอร์ มีค่าเท่าไรแล้วนำไปคำนวณด้วยสูตรเอนโทรปี สามารถเขียนได้เป็นดังสมการที่ 4.2 [11]

$$\text{Entropy}(\Omega, \phi) = -\frac{1}{K}\sum_k \sum_j \frac{|\omega_k \cap C_j|}{|\omega_k|} \log_2 \frac{|\omega_k \cap C_j|}{|\omega_k|} \qquad (4.2)$$

โดย $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$ คือ เซตของคลัสเตอร์ทั้งหมด

$\phi = \{C_1, C_2, \ldots, C_J\}$ คือ เซตของคลาสทั้งหมด

K คือ จำนวนคลัสเตอร์ทั้งหมด

J คือ จำนวนคลาสทั้งหมด

N คือ จำนวนเซตข้อมูลทั้งหมด

จากรูปที่ 4.1 สามารถคำนวณ Entropy ได้ดังนี้

$$\text{Entropy} = -\left(\left(\frac{5}{6}\log_2\frac{5}{6}+\frac{1}{6}\log_2\frac{1}{6}+0\right)+\left(\frac{1}{6}\log_2\frac{1}{6}+\frac{4}{6}\log_2\frac{4}{6}+\frac{1}{6}\log_2\frac{1}{6}\right)+\left(\frac{2}{5}\log_2\frac{2}{5}+0+\frac{3}{5}\log_2\frac{3}{5}\right)\right)/3$$

$$\text{Entropy} = -\frac{(-0.65-1.2516-0.9710)}{3} = 0.9575$$

ถ้าค่า Entropy มีค่า 0 แสดงว่าการทำคลัสเตอร์มีคุณภาพสูงสุด ถ้าค่า Entropy สูงแสดงว่าการทำคลัสเตอร์มีคุณภาพต่ำ

4.1.3 NMI (Normalized Mutual Information) ในการวัดประสิทธิภาพของ Purity ไม่สามารถวัดประสิทธิตามแนวโน้มจำนวนของคลัสเตอร์ ในการทำคลัสเตอร์ถ้ามีการจำนวนคลัสเตอร์มากเกินอาจจะทำให้การทำคลัสเตอร์ไม่มีประสิทธิภาพได้ NMI เป็นตัววัดประสิทธิภาพที่มีแนวโน้มตามจำนวนคลัสเตอร์ สามารถคำนวณได้ตามสมการ 4.3 – 4.6 [11]

$$\text{NMI}(\Omega,\phi) = \frac{I(\Omega,\phi)}{(H(\Omega)+H(\phi))/2} \tag{4.3}$$

$$I(\Omega,\phi) = \sum_k\sum_j \frac{|\omega_k\cap C_j|}{N}\log_2\frac{N|\omega_k\cap C_j|}{|\omega_k||C_j|} \tag{4.4}$$

$$H(\Omega) = -\sum_k \frac{|\omega_k|}{N}\log_2\frac{|\omega_k|}{N} \tag{4.5}$$

$$H(\phi) = -\sum_j \frac{|C_j|}{N}\log_2\frac{|C_j|}{N} \tag{4.6}$$

โดย $\Omega = \{\omega_1, \omega_2, \dots, \omega_K\}$ คือ เซตของคลัสเตอร์ทั้งหมด

$\phi = \{C_1, C_2, \dots, C_J\}$ คือ เซตของคลาสทั้งหมด

K คือ จำนวนคลัสเตอร์ทั้งหมด

J คือ จำนวนคลาสทั้งหมด

N คือ จำนวนเซตข้อมูลทั้งหมด

จากรูปที่ 4.1 สามารถคำนวณ NMI ได้ดังนี้

$$I(\Omega,\phi)=\left(\left(\frac{5}{17}\log_2\frac{17\times5}{6\times8}+\frac{1}{17}\log_2\frac{17\times1}{6\times5}+0\right)+\left(\frac{1}{17}\log_2\frac{17\times1}{6\times8}+\frac{4}{17}\log_2\frac{17\times4}{6\times5}+\frac{1}{17}\log_2\frac{17\times1}{6\times4}\right)+\left(\frac{2}{17}\log_2\frac{17\times2}{5\times8}+0+\frac{3}{17}\log_2\frac{17\times3}{5\times4}\right)\right)$$

$$I(\Omega,\phi)=(0.1943+0.1604+0.2107)=0.5654$$

$$H(\Omega)=-\left(\frac{6}{17}\log_2\frac{6}{17}+\frac{6}{17}\log_2\frac{6}{17}+\frac{5}{17}\log_2\frac{5}{17}\right)=1.5799$$

$$H(\phi)=-\left(\frac{8}{17}\log_2\frac{8}{17}+\frac{5}{17}\log_2\frac{5}{17}+\frac{4}{17}\log_2\frac{4}{17}\right)=1.5222$$

$$NMI=\frac{0.5654}{(1.5799+1.5222)/2}=0.3645$$

การคำนวณ NMI จะมีเทอม $(H(\Omega)+H(\phi))/2$ ที่มีแนวโน้มตามจำนวนคลัสเตอร์เพราะยิ่งจำนวนคลัสเตอร์มากก็จะทำให้ค่าในเทอมนี้มากและทำให้ NMI มีค่าน้อยลงด้วย
ถ้าค่า NMI มีค่า 0 แสดงว่าการทำคลัสเตอร์มีคุณภาพต่ำสุด แต่ถ้าค่า NMI มีค่า 1 แสดงว่าการทำคลัสเตอร์มีคุณภาพสูงสุด

4.1.4 F measure เป็นการวัดประสิทธิภาพที่ตั้งสมมุติฐานไว้ว่าจะต้องมีคลาสเดียวอยู่คลัสเตอร์เดียวกัน โดยจะทำการพิจารณาเป็นคู่ของข้อมูล เราจะกำหนดให้ 2 จุดที่อยู่คลัสเตอร์เดียวกันมีคลาสเหมือนหรือไม่เหมือนกันและดูว่าเหมือนกันจริงหรือไม่ จะทำให้เกิดกรณี 4 กรณีได้ดังนี้

(1) True Positive (TP) คือ เราตัดสินว่า 2 จุดเป็นคลาสเดียวกันและ 2 จุดนั้นเป็นคลาสเดียวกันจริง

(2) True Negative (TN) คือ เราตัดสินว่า 2 จุดเป็นไม่เป็นคลาสเดียวกันและ 2 จุดนั้นไม่เป็นคลาสเดียวกันจริง

(3) False Positive (FP) คือ เราตัดสินว่า 2 จุดเป็นไม่เป็นคลาสเดียวกันแต่ 2 จุดนั้นเป็นคลาสเดียวกัน

(4) False Negative (FN) คือ เราตัดสินว่า 2 จุดเป็นคลาสเดียวกันแต่ 2 จุดนั้นไม่เป็นคลาสเดียวกัน

ในการคำนวณ F measure เราต้องนับจำนวนที่เกิดขึ้นของทั้ง 4 กรณีแล้วนำมาคำนวณตามสมการที่ 4.7 – 4.9 [11]

$$P = \frac{TP}{TP+FP} \tag{4.7}$$

$$R = \frac{TP}{TP+FN} \tag{4.8}$$

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P+R} \tag{4.9}$$

ถ้า $\beta > 1$ จะให้ค่า FN มีผลมากกว่าค่าของ FP

จากรูปที่ 4.1 สามารถคำนวณ F measure โดยให้ $\beta = 1$ ได้ดังนี้

ตารางที่ 4.1 แสดงจำนวนของคลาสที่นับได้ในแต่ละคลัสเตอร์

Class\Cluster	คลัสเตอร์ 1	คลัสเตอร์ 2	คลัสเตอร์ 3	รวม
X	5	1	2	8
O	1	4	0	5
◊	0	1	3	4
รวม	6	6	5	17

$$TP = \binom{5}{2} + \binom{2}{2} + \binom{4}{2} + \binom{3}{2} = 20$$

$$FN = \binom{8}{2} + \binom{5}{2} + \binom{4}{2} - TP = 24$$

$$FP = \binom{6}{2} + \binom{6}{2} + \binom{5}{2} - TP = 20$$

$$TN = \binom{17}{2} - TP - FP - FN = 72$$

$$P = \frac{20}{20+20} = 0.5$$

$$R = \frac{20}{20+24} = 0.4545$$

$$F = \frac{(1^2+1)0.5 \times 0.4545}{1^2(0.5)+0.4545} = 0.4762$$

ถ้าค่า F measure มีค่า 0 แสดงว่าการทำคลัสเตอร์มีคุณภาพต่ำสุด แต่ถ้าค่า F measure มีค่า 1 แสดงว่าการทำคลัสเตอร์มีคุณภาพสูงสุด

4.2 ข้อมูลที่ใช้ในการทดลอง

4.2.1 ข้อมูลมาตรฐาน

1) Dermatology

ข้อมูล Dermatology เป็นชุดข้อมูลประกอบด้วยข้อมูลที่ได้จากการสุ่มตัวอย่างจำนวน 366 ตัวอย่าง แต่ละตัวอย่างมีแอตทริบิวจำนวน 34 แอตทริบิวและคลาสแอตทริบิวจำนวน 1 แอตทริบิว โดยจำแนกคลาสออกเป็น 6 ประเภท คือ psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis และ pityriasis rubra pilaris แต่ละประเภทประกอบด้วยข้อมูลตัวอย่างจำนวน 112, 61, 72,49,52 และ 20 ตามลำดับ มี Missing Values แสดงเป็นสัญลักษณ์ '?' ให้แทนทุกๆ Missing Values เป็นเลข 8

2) Libras Moverment

Libras Moverment เป็นชุดข้อมูลประกอบด้วยข้อมูลที่ได้จากการสุ่มตัวอย่างจำนวน 360 ตัวอย่าง แต่ละตัวอย่างมีแอตทริบิวจำนวน 90 แอตทริบิวและคลาสแอตทริบิวจำนวน 1 แอตทริบิว โดยจำแนกคลาสออกเป็น 15 ประเภท คือ curved swing, horizontal swing, vertical swing, anti-clockwise arc, clockwise arc, circle, horizontal straight-line, vertical straight-line, horizontal zigzag, vertical zigzag, horizontal wavy, vertical wavy, face-up curve, face-down curve และ tremble แต่ละประเภทประกอบด้วยข้อมูลตัวอย่าง 24 ตัวอย่าง ไม่มี Missing Values

3) Large Soybean

Large Soybean เป็นชุดข้อมูลประกอบด้วยข้อมูลที่ได้จากการสุ่มตัวอย่างจำนวน 307 ตัวอย่าง แต่ละตัวอย่างมีแอตทริบิวจำนวน 35 แอตทริบิวและคลาสแอตทริบิวจำนวน 1 แอตทริบิว โดยจำแนกคลาสออกเป็น 19 ประเภท คือ diaporthe-stem-canker, charcoal-rot, rhizoctonia-root-rot, phytophthora-rot, brown-stem-rot, powdery-mildew, downy-mildew, brown-spot, bacterial-blight, bacterial-pustule, purple-seed-stain, anthracnose, phyllosticta-leaf-spot, alternarialeaf-spot, frog-eye-leaf-spot, diaporthe-pod-&-stem-blight, cyst-nematode, 2-4-d-injury และ herbicide-injury แต่ละประเภทประกอบด้วยข้อมูลตัวอย่างจำนวน 10, 10, 10, 40, 20, 10, 10, 40, 10, 10, 10, 10, 20, 10, 40, 40, 6, 6, 1 และ 4 ตามลำดับ มี Missing Values แสดงเป็นสัญลักษณ์ '?' ให้ Missing Values ของแต่ละแอตทริบิวตามตารางที่ 4.2

ตารางที่ 4.2 แสดงเลขที่ต้องแทนในค่าที่เป็น Missing Values

ลำดับแอตทริบิว	ชื่อแอตทริบิว	จำนวนที่แทนใน Missing Values
1	date	0
2	plant-stand	1
3	precip	8
4	temp	11
5	hail	7
6	crop-hist	41
7	area-damaged	1
8	severity	1
9	seed-tmt	41
10	germination	41
11	plant-growth	36
12	leaves	1
13	leafspots-halo	0
14	leafspots-marg	25
15	leafspot-size	25
16	leaf-shread	25
17	leaf-malf	26
18	leaf-mild	25
19	stem	30
20	lodging	1
21	stem-cankers	41
22	canker-lesion	11
23	fruiting-bodies	11
24	external decay	35
25	mycelium	11
26	int-discolor	11
27	sclerotia	11

ตารางที่ 4.2 (ต่อ) แสดงเลขที่ต้องแทนในค่าที่เป็น Missing Values

ลำดับแอตทริบิว	ชื่อแอตทริบิว	จำนวนที่แทนใน Missing Values
28	fruit-pods	11
29	fruit spots	25
30	seed	35
31	mold-growth	29
32	seed-discolor	29
33	seed-size	35
34	shriveling	29
35	roots	35

4) Wine Recognition

Wine Recognition เป็นชุดข้อมูลประกอบด้วยข้อมูลที่ได้จากการสุ่มตัวอย่างจำนวน 178 ตัวอย่าง แต่ละตัวอย่างมีแอตทริบิวจำนวน 13 แอตทริบิวและคลาสแอตทริบิวจำนวน 1 แอตทริบิว โดยจำแนกคลาสออกเป็น 3 ประเภท คือ 0, 2 และ 3 แต่ละประเภทประกอบด้วยข้อมูลตัวอย่างจำนวน 59, 71 และ 48 ตามลำดับ ไม่มี Missing Values

5) Iris Plants

Iris Plants เป็นชุดข้อมูลประกอบด้วยข้อมูลที่ได้จากการสุ่มตัวอย่างจำนวน 150 ตัวอย่าง แต่ละตัวอย่างมีแอตทริบิวจำนวน 4 แอตทริบิวและคลาสแอตทริบิวจำนวน 1 แอตทริบิวโดยจำแนกคลาสออกเป็น 3 ประเภท คือ Setosa, Versicolour และ Virginica แต่ละประเภทประกอบด้วยข้อมูลตัวอย่างจำนวน 50 ตัวอย่าง ไม่มี Missing Values

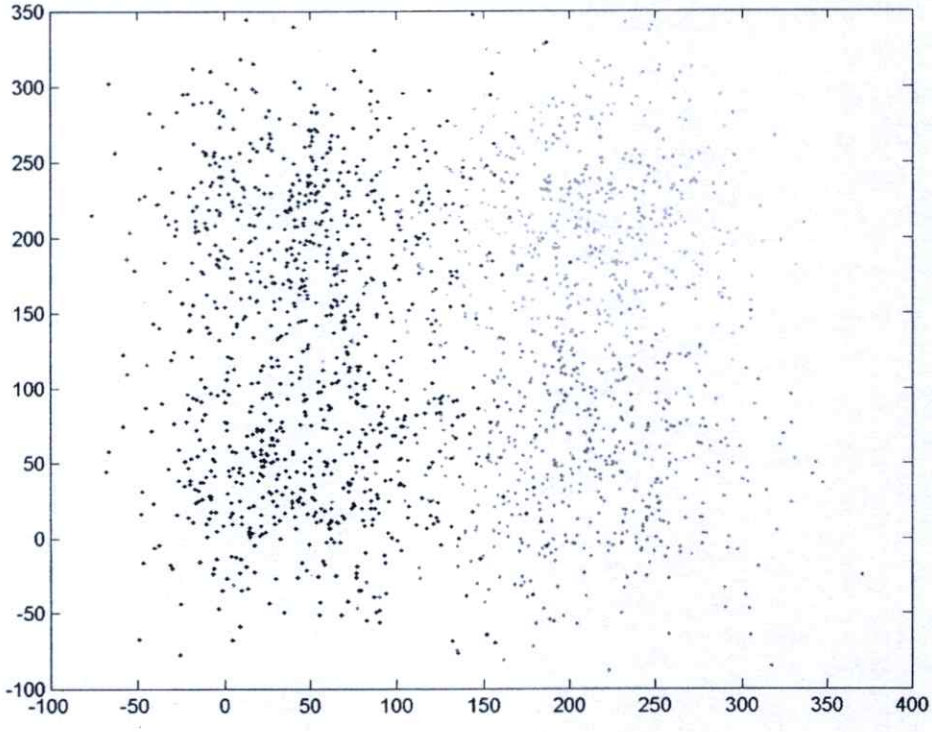
6) BUPA liver disorders

BUPA liver disorders เป็นชุดข้อมูลประกอบด้วยข้อมูลที่ได้จากการสุ่มตัวอย่างจำนวน 346 ตัวอย่าง แต่ละตัวอย่างมีแอตทริบิวจำนวน 6 แอตทริบิวและคลาสแอตทริบิวจำนวน 1 แอตทริบิวโดยจำแนกคลาสออกเป็น 2 ประเภท คือ 1 และ 2 ไม่มี Missing Values

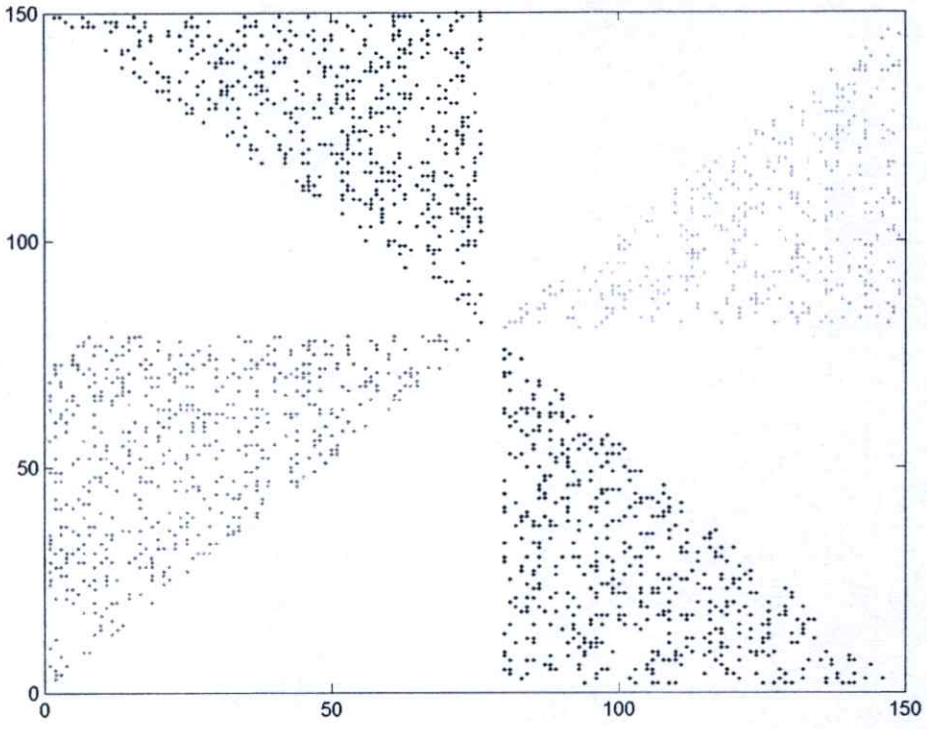
4.2.2 ข้อมูลที่ผู้วิจัยสร้างขึ้นเอง

ข้อมูลที่ผู้ทำวิจัยสร้างขึ้นเองมีทั้งหมด 4 ชุดข้อมูลดังรูปที่ 4.2 – 4.5 ข้อมูลแต่ละรูปมีขนาดแตกต่างกันดังนี้ Normal Distribution ขนาด 500 x 500 pixel, Fan ขนาด 150 x 150 pixel, Ring

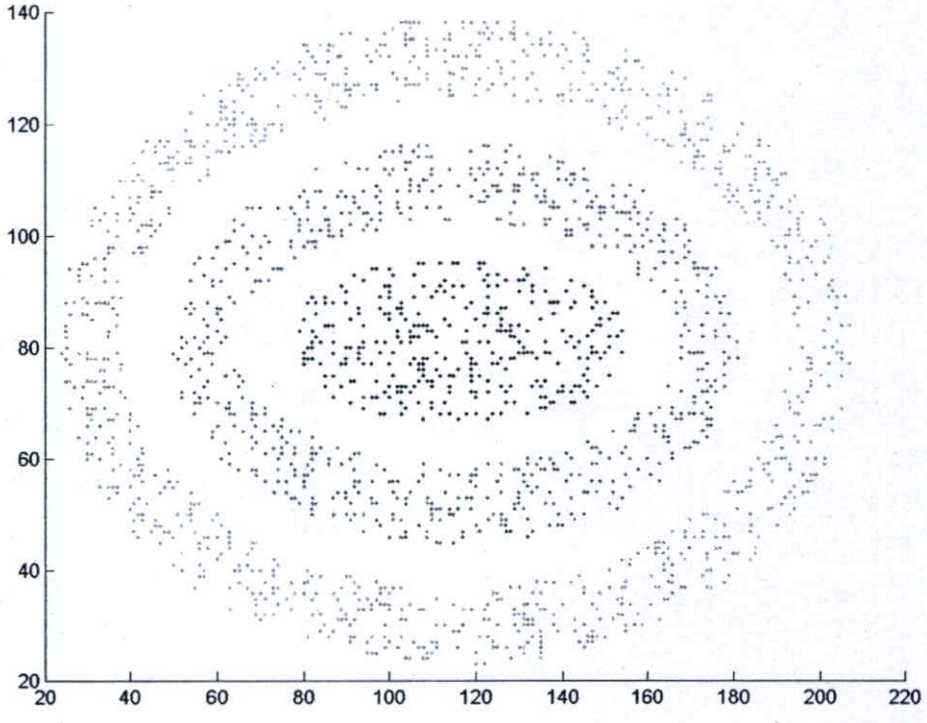
ขนาด 245 x 163 pixel และ Shape ขนาด 284 x 176 pixel โดยแต่ละรูปแบบอินพุทได้จากการสุ่มคู่อันดับ (x, y) จากรูปที่สร้างขึ้น โดยมีการสุ่มรูปแบบอินพุทขึ้นมาทั้งหมด 2000 รูปแบบอินพุท



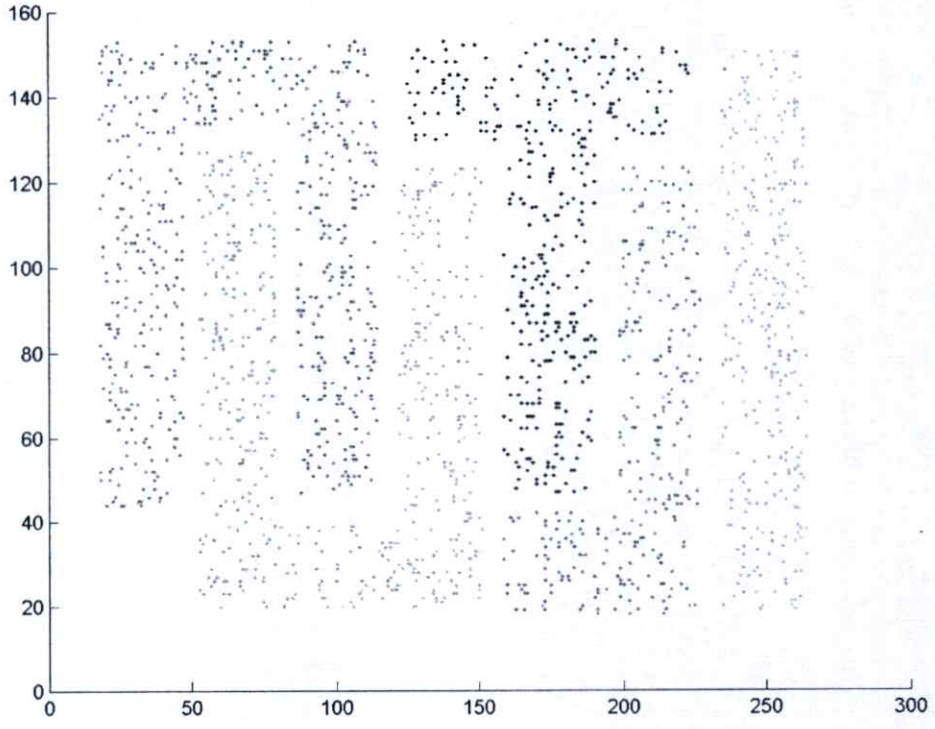
รูปที่ 4.2 แสดงลักษณะของข้อมูล Normal Distribution ที่ผู้ทำวิจัยสร้างขึ้นเอง



รูปที่ 4.3 แสดงลักษณะของข้อมูล Fan ที่ผู้ทำวิจัยสร้างขึ้นเอง



รูปที่ 4.4 แสดงลักษณะของข้อมูล Ring ที่ผู้ทำวิจัยสร้างขึ้นเอง



รูปที่ 4.5 แสดงลักษณะของข้อมูล Shape ที่ผู้ทำวิจัยสร้างขึ้นเอง

4.3 การกำหนดพารามิเตอร์

4.3.1 พารามิเตอร์ในงานวิจัยที่นำเสนอ

1. ค่าที่กำหนดให้จุดศูนย์กลางหยุดการเคลื่อนที่ (threshold) = 0.001 ในทุกๆปัญหาที่ทำการทดลอง

2. ค่าเกณฑ์ในการลบคลัสเตอร์ (γ) จะต้องทำหาค่า γ ที่เหมาะสมที่สุดในแต่ละปัญหา ในการทดลองจะทำการเรียนรู้โมเดลของงานวิจัยที่นำเสนอโดยเลือกค่า γ ในช่วงของจำนวน 0.01 ถึง 0.3 โดยเพิ่มทีละ 0.01 แล้วเลือกค่า γ ที่ให้ค่าประสิทธภาพดีที่สุด

4.3.2 พารามิเตอร์ใน K-means และ C-means

ค่าจำนวนกลุ่มที่ต้องการ K กลุ่ม เป็นพารามิเตอร์ของ K-means และ C-means จะต้องทำการหาค่า K ที่เหมาะสมที่สุดในแต่ละปัญหา ในการทดลองนี้จะทำการเรียนรู้โมเดลของ K-means และ C-means โดยเลือกค่า K ในช่วงของจำนวนคลาสทั้งหมดในเซตของข้อมูลไปถึง 10 เปอร์เซ็นของ

จำนวนตัวอย่างของเซตข้อมูลทั้งหมดโดยเพิ่มทีละ 1 แล้วเลือก K ที่ทำให้ได้ค่าของตัววัดประสิทธิภาพแต่ละแบบมีค่าดีที่สุดนำไปเปรียบเทียบกับงานวิจัยที่นำเสนอ ซึ่งในแต่ละปัญหามีช่วงจำนวนที่ต้องเลือกในตารางที่ 4.3 กับตารางที่ 4.4

ตารางที่ 4.3 แสดงช่วงพารามิเตอร์จำนวนกลุ่มในแต่ละปัญหาของ K-mean และ C-mean ของชุดข้อมูลมาตรฐาน

ชุดข้อมูล	ช่วงที่เลือกค่าจำนวนกลุ่ม
Dermatology	6 - 37
Libras Movement	15 - 36
Large Soybean	19 - 31
Wine Recognition	3 - 18
Iris Plants	3 - 15
BUPA liver disorders	2 - 35

ตารางที่ 4.4 แสดงช่วงพารามิเตอร์จำนวนกลุ่มในแต่ละปัญหาของ K-mean และ C-mean ของข้อมูลที่ผู้วิจัยสร้างขึ้นเอง

ชุดข้อมูล	ช่วงที่เลือกค่าจำนวนกลุ่ม
Normal Distribution	4 - 200
Fan	4 - 200
Ring	3 - 200
Shape	3 - 200

4.3.3 พารามิเตอร์ใน DBSCAN

จำนวนขั้นต่ำของจุดภายในกลุ่ม (MinPts) เป็นพารามิเตอร์ของ DBSCAN จะต้องทำการหาค่า MinPts ที่เหมาะสมที่สุด ในการทดลองนี้จะทำการเรียนรู้โมเดล DBSCAN โดยเลือกค่า MinPts ในช่วง 2 ถึง 20 โดยเพิ่มทีละ 1 แล้วเลือก MinPts ที่ทำให้ได้ค่าของตัววัดประสิทธิภาพแต่ละแบบมีค่าดีที่สุดนำไปเปรียบเทียบกับงานวิจัยที่นำเสนอ

4.4 ผลการทดลอง

4.4.1 การทดลองที่ 1

ในการทดลองนี้จะนำงานวิจัยนี้เปรียบเทียบกับอัลกอริทึม K-means, C-means และ DBSACN โดยเลือกค่าของพารามิเตอร์ที่เหมาะสมที่สุดของแต่ละอัลกอริทึมมาทดลอง ผลที่ได้จากการทดลองจะแสดงในลักษณะค่าเฉลี่ยของแต่ละตัววัดประสิทธิภาพ ซึ่งเป็นค่าเฉลี่ยที่ได้จากการทดลองที่มีการกำหนดค่าพารามิเตอร์และชุดข้อมูลที่เหมือนกันทั้งหมด 15 ครั้ง

4.4.1.1 ชุดข้อมูลมาตรฐาน

1) ชุดข้อมูล Dermatology

ค่าของตัวแปรที่ใช้ในงานวิจัยที่นำเสนอ $\gamma =$ 0.06 ค่าของตัวแปรที่ใช้ใน DBSCAN คือ MinPts มีค่าดังนี้ ตัววัดความบริสุทธิ์เท่ากับ 2, ตัววัดเอนโทรปีเท่ากับ 2, ตัววัด NMI เท่ากับ 17 และตัววัด F measure เท่ากับ 17 โดยผลการทดลองแสดงได้ดังตารางที่ 4.5 – 4.6

ตารางที่ 4.5 แสดงผลการทดลองของข้อมูล Dermatology เมื่อเปรียบเทียบกับ K-means, C-means และ DBSCAN โดยใช้ตัววัดประสิทธิภาพความบริสุทธิ์และเอนโทรปี

โมเดล	ความบริสุทธิ์			เอนโทรปี		
	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.8056	40	4.9176	0.3252	40	4.9176
K-means	0.7095	37	0.1030	0.5223	36	0.0749
C-means	0.6033	37	53.3244	0.9170	35	20.3059
DBSCAN	0.5710	28	0.0627	0.2016	28	0.0793

ตารางที่ 4.6 แสดงผลการทดลองของข้อมูล Dermatology เมื่อเปรียบเทียบกับ K-means, C-means และ DBSCAN โดยใช้ตัววัดประสิทธิภาพ NMI และ F measure

โมเดล	NMI			F measure		
	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.4913	40	4.9176	0.1359	40	4.9176
K-means	0.4394	37	0.0759	0.2186	6	0.0677
C-means	0.3260	37	35.0874	0.2049	6	0.4848
DBSCAN	0.6836	5	0.0862	0.5320	5	0.0970

2) ชุดข้อมูล Libras Movement

ค่าของตัวแปรที่ใช้ในงานวิจัยที่นำเสนอ $\gamma =$ 0.06 ค่าของตัวแปรที่ใช้ใน DBSCAN คือ MinPts มีค่าดังนี้ ตัววัดความบริสุทธิ์เท่ากับ 2, ตัววัดเอนโทรปีเท่ากับ 2, ตัววัด NMI เท่ากับ 2 และตัววัด F measure เท่ากับ 2 โดยผลการทดลองแสดงได้ดังตารางที่ 4.7 – 4.8

ตารางที่ 4.7 แสดงผลการทดลองของข้อมูล Libras Movement เมื่อเปรียบเทียบกับ K-means, C-means และ DBSCAN โดยใช้ตัววัดประสิทธิภาพ NMI และ F measure

โมเดล	ความบริสุทธิ์			เอนโทรปี		
	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.8718	48	14.7142	0.2935	48	14.7142
K-means	0.5341	36	0.1544	0.6816	36	0.1072
C-means	0.1448	22	3.0870	0.3030	36	3.3121
DBSCAN	0.0667	1	0.1146	3.9069	1	0.1369

ตารางที่ 4.8 แสดงผลการทดลองของข้อมูล Libras Movement เมื่อเปรียบเทียบกับ K-means, C-means และ DBSCAN โดยใช้ตัววัดประสิทธิภาพ NMI และ F measure

โมเดล	NMI			F measure		
	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.6728	48	14.7142	0.1192	48	14.7142
K-means	0.6120	35	0.1066	0.3850	16	0.0886
C-means	0.1279	22	1.9555	0.2542	18	1.7385
DBSCAN	0.0000	1	0.1438	0.2550	1	0.1534

3) ชุดข้อมูล Large Soybean

ค่าของตัวแปรที่ใช้ในงานวิจัยที่นำเสนอ $\gamma =$ 0.07 ค่าของตัวแปรที่ใช้ใน DBSCAN คือ MinPts มีค่าดังนี้ ตัววัดความบริสุทธิ์เท่ากับ 2, ตัววัดเอนโทรปีเท่ากับ 2, ตัววัด NMI เท่ากับ 2 และตัววัด F measure เท่ากับ 12 โดยผลการทดลองแสดงได้ดังตารางที่ 4.9 – 4.10

ตารางที่ 4.9 แสดงผลการทดลองของข้อมูล Large Soybean เมื่อเปรียบเทียบกับ K-means, C-means และ DBSCAN โดยใช้ตัววัดประสิทธิภาพความบริสุทธิ์และเอนโทรปี

โมเดล	ความบริสุทธิ์			เอนโทรปี		
	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.8684	42	5.4320	0.4235	42	5.4320
K-means	0.2862	30	0.0324	0.1537	31	0.0207
C-means	0.5199	31	21.0647	0.2947	30	16.7452
DBSCAN	0.2573	6	0.0536	0.6067	6	0.0527

ตารางที่ 4.10 แสดงผลการทดลองของข้อมูล Large Soybean เมื่อเปรียบเทียบกับ K-means, C-means และ DBSCAN โดยใช้ตัววัดประสิทธิภาพ NMI และ F measure

โมเดล	NMI			F measure		
	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.7233	42	5.4320	0.2217	42	5.4320
K-means	0.3666	26	0.0215	0.3850	26	0.0213
C-means	0.5774	25	14.7703	0.4391	29	17.2343
DBSCAN	0.2727	6	0.0650	0.3360	2	0.0425

4) ชุดข้อมูล Wine Recognition

ค่าของตัวแปรที่ใช้ในงานวิจัยที่นำเสนอ $\gamma =$ 0.17 ค่าของตัวแปรที่ใช้ใน DBSCAN คือ MinPts มีค่าดังนี้ ตัววัดความบริสุทธิ์เท่ากับ 2, ตัววัดเอนโทรปีเท่ากับ 2, ตัววัด NMI เท่ากับ 3 และตัววัด F measure เท่ากับ 3 โดยผลการทดลองแสดงได้ดังตารางที่ 4.11 – 4.12

ตารางที่ 4.11 แสดงผลการทดลองของข้อมูล Wine Recognition เมื่อเปรียบเทียบกับ K-means, C-means และ DBSCAN โดยใช้ตัววัดประสิทธิภาพความบริสุทธิ์และเอนโทรปี

โมเดล	ความบริสุทธิ์			เอนโทรปี		
	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.6854	2	1.4324	0.9380	2	1.4324
K-means	0.7307	18	0.0405	0.4497	14	0.0261
C-means	0.7311	7	0.5811	0.4967	18	1.9644
DBSCAN	0.1236	11	0.0167	0.0909	11	0.0192

ตารางที่ 4.12 แสดงผลการทดลองของข้อมูล Wine Recognition เมื่อเปรียบเทียบกับ K-means, C-means และ DBSCAN โดยใช้ตัววัดประสิทธิภาพ NMI และ F measure

โมเดล	NMI			F measure		
	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.4536	2	1.4324	0.7766	2	1.4324
K-means	0.4248	3	0.0283	0.6240	3	0.0300
C-means	0.4168	3	0.1610	0.5736	3	0.1742
DBSCAN	0.8891	2	0.0209	0.4348	2	0.0192

5) ชุดข้อมูล Iris Plants

ค่าของตัวแปรที่ใช้ในงานวิจัยที่นำเสนอ $\gamma =$ 0.21 ค่าของตัวแปรที่ใช้ใน DBSCAN คือ MinPts มีค่าดังนี้ ตัววัดความบริสุทธิ์เท่ากับ 2, ตัววัดเอนโทรปีเท่ากับ 2, ตัววัด NMI เท่ากับ 3 และตัววัด F measure เท่ากับ 3 โดยผลการทดลองแสดงได้ดังตารางที่ 4.13 – 4.14

ตารางที่ 4.13 แสดงผลการทดลองของข้อมูล Iris Plants เมื่อเปรียบเทียบกับ K-means, C-means และ DBSCAN โดยใช้ตัววัดประสิทธิภาพความบริสุทธิ์และเอนโทรปี

โมเดล	ความบริสุทธิ์			เอนโทรปี		
	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.8258	3	0.5450	0.3921	3	0.5450
K-means	0.9618	13	0.0249	0.0803	15	0.0184
C-means	0.9787	15	0.5733	0.0814	15	0.3779
DBSCAN	0.6800	3	0.0113	0.3332	3	0.0137

ตารางที่ 4.14 แสดงผลการทดลองของข้อมูล Iris Plants เมื่อเปรียบเทียบกับ K-means, C-means และ DBSCAN โดยใช้ตัววัดประสิทธิภาพ NMI และ F measure

โมเดล	NMI			F measure		
	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.7597	3	0.5450	0.8558	3	0.5450
K-means	0.7449	3	0.0142	0.8272	3	0.0145
C-means	0.7496	3	0.0529	0.8263	3	0.0551
DBSCAN	0.7337	2	0.0176	0.8802	2	0.0146

6) ชุดข้อมูล Bupa Liver Disorders

ค่าของตัวแปรที่ใช้ในงานวิจัยที่นำเสนอ $\gamma =$ 0.05 ค่าของตัวแปรที่ใช้ใน DBSCAN คือ MinPts มีค่าดังนี้ ตัววัดความบริสุทธิ์เท่ากับ 4, ตัววัดเอนโทรปีเท่ากับ 2, ตัววัด NMI เท่ากับ 2 และตัววัด F measure เท่ากับ 4 โดยผลการทดลองแสดงได้ดังตารางที่ 4.15 – 4.16

ตารางที่ 4.15 แสดงผลการทดลองของข้อมูล Bupa Liver Disorders เมื่อเปรียบเทียบกับ K-means, C-means และ DBSCAN โดยใช้ตัววัดประสิทธิภาพความบริสุทธิ์และเอนโทรปี

โมเดล	ความบริสุทธิ์			เอนโทรปี		
	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.8303	53	4.3526	0.1609	53	4.3526
K-means	0.6286	33	0.1056	0.2951	31	0.0932
C-means	0.6895	35	52.5367	0.7509	35	53.2899
DBSCAN	0.5565	2	0.0403	0.7466	4	0.0444

ตารางที่ 4.16 แสดงผลการทดลองของข้อมูล Bupa Liver Disorders เมื่อเปรียบเทียบกับ K-means, C-means และ DBSCAN โดยใช้ตัววัดประสิทธิภาพ NMI และ F measure

โมเดล	NMI			F measure		
	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.1399	53	4.3526	0.0247	53	4.3526
K-means	0.0348	33	0.0761	0.7212	2	0.0300
C-means	0.0362	20	5.6655	0.6736	2	0.1833
DBSCAN	0.0900	4	0.0325	0.8251	2	0.0320

4.4.1.2 ชุดข้อมูลที่ผู้ทำวิจัยสร้างขึ้นเอง

1) ชุดข้อมูล Normal Distribution

ค่าของตัวแปรที่ใช้ในงานวิจัยที่นำเสนอ $\gamma =$ 0.09 ค่าของตัวแปรที่ใช้ใน DBSCAN คือ MinPts มีค่าดังนี้ ตัววัดความบริสุทธิ์เท่ากับ 2, ตัววัดเอนโทรปีเท่ากับ 3, ตัววัด NMI เท่ากับ 2 และตัววัด F measure เท่ากับ 8 โดยผลการทดลองแสดงได้ดังตารางที่ 4.17 – 4.18

ตารางที่ 4.17 แสดงผลการทดลองของข้อมูล Normal Distribution เมื่อเปรียบเทียบกับ K-means, C-means และ DBSCAN โดยใช้ตัววัดประสิทธิภาพความบริสุทธิ์และเอนโทรปี

โมเดล	ความบริสุทธิ์			เอนโทรปี		
	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.8786	4	28.3221	0.6659	4	28.3221
K-means	0.8793	4	0.6110	0.4725	30	1.4403
C-means	0.8805	4	2.1151	0.4715	30	114.5556
DBSCAN	0.8020	138	0.4042	0.2929	40	0.3591

ตารางที่ 4.18 แสดงผลการทดลองของข้อมูล Normal Distribution เมื่อเปรียบเทียบกับ K-means, C-means และ DBSCAN โดยใช้ตัววัดประสิทธิภาพ NMI และ F measure

โมเดล	NMI			F measure		
	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.6659	4	28.3221	0.7793	4	28.3221
K-means	0.6637	4	0.2863	0.7796	4	0.2897
C-means	0.6660	4	1.1647	0.7816	4	1.1674
DBSCAN	0.4602	138	0.4110	0.6217	2	0.3642

2) ชุดข้อมูล Fan

ค่าของตัวแปรที่ใช้ในงานวิจัยที่นำเสนอ $\gamma =$ 0.01 ค่าของตัวแปรที่ใช้ใน DBSCAN คือ MinPts มีค่าดังนี้ ตัววัดความบริสุทธิ์เท่ากับ 4, ตัววัดเอนโทรปีเท่ากับ 2, ตัววัด NMI เท่ากับ 4 และตัววัด F measure เท่ากับ 4 โดยผลการทดลองแสดงได้ดังตารางที่ 4.19 – 4.20

ตารางที่ 4.19 แสดงผลการทดลองของข้อมูล Fan เมื่อเปรียบเทียบกับ K-means, C-means และ DBSCAN โดยใช้ตัววัดประสิทธิภาพความบริสุทธิ์และเอนโทรปี

โมเดล	ความบริสุทธิ์			เอนโทรปี		
	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.9980	4	15.6151	0.0197	4	15.6151
K-means	0.9877	25	1.7381	0.0378	30	1.1899
C-means	0.9975	4	1.3146	0.0166	12	13.1596
DBSCAN	0.9985	6	0.3564	0.0000	95	0.3552

ตารางที่ 4.20 แสดงผลการทดลองของข้อมูล Fan เมื่อเปรียบเทียบกับ K-means, C-means และ DBSCAN โดยใช้ตัววัดประสิทธิภาพ NMI และ F measure

โมเดล	NMI			F measure		
	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.9917	4	15.6151	0.9962	4	15.6151
K-means	0.9693	4	0.2339	0.9715	4	0.2359
C-means	0.9887	4	0.6747	0.9949	4	0.6794
DBSCAN	0.9835	6	0.3522	0.9840	6	0.3552

3) ชุดข้อมูล Ring

ค่าของตัวแปรที่ใช้ในงานวิจัยที่นำเสนอ $\gamma =$ 0.09 ค่าของตัวแปรที่ใช้ใน DBSCAN คือ MinPts มีค่าดังนี้ ตัววัดความบริสุทธิ์เท่ากับ 6, ตัววัดเอนโทรปีเท่ากับ 2, ตัววัด NMI เท่ากับ 5 และตัววัด F measure เท่ากับ 5 โดยผลการทดลองแสดงได้ดังตารางที่ 4.21 – 4.22

ตารางที่ 4.21 แสดงผลการทดลองของข้อมูล Ring เมื่อเปรียบเทียบกับ K-means, C-means และ DBSCAN โดยใช้ตัววัดประสิทธิภาพความบริสุทธิ์และเอนโทรปี

โมเดล	ความบริสุทธิ์			เอนโทรปี		
	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.9665	27	15.6546	0.1699	27	15.6546
K-means	0.9736	30	1.7475	0.0764	30	1.3754
C-means	0.9699	30	92.0679	0.1176	30	85.4087
DBSCAN	1.0000	3	0.3542	0.0000	178	0.3606

ตารางที่ 4.22 แสดงผลการทดลองของข้อมูล Ring เมื่อเปรียบเทียบกับ K-means, C-means และ DBSCAN โดยใช้ตัววัดประสิทธิภาพ NMI และ F measure

โมเดล	NMI			F measure		
	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.4140	27	15.6546	0.1100	27	15.6546
K-means	0.4324	30	0.9765	0.3568	3	0.5874
C-means	0.4191	30	55.0363	0.3651	3	2.8076
DBSCAN	1.0000	3	0.3555	1.0000	3	0.3585

4) ชุดข้อมูล Shape

ค่าของตัวแปรที่ใช้ในงานวิจัยที่นำเสนอ $\gamma =$ 0.09 ค่าของตัวแปรที่ใช้ใน DBSCAN คือ MinPts มีค่าดังนี้ ตัววัดความบริสุทธิ์เท่ากับ 4, ตัววัดเอนโทรปีเท่ากับ 2, ตัววัด NMI เท่ากับ 5 และตัววัด F measure เท่ากับ 10 โดยผลการทดลองแสดงได้ดังตารางที่ 4.23 – 4.24

ตารางที่ 4.23 แสดงผลการทดลองของข้อมูล Shape เมื่อเปรียบเทียบกับ K-means, C-means และ DBSCAN โดยใช้ตัววัดประสิทธิภาพความบริสุทธิ์และเอนโทรปี

โมเดล	ความบริสุทธิ์			เอนโทรปี		
	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.9738	24	17.9276	0.1497	24	17.9276
K-means	0.9763	30	1.7363	0.0684	30	1.1194
C-means	0.9663	28	46.9909	0.1300	29	60.6791
DBSCAN	0.9640	42	0.3712	0.0000	263	0.3732

ตารางที่ 4.24 แสดงผลการทดลองของข้อมูล Shape เมื่อเปรียบเทียบกับ K-means, C-means และ DBSCAN โดยใช้ตัววัดประสิทธิภาพ NMI และ F measure

โมเดล	NMI			F measure		
	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)	ถูกต้อง	คลัสเตอร์	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.4471	24	17.9276	0.1340	24	17.9276
K-means	0.4503	30	0.7940	0.4020	3	0.8120
C-means	0.4322	29	39.0475	0.3922	4	11.7023
DBSCAN	0.6592	10	0.3655	0.6577	2	0.3553

4.4.2 การทดลองที่ 2

ในการทดลองนี้จะนำงานวิจัยนี้เปรียบเทียบกับอัลกอริทึม K-means และ C-means โดยจะทำการเรียนรู้โมเดลงานวิจัยที่นำเสนอก่อน จากนั้นนำค่าจำนวนคลัสเตอร์เฉลี่ยที่ได้ไปเป็นพารามิเตอร์ให้กับอัลกอริทึม K-means และ C-means แล้วทำการทดลอง ผลที่ได้จากการทดลองจะแสดงในลักษณะค่าเฉลี่ยของแต่ละตัววัดประสิทธิภาพ ซึ่งเป็นค่าเฉลี่ยที่ได้จากการทดลองที่มีการกำหนดค่าพารามิเตอร์และชุดข้อมูลที่เหมือนกันทั้งหมด 15 ครั้ง

4.4.2.1 ชุดข้อมูลมาตรฐาน

1) ชุดข้อมูล Dermatology

ค่าของตัวแปรที่ใช้ในงานวิจัยที่นำเสนอ $\gamma =$ 0.06 ค่าจำนวนคลัสเตอร์เท่ากับ 40 โดยผลการทดลองแสดงได้ดังตารางที่ 4.25

ตารางที่ 4.25 แสดงผลการทดลองของข้อมูล Dermatology เมื่อเปรียบเทียบกับ K-means และ C-means โดยใช้ตัววัดประสิทธิภาพความบริสุทธิ์, เอนโทรปี, NMI และ F measure

โมเดล	ความบริสุทธิ์	เอนโทรปี	NMI	F measure	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.8056	0.3253	0.4913	0.1359	4.9176
K-means	0.7097	0.5021	0.4392	0.1481	0.1321
C-means	0.6153	0.7882	0.3352	0.1149	41.6252

2) ชุดข้อมูล Libras Movement

ค่าของตัวแปรที่ใช้ในงานวิจัยที่นำเสนอ $\gamma =$ 0.06 ค่าจำนวนคลัสเตอร์เท่ากับ 48 โดยผลการทดลองแสดงได้ดังตารางที่ 4.26

ตารางที่ 4.26 แสดงผลการทดลองของข้อมูล Libras Movement เมื่อเปรียบเทียบกับ K-means และ C-means โดยใช้ตัววัดประสิทธิภาพความบริสุทธิ์, เอนโทรปี, NMI และ F measure

โมเดล	ความบริสุทธิ์	เอนโทรปี	NMI	F measure	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.8719	0.2935	0.6728	0.1193	14.7142
K-means	0.5537	0.5560	0.6149	0.3168	0.1898
C-means	0.1443	0.2342	0.1205	0.2475	7.5423

3) ชุดข้อมูล Large Soybean

ค่าของตัวแปรที่ใช้ในงานวิจัยที่นำเสนอ $\gamma =$ 0.07 ค่าจำนวนคลัสเตอร์เท่ากับ 42 โดยผลการทดลองแสดงได้ดังตารางที่ 4.27

ตารางที่ 4.27 แสดงผลการทดลองของข้อมูล Large Soybean เมื่อเปรียบเทียบกับ K-means และ C-means โดยใช้ตัววัดประสิทธิภาพความบริสุทธิ์, เอนโทรปี, NMI และ F measure

โมเดล	ความบริสุทธิ์	เอนโทรปี	NMI	F measure	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.8684	0.4235	0.7233	0.2217	5.4320
K-means	0.2884	0.1127	0.3548	0.3741	0.0447
C-means	0.5181	0.2533	0.5749	0.4140	38.1557

4) ชุดข้อมูล Wine Recognition

ค่าของตัวแปรที่ใช้ในงานวิจัยที่นำเสนอ $\gamma = 0.17$ ค่าจำนวนคลัสเตอร์เท่ากับ 2 โดยผลการทดลองแสดงได้ดังตารางที่ 4.28

ตารางที่ 4.28 แสดงผลการทดลองของข้อมูล Wine Recognition เมื่อเปรียบเทียบกับ K-means และ C-means โดยใช้ตัววัดประสิทธิภาพความบริสุทธิ์, เอนโทรปี, NMI และ F measure

โมเดล	ความบริสุทธิ์	เอนโทรปี	NMI	F measure	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.6854	0.9380	0.4536	0.7766	1.4324
K-means	0.6573	0.9056	0.4210	0.7512	0.0287
C-means	0.6573	0.9362	0.4073	0.7469	0.1016

5) ชุดข้อมูล Iris Plants

ค่าของตัวแปรที่ใช้ในงานวิจัยที่นำเสนอ $\gamma = 0.21$ ค่าจำนวนคลัสเตอร์เท่ากับ 3 โดยผลการทดลองแสดงได้ดังตารางที่ 4.29

ตารางที่ 4.29 แสดงผลการทดลองของข้อมูล Iris Plants เมื่อเปรียบเทียบกับ K-means และ C-means โดยใช้ตัววัดประสิทธิภาพความบริสุทธิ์, เอนโทรปี, NMI และ F measure

โมเดล	ความบริสุทธิ์	เอนโทรปี	NMI	F measure	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.8258	0.3921	0.7597	0.8558	0.5450
K-means	0.8169	0.3951	0.7131	0.8193	0.0223
C-means	0.8933	0.3794	0.7496	0.8263	0.0964

6) ชุดข้อมูล Bupa Liver Disorders

ค่าของตัวแปรที่ใช้ในงานวิจัยที่นำเสนอ $\gamma =$ 0.05 ค่าจำนวนคลัสเตอร์เท่ากับ 53 โดยผลการทดลองแสดงได้ดังตารางที่ 4.30

ตารางที่ 4.30 แสดงผลการทดลองของข้อมูล Bupa Liver Disorders เมื่อเปรียบเทียบกับ K-means และ C-means โดยใช้ตัววัดประสิทธิภาพความบริสุทธิ์, เอนโทรปี, NMI และ F measure

โมเดล	ความบริสุทธิ์	เอนโทรปี	NMI	F measure	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.8303	0.1609	0.1399	0.0247	4.3526
K-means	0.6280	0.2245	0.0361	0.1964	0.1498
C-means	0.7121	0.6014	0.0633	0.0383	125.4623

4.4.2.2 ชุดข้อมูลที่ผู้ทำวิจัยสร้างขึ้นเอง

1) ชุดข้อมูล Normal Distribution

ค่าของตัวแปรที่ใช้ในงานวิจัยที่นำเสนอ $\gamma = 0.09$ ค่าจำนวนคลัสเตอร์เท่ากับ 4 โดยผลการทดลองแสดงได้ดังตารางที่ 4.31

ตารางที่ 4.31 แสดงผลการทดลองของข้อมูล Normal Distribution เมื่อเปรียบเทียบกับ K-means และ C-means โดยใช้ตัววัดประสิทธิภาพความบริสุทธิ์, เอนโทรปี, NMI และ F measure

โมเดล	ความบริสุทธิ์	เอนโทรปี	NMI	F measure	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.8786	0.6659	0.6659	0.7793	28.3221
K-means	0.8793	0.6723	0.6638	0.7798	0.4544
C-means	0.8805	0.6678	0.6660	0.7817	1.5738

2) ชุดข้อมูล Fan

ค่าของตัวแปรที่ใช้ในงานวิจัยที่นำเสนอ $\gamma = 0.01$ ค่าจำนวนคลัสเตอร์เท่ากับ 4 โดยผลการทดลองแสดงได้ดังตารางที่ 4.32

ตารางที่ 4.32 แสดงผลการทดลองของข้อมูล Fan เมื่อเปรียบเทียบกับ K-means และ C-means โดยใช้ตัววัดประสิทธิภาพความบริสุทธิ์, เอนโทรปี, NMI และ F measure

โมเดล	ความบริสุทธิ์	เอนโทรปี	NMI	F measure	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.9980	0.0197	0.9917	0.9962	15.6151
K-means	0.9856	0.0247	0.9831	0.9852	0.4386
C-means	0.9975	0.0246	0.9887	0.9949	0.9631

3) ชุดข้อมูล Ring

ค่าของตัวแปรที่ใช้ในงานวิจัยที่นำเสนอ $\gamma = 0.09$ ค่าจำนวนคลัสเตอร์เท่ากับ 27 โดยผลการทดลองแสดงได้ดังตารางที่ 4.33

ตารางที่ 4.33 แสดงผลการทดลองของข้อมูล Ring เมื่อเปรียบเทียบกับ K-means และ C-means โดยใช้ตัววัดประสิทธิภาพความบริสุทธิ์, เอนโทรปี, NMI และ F measure

โมเดล	ความบริสุทธิ์	เอนโทรปี	NMI	F measure	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.9665	0.1699	0.4140	0.1100	15.6546
K-means	0.9601	0.1275	0.4223	0.1128	1.1806
C-means	0.9613	0.1654	0.4135	0.1099	74.1465

4) ชุดข้อมูล Shape

ค่าของตัวแปรที่ใช้ในงานวิจัยที่นำเสนอ $\gamma =$ 0.09 ค่าจำนวนคลัสเตอร์เท่ากับ 24 โดยผลการทดลองแสดงได้ดังตารางที่ 4.34

ตารางที่ 4.34 แสดงผลการทดลองของข้อมูล Shape เมื่อเปรียบเทียบกับ K-means และ C-means โดยใช้ตัววัดประสิทธิภาพความบริสุทธิ์, เอนโทรปี, NMI และ F measure

โมเดล	ความบริสุทธิ์	เอนโทรปี	NMI	F measure	เวลา (วินาที)
งานวิจัยที่นำเสนอ	0.9738	0.1497	0.4471	0.1340	17.9276
K-means	0.9524	0.1559	0.4381	0.1318	1.2642
C-means	0.9398	0.2337	0.4148	0.1266	46.5891

4.4.3 การทดลองที่ 3

ในการทดลองนี้จะทำการวิเคราะห์ Sensitivity โดยการเปลี่ยนพารามิเตอร์ γ เป็นค่าต่างๆ เพื่อจะแสดงให้เห็นว่าพารามิเตอร์ γ มีผลต่องานวิจัยที่นำเสนอมากน้อยเพียงใด โดยค่าพารามิเตอร์ γ จะแบ่งการทดลองออกเป็น 5 รูปแบบ คือ

Model 1 งานวิจัยที่นำเสนอโดยใส่ค่า γ ที่เหมาะสมที่สุดเหมือนกับในการทดลองที่ 1 และ 2

Model 2 งานวิจัยที่นำเสนอโดยใส่ค่า γ เพิ่ม 10% จากค่า γ ที่เหมาะสมที่สุด

Model 3 งานวิจัยที่นำเสนอโดยใส่ค่า γ ลด 10% จากค่า γ ที่เหมาะสมที่สุด

Model 4 งานวิจัยที่นำเสนอโดยใส่ค่า γ เพิ่ม 20% จากค่า γ ที่เหมาะสมที่สุด

Model 5 งานวิจัยที่นำเสนอโดยใส่ค่า γ ลด 20% จากค่า γ ที่เหมาะสมที่สุด

4.4.3.1 ชุดข้อมูลมาตรฐาน

1) ชุดข้อมูล Dermatology

ผลการทดลองแสดงได้ดังตารางที่ 4.35 และอัตราร้อยละความแตกต่างตัววัดประสิทธิภาพแต่ละตัวเมื่อเปรียบเทียบกับ Model 1 แสดงดังตารางที่ 4.36

ตารางที่ 4.35 แสดงผลการทดลอง Sensitivity ของข้อมูล Dermatology

โมเดล	γ	ความบริสุทธิ์	เอนโทรปี	NMI	F measure	จำนวนคลัสเตอร์	เวลา (วินาที)
Model 1	0.0600	0.8056	0.3253	0.4913	0.1359	40	4.9176
Model 2	0.0666	0.8022	0.3275	0.4866	0.1345	41	4.9353
Model 3	0.0540	0.8151	0.2168	0.4868	0.1265	50	4.9402
Model 4	0.0720	0.7911	0.4408	0.4871	0.1433	30	4.9691
Model 5	0.0480	0.8140	0.2150	0.4839	0.1244	51	4.9478

ตารางที่ 4.36 แสดงอัตราร้อยละความแตกต่างตัววัดประสิทธิภาพแต่ละตัวของข้อมูล Dermatology

โมเดล	γ	ความบริสุทธิ์	เอนโทรปี	NMI	F measure
Model 2	+10%	0.4220%	-0.6763%	0.9566%	1.0302%
Model 3	-10%	-1.1792%	33.3538%	0.9159%	6.9169%
Model 4	+20%	1.7999%	-35.5057%	0.8549%	-5.4452%
Model 5	-20%	-1.0427%	33.9072%	1.5062%	8.4621%

2) ชุดข้อมูล Libras Moverment

ผลการทดลองแสดงได้ดังตารางที่ 4.37 และอัตราร้อยละความแตกต่างตัววัดประสิทธิภาพแต่ละตัวเมื่อเปรียบเทียบกับ Model 1 แสดงดังตารางที่ 4.38

ตารางที่ 4.37 แสดงผลการทดลอง Sensitivity ของข้อมูล Libras Moverment

โมเดล	γ	ความบริสุทธิ์	เอนโทรปี	NMI	F measure	จำนวนคลัสเตอร์	เวลา (วินาที)
Model 1	0.0600	0.8719	0.2935	0.6728	0.1193	48	14.7142
Model 2	0.0660	0.8724	0.2866	0.6716	0.1159	46	14.9639
Model 3	0.0540	0.9885	0.0103	0.6665	0.0554	96	16.0886
Model 4	0.0720	0.7883	0.5352	0.6664	0.1688	29	14.0253
Model 5	0.0480	0.9889	0.0097	0.6667	0.0555	97	16.3811

ตารางที่ 4.38 แสดงอัตราร้อยละความแตกต่างตัววัดประสิทธิภาพแต่ละตัวของข้อมูล Libras Moverment

โมเดล	γ	ความบริสุทธิ์	เอนโทรปี	NMI	F measure
Model 2	+10%	-0.0573%	2.3509%	0.1784%	2.8500%
Model 3	-10%	-13.3731%	96.4906%	0.9364%	53.5624%
Model 4	+20%	9.5883%	-82.3509%	0.9512%	-41.4920%
Model 5	-20%	-13.4190%	96.6951%	0.9067%	53.4786%

3) ชุดข้อมูล Large Soybean

ผลการทดลองแสดงได้ดังตารางที่ 4.39 และอัตราร้อยละความแตกต่างตัววัดประสิทธิภาพแต่ละตัวเมื่อเปรียบเทียบกับ Model 1 แสดงดังตารางที่ 4.40

ตารางที่ 4.39 แสดงผลการทดลอง Sensitivity ของข้อมูล Large Soybean

โมเดล	γ	ความบริสุทธิ์	เอนโทรปี	NMI	F measure	จำนวนคลัสเตอร์	เวลา (วินาที)
Model 1	0.0700	0.8684	0.4235	0.7233	0.2217	42	5.4320
Model 2	0.0770	0.8026	0.6117	0.6993	0.2497	30	5.2999
Model 3	0.0630	0.8932	0.2952	0.7097	0.1695	59	5.5184
Model 4	0.0840	0.7581	0.7695	0.6838	0.2762	22	5.1606
Model 5	0.0560	0.9474	0.0702	0.6916	0.1258	82	5.7812

ตารางที่ 4.40 แสดงอัตราร้อยละความแตกต่างตัววัดประสิทธิภาพแต่ละตัวของข้อมูล Large Soybean

โมเดล	γ	ความบริสุทธิ์	เอนโทรปี	NMI	F measure
Model 2	+10%	7.5772%	-44.4392%	3.3181%	-12.6297%
Model 3	-10%	-2.8558%	30.2952%	1.8803%	23.5453%
Model 4	+20%	12.7015%	-81.7001%	5.4611%	-24.5828%
Model 5	-20%	-9.0972%	83.4238%	4.3827%	43.2567%

4) ชุดข้อมูล Wine Recognition

ผลการทดลองแสดงได้ดังตารางที่ 4.41 และอัตราร้อยละความแตกต่างตัววัดประสิทธิภาพแต่ละตัวเมื่อเปรียบเทียบกับ Model 1 แสดงดังตารางที่ 4.42

ตารางที่ 4.41 แสดงผลการทดลอง Sensitivity ของข้อมูล Wine Recognition

โมเดล	γ	ความบริสุทธิ์	เอนโทรปี	NMI	F measure	จำนวนคลัสเตอร์	เวลา (วินาที)
Model 1	0.1700	0.6854	0.9380	0.4536	0.7766	2	1.4324
Model 2	0.1870	0.6854	0.9380	0.4536	0.7766	2	1.4523
Model 3	0.1530	0.6854	0.9380	0.4536	0.7766	2	1.5334
Model 4	0.2040	0.6854	0.9380	0.4536	0.7766	2	1.5442
Model 5	0.1360	0.6854	0.7092	0.4247	0.6742	3	1.4934

ตารางที่ 4.42 แสดงอัตราร้อยละความแตกต่างตัววัดประสิทธิภาพแต่ละตัวของข้อมูล Wine Recognition

โมเดล	γ	ความบริสุทธิ์	เอนโทรปี	NMI	F measure
Model 2	+10%	0%	0%	0%	0%
Model 3	-10%	0%	0%	0%	0%
Model 4	+20%	0%	0%	0%	0%
Model 5	-20%	0%	24.3923%	6.3713%	13.1857%

5) ชุดข้อมูล Iris Plants

ผลการทดลองแสดงได้ดังตารางที่ 4.43 และอัตราร้อยละความแตกต่างตัววัดประสิทธิภาพแต่ละตัวเมื่อเปรียบเทียบกับ Model 1 แสดงดังตารางที่ 4.44

ตารางที่ 4.43 แสดงผลการทดลอง Sensitivity ของข้อมูล Iris Plants

โมเดล	γ	ความบริสุทธิ์	เอนโทรปี	NMI	F measure	จำนวนคลัสเตอร์	เวลา (วินาที)
Model 1	0.2100	0.8258	0.3921	0.7597	0.8558	3	0.5450
Model 2	0.2310	0.8076	0.4065	0.7542	0.8556	3	0.5872
Model 3	0.1890	0.8360	0.3938	0.7341	0.8003	3	0.5558
Model 4	0.2520	0.6987	0.4927	0.7319	0.8726	2	0.5253
Model 5	0.1680	0.8569	0.3784	0.7174	0.7660	3	0.4798

ตารางที่ 4.44 แสดงอัตราร้อยละความแตกต่างตัววัดประสิทธิภาพแต่ละตัวของข้อมูล Iris Plants

โมเดล	γ	ความบริสุทธิ์	เอนโทรปี	NMI	F measure
Model 2	+10%	2.2039%	-3.6725%	0.7240%	0.0234%
Model 3	-10%	-1.2352%	-0.4336%	3.3698%	6.4852%
Model 4	+20%	15.3911%	-25.6567%	3.6593%	-1.9631%
Model 5	-20%	-3.7660%	3.4940%	5.5680%	10.4931%

6) ชุดข้อมูล BUPA liver disorders

ผลการทดลองแสดงได้ดังตารางที่ 4.45 และอัตราร้อยละความแตกต่างตัววัดประสิทธิภาพแต่ละตัวเมื่อเปรียบเทียบกับ Model 1 แสดงดังตารางที่ 4.46

ตารางที่ 4.45 แสดงผลการทดลอง Sensitivity ของข้อมูล BUPA liver disorders

โมเดล	γ	ความบริสุทธิ์	เอนโทรปี	NMI	F measure	จำนวนคลัสเตอร์	เวลา (วินาที)
Model 1	0.0500	0.8303	0.1609	0.1399	0.0247	53	4.3526
Model 2	0.0550	0.8263	0.1650	0.1390	0.0243	52	4.2421
Model 3	0.0450	0.8278	0.1632	0.1392	0.0245	52	4.3380
Model 4	0.0600	0.7411	0.5784	0.0886	0.0302	35	4.2066
Model 5	0.0400	0.8261	0.1649	0.1387	0.0231	55	4.3070

ตารางที่ 4.46 แสดงอัตราร้อยละความแตกต่างตัววัดประสิทธิภาพแต่ละตัวของข้อมูล BUPA liver disorders

โมเดล	γ	ความบริสุทธิ์	เอนโทรปี	NMI	F measure
Model 2	+10%	0.4818%	-2.5482%	0.6433%	1.6194%
Model 3	-10%	0.3011%	-1.4295%	0.5004%	0.8097%
Model 4	+20%	10.7431%	-259.4779%	36.6690%	-22.2672%
Model 5	-20%	0.5058%	-2.4860%	0.8578%	6.4777%

4.4.3.2 ชุดข้อมูลที่ผู้ทำวิจัยสร้างขึ้นเอง

1) ชุดข้อมูล Normal Distribution

ผลการทดลองแสดงได้ดังตารางที่ 4.47 และอัตราร้อยละความแตกต่างตัววัดประสิทธิภาพแต่ละตัวเมื่อเปรียบเทียบกับ Model 1 แสดงดังตารางที่ 4.48

ตารางที่ 4.47 แสดงผลการทดลอง Sensitivity ของข้อมูล Normal Distribution

โมเดล	γ	ความบริสุทธิ์	เอนโทรปี	NMI	F measure	จำนวนคลัสเตอร์	เวลา (วินาที)
Model 1	0.0900	0.8786	0.6659	0.6659	0.7793	4	28.3221
Model 2	0.0990	0.8786	0.6659	0.6659	0.7793	4	28.0638
Model 3	0.0810	0.8786	0.5389	0.6538	0.7577	5	27.7408
Model 4	0.1080	0.8786	0.6660	0.6658	0.7793	4	27.6682
Model 5	0.0720	0.8786	0.5391	0.6537	0.7576	5	27.2983

ตารางที่ 4.48 แสดงอัตราร้อยละความแตกต่างตัววัดประสิทธิภาพแต่ละตัวของข้อมูล Normal Distribution

โมเดล	γ	ความบริสุทธิ์	เอนโทรปี	NMI	F measure
Model 2	+10%	0%	0%	0%	0%
Model 3	-10%	0%	19.0719%	1.8171%	2.7717%
Model 4	+20%	0%	-0.0150%	0.0150%	0%
Model 5	-20%	0%	19.0419%	1.8321%	2.7846%

2) ชุดข้อมูล Fan

ผลการทดลองแสดงได้ดังตารางที่ 4.49 และอัตราร้อยละความแตกต่างตัววัดประสิทธิภาพแต่ละตัวเมื่อเปรียบเทียบกับ Model 1 แสดงดังตารางที่ 4.50

ตารางที่ 4.49 แสดงผลการทดลอง Sensitivity ของข้อมูล Fan

โมเดล	γ	ความบริสุทธิ์	เอนโทรปี	NMI	F measure	จำนวนคลัสเตอร์	เวลา (วินาที)
Model 1	0.0100	0.9980	0.0197	0.9917	0.9962	4	15.6151
Model 2	0.0110	0.9980	0.0197	0.9917	0.9962	4	15.4576
Model 3	0.0090	0.9980	0.0197	0.9917	0.9962	4	14.2188
Model 4	0.0120	0.9980	0.0197	0.9917	0.9962	4	15.2365
Model 5	0.0080	0.9980	0.0197	0.9917	0.9962	4	14.7124

ตารางที่ 4.50 แสดงอัตราร้อยละความแตกต่างตัววัดประสิทธิภาพแต่ละตัวของข้อมูล Fan

โมเดล	γ	ความบริสุทธิ์	เอนโทรปี	NMI	F measure
Model 2	+10%	0%	0%	0%	0%
Model 3	-10%	0%	0%	0%	0%
Model 4	+20%	0%	0%	0%	0%
Model 5	-20%	0%	0%	0%	0%

3) ชุดข้อมูล Ring

ผลการทดลองแสดงได้ดังตารางที่ 4.51 และอัตราร้อยละความแตกต่างตัววัดประสิทธิภาพแต่ละตัวเมื่อเปรียบเทียบกับ Model 1 แสดงดังตารางที่ 4.52

ตารางที่ 4.51 แสดงผลการทดลอง Sensitivity ของข้อมูล Ring

โมเดล	γ	ความบริสุทธิ์	เอนโทรปี	NMI	F measure	จำนวนคลัสเตอร์	เวลา (วินาที)
Model 1	0.0900	0.9665	0.1699	0.4140	0.1100	27	15.6546
Model 2	0.0990	0.8994	0.3075	0.3788	0.1318	20	14.7550
Model 3	0.0810	0.9624	0.1676	0.4033	0.1013	30	15.2108
Model 4	0.1080	0.8169	0.6074	0.2997	0.1465	15	14.8715
Model 5	0.0720	0.9662	0.1607	0.4067	0.1019	30	15.0283

ตารางที่ 4.52 แสดงอัตราร้อยละความแตกต่างตัววัดประสิทธิภาพแต่ละตัวของข้อมูล Ring

โมเดล	γ	ความบริสุทธิ์	เอนโทรปี	NMI	F measure
Model 2	+10%	6.9426%	-80.9888%	8.5024%	-19.8182%
Model 3	-10%	0.4242%	1.3537%	2.5845%	7.9091%
Model 4	+20%	15.4785%	-257.5044%	27.6087%	-33.1818%
Model 5	-20%	0.0310%	5.4149%	1.7633%	7.3636%

4) ชุดข้อมูล Shape

ผลการทดลองแสดงได้ดังตารางที่ 4.53 และอัตราร้อยละความแตกต่างตัววัดประสิทธิภาพแต่ละตัวเมื่อเปรียบเทียบกับ Model 1 แสดงดังตารางที่ 4.54

ตารางที่ 4.53 แสดงผลการทดลอง Sensitivity ของข้อมูล Shape

โมเดล	γ	ความบริสุทธิ์	เอนโทรปี	NMI	F measure	จำนวนคลัสเตอร์	เวลา (วินาที)
Model 1	0.0900	0.9738	0.1497	0.4471	0.1340	24	17.9276
Model 2	0.0990	0.9452	0.2336	0.4313	0.1480	21	16.7069
Model 3	0.0810	0.9723	0.1459	0.4442	0.1329	25	17.6350
Model 4	0.1080	0.9129	0.3403	0.4131	0.1715	17	17.0059
Model 5	0.0720	0.9738	0.1455	0.4457	0.1331	25	16.8316

ตารางที่ 4.54 แสดงอัตราร้อยละความแตกต่างตัววัดประสิทธิภาพแต่ละตัวของข้อมูล Shape

โมเดล	γ	ความบริสุทธิ์	เอนโทรปี	NMI	F measure
Model 2	+10%	2.9369%	-56.0454%	3.5339%	-10.4478%
Model 3	-10%	0.1540%	2.5384%	0.6486%	0.8209%
Model 4	+20%	6.2539%	-127.3213%	7.6046%	-27.9851%
Model 5	-20%	0.0000%	2.8056%	0.3131%	0.6716%

4.5 สรุปผลการทดลอง

4.5.1 สรุปผลการทดลอง 1

4.5.1.1 สรุปผลการทดลองของชุดข้อมูลมาตรฐาน

1) ข้อมูล Dermatology ค่าความบริสุทธิ์เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.0961 สูงกว่า C-means 0.2023 และสูงกว่า DBSCAN 0.2346 ค่าเอนโทรปีเฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.1971 สูงกว่า C-means 0.5918 และต่ำกว่า DBSCAN 0.1236 ค่า NMI เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.0519 สูงกว่า C-means 0.1653 และต่ำกว่า DBSCAN 0.1923 ค่า F measure เฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.0827 ต่ำกว่า C-means 0.0690 และต่ำกว่า DBSCAN 0.3961

2) ข้อมูล Libras Movement ค่าความบริสุทธิ์เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.3377 สูงกว่า C-means 0.7270 และสูงกว่า DBSCAN 0.8051 ค่าเอนโทรปีเฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.3881 สูงกว่า C-means 0.0095 และต่ำกว่า DBSCAN 3.6134 ค่า NMI เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.0608 สูงกว่า C-means 0.5449 และต่ำกว่า DBSCAN 0.6728 ค่า F measure เฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.2658 ต่ำกว่า C-means 0.1350 และต่ำกว่า DBSCAN 0.1358

3) ข้อมูล Large Soybean ค่าความบริสุทธิ์เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.5822 สูงกว่า C-means 0.3485 และสูงกว่า DBSCAN 0.6111 ค่าเอนโทรปีเฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.2698 ต่ำกว่า C-means 0.1288 และสูงกว่า DBSCAN 0.1832 ค่า NMI เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.3567 สูงกว่า C-mea 0.1459 และสูงกว่า DBSCAN 0.4506 ค่า F measure เฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.1633 ต่ำกว่า C-means 0.2174 และต่ำกว่า DBSCAN 0.1143

4) ข้อมูล Wine Recognition ค่าความบริสุทธิ์เฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.0453 ต่ำกว่า C-means 0.0457 และสูงกว่า DBSCAN 0.5618 ค่าเอนโทรปีเฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.4883 ต่ำกว่า C-means 0.4413 และต่ำกว่า DBSCAN 0.8471 ค่า NMI เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.0288 สูงกว่า C-means 0.0368 และต่ำกว่า DBSCAN 0.4355 ค่า F measure เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.1526 สูงกว่า C-means 0.2030 และสูงกว่า DBSCAN 0.3418

5) ข้อมูล Iris Plants ค่าความบริสุทธิ์เฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.1360 ต่ำกว่า C-means 0.1529 และสูงกว่า DBSCAN 0.1458 ค่าเอนโทรปีเฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.3118 ต่ำกว่า C-means 0.3107 และต่ำกว่า DBSCAN 0.0589 ค่า NMI เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.0148 สูงกว่า C-means 0.0101 และสูงกว่า DBSCAN 0.0260 ค่า F measure เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.0286 สูงกว่า C-means 0.0295 และต่ำกว่า DBSCAN 0.0244

6) ข้อมูล Bupa Liver Disorders ค่าความบริสุทธิ์เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.2017 สูงกว่า C-means 0.1408 และสูงกว่า DBSCAN 0.2738 ค่าเอนโทรปีเฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.342 สูงกว่า C-means 0.59 และสูงกว่า DBSCAN 0.5857 ค่า NMI เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.1051 สูงกว่า C-means 0.1037 และสูงกว่า DBSCAN 0.0499 ค่า F measure เฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.6965 ต่ำกว่า C-means 0.6489 และต่ำกว่า DBSCAN 0.8004

4.5.1.2 สรุปผลการทดลองของชุดข้อมูลที่ผู้ทำวิจัยสร้างขึ้นเอง

1) ข้อมูล Normal Distribution ค่าความบริสุทธิ์เฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.0007 ต่ำกว่า C-means 0.0019 และสูงกว่า DBSCAN 0.0766 ค่าเอนโทรปีเฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.1934 ต่ำกว่า C-means 0.1944 และต่ำกว่า DBSCAN 0.373 ค่า NMI เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.0022 ต่ำกว่า C-means 0.0001 และสูงกว่า DBSCAN 0.2057 ค่า F measure เฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.0003 ต่ำกว่า C-means 0.0023 และต่ำกว่า DBSCAN 0.1576

2) ข้อมูล Fan ค่าความบริสุทธิ์เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.0103 สูงกว่า C-means 0.0005 และต่ำกว่า DBSCAN 0.0005 ค่าเอนโทรปีเฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.0181 สูงกว่า C-means 0.0031 และสูงกว่า DBSCAN 0.0197 ค่า NMI เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.0224 สูงกว่า C-means 0.003 และสูงกว่า DBSCAN 0.0082 ค่า F measure เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.0247 สูงกว่า C-means 0.0013 และสูงกว่า DBSCAN 0.0122

3) ข้อมูล Ring ค่าความบริสุทธิ์เฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.0071 ต่ำกว่า C-means 0.0034 และต่ำกว่า DBSCAN 0.0335 ค่าเอนโทรปีเฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.0935 ต่ำกว่า C-means 0.0523 และต่ำกว่า DBSCAN 0.1699 ค่า NMI เฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.0184 ต่ำกว่า C-means 0.0051 และต่ำกว่า DBSCAN 0.586 ค่า F measure เฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.2468 ต่ำกว่า C-means 0.2551 และต่ำกว่า DBSCAN 0.89

4) ข้อมูล Shape ค่าความบริสุทธิ์เฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.0025 สูงกว่า C-means 0.0075 และสูงกว่า DBSCAN 0.0098 ค่าเอนโทรปีเฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.0813 ต่ำกว่า C-means 0.0197 และต่ำกว่า DBSCAN 0.1497 ค่า NMI เฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.0032 สูงกว่า C-means 0.0149 และต่ำกว่า DBSCAN 0.2121 ค่า F measure เฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.268 ต่ำกว่า C-means 0.2582 และต่ำกว่า DBSCAN 0.5237

4.5.2 สรุปผลการทดลอง 2

4.5.2.1 สรุปผลการทดลองของชุดข้อมูลมาตรฐาน

1) ข้อมูล Dermatology ค่าความบริสุทธิ์เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.0959 และสูงกว่า C-means 0.1903 ค่าเอนโทรปีเฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.1768 และสูงกว่า C-means 0.4629 ค่า NMI เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.0521 และสูงกว่า C-means 0.1561 ค่า F measure เฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.0122 และสูงกว่า C-means 0.021

2) ข้อมูล Libras Movement ค่าความบริสุทธิ์เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.3182 และสูงกว่า C-means 0.7276 ค่าเอนโทรปีเฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.2625 และต่ำกว่า C-means 0.0593 ค่า NMI เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.0579 และสูงกว่า C-means 0.5523 ค่า F measure เฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.1975 และต่ำกว่า C-means 0.1282

3) ข้อมูล Large Soybean ค่าความบริสุทธิ์เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.58 และสูงกว่า C-means 0.3503 ค่าเอนโทรปีเฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.3108 และต่ำกว่า C-means 0.1702 ค่า NMI เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.3685 และสูงกว่า C-means 0.1484 ค่า F measure เฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.1524 และต่ำกว่า C-means 0.1923

4) ข้อมูล Wine Recognition ค่าความบริสุทธิ์เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.0281 และสูงกว่า C-means 0.0281 ค่าเอนโทรปีเฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.0324 และต่ำกว่า C-means 0.0018 ค่า NMI เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.0326 และสูงกว่า C-means 0.0463 ค่า F measure เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.0254 และสูงกว่า C-means 0.0297

5) ข้อมูล Iris Plants ค่าความบริสุทธิ์เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.0089 และต่ำกว่า C-means 0.0675 ค่าเอนโทรปีเฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.003 และต่ำกว่า C-means 0.0127 ค่า NMI เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.0466 และสูงกว่า C-means 0.0101 ค่า F measure เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.0365 และสูงกว่า C-means 0.0295

6) ข้อมูล Bupa Liver Disorders ค่าความบริสุทธิ์เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.2023 และสูงกว่า C-means 0.1182 ค่าเอนโทรปีเฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.0636 และสูงกว่า C-means 0.4405 ค่า NMI เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.1038 และสูงกว่า C-means 0.0766 ค่า F measure เฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.1717 และต่ำกว่า C-means 0.0136

4.5.2.2 สรุปผลการทดลองของชุดข้อมูลที่ผู้ทำวิจัยสร้างขึ้นเอง

1) ข้อมูล Normal Distribution ค่าความบริสุทธิ์เฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.0007 และต่ำกว่า C-means 0.0019 ค่าเอนโทรปีเฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.0064 และสูงกว่า C-means 0.0019 ค่า NMI เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.0021 และต่ำกว่า C-means 0.0001 ค่า F measure เฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.0005 และต่ำกว่า C-means 0.0024

2) ข้อมูล Fan ค่าความบริสุทธิ์เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.0124 และสูงกว่า C-means 0.0005 ค่าเอนโทรปีเฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.005 และสูงกว่า C-means 0.0049 ค่า NMI เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.0086 และสูงกว่า C-means 0.003 ค่า F measure เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.011 และสูงกว่า C-means 0.0013

3) ข้อมูล Ring ค่าความบริสุทธิ์เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.0064 และสูงกว่า C-means 0.0052 ค่าเอนโทรปีเฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.0424 และต่ำกว่า C-means 0.0045 ค่า NMI เฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.0083 และสูงกว่า C-means 0.0005 ค่า F measure เฉลี่ยของงานวิจัยที่นำเสนอมีค่าต่ำกว่า K-means 0.0028 และสูงกว่า C-means 0.0001

4) ข้อมูล Shape ค่าความบริสุทธิ์เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.0214 และสูงกว่า C-means 0.034 ค่าเอนโทรปีเฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.0062 และสูงกว่า C-means 0.084 ค่า NMI เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.009 และสูงกว่า C-means 0.0323 ค่า F measure เฉลี่ยของงานวิจัยที่นำเสนอมีค่าสูงกว่า K-means 0.0022 และสูงกว่า C-means 0.0074

4.5.3 สรุปผลการทดลอง 3

ในการสรุปจะบอกถึงอัตราร้อยละเฉลี่ย ของการเปลี่ยนแปลงของค่าประสิทธิภาพของทุกๆ Model เมื่อพารามิเตอร์ γ เปลี่ยนเป็นค่าต่างๆ ตามที่กำหนดไว้ในการทดลอง

4.5.3.1 สรุปผลการทดลองของชุดข้อมูลมาตรฐาน

1) ข้อมูล Dermatology ค่าความบริสุทธิ์เฉลี่ยของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 1.111% ค่าเอนโทรปีของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 25.8607% ค่า NMI ของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 1.0584% ค่า F measure ของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 5.4636%

2) ข้อมูล Libras Moverment ค่าความบริสุทธิ์เฉลี่ยของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 9.1094% ค่าเอนโทรปีของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 69.4719% ค่า NMI ของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 0.7432% ค่า F measure ของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 37.8458%

3) ข้อมูล Large Soybean ค่าความบริสุทธิ์เฉลี่ยของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 8.0579% ค่าเอนโทรปีของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 59.9646% ค่า NMI ของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 3.7605% ค่า F measure ของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 36.0036%

4) ข้อมูล Wine Recognition ค่าความบริสุทธิ์เฉลี่ยของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 0% ค่าเอนโทรปีของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 6.0981% ค่า NMI ของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 1.5928% ค่า F measure ของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 3.2964%

5) ข้อมูล Iris Plants ค่าความบริสุทธิ์เฉลี่ยของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 5.6491% ค่าเอนโทรปีของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 8.3142% ค่า NMI ของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 3.3303% ค่า F measure ของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 4.7412%

6) ข้อมูล BUPA liver disorders ค่าความบริสุทธิ์เฉลี่ยของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 3.0079% ค่าเอนโทรปีของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย

66.4854% ค่า NMI ของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 9.6676% ค่า F measure ของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 7.7935%

4.5.3.2 สรุปผลการทดลองของชุดข้อมูลที่ผู้ทำวิจัยสร้างขึ้นเอง

1) ข้อมูล Normal Distribution ค่าความบริสุทธิ์เฉลี่ยของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 0% ค่าเอนโทรปีของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 9.5322% ค่า NMI ของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 0.9161% ค่า F measure ของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 1.3891%

2) ข้อมูล Fan ค่าความบริสุทธิ์เฉลี่ยของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 0% ค่าเอนโทรปีของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 0% ค่า NMI ของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 0% ค่า F measure ของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 0%

3) ข้อมูล Ring ค่าความบริสุทธิ์เฉลี่ยของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 5.7191% ค่าเอนโทรปีของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 86.3155% ค่า NMI ของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 10.1147% ค่า F measure ของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 17.0682%

4) ข้อมูล Shape ค่าความบริสุทธิ์เฉลี่ยของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 2.3362% ค่าเอนโทรปีของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 47.1777% ค่า NMI ของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 3.0251% ค่า F measure ของแต่ละ Model แตกต่างกันเป็นอัตราร้อยละเฉลี่ย 9.9813%

บทที่ 5
สรุปผลการวิจัย และข้อเสนอแนะ

5.1 สรุปผลงานวิจัย

งานวิจัยนี้ได้นำเสนอวิธีการจัดแบ่งกลุ่มข้อมูล (Clustering) แบบใหม่ที่ใช้หลักการที่เลียนแบบปรากฏการณ์ทางฟิสิกส์ โดยงานวิจัยที่นำเสนอจะพยายามสร้างโมเดลที่เลียนแบบกฎแรงโน้มถ่วงและกฎการเคลื่อนที่ของนิวตัน เคลื่อนที่ไปยังบริเวณจุดที่มีข้อมูลรวมกันอยู่แน่นหนาสูงและจุดนั้นเราจะถือว่าเหมาะที่จะเป็นจุดศูนย์กลางหรือเป็นตัวแทนกลุ่ม

หลักการทำงานของงานวิจัยที่นำเสนอเริ่มจากทำการสุ่มรูปแบบอินพุตหนึ่งตัวมาเป็นจุดศูนย์กลางเริ่มต้น ทำการปรับค่าจุดศูนย์กลางของคลัสเตอร์ โดยดูจากจุดที่อยู่รอบข้างซึ่งออกแรงกระทำต่อจุดศูนย์กลาง ซึ่งในการคำนวณแรงเราจะใช้สมการแรงโน้มถ่วงคำนวณแรงที่กระทำบนจุดศูนย์กลางนั้น จากนั้นก็หาแรงรวมที่เกิดขึ้นบนจุดศูนย์กลางแล้วไปคำนวณหาการเคลื่อนที่โดยใช้สมการการเคลื่อนของนิวตัน ตัวแปรที่มีผลในการปรับค่าจุดศูนย์กลางคือ ค่าควบคุมอัตราเร่ง(β) โดยเมื่อจุดศูนย์กลางที่กำลังเคลื่อนที่เข้าใกล้จุดศูนย์ที่เคยบันทึกไว้ก่อนหน้าจะทำให้ค่าควบคุมอัตราเร่งลดลงส่งผลทำให้อัตราเร่งช้าลงและทำให้การเคลื่อนที่ช้าลงด้วย เมื่อทำการปรับค่าจุดศูนย์กลางไปแต่ละรอบแล้วเราจะทำการตรวจสอบว่าตำแหน่งจุดศูนย์ที่ปรับไปนั้นมีค่าเปลี่ยนแปลงไปจากรอบที่แล้วหรือไม่ ถ้าเริ่มไม่เปลี่ยนแปลงแล้วก็บันทึกตำแหน่งจุดศูนย์กลางใหม่ที่ได้ หลังจากนั้นก็นำตำแหน่งจุดศูนย์กลางใหม่ที่ได้มาทำการตรวจสอบว่าอยู่ใกล้กันหรือไม่ ถ้าอยู่ใกล้ก็จะทำการรวมจุดกลางศูนย์ที่อยู่ใกล้ชิดกัน จากนั้นก็ทำการตรวจสอบการลบคลัสเตอร์โดยจะนับจำนวนสมาชิกของแต่ละคลัสเตอร์แล้วดูว่าจำนวนสมาชิกในแต่ละคลัสเตอร์มีค่าผ่านเกณฑ์ที่กำหนดไว้หรือไม่ ถ้าไม่ผ่านก็จะทำการลบคลัสเตอร์นั้น ขั้นตอนทั้งหมดจะถูกทำซ้ำจนกระทั่งข้อมูลทุกจุดในเซตข้อมูลถูกกระทำทั้งหมด

ข้อมูลที่นำมาใช้ในขั้นตอนการเรียนรู้นำมาจากฐานข้อมูล UCI จำนวน 6 ชุดข้อมูลและเป็นข้อมูลที่ผู้วิจัยสร้างขึ้นเองจำนวน 4 ชุดข้อมูล

5.2 ข้อดีของงานวิจัย

1. จากการทดลองอัลกอริทึมที่นำเสนอได้จำนวนกลุ่มเองอัตโนมัติ การเรียนรู้กระทำเพียง 15 รอบต่อหนึ่งปัญหาเท่านั้น ในส่วนของ K-means และ C-means จะต้องทำการเรียนรู้ 15 รอบต่อหนึ่งจำนวนกลุ่ม ซึ่งในแต่ละปัญหาจะต้องเรียนรู้จำนวนกลุ่มในช่วงของจำนวนคลาสทั้งหมดใน

เซตของข้อมูลไปถึง 10 เปอร์เซ็นของจำนวนตัวอย่างของเซตข้อมูลทั้งหมด เพื่อหาจำนวนกลุ่มที่เหมาะสมกับเซตของข้อมูล ทำให้ K-means และ C-means ใช้เวลารวมทั้งหมดในการเรียนรู้นานกว่าอัลกอริทึมที่นำเสนอ

2. จากข้อมูลที่ทำการทดลองที่ 1 จำนวน 10 ชุดข้อมูล พบว่าอัลกอริทึมที่นำเสนอมีความถูกต้องในตัววัดความบริสุทธิ์ดีกว่า K-means 5 ชุดข้อมูล ดีกว่า C-means 6 ชุดข้อมูล และดีกว่า DBSCAN 5 ชุดข้อมูล ตัววัดเอนโทรปีดีกว่า K-means 4 ชุดข้อมูล ดีกว่า C-means 3 ชุดข้อมูล และดีกว่า DBSCAN 4 ชุดข้อมูล ตัววัด NMI ดีกว่า K-means 8 ชุดข้อมูล ดีกว่า C-means 8 ชุดข้อมูล และดีกว่า DBSCAN 6 ชุดข้อมูล ส่วนตัววัดประสิทธิภาพ F measure ดีกว่า K-means 3 ชุดข้อมูล ดีกว่า C-means 3 ชุดข้อมูล และดีกว่า DBSCAN 3 ชุดข้อมูล

3. จากข้อมูลที่ทำการทดลองที่ 2 จำนวน 10 ชุดข้อมูล พบว่าอัลกอริทึมที่นำเสนอมีความถูกต้องในตัววัดความบริสุทธิ์ดีกว่า K-means 9 ชุดข้อมูล และดีกว่า C-means 8 ชุดข้อมูล ตัววัดเอนโทรปีดีกว่า K-means 7 ชุดข้อมูล และดีกว่า C-means 6 ชุดข้อมูล ตัววัด NMI ดีกว่า K-means 9 ชุดข้อมูล และดีกว่า C-means 9 ชุดข้อมูล ส่วนตัววัดประสิทธิภาพ F measure ดีกว่า K-means 4 ชุดข้อมูล และดีกว่า C-means 6 ชุดข้อมูล

4. ในส่วนของการเปรียบเทียบกับอัลกอริทึม DBSCAN ซึ่งเป็นอัลกอริทึมที่ดูจากลักษณะหนาแน่นของข้อมูลและให้จำนวนกลุ่มอัตโนมัติเหมือนอัลกอริทึมที่นำเสนอ อัลกอริทึมที่นำเสนอให้ผลลัพธ์ที่นำไปใช้ประโยชน์ได้มากเพราะโมเดลที่ได้จากการเรียนรู้เป็นเซตจุดศูนย์กลางที่เป็นตัวแทนของกลุ่ม เมื่อมีข้อมูลรูปแบบใหม่สามารถนำเซตจุดศูนย์กลางไปใช้ได้เลย แต่ในส่วน DBSCAN ต้องนำข้อมูลรูปแบบใหม่ไปเรียนรู้โมเดลถึงจะบอกได้ว่าอยู่กลุ่มไหน

5.3 ปัญหาที่พบในงานวิจัย

ในกรณีที่ทำการเรียนรู้โมเดลของอัลกอริทึมที่นำเสนอและทำการเรียนรู้โมเดลของ K-means, C-means และ DBSCAN ในจำนวนรอบที่เท่ากันและชุดข้อมูลที่เหมือนกัน เมื่อเปรียบเทียบเวลาแล้วอัลกอริทึมที่นำเสนอใช้เวลาในการเรียนรู้นานกว่า K-means, C-means และ DBSCAN เพราะมีการคำนวณที่ซับซ้อนมากกว่าทำให้ใช้เวลาในการประมวลผลนานกว่า

5.4 แนวทางการพัฒนาในอนาคต

1. พัฒนาอัลกอริทึมที่นำเสนอให้สามารถทำการจัดกลุ่มข้อมูลแบบรู้เป้าหมาย(Classification) ได้

2. พัฒนาอัลกอริทึมที่ใช้ในการจัดแบ่งกลุ่มข้อมูล (Clustering) โดยอ้างอิงจากทฤษฎีทางฟิสิกส์ในด้านอื่นๆ ที่สามารถนำมาประยุกต์ใช้ในการจัดกลุ่มข้อมูล (Clustering) ได้

บรรณานุกรม

[1] Matteo Matteucci. **A Tutorial on Clustering Algorithms**. [Online]. Available: http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/.

[2] W. Liao. 2005 **The Software Package of Parallel Kmeans.** [Online]. Available: http://www.ece.northwestern.edu/~wkliao/Kmeans/index.html.

[3] Dr. Wararat Rungworawut. **Chapter 9: Clustering Analysis.** [Online]. Available: 202.28.94.51/users/wararat/322475/DM_ch9.ppt.

[4] Somchai Champathong. **Alternative Adaptive Fuzzy C-Means Clustering.** Department of Computer Science, Faculty of Science, Khon Kaen University, Thailand , 2006.

[5] K. Fukinaga, L. D. Hostetler. **The estimation of the gradient of a density function, with application in pattern recognition.** IEEE Trans. Information Theory, vol. 21, pp. 32-40, 1975.

[6] M. Ester, H. Kriegel, J. Sander, X. Xu. **A density-based algorithm for discovering clusters in large spatial databases with noise.** proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, OR, 1996

[7] D. Holliday, R. Resnick, J. Walker. **Fundamentals of physics.** John Wiley and Sons, 1993.

[8] Bin Wang. **Bandwidth Selection for Weighted Kernel Density Estimation.** Mathematics and Statistics Department, University of South Alabama, Mobile, AL 36688, 2007.

[9] E. Rashedi, H. Nezamabadi-pour, and S. Saryazdi. **GSA: A Gravitational Search Algorithm.** Information Sciences, Elsevier, pp. 2232-2248, 2009.

[10] A. Frank and A. Asuncion, **UCI Machine Learning Repository**. [http://archive.ics.uci.edu/ml], University of California, School of Information and Computer Science, 2010.

[11] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. **Introduction to Information Retrieval - Chapter 16.** Cambridge University Press. 2008.

[12] Jonatan Gomez, Dipankar Dasgupta and Olfa Nasraoui. **A New Gravitational Clustering Algorithm.** The University of Memphis – Department of Mathematical Sciences and Department of Electrical & Computer Engineering, 2003.

[13] W.E Wright. **Gravitational Clustering**. Pattern Recognition, 9:151-166, Pergamon pass 1977.

[14] T. Cormer, C. Leiserson, and R. Rivest, R. **Introduction to Algorithms.** Plenun Press, 1990.

ภาคผนวก

ภาคผนวก ก.

ผลงานวิจัยที่ได้รับการตีพิมพ์เผยแพร่

Arit Thammano and Ponglert Sangkapas. "Gravitational Clustering Algorithm: GCA.". **Proceedings of the 26th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC2011).** Gyeongju, Korea, 2011.

CALL FOR PAPERS

ITC-CSCC 2011

The 26th International Technical Conference on Circuits/Systems, Computers and Communications

June 19-22 2011 | Hyundai Hotel, Gyeongju, Korea

WELCOME TO ITC-CSCC

The 26th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC 2011) will be held on June 19-22 at Hyundai Hotel in Gyeongju, Korea. We would like to invite all the scholars and experts around the world to attend the conference to be hosted at the wonderful city of Korea. The meetings at Shimonoseki in 2008, Jeju in 2009, Pattaya in 2010 were successfully held in this world leading conference devoted to the advancement of high technologies in Circuits/Systems, Computers and Communications.

General Chair
- Byung-Gook Park (Seoul National University, Korea)

General Co-Chairs
- Hisakazu Kikuchi (Niigata University, Japan)
- Prabhas Chongstivatana (Chulalongkorn University, Thailand)

Program Committee Chairs
- Joonki Paik (Chung-Ang University, Korea)
- Makoto Nakashizuka (Osaka University, Japan)
- Prayoot Akkaraekthalin (King Mongkut's Institute of Technology North Bangkok, Thailand)

TOPICS

The conference is open to researchers from all regions of the world. Participation from Asia Pacific region is particularly encouraged. Proposals for special sessions are welcome. Papers with original works in all aspects of Circuits/Systems, Computers and Communications are invited. Topics include, but not limited to, the followings:

Circuits & Systems
- Computer Aided Design
- Intelligent Transportation Systems & Technology
- Linear / Nonlinear Systems
- Medical Electronics & Circuits
- Semiconductor Devices and Technology
- Power Electronics & Circuits
- Analog Circuits
- RF Circuits
- Modern Control
- Neural Networks
- VLSI Design
- Verification and Testing

Computers
- Artificial Intelligence
- Biocomputing
- Computer Systems & Applications
- Computer Vision
- Face Detection & Recognition
- Image Coding & Analysis
- Image Processing
- Internet Technology & Applications
- Motion Analysis
- Multimedia Service & Technology
- Object Extraction & Technology
- Security
- Watermarking

Communications
- Antenna & Wave Propagation
- Audio / Speech Signal Processing
- Circuits & Components for Communications
- IP Networks & QoS
- MIMO & Space-Time Codes
- Multimedia Communications
- Mobile & Wireless Communications
- Network Management & Design
- Optical Communications and Components
- Radar / Remote Sensing
- Communication Signal Processing
- Ubiquitous Networks
- UWB
- Visual Communications

PLENARY SPEAKERS

Prof. Aggelos K. Katsaggelos (Northwestern University, USA)
- Applications of Sparse and Redundant Representations in Digital Processing

Prof. Hideki Asai (Shizuoka University, Japan)
- Present Status and Future Trend of PI/SI/EMI Simulation Technology for High-Speed Electronic Design

Prof. Monai Krairiksh (King Mongkut's Institute of Technology Ladkrabang, Thailand)
- Microwave for Agricultural Applications

Dr. Jaemoon Jo (Samsung Electronics, Korea)
- Future Technology of Medical Imaging Devices

PROCEEDINGS

All registered participants are provided with conference proceedings. Moreover, authors of the accepted papers are encouraged to submit full-length manuscripts to IEICE Transactions or IEEK JSTS (Journal of Semiconductor Technology and Science). Papers passed through the standard editing procedures of the IEICE Transactions or IEEK JSTS will be published in regular issues. The authors (or their institute) are requested to pay the publication charge for the IEICE Transactions when their paper is accepted.

SUBMISSION OF PAPERS

Prospective authors are invited to submit original papers of either MS Word or PDF format written in English. Abstracts are limited to two pages of text and figures. Only on-line abstract can be submitted to http://www.itc-cscc2011.org. If you have any trouble in preparing papers and on-line submission, please contact the conference secretariat.

AUTHOR'S SCHEDULE

Deadline for abstract submission: Apr. 14, 2011
Notification of acceptance: Apr. 30, 2011
Camera-ready manuscripts: May 15, 2011

CONFERENCE VENUE AND CITY

Gyeongju is a coastal city in the far southeastern corner of North Gyeongsang province in South Korea. It is the second largest city by area in the province, covering 1,300㎢ with a population of 270,000 people. Gyeongju is 370㎞ southeast of Seoul, and 60㎞ north of Busan. Numerous low mountains are scattered around the city. Gyeongju was the capital of the ancient kingdom of Silla (57 BC - 935 AD) which ruled most of the Korean Peninsula between the 7th and 9th centuries. A vast number of archaeological sites and cultural heritages from this period remain in the city. Gyeongju is often referred to as "the museum without walls." Among such historical treasures, Seokguram grotto, Bulguksa temple and Gyeongju Historic Areas are designated as World Heritage Sites by UNESCO. Many major historical sites have helped Gyeongju become one of the most popular tourist destinations in South Korea.

SPONSORED BY

The Institute of Electronics Engineers of Korea (IEEK), Korea
The Institute of Electronics, Information and Communication Engineers (IEICE), Japan
The Electrical Engineering / Electronics, Computer, Telecommunications and Information Association, Thailand

CONTACT POINT

- Website : http://www.itc-cscc2011.org
- E-mail : itc2011@jcnter.co.kr
- Phone : +82-2-571-2721

ITC-CSCC 2011, June 19-22, 2011, Gyeongju, Korea

Gravitational Clustering Algorithm: GCA

Arit Thammano[1] and Ponglert Sangkapas[2]
Computational Intelligence Laboratory
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok, 10520 Thailand
E-mail: [1]arit@it.kmitl.ac.th and [2]s1066414@kmitl.ac.th

Abstract

In the past decades, various clustering methods have been developed, for example K-means, K-medoid, and fuzzy C-means. Most of these algorithms are faced with the problem of selecting an appropriate number of clusters. Therefore, the main objective of this paper is to overcome the above problem by letting the clustering algorithm to automatically identify the number of clusters itself. This new clustering algorithm is based on Newton's law of gravity and Newton's laws of motion. The concepts of both Newton's laws are employed to move the cluster centers to the areas where the data density is high. The proposed method has been evaluated and compared to two of the most widely used clustering algorithms, K-means and fuzzy C-means. The obtained results confirm the high performance of the proposed method in clustering various benchmark problems.

Keywords: Clustering; Naturally Inspired Algorithm; Newton's Law of Gravity.

1. Introduction

Clustering is the process of grouping the data into clusters so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters [1]. Among the existing clustering algorithms, K-means and fuzzy C-means are the most widely known; they are rather simple and remarkably effective. Beside K-means and fuzzy C-means, recently, more and more attention has been focused on using naturally inspired algorithms to solve the clustering problem [2, 3]. However, the performance of all the above mentioned algorithms depends highly on the user's ability to specify an appropriate number of clusters. Therefore, the main objective of this paper is to introduce a new clustering algorithm based on the concepts of Newton's law of gravity and Newton's laws of motion. The proposed algorithm has the ability to automatically identify the number of clusters itself.

This paper is divided into 4 sections. Following this introduction, section 2 presents the proposed Gravitational Clustering Algorithm. A brief description of the experimental data and the experimental results are given in section 3. Finally, section 4 is the conclusion.

2. Gravitational Clustering Algorithm

The proposed Gravitational Clustering Algorithm (GCA) is a naturally inspired algorithm based on the law of gravity and the laws of motion [4, 5]. The main idea is that every particle in the universe attracts every other particle with a force which varies directly as the product of the masses of the particles and inversely as the square of the distance between them. The area with higher particle density (more particles in the same amount of space) has more attractive force than the one with lower particle density. That is, all particles scattered throughout the universe are drawn together by attractive forces of high particle density areas. This paper employs the above concept to move the cluster centers to the areas where the data density is high. The steps of the GCA are described as follows:

A. Define the size of the cluster by using the following equation:

$$S = \frac{2\sigma}{\sqrt[3]{N}} \tag{1}$$

where N is the total number of the dataset. σ is the standard deviation of the dataset.

B. Randomly select from the dataset the center of the first cluster.

$$C_1 = (c_{11}, \ldots, c_{1d}, \ldots, c_{1D}) \tag{2}$$

where D is the dimension of the data.

C. Update the cluster center by using the law of gravity and the laws of motion. This updating step is done as follows:

C.1. Locate the data samples $X_j = (x_{j1}, \ldots, x_{jd}, \ldots, x_{jD})$ within a radius of S from the cluster center. Then mark them as used.

$$\left\| X_j - C_1 \right\| \le S \tag{3}$$

where j = 1, 2, ..., P. P is the total number of the data within a radius of S from the cluster center.

C.2. Calculate the total force acting on the cluster center C_i.

$$F_c(t) = \sum_{j=1}^{P} F_{cj}(t) \quad (4)$$

$$F_{cj}(t) = G(t)\frac{M_c \times M_j}{R_{cj} + \varepsilon}(X_j(t) - C_i(t)) \quad (5)$$

$$M_* = \frac{m_*(t)}{\sum_{q=1}^{P} m_q(t)} \quad (6)$$

$$m_*(t) = \frac{density_*(t) - min\_density}{max\_density - min\_density} \quad (7)$$

$$min\_density = \min_{q\in\{1,2,...,P\}} density_q(t) \quad (8)$$

$$max\_density = \max_{q\in\{1,2,...,P\}} density_q(t) \quad (9)$$

where $F_{cj}(t)$ is the force acting on the cluster center due to the data X_j. M_j is the active gravitational mass related to the j^{th} data. M_c is the passive gravitational mass related to the cluster center C_i. $R_{cj}(t)$ is the Euclidean distance between C_i and X_j. G(t) is the gravitational constant, whose value decreases over time. In this paper, the value of G(t) is calculated according to the following equation:

$$G(t) = S \times e^{-\frac{t}{100}} \quad (10)$$

$density_c(t)$ is the density of the area surrounding the cluster center C_i, which can be calculated by the following equation:

$$density_c(t) = \left(-\frac{1}{2\log_2\left(\frac{P}{N}\right) - \varepsilon}\right) \times \frac{1}{P}\left(\sum_{j=1}^{P} e^{-\frac{\|X_j - C_i\|^2}{S^2}}\right) \quad (11)$$

C.3. Calculate the acceleration of the cluster center C_i by using Newton's second law of motion.

$$a_c(t) = \frac{F_c(t)}{M_{cc}} \quad (12)$$

$$M_{cc} = M_c \quad (13)$$

where M_{cc} is the inertial mass of the center C_i.

C.4. Calculate the velocity of the cluster center C_i according to the following equation:

$$v_c(t+1) = rand \times v_c(t) + \beta a_c(t) \quad (14)$$

where rand is a random number in the interval [0, 1]. β is a decay variable that is equal to 1 at the beginning, and decreases linearly to 0 as the center C_i moves closer to the other centers.

C.5. Update the cluster center as follows:

$$C_i(t+1) = C_i(t) + v_c(t+1) \quad (15)$$

D. Examine the criterion in (16). If the Euclidean distance between $C_i(t+1)$ and $C_i(t)$ is less than the threshold, continue to step E. Otherwise, go back to step C.

$$\|C_i(t+1) - C_i(t)\| < threshold \quad (16)$$

E. Save the cluster center obtained from step D. Then check this new center against the previously saved ones. If the Euclidean distance between the new cluster center C_i and the previously saved center C_k is less than S/2 as shown in equation (17), merge the two centers together by using equation (18).

$$\|C_i - C_k\| < \frac{S}{2} \quad (17)$$

$$C_i = \frac{density_{C_i}}{density_{C_i} + density_{C_k}} C_i + \frac{density_{C_k}}{density_{C_i} + density_{C_k}} C_k \quad (18)$$

where k = 1, 2, 3, ..., i-1. This step is repeated until the distance between the new cluster center C_i resulting from the merge operation and the previously saved center C_k is greater than S/2.

F. This step is the cluster reduction step. The center whose cluster satisfies the following condition will be deleted.

$$-\frac{1}{2\log_2\left(\frac{s_i}{N}\right) - \varepsilon} < \gamma \quad (19)$$

where s_i is the number of members of the i^{th} cluster. N is the total number of data in the dataset. γ is a vigilance parameter.

G. Check whether there is at least one data point that has not been marked as used. If so, randomly select from the unmarked data the center of the next cluster, and go back to step C. If not, stop the loop.

3. Experimental Results

The performance of the GCA is compared to that of K-means and fuzzy C-means algorithms. In order to compare the three algorithms, the experiments were conducted on 7 benchmark datasets from UCI machine learning repository [6]: "Credit approval," "Haberman's survival," "Heart disease," "Liver disorders," "Pima Indians diabetes," "Zoo," "Ionosphere." The performance of the algorithms is measured in terms of the ability to cluster data from the same class together. Brief descriptions of the data sets are given below:

1. The first data set is the Credit approval data set. This data set contains 690 instances. Each instance is described by 15 attributes and belongs to one of two classes. There are 307 instances of class "+" and 383 instances of class "-".
2. The second data set is the Haberman's survival data. This data set contains 306 instances. Three numerical attributes are used to predict the output class (class 1 or class 2).
3. The third data set is the Heart disease data set. This data set contains 13 attributes and 303 records. There are 164 records of class 0, 55 records of class 1, 36 records of class 2, 35 records of class 3, and 13 records of class 4.
4. The fourth data set is the Liver disorders data set. There are 345 patterns in this data set. Six numerical attributes are used to predict whether or not an unmarried man has a liver disorder.
5. The fifth data set is the Pima Indians diabetes database. This database contains 768 examples. Each example is described by 8 numerical attributes and belongs to one of two classes (class 0 or class 1). There are 500 examples of class 0 and 268 examples of class 1.
6. The sixth is the Zoo data set. Sixteen attributes are used to determine the type of animals (class 1 – 7). This data set contains 101 instances. There are 41 instances of class 1, 20 instances of class 2, 5 instances of class 3, 13 instances of class 4, 4 instances of class 5, 8 instances of class 6, and 10 instances of class 7.
7. The seventh data set is the Ionosphere data. This data set has 351 instances. Each instance is described by 34 continuous attributes and belongs to one of two classes ("good" or "bad").

Table 1 summarizes the results obtained from the three clustering algorithms. The figures which are bold represent the best results among the three methods.

4. Conclusion

Newton's law of gravity and Newton's second law of motion are used in the development of the GCA presented in this paper. The GCA algorithm consists of 2 main steps: (1) the process of moving the cluster centers to areas where the data density is high by using the law of gravity and the laws of motion, and (2) the process of reducing the number of redundant clusters. Unlike K-means and fuzzy C-means whose performances depend on the user's ability in defining the number of clusters, GCA has the ability to identify the proper number of clusters for the problem at hand. The experimental results show that GCA comfortably outperforms both K-means and fuzzy C-means algorithms.

Table 1: Experimental results

Data Set	Accuracy		
	GCA	K-means	Fuzzy C-means
Credit Approval	**0.6522**	0.6014	0.6014
Haberman's Survival	0.7049	0.7213	**0.7541**
Heart Disease	0.4590	**0.4754**	0.4590
Liver Disorders	**0.6232**	**0.6232**	0.5652
Pima Indians Diabetes	**0.6234**	0.6169	0.5909
Zoo	**0.8500**	0.8000	0.8000
Ionosphere	**0.9000**	0.8429	0.7857

References

[1] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2001.

[2] U. Maulik and S. Bandyopadhyay, "Genetic Algorithm-based Clustering Technique," *Pattern Recognition*, vol. 33, Elsevier, pp. 1455-1465, 2000.

[3] A. Thammano and U. Kakulphimp, "Genetic Algorithm-based Clustering and Its New Mutation Operator," *Lecture Notes in Computer Science*, vol. 4113, Springer-Verlag, pp. 703-708, 2006.

[4] D. Holliday, R. Resnick, and J. Walker, *Fundamentals of Physics*, John Wiley and Sons, 1993.

[5] E. Rashedi, H. Nezamabadi-pour, and S. Saryazdi, "GSA: A Gravitational Search Algorithm," *Information Sciences*, Elsevier, pp. 2232-2248, 2009.

[6] A. Frank and A. Asuncion, *UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]*, University of California, School of Information and Computer Science, 2010.

ประวัติผู้เขียน

ชื่อ-นามสกุล	นายพงศ์เลิศ สังกะเพศ
วันเดือนปีเกิด	2 เมษายน 2529
ที่อยู่	173 ถนนพิชิตรังสรรค์ ต.ในเมือง
	อ.เมือง จ.อุบลราชธานี 34000
ประวัติการศึกษา	2550 มหาวิทยาลัยเกษตรศาสตร์
	คณะวิศวกรรมศาสตร์ สาขาวิศวกรรมคอมพิวเตอร์
ประสบการณ์การทำงาน	-