วิธีการใหม่ในการกำหนดจุดศูนย์กลางเริ่มต้นสำหรับการจัดกลุ่มแบบเคมีนส์

A NEW METHOD FOR FINDING THE INITIAL CENTROIDS FOR K-MEANS CLUSTERING

พัธนี ฉัตรธเนศ

PATHCHARNEE CHATTANES

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2557

KMITL-2014-IT-M-001-002

วิธีการใหม่ในการกำหนดจุดศูนย์กลางเริ่มต้นสำหรับการจัดกลุ่มแบบเคมีนส์

A NEW METHOD FOR FINDING THE INITIAL CENTROIDS FOR K-MEANS CLUSTERING

พัชนี ฉัตรธเนศ

PATHCHARNEE CHATTANES

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2557

KMITL-2014-IT-M-001-002

A NEW METHOD FOR FINDING THE INITIAL CENTROIDS FOR K-MEANS CLUSTERING

PATHCHARNEE CHATTANES

A THESIS SUBMITTED IN PATIAL 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

2014

KMITL-2014-IT-M-001-002

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ใบรับรองวิทยานิพนธ์

หัวข้อวิทยานิพนธ์ วิธีการใหม่ในการกำหนดจุดศูนย์กลางเริ่มต้นสำหรับการจัดกลุ่มแบบเคมีนส์

A NEW METHOD FOR FINDING THE INITIAL CENTROIDS FOR K-MEANS CLUSTERING

นักศึกษา นางสาวพัชนี ฉัตรธเนศ

รหัสประจำตัว ๕๔๖๖๐๔๐๔

ปริญญา วิทยาศาสตรมหาบัณฑิต

สาขาวิชา เทคโนโลยีสารสนเทศ

อาจารย์ที่ปรึกษาวิทยานิพนธ์ รองศาสตราจารย์ ดร.อาริต ธรรมโน

คณะกรรมการสอบวิทยานิพนธ์	ลายมือชื่อ
รองศาสตราจารย์ ดร.วรพจน์ กรีสุระเดช	
รองศาสตราจารย์ ดร.พีระพนธ์ โสพัศสถิตย์	
รองศาสตราจารย์ ดร.อาริต ธรรมโน	
รองศาสตราจารย์ ดร.พรฤดี เนติโสภากุล	
ดร.กิติ์สุชาต พสุภา	

วัน / เดือน/ ปี ที่สอบ วันพุธที่ ๑๔ พฤษภาคม ๒๕๕๗ เวลา ๑๓.๐๐ น.

สถานที่สอบ ณ ห้อง ๓๓๕ ชั้น ๓ คณะเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศรับรองแล้ว

(รองศาสตราจารย์ ดร.จันทร์บูรณ์ สถิตวิริยวงศ์)

คณบดีคณะเทคโนโลยีสารสนเทศ

วันที่ ๒๒ **เดือน** พฤษภาคม **พ.ศ.** ๒๕๕๗

หัวข้อวิทยานิพนธ์	วิธีการใหม่ในการกำหนดจุดศูนย์กลางเริ่มต้นสำหรับการจัดกลุ่มแบบเคมีนส์
นักศึกษา	นางสาวพัชนี ฉัตรธเนศ
รหัสประจำตัว	54660404
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
พ.ศ.	2557
อาจารย์ผู้ควบคุมวิทยานิพนธ์	รศ.ดร. อาริต ธรรมโน

บทคัดย่อ

วิทยานิพนธ์นี้นำเสนอการปรับปรุงกระบวนการจัดกลุ่มแบบเคมีนส์ เพื่อแก้ไขข้อจำกัดบางประการของการจัดกลุ่มแบบเคมีนส์ ข้อจำกัดดังกล่าวคือ ผลลัพธ์ในการจัดกลุ่มแบบเคมีนส์จะขึ้นอยู่กับจุดศูนย์กลางเริ่มต้น หากเลือกจุดศูนย์กลางเริ่มต้นได้ไม่ดี ก็จะส่งผลให้ผลลัพธ์ในการจัดกลุ่มที่ได้ไม่มีประสิทธิภาพเท่าที่ควร โดยวิทยานิพนธ์นี้นำเสนอวิธีการคำนวณหาจุดศูนย์กลางเริ่มต้นแทนการใช้วิธีการสุ่มข้อมูลของการจัดกลุ่มแบบเคมีนส์ดั้งเดิม ซึ่งวิธีการนี้จะเป็นกระบวนการหาจุดศูนย์กลางเริ่มต้นก่อนที่จะจัดกลุ่มด้วยเคมีนส์ตามปกติ ซึ่งวิธีการคำนวณหาจุดศูนย์กลางเริ่มต้นที่วิทยานิพนธ์นี้นำเสนอจะใช้แนวความคิดหลักคือ ข้อมูลที่อยู่ในกลุ่มหรือคลัสเตอร์เดียวกันจะมีระยะทางใกล้กันและข้อมูลที่อยู่ต่างกลุ่มหรือต่างคลัสเตอร์กันจะมีระยะทางห่างกัน ซึ่งจากการทดลองสามารถแสดงให้เห็นว่าวิธีการคำนวณหาจุดศูนย์กลางเริ่มต้นที่วิทยานิพนธ์นี้นำเสนอให้ผลลัพธ์การจัดกลุ่มที่มีประสิทธิภาพมากกว่าการใช้วิธีการสุ่มข้อมูลเพื่อใช้เป็นจุดศูนย์กลางเริ่มต้นของการจัดกลุ่มแบบเคมีนส์ดั้งเดิม

Thesis	A New Method for Finding the Initial Centroids for K-means Clustering
Student	Ms. Pathcharnee Chattanes
Student ID	54660404
Degree	Master of Science
Program	Information Technology
Year	2014
Thesis Advisor	Assoc. Prof. Dr. Arit Thammano

ABSTRACT

This thesis proposes an improved k-means clustering algorithm to overcome some shortcoming: sensitive to the initial centroid which leads to different results. To solve such problems, this thesis proposes a method for finding the appropriate initial centroids for K-means algorithm instead of random selection, which is preprocessing process of k-means algorithm. The main idea used for findind the initial centroids is the objects in the same cluster are more similar to one another than to the objects in other clusters. The experimental results show that the proposed initialization method produces more accurate clusters than the original K-means algorithm for most of the datasets.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี เนื่องจากได้รับคำแนะนำและความช่วยเหลือในด้านต่างๆ เป็นอย่างดีตลอดระยะเวลาในการทำงาน ข้าพเจ้าขอกราบขอบพระคุณอาจารย์ที่ปรึกษา รศ.ดร.อาริต ธรรมโน ที่ให้ความกรุณา คอยเอาใจใส่ ให้ความรู้และคำแนะนำต่างๆ มาโดยตลอด

ขอกราบขอบพระคุณคณาจารย์คณะเทคโนโลยีสารสนเทศทุกท่าน ตลอดจนครูอาจารย์ทุกท่านที่เคยสอนข้าพเจ้ามาตั้งแต่อดีตจนถึงปัจจุบัน ที่ได้ประสิทธิประสาทวิชาความรู้ให้แก่ข้าพเจ้า

ขอขอบคุณพี่ๆ ในห้องปฏิบัติการและเพื่อนๆ ทุกคน ที่ได้ให้คำแนะนำที่ดีในการทำวิจัยและคอยเป็นกำลังใจให้กับข้าพเจ้า

สุดท้ายนี้ ข้าพเจ้าขอกราบขอบพระคุณบิดา มารดา และครอบครัวของข้าพเจ้า ที่ให้การเลี้ยงดู อบรมสั่งสอน รวมทั้งคอยให้กำลังใจและได้ให้การสนับสนุนในทุกๆ ด้าน

คุณงามความดีอันใดที่เกิดจากวิทยานิพนธ์ฉบับนี้ เป็นผลมาจากความกรุณาของทุกท่านที่กล่าวมาข้างต้น ข้าพเจ้ารู้สึกซาบซึ้งเป็นอย่างยิ่ง จึงใคร่ขอขอบพระคุณไว้ ณ ที่นี้

พัชนี ฉัตรธเนศ

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

สารบัญรูป

สารบัญรูป (ต่อ)

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในโลกยุคโลกาภิวัตน์ คงไม่มีผู้ใดปฏิเสธว่า ระบบสารสนเทศได้เข้ามามีบทบาทต่อชีวิตของผู้คนเป็นอย่างมาก ไม่ว่าจะเพื่อความบันเทิงหรือเพื่อการดำเนินธุรกิจ การสร้างระบบสารสนเทศจำเป็นต้องใช้ทรัพยากรที่สำคัญอย่างหนึ่งคือ ข้อมูล แม้ว่าข้อมูลจะมีความสำคัญต่อการสร้างระบบสารสนเทศเพียงใด หากไม่มีเครื่องมือหรือวิธีการที่จะจัดการกับข้อมูลอย่างมีประสิทธิภาพ ข้อมูลที่มีอยู่เหล่านั้นก็อาจจะไม่ก่อให้เกิดประโยชน์ใดๆ

การจัดกลุ่ม (Clustering) ถือเป็นวิธีการจัดการข้อมูลที่เริ่มมีบทบาทมากขึ้นในปัจจุบัน ซึ่งอัลกอริทึมในการจัดกลุ่มมีด้วยกันหลายวิธี โดยวิธีพื้นฐานที่ใช้ในการจัดกลุ่ม ได้แก่ การจัดกลุ่มแบบเคมีนส์ (K-means) แม้ว่าการจัดกลุ่มแบบเคมีนส์จะเป็นวิธีพื้นฐานที่ง่ายต่อความเข้าใจ แต่การจัดกลุ่มแบบเคมีนส์ก็ยังมีข้อจำกัดบางประการ โดยวิทยานิพนธ์นี้จะมุ่งเน้นการปรับปรุงประสิทธิภาพและแก้ไขข้อจำกัดต่างๆ ในการจัดกลุ่มแบบเคมีนส์

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

1. เพื่อศึกษาทฤษฎีเกี่ยวกับการจัดกลุ่มแบบเคมีนส์
2. เพื่อศึกษางานวิจัยที่เกี่ยวของกับการปรับปรุงประสิทธิภาพการจัดกลุ่มแบบเคมีนส์
3. นำเสนอวิธีการในการปรับปรุงกระบวนการของการจัดกลุ่มแบบเคมีนส์ อันจะนำไปสู่ประสิทธิภาพในการจัดกลุ่มที่ดีขึ้น

1.3 ขอบเขตการวิจัย

เคมีนส์เป็นอัลกอริทึมในการจัดกลุ่มที่ง่ายต่อความเข้าใจ แต่ถึงอย่างไรก็ยังมีปัญหาอยู่หลายประการ ได้แก่

- ผู้ใช้จะต้องระบุจำนวนกลุ่ม (K) ล่วงหน้า ซึ่งในทางปฏิบัติผู้ใช้จะไม่สามารถทราบได้ว่าจำนวนกลุ่มที่เหมาะสมควรจะเป็นเท่าใด
- ในการจัดกลุ่มแบบเคมีนส์ ในขั้นตอนแรกจะต้องทำการสุ่มค่าใดๆ เพื่อใช้เป็นจุดศูนย์กลางเริ่มต้นของกลุ่มแต่ละกลุ่ม ค่าเริ่มต้นที่ได้จากการสุ่มแต่ละครั้งจะมีค่าแตกต่างกัน ซึ่งผลลัพธ์ในการจัดกลุ่มแบบเคมีนส์จะขึ้นอยู่กับจุดศูนย์กลางเริ่มต้น ดังนั้นหากทดลองจัดกลุ่มข้อมูลแบบเคมีนส์หลายๆ ครั้ง อาจได้ผลลัพธ์การจัดกลุ่มที่แตกต่างกัน

ในงานวิจัยนี้จะมุ่งเน้นการแก้ไขปัญหาในข้อหลังเท่านั้น โดยมีขอบเขตการวิจัย ดังนี้

1. คิดค้นวิธีการหาจุดศูนย์กลางเริ่มต้นแทนการใช้วิธีการสุ่มข้อมูล โดยวิธีการนี้จะเป็นกระบวนการหาจุดศูนย์กลางเริ่มต้นก่อนที่จะจัดกลุ่มด้วยเคมีนส์ตามปกติ (Preprocessing) เพื่อให้ผลลัพธ์ในการจัดกลุ่มที่ได้มีความเสถียรภาพและประสิทธิภาพมากยิ่งขึ้น

2. งานวิจัยนี้ได้ทดสอบประสิทธิภาพในการจัดกลุ่มโดยใช้ข้อมูลที่เป็นชุดข้อมูลมาตรฐาน

1.4 ขั้นตอนการดำเนินงานวิจัย

1. ศึกษาทฤษฎีที่เกี่ยวข้องกับการจัดกลุ่ม
2. ศึกษากระบวนการทำงาน ข้อดี และข้อจำกัดของวิธีการจัดกลุ่มแบบเคมีนส์
3. ศึกษางานวิจัยที่เกี่ยวข้องกับการปรับปรุงประสิทธิภาพการจัดกลุ่มแบบเคมีนส์ตามขอบเขตการวิจัยที่กำหนด
4. คิดค้นวิธีการปรับปรุงประสิทธิภาพการจัดกลุ่มแบบเคมีนส์
5. เขียนโปรแกรมเพื่อทดสอบประสิทธิภาพของอัลกอริทึมที่คิดค้น
6. วิเคราะห์และประเมินผลการทดลอง
7. สรุปผลการดำเนินการวิจัยและข้อเสนอแนะ
8. จัดทำเอกสารที่เกี่ยวข้องกับงานวิจัย

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. ผู้วิจัยมีความรู้ความเข้าใจเกี่ยวกับการจัดกลุ่มแบบเคมีนส์มากขึ้น
2. การปรับปรุงกระบวนการจัดกลุ่มแบบเคมีนส์ทำให้ผลลัพธ์ในการจัดกลุ่มมีประสิทธิภาพมากขึ้น และสามารถนำไปประยุกต์ใช้ในการสร้างระบบสารสนเทศที่ดีได้
3. วิทยานิพนธ์นี้จะสามารถเป็นแหล่งค้นคว้าหรือแหล่งอ้างอิงให้กับผู้ที่สนใจได้

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้ได้แบ่งเนื้อหาออกเป็น 2 ส่วน ดังนี้ ส่วนแรกอธิบายถึงทฤษฎีพื้นฐานที่เกี่ยวข้องกับงานวิจัยนี้และส่วนที่ 2 เป็นรายละเอียดของงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 การจัดกลุ่ม

การจัดกลุ่ม คือกระบวนการจัดข้อมูลออกเป็นกลุ่มๆ แต่ละกลุ่มอยู่ในรูปแบบของคลาสหรือคลัสเตอร์ (Cluster) ข้อมูลที่อยู่ภายในกลุ่มเดียวกันจะมีความคล้ายคลึงกัน ส่วนข้อมูลที่อยู่คนละกลุ่มจะมีความแตกต่างกัน การจัดกลุ่มแตกต่างกับการแบ่งประเภท (Classification) คือ การจัดกลุ่มจะไม่ทราบจำนวนกลุ่มล่วงหน้า แต่การแบ่งประเภทจะทราบจำนวนประเภทล่วงหน้า โดยการจัดกลุ่มสามารถแบ่งออกเป็น 2 ประเภทหลักๆ ได้แก่

2.1.1.1 การจัดกลุ่มแบบแบ่งส่วน (Partitioning Methods)

การจัดกลุ่มแบบแบ่งส่วน จะแบ่งข้อมูลทั้งหมดออกเป็น K กลุ่ม โดยกลุ่มแต่ละกลุ่มจะต้องประกอบด้วยข้อมูลอย่างน้อย 1 ข้อมูล และข้อมูลแต่ละตัวจะต้องถูกจัดให้อยู่ในกลุ่มใดกลุ่มหนึ่งเพียงกลุ่มเดียวเท่านั้น การจัดกลุ่มข้อมูลวิธีนี้ผู้ใช้จะต้องเป็นผู้ระบุค่า K หรือจำนวนกลุ่มที่ต้องการ อัลกอริทึมในการจัดกลุ่มแบบแบ่งส่วนที่นิยมใช้คือ การจัดกลุ่มแบบเคมีนส์ (K-means)

2.1.1.2 การจัดกลุ่มแบบลำดับชั้น (Hierarchical Methods)

การจัดกลุ่มแบบลำดับชั้นจะแบ่งข้อมูลออกเป็นลำดับชั้นคล้ายกับต้นไม้ ซึ่งวิธีการแบ่งกลุ่มข้อมูลแบบลำดับชั้นสามารถแบ่งออกเป็น 2 แนวทางตามลักษณะการสร้างลำดับชั้นคือ

1) การจัดกลุ่มแบบลำดับชั้นโดยการรวมเป็นกลุ่มก้อน (Agglomerative Approach) เป็นการจัดการข้อมูลจากล่างขึ้นบน (Bottom-up Approach) เริ่มต้นโดยการให้ข้อมูลแต่ละตัวอยู่ในกลุ่มที่ต่างกัน หรืออีกนัยหนึ่งกลุ่มทุกกลุ่มจะมีข้อมูลอยู่ภายในเพียง 1 ข้อมูล จากนั้นจึงทำการวนซ้ำเพื่อรวมกลุ่มที่มีความคล้ายคลึงกันเข้าด้วยกัน จนกระทั่งกลุ่มทุกกลุ่มถูกรวมเข้าเป็นกลุ่มเดียว หรือเมื่อรวมได้จำนวนกลุ่มที่ต้องการ

2) การจัดกลุ่มแบบลำดับชั้นโดยการแบ่งแยกแตกออก (Divisive Approach) เป็นการจัดการข้อมูลจากบนลงล่าง (Top-down Approach) เริ่มต้นโดยให้ข้อมูลทุกตัวอยู่ในกลุ่มเดียวกัน จากนั้นจึงหาคู่ของข้อมูลที่มีความแตกต่างกันมากที่สุดภายในกลุ่มนั้น แล้วแยกข้อมูลทั้งคู่

ให้อยู่ต่างกลุ่มกัน ค่อยๆ แบ่งกลุ่มให้เล็กลงมาเรื่อยๆ จนกระทั่งข้อมูลทุกตัวถูกแยกออกจากกันทั้งหมด หรือเมื่อแบ่งได้จำนวนกลุ่มที่ต้องการ

งานวิจัยนี้จะมุ่งเน้นการปรับปรุงประสิทธิภาพของการจัดกลุ่มแบบเคมีนส์ เนื่องจากการจัดกลุ่มแบบเคมีนส์เป็นวิธีที่ไม่ซับซ้อนและเป็นที่นิยมใช้ แต่การจัดกลุ่มแบบเคมีนส์ก็ยังมีข้อจำกัดบางประการที่ทำให้ประสิทธิภาพการจัดกลุ่มลดน้อยลง สำหรับรายละเอียดของการจัดกลุ่มแบบเคมีนส์จะกล่าวถึงในหัวข้อถัดไป

2.1.2 การจัดกลุ่มแบบเคมีนส์

การจัดกลุ่มแบบเคมีนส์จัดเป็นการจัดกลุ่มแบบแบ่งส่วน โดยผู้ใช้จะต้องเป็นผู้กำหนดค่า K หรือจำนวนกลุ่มเอง โดยอัลกอริทึมการจัดกลุ่มแบบเคมีนส์สรุปได้ดังนี้

1) สุ่มข้อมูลจำนวน K ตัว เพื่อใช้เป็นจุดศูนย์กลางเริ่มต้นในแต่ละกลุ่ม

2) นำข้อมูลทั้งหมดมาจัดเข้ากลุ่ม ซึ่งการพิจารณาว่าข้อมูลแต่ละตัวควรจะถูกจัดเข้าไปอยู่ในกลุ่มใดจะใช้การวัดความใกล้เคียงหรือความคล้ายคลึงกันระหว่างข้อมูลนั้นๆ กับจุดศูนย์กลางของกลุ่มแต่ละกลุ่ม ซึ่งข้อมูลแต่ละตัวจะถูกจัดเข้าไปอยู่ในกลุ่มที่มีความคล้ายคลึงกันมากที่สุด โดยวิธีที่นิยมใช้ในการวัดความคล้ายคลึงกันระหว่างข้อมูลกับจุดศูนย์กลางของกลุ่ม คือ การวัดระยะทางแบบยุคลิด (Euclidean Distance)

$$d(x_i,x_j) = \sqrt{\sum_{k=1}^{m} (x_{ik}-x_{jk})^2} \qquad (2.1)$$

เมื่อ $d(x_i,x_j)$ คือระยะทางแบบยุคลิดระหว่างข้อมูล x_i และข้อมูล x_j

m คือจำนวนมิติของข้อมูล

3) คำนวณจุดศูนย์กลางของทุกกลุ่มใหม่ โดยคำนวณจากค่าเฉลี่ย (Mean) ของข้อมูลภายในกลุ่มแต่ละกลุ่ม

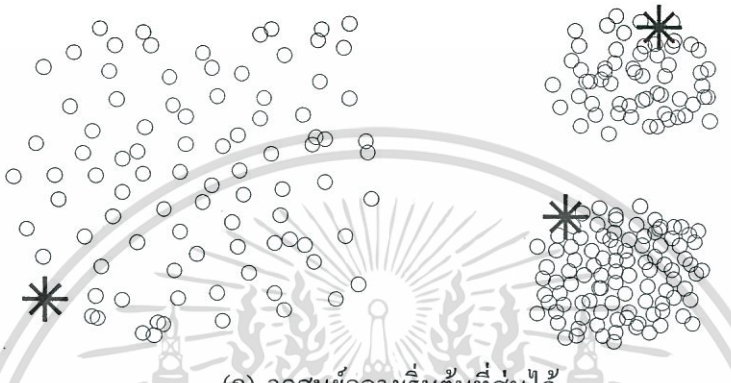
$$c_i = \frac{1}{n_i}\sum_{x \in Ci} x \qquad (2.2)$$

เมื่อ c_i คือจุดศูนย์กลางของกลุ่ม C_i

n_i คือจำนวนข้อมูลทั้งหมดภายในกลุ่ม

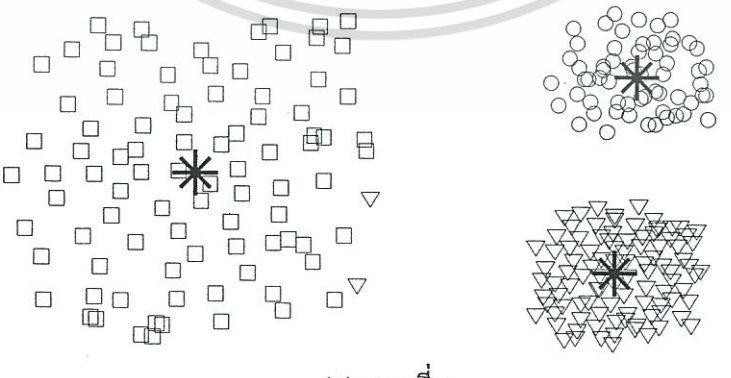
4) ทำซ้ำขั้นตอนที่ 2-3 จนกระทั่งจุดศูนย์กลางไม่มีการเปลี่ยนแปลง

ตัวอย่างการจัดกลุ่มแบบเคมีนส์แสดงได้ดังรูปที่ 2.1 ซึ่งข้อมูลตัวอย่างนี้สามารถจัดกลุ่มได้ทั้งหมด 3 กลุ่ม ในขั้นตอนแรกจะเป็นการสุ่มข้อมูลจำนวน 3 ตัวเพื่อใช้เป็นจุดศูนย์กลางเริ่มต้นของแต่ละกลุ่ม ดังรูปที่ 2.1(ก) หลังจากนั้นก็นำข้อมูลทั้งหมดมาจัดกลุ่มตามอัลกอริทึมของเคมีนส์ทำซ้ำไปเรื่อยๆ จนกระทั่งกลุ่มของข้อมูลและจุดศูนย์กลางของแต่ละกลุ่มไม่มีการเปลี่ยนแปลง ดังรูปที่ 2.1(ข) – 2.1(ง)

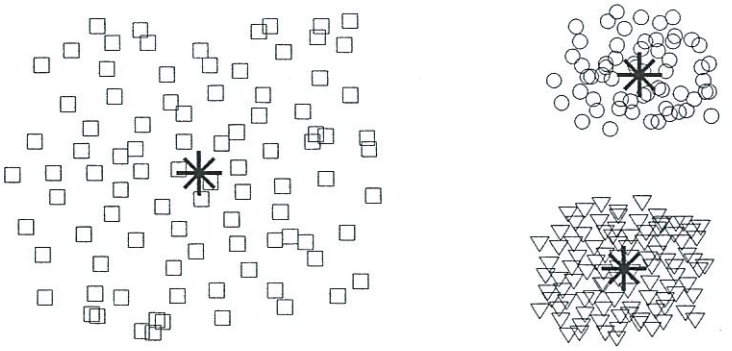


(ก) จุดศูนย์กลางเริ่มต้นที่สุ่มได้

(ข) รอบที่ 1



(ค) รอบที่ 2



(ง) รอบที่ 3

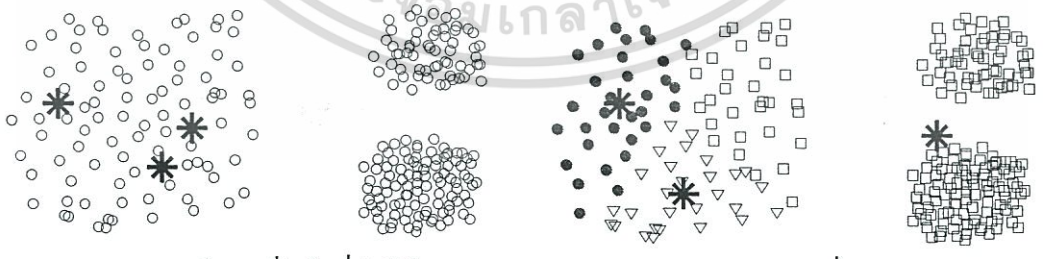
รูปที่ 2.1 ตัวอย่างการจัดกลุ่มแบบเคมีนส์

การจัดกลุ่มแบบเคมีนส์เป็นวิธีที่ง่ายต่อความเข้าใจและใช้เวลาในการประมวลผลเร็ว แต่อย่างไรก็ตาม การจัดกลุ่มแบบเคมีนส์ก็ยังมีข้อจำกัดบางประการที่ทำให้ผลลัพธ์การจัดกลุ่มที่ได้ไม่มีประสิทธิภาพ ซึ่งข้อจำกัดของการจัดกลุ่มแบบเคมีนส์ ได้แก่

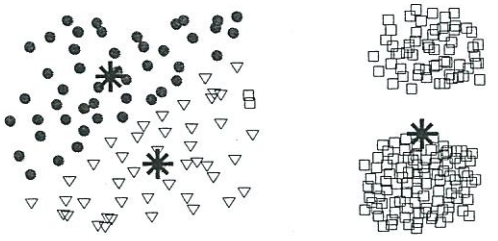
1) การจัดกลุ่มแบบเคมีนส์ ผู้ใช้จะต้องเป็นผู้กำหนดค่า K หรือจำนวนกลุ่มเอง ซึ่งในทางปฏิบัติจะไม่สามารถรู้ได้ว่าจำนวนกลุ่มที่เหมาะสมควรจะเป็นเท่าใด

2) ในกรณีที่กลุ่มของข้อมูลมีรูปร่างซับซ้อน หรือกลุ่มแต่ละกลุ่มมีขนาดหรือความหนาแน่นแตกต่างกัน ประสิทธิภาพในการจัดกลุ่มแบบเคมีนส์จะลดลง

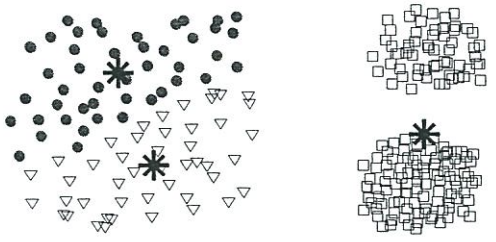
3) ประสิทธิภาพในการจัดกลุ่มแบบเคมีนส์จะขึ้นอยู่กับจุดศูนย์กลางเริ่มต้นที่ถูกสุ่มขึ้นมา หากสุ่มข้อมูลได้ไม่ดีก็จะทำให้ได้ผลลัพธ์ที่ดีที่สุดเฉพาะที่ (Local Optimum) และทำให้ประสิทธิภาพในการจัดกลุ่มลดลง ตัวอย่างผลลัพธ์การจัดกลุ่มเมื่อสุ่มข้อมูลเพื่อใช้เป็นจุดศูนย์กลางเริ่มต้นได้ไม่ดี แสดงได้ดังรูปที่ 2.2



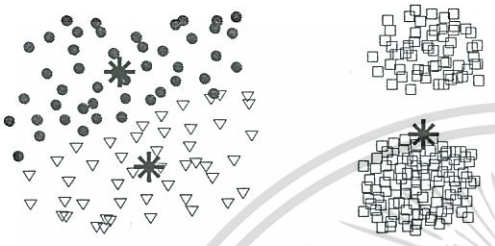
(ก) จุดศูนย์กลางเริ่มต้นที่สุ่มได้ (ข) รอบที่ 1



(ค) รอบที่ 2



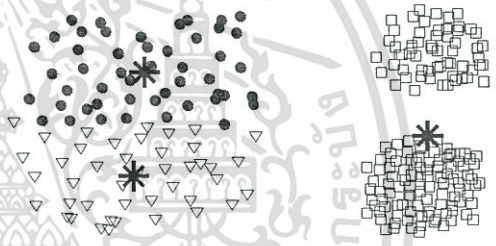
(ง) รอบที่ 3



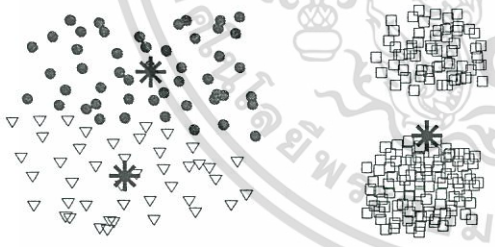
(จ) รอบที่ 4

(ฉ) รอบที่ 5

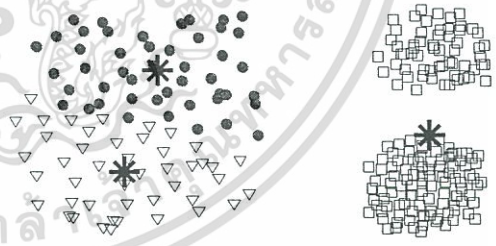
(ช) รอบที่ 6



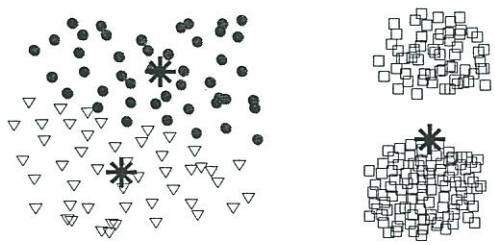
(ซ) รอบที่ 7



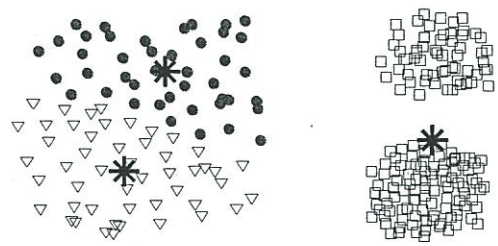
(ฌ) รอบที่ 8



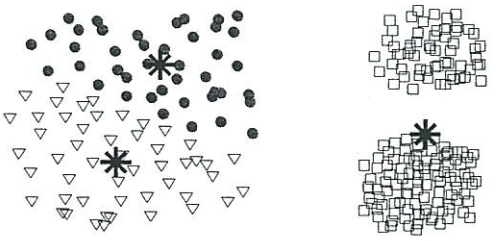
(ญ) รอบที่ 9



(ฎ) รอบที่ 10



(ฏ) รอบที่ 11



(ฐ) รอบที่ 12

รูปที่ 2.2 ตัวอย่างผลลัพธ์การจัดกลุ่มเมื่อสุ่มข้อมูลเพื่อใช้เป็นจุดศูนย์กลางเริ่มต้นไม่ดี

2.2 งานวิจัยที่เกี่ยวข้อง

2.2.1 Enhancing the K-means Clustering Algorithm by Using a O(n logn) Heuristic Method for Finding Better Initial Centroids (Nazeer, Kumar and Sebastian. 2011)

งานวิจัยนี้นำเสนออัลกอริทึมในการกำหนดข้อมูลจุดศูนย์กลางเริ่มต้นแทนที่การสุ่มข้อมูลตามอัลกอริทึมของเคมีนส์แบบดั้งเดิม แนวความคิดที่ใช้คือการเรียงลำดับข้อมูลและแบ่งข้อมูลออกเป็นส่วนๆ โดยมีขั้นตอนหลักๆ คือ ในขั้นตอนแรกจะต้องเรียงลำดับข้อมูลทั้งหมด หลังจากนั้นจึงแบ่งข้อมูลที่เรียงลำดับไว้แล้วออกเป็น K ส่วน แล้วนำค่าเฉลี่ยของแต่ละส่วนมาใช้เป็นจุดศูนย์กลางเริ่มต้นของการจัดกลุ่มแบบเคมีนส์ โดยมีรายละเอียดของอัลกอริทึมดังนี้

1) หาค่าพิสัย (Range) ของทุกคอลัมน์ โดยค่าพิสัยหาได้จาก

$$\text{พิสัย} = \text{ค่าสูงสุดของข้อมูล} - \text{ค่าต่ำสุดของข้อมูล} \qquad (2.3)$$

2) เรียงลำดับข้อมูลโดยใช้การเรียงลำดับข้อมูลแบบฮีพ (Heap Sort) โดยเรียงลำดับตามคอลัมน์ที่มีค่าพิสัยสูงสุด

3) แบ่งข้อมูลที่เรียงลำดับแล้วออกเป็น K ส่วน

4) หาค่าเฉลี่ยของข้อมูลแต่ละส่วน นำค่าเฉลี่ยที่ได้ไปใช้เป็นจุดศูนย์กลางเริ่มต้นของการจัดกลุ่มแบบเคมีนส์

5) จัดกลุ่มโดยใช้เคมีนส์ตามปกติ

เมื่อ n คือจำนวนข้อมูลทั้งหมด และ m คือจำนวนคอลัมน์ทั้งหมด จะสามารถวิเคราะห์ความซับซ้อนทางเวลาในการกำหนดจุดศูนย์กลางเริ่มต้นได้ดังนี้

- เวลาที่ใช้ในการหาค่าต่ำสุดและค่าสูงสุดของทุกคอลัมน์เท่ากับ O(n)
- เวลาที่ใช้ในการหาคอลัมน์ที่มีค่าพิสัยสูงสุดเท่ากับ O(m)
- เวลาที่ใช้ในการเรียงลำดับข้อมูลแบบฮีพเท่ากับ O(n logn)

- เวลาที่ใช้ในการแบ่งข้อมูลออกเป็น K ส่วน และหาค่าเฉลี่ยในแต่ละส่วนเท่ากับ O(n)

ดังนั้น ความซับซ้อนทางเวลาในการกำหนดจุดศูนย์กลางเริ่มต้นเท่ากับ O(n logn)

งานวิจัยนี้ได้ทดสอบประสิทธิภาพในการจัดกลุ่มโดยใช้ข้อมูลจากคลังข้อมูลยูซีไอ (UCI Repository) (Bache and Lichman. 2013) ชุดข้อมูลที่ใช้ ได้แก่ ชุดข้อมูลเอสเชอริเชีย โคไล (E-coli) ชุดข้อมูลโรคมะเร็งเต้านม (Breast Cancer-Wisconsin) และชุดข้อมูลโรคไทรอยด์ (Thyroid) โดยใช้อัลกอริทึมของงานวิจัยนี้มาจัดกลุ่มข้อมูล เปรียบเทียบกับผลลัพธ์การจัดกลุ่มที่ได้จากเคมีนส์ดั้งเดิม และเคมีนส์ที่ได้รับการปรับปรุงโดย Fahim et.al. (2006) สำหรับค่าความถูกต้อง (Accuracy) และเวลาที่ใช้ในการจัดกลุ่มเป็นมิลลิเซก (ms) ของอัลกอริทึมทั้ง 3 วิธีแสดงได้ดังตารางที่ 2.1

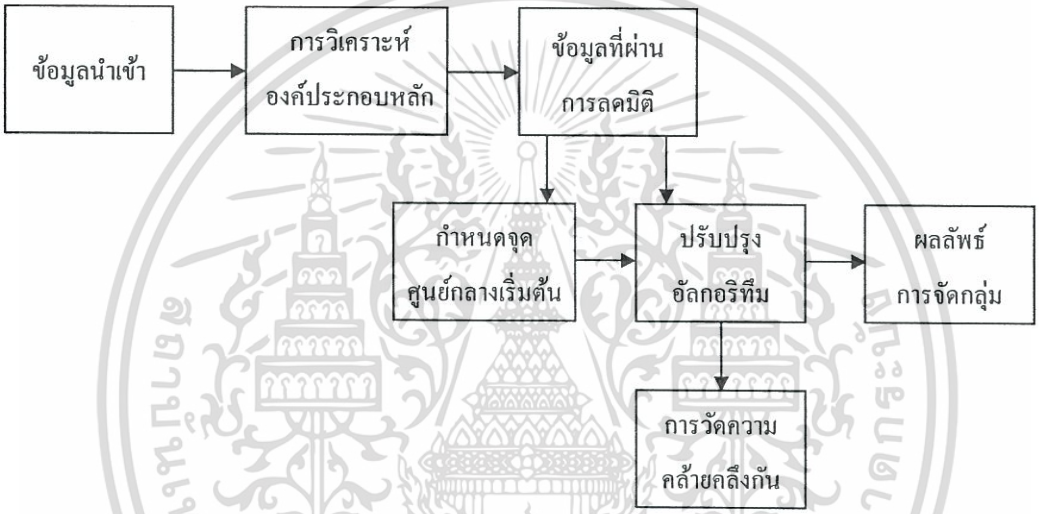
ตารางที่ 2.1 ความถูกต้องและเวลาที่ใช้ในการจัดกลุ่มของอัลกอริทึม 3 วิธี

Data Sets	Algorithms					
	K-means		Improved K-means (Fahim et.al, 2006)		Improved K-means (Nazeer, Kumar and Sebastian. 2011)	
	Accuracy (%)	Time Taken(ms)	Accuracy (%)	Time Taken(ms)	Accuracy (%)	Time Taken(ms)
E-coli	79.7	64	81.5	48	81.5	40
Breast Cancer	96.0	68	96.2	56	96.2	42
Thyroid	75.0	60	82.3	56	86.0	52

จากการศึกษางานวิจัยนี้ สามารถวิเคราะห์จุดเด่นของอัลกอริทึมที่งานวิจัยนี้นำเสนอได้ว่า อัลกอริทึมที่งานวิจัยนี้นำเสนอสามารถกำหนดจุดศูนย์กลางเริ่มต้นแทนที่การสุ่มข้อมูลแบบเคมีนส์ดั้งเดิม ทำให้การจัดกลุ่มที่ได้มีประสิทธิภาพมากขึ้นทั้งในแง่ของความถูกต้องและเวลาที่ใช้ในการประมวลผล แต่อย่างไรก็ตาม เมื่อเปรียบเทียบผลลัพธ์การจัดกลุ่มที่ได้จากอัลกอริทึมของงานวิจัยนี้กับผลลัพธ์การจัดกลุ่มที่ได้จากเคมีนส์ดั้งเดิมและเคมีนส์ที่ได้รับการปรับปรุงโดย Fahim et.al. (2006) ประสิทธิภาพในแง่ของความถูกต้องก็ยังเพิ่มขึ้นไม่มากเท่าที่ควร และหากจำนวนข้อมูลภายในกลุ่มแต่ละกลุ่มแตกต่างกัน จุดศูนย์กลางเริ่มต้นที่หาได้จะไม่มีประสิทธิภาพ

2.2.2 An Efficient Method to Improve the Clustering Performance for High Dimensional Data by Principal Component Analysis and Modified K-means (Tajunisha and Saravanan. 2011)

งานวิจัยนี้นำเสนอการประยุกต์ใช้การวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis : PCA) ในการกำหนดข้อมูลจุดศูนย์กลางเริ่มต้นและลดมิติของข้อมูลเพื่อเพิ่มความถูกต้องในการจัดกลุ่มและลดเวลาที่ใช้ในการประมวลผล และนำเสนอการปรับปรุงแก้ไขอัลกอริทึมของเคมีนส์บางขั้นตอนเพื่อลดการคำนวณระยะทางระหว่างข้อมูล โดยขั้นตอนในภาพรวมของงานวิจัยนี้แสดงได้ดังรูปที่ 2.3



รูปที่ 2.3 ขั้นตอนในภาพรวมของงานวิจัย

งานวิจัยนี้ประกอบด้วยขั้นตอนหลัก 2 ขั้นตอน ขั้นตอนแรกคือการลดมิติข้อมูลและการกำหนดจุดศูนย์กลางเริ่มต้นโดยใช้การวิเคราะห์องประกอบหลัก และขั้นตอนที่ 2 คือการนำข้อมูลทั้งหมดมาจัดเข้ากลุ่มโดยใช้อัลกอริทึมเคมีนส์ที่ปรับปรุงแล้ว

ขั้นตอนการลดมิติข้อมูลและการกำหนดจุดศูนย์กลางเริ่มต้นมีรายละเอียดดังนี้

1) ลดมิติของข้อมูลโดยใช้การวิเคราะห์องค์ประกอบหลัก
2) หาแกนองค์ประกอบหลักแกนแรก ซึ่งเป็นแกนที่มีค่าความแปรปรวน (Variance) สูงสุด
3) เรียงลำดับข้อมูลที่ผ่านการลดมิติมาแล้วตามแกนองค์ประกอบหลัก
4) แบ่งข้อมูลที่เรียงลำดับแล้วออกเป็น K ส่วน
5) หาค่ามัธยฐาน (Median) ของข้อมูลแต่ละส่วน
6) ใช้ค่ามัธยฐานที่ได้เป็นจุดศูนย์กลางเริ่มต้น

ขั้นตอนการนำข้อมูลทั้งหมดมาจัดเข้ากลุ่มมีรายละเอียดดังนี้

1) คำนวณระยะทางระหว่างข้อมูล x_i (1<=i<=n) กับจุดศูนย์กลางเริ่มต้น c_j (1<=j<=k) โดยใช้การคำนวณระยะทางแบบยุคลิดตามสมการที่ 2.1

2) นำข้อมูลทั้งหมดจัดเข้ากลุ่มที่มีระยะทางใกล้ที่สุด แล้วเก็บกลุ่มที่ใกล้ที่สุดนั้นลงในอาร์เรย์ Cluster[] และเก็บระยะทางระหว่างข้อมูลกับจุดศูนย์กลางของกลุ่มที่อยู่ใกล้ที่สุดนั้นลงในอาร์เรย์ Dist[]

3) คำนวณจุดศูนย์กลางของแต่ละกลุ่มใหม่โดยคำนวณจากค่าเฉลี่ย

4) คำนวณระยะทางระหว่างข้อมูลนั้นๆ กับจุดศูนย์กลางใหม่ของกลุ่มที่ข้อมูลนั้นอยู่ หากระยะทางที่ได้มีค่าน้อยกว่าหรือเท่ากับระยะทางเดิม ข้อมูลนั้นจะอยู่ในกลุ่มเดิม แต่หากระยะทางที่ได้มีค่ามากกว่าระยะทางเดิม จะต้องคำนวณระยะทางระหว่างข้อมูลนั้นกับจุดศูนย์กลางของทุกกลุ่มใหม่ ระยะทางระหว่างข้อมูลกับจุดศูนย์กลางของกลุ่มใดใกล้ที่สุด ข้อมูลจะถูกจัดเข้าไปอยู่ในกลุ่มนั้น แล้วเก็บกลุ่มที่ใกล้ที่สุดนั้นลงในอาร์เรย์ Cluster[] และเก็บระยะทางระหว่างข้อมูลกับจุดศูนย์กลางของกลุ่มที่อยู่ใกล้ที่สุดนั้นลงในอาร์เรย์ Dist[]

5) ทำซ้ำตั้งแต่ขั้นตอนที่ 3-4 จนกระทั่งไม่มีการเปลี่ยนแปลง

งานวิจัยนี้ได้ทดสอบประสิทธิภาพในการจัดกลุ่มโดยใช้ข้อมูลจากคลังข้อมูลยูซีไอ ชุดข้อมูลที่ใช้ ได้แก่ ชุดข้อมูลดอกไอริส (Iris) ชุดข้อมูลกระจก (Glass) ชุดข้อมูลไวน์ (Wine) และชุดข้อมูลการแบ่งส่วนข้อมูล (Image Segmentation) โดยเปรียบเทียบอัลกอริทึมที่งานวิจัยนี้นำเสนอกับการจัดกลุมแบบเคมีนส์ดั้งเดิม การจัดกลุ่มแบบเคมีนส์โดยใช้การวิเคราะห์องค์ประกอบหลักเพื่อลดมิติข้อมูลแต่จุดศูนย์กลางเริ่มต้นยังใช้การสุ่มข้อมูล และการจัดกลุ่มแบบเคมีนส์ที่ใช้การวิเคราะห์องค์ประกอบหลักเพื่อลดมิติข้อมูลและเพื่อกำหนดจุดศูนย์กลางเริ่มต้น

จากการทดสอบประสิทธิภาพในการจัดกลุ่มของอัลกอริทึมทั้ง 4 วิธี สามารถสรุปได้ว่าเมื่อนำการวิเคราะห์องค์ประกอบหลักมาใช้เพื่อลดมิติของข้อมูลร่วมกับการจัดกลุ่มแบบเคมีนส์ดั้งเดิม และการนำการวิเคราะห์องค์ประกอบหลักและการกำหนดจุดศูนย์กลางเริ่มต้นร่วมกับการจัดกลุ่มแบบเคมีนส์ดั้งเดิม ประสิทธิภาพในการจัดกลุ่มทั้งในแง่ของความถูกต้องและเวลาที่ใช้ในการประมวลผลดีกว่าการจัดกลุ่มแบบเคมีนส์ดั้งเดิมเพียงอย่างเดียว และเมื่อนำการนำการวิเคราะห์องค์ประกอบหลักและการกำหนดจุดศูนย์กลางเริ่มต้นมาใช้ร่วมกับอัลกอริทึมของเคมีนส์ที่ผ่านการปรับปรุงเพื่อลดการคำนวณระยะทางแล้ว ความถูกต้องที่ได้จะเท่าเดิม แต่ในแง่ของเวลาที่ใช้ในการประมวลผลจะมีความรวดเร็วมากขึ้น

จากการศึกษางานวิจัยนี้ สามารถวิเคราะห์จุดเด่นของอัลกอริทึมที่งานวิจัยนี้นำเสนอได้ว่าการนำการวิเคราะห์องค์ประกอบหลักและการกำหนดจุดศูนย์กลางเริ่มต้นมาใช้ร่วมกับการปรับปรุงอัลกอริทึมของเคมีนส์เพื่อลดการคำนวณระยะทาง จะทำให้ผลลัพธ์ในการจัดกลุ่มมีความถูกต้องและมีความรวดเร็วในการประมวลผลมากขึ้น แต่อย่างไรก็ตามหากจำนวนข้อมูลภายในกลุ่มแต่ละกลุ่มแตกต่างกัน จุดศูนย์กลางเริ่มต้นที่หาได้จะไม่มีประสิทธิภาพ

2.2.3 A Clustering Method Based on K-Means Algorithm (Li and Wu. 2012)

งานวิจัยนี้นำเสนออัลกอริทึมในการกำหนดข้อมูลจุดศูนย์กลางเริ่มต้นแทนที่การสุ่มข้อมูลตามอัลกอริทึมของเคมีนส์แบบดั้งเดิม แนวความคิดหลักของงานวิจัยนี้คือการพยายามให้ข้อมูลที่ถูกเลือกขึ้นมาเป็นจุดศูนย์กลางเริ่มต้นทุกตัวอยู่ห่างกันมากที่สุด โดยมีรายละเอียดของอัลกอริทึมดังนี้

กำหนดให้ข้อมูลทั้งหมดคือ $\{x_1,x_2,\ldots,x_N\}$

1) สุ่มเลือกข้อมูลขึ้นมา 1 ตัวจาก $\{x_1,x_2,\ldots,x_N\}$ เพื่อใช้เป็นจุดศูนย์กลางเริ่มต้นกลุ่มแรก (z_1)

2) คำนวณหาระยะทางระหว่างข้อมูลแต่ละตัวกับ z_1 เลือกข้อมูลที่มีระยะทางห่างจาก z_1 มากที่สุดมาเป็นจุดศูนย์กลางเริ่มต้นของกลุ่มที่ 2

$$\| x_j - z_1\| = \max_i \{ \| x_i - z_1\| \},\ i=1, 2, \ldots, N \tag{2.4}$$

จะได้จุดศูนย์กลางของกลุ่มที่ 2 $(z_2) = x_j$

3) คำนวณหาระยะทางระหว่างข้อมูลแต่ละตัวกับ *z1* และ *z2*

$$d_{i1} = \| x_i - z_1\|,\ i=1, 2, \ldots, N \tag{2.5}$$

$$d_{i2} = \| x_i - z_2\|,\ i=1, 2, \ldots, N \tag{2.6}$$

เปรียบเทียบ z_1 และ z_2 ของข้อมูลแต่ละตัว แล้วนำค่าที่น้อยกว่าของข้อมูลทุกตัวมาเปรียบเทียบเพื่อหาค่าที่มากที่สุดอีกครั้งหนึ่ง ค่าใดมากที่สุด จะใช้ข้อมูลตัวนั้นเป็นจุดศูนย์กลางเริ่มต้นของกลุ่มที่ 3

$$\min(d_{j1}, d_{j2}) = \max_i \{ \min(d_{i1},d_{i2}) \},\ i =1, 2, \ldots, N \tag{2.7}$$

จะได้จุดศูนย์กลางของกลุ่มที่ 3 $(z_3) = x_j$

4) หาจุดศูนย์กลางเริ่มต้นให้ครบตามจำนวนกลุ่ม (k) สมมุติว่าได้จุดศูนย์กลางมาแล้วทั้งหมด r จุด ($r<k$) แต่ละจุดคือ $\{z_i,\ I = 1,2,\ldots, r\}$ จะสามารถหาจุดศูนย์กลางเริ่มต้นของจุดที่ r+1 ได้ดังนี้

$$\min(d_{j1}, d_{j2}, \ldots, d_{jr}) = \max_i \{ \min(d_{i1},d_{i2}, \ldots, d_{ir}) \},\ i =1, 2, \ldots, N \tag{2.8}$$

จะได้จุดศูนย์กลางของกลุ่มที่ r+1 $(z_{r+1}) = x_j$ ทำซ้ำจนกระทั่งได้จุดศูนย์กลางครบทั้งหมด k กลุ่ม

5) จัดกลุ่มโดยใช้เคมีนส์ตามปกติ

งานวิจัยนี้ได้ทดสอบประสิทธิภาพในการจัดกลุ่มโดยสร้างข้อมูล 2 มิติขึ้นมา 20 ตัว โดยข้อมูลแบ่งได้เป็น 5 ประเภท แล้วเปรียบเทียบประสิทธิภาพที่ได้จากการจัดกลุ่มแบบเคมีนส์ดั้งเดิมและการจัดกลุ่มโดยใช้อัลกอริทึมที่งานวิจัยนี้นำเสนอ โดยข้อมูลที่สร้างขึ้นมาเพื่อใช้ในการทดลองแสดงดังตารางที่ 2.2

ตารางที่ 2.2 ข้อมูลที่ใช้ในการทดลองการจัดกลุ่ม

Pattern	Abscissa	Ordinate	Class
X1	1	1	1
X2	1.5	1.5	1
X3	1.5	1.1	1
X4	81	80	4
X5	7.3	8	2
X6	35.7	33.4	5
X7	8	7.3	2
X8	21.2	20	3
X9	81	73	4
X10	6.9	7.6	2
X11	1.69	0.93	1
X12	0.3	1.1	1
X13	7	7.4	2
X14	6.9	6.9	2
X15	22.2	20.5	3
X16	23	21	3
X17	80.6	73.2	4
X18	36.7	38.55	5
X19	34.76	33.6	5
X20	81	73.6	4

งานวิจัยนี้มีการทดสอบประสิทธิภาพโดยพิจารณาจากฟังก์ชันเป้าหมาย (Criterion Function) J_j ซึ่งการจัดกลุ่มที่ดีจะมีค่าฟังก์ชันเป้าหมายที่น้อย โดยฟังก์ชันเป้าหมายสามารถคำนวณได้จาก

$$J_j = \sum_{x \in s_j(t)} \| x - z_j(t+1) \|^2, \quad j=1, 2, \ldots, k \qquad (2.9)$$

เมื่อ s_j คือข้อมูลกลุ่มที่ j

t คือจำนวนรอบในการประมวลผล

ผลการทดสอบประสิทธิภาพในการจัดกลุ่มแบบเคมีนส์ดั้งเดิมและอัลกอริทึมที่งานวิจัยนี้นำเสนอแสดงดังตารางที่ 2.3

ตารางที่ 2.3 ผลการทดสอบประสิทธิภาพในการจัดกลุ่มของอัลกอริทึมทั้ง 2 วิธี

	Standard K-means	Improved K-means (Li and Wu. 2012)
Iterations	9	6
Cluster Criterion Function *J*	657.603	58.3263
First class	X1,X12	X1,X2,X3,X11,X12
Second class	X5,X7,X10,X13,X14	X5,X7,X10,X13,X14
Third class	X2,X3,X11	X8,X15,X16
Forth class	X4,X9,X17,X20	X4,X9,X17,X20
Fifth class	X6,X8,X15,X16,X18,19	X6,X18,X19

จากตารางที่ 2.3 จะเห็นว่าอัลกอริทึมที่งานวิจัยนี้ทำเสนอใช้จำนวนรอบในการลู่เข้าสู่จุดที่ดีที่สุดน้อยกว่าการจัดกลุ่มแบบเคมีนส์ดั้งเดิม นอกจากนี้ยังได้ค่าฟังก์ชันเป้าหมายที่น้อยกว่า นั่นหมายถึงมีประสิทธิภาพในการจัดกลุ่มที่ดีกว่า และจะเห็นว่าข้อมูลที่อยู่ภายในกลุ่มมีความถูกต้องมากกว่า

จากการศึกษางานวิจัยนี้ สามารถวิเคราะห์จุดเด่นของอัลกอริทึมที่งานวิจัยนี้นำเสนอได้ว่า อัลกอริทึมที่งานวิจัยนี้นำเสนอสามารถกำหนดจุดศูนย์กลางเริ่มต้นแทนที่การสุ่มข้อมูลแบบเคมีนส์ดั้งเดิม ทำให้การจัดกลุ่มที่ได้มีประสิทธิภาพมากขึ้น นอกจากนี้ยังใช้เวลาในการลู่เข้าสู่จุดที่ดีที่สุดได้รวดเร็ว แต่อย่างไรก็ตาม อัลกอริทึมที่งานวิจัยนี้นำเสนอยังต้องสุ่มข้อมูลขึ้นมา 1 ตัวเพื่อใช้เป็นจุดศูนย์กลางเริ่มต้นของกลุ่มแรก ดังนั้น ผลลัพธ์ในการจัดกลุ่มจึงอาจจะยังขึ้นกับจุดศูนย์กลางเริ่มต้นของกลุ่มแรกที่สุ่มขึ้นมา หากสุ่มข้อมูลตัวแรกขึ้นมาได้ไม่ดี ก็อาจจะทำให้ประสิทธิภาพที่ได้จากการจัดกลุ่มลดลง

2.2.4 A new algorithm for initial cluster centers in K-means algorithm (Erisoglu, Calis and Sakallioglu. 2011)

งานวิจัยนี้นำเสนอวิธีการคำนวณหาจุดศูนย์กลางเริ่มต้นให้กับการจัดกลุ่มแบบเคมีนส์โดยการเลือกข้อมูลขึ้นมา 2 มิติที่สามารถใช้เป็นตัวแทนของข้อมูลทั้งหมดได้ แล้วทำการคำนวณหาจุดศูนย์กลางเริ่มต้นโดยใช้ข้อมูล 2 มิติที่เลือกมาเป็นหลัก โดยหลักการของการคำนวณหาจุดศูนย์กลางเริ่มต้นที่งานวิจัยนี้นำเสนอคือพยายามให้ข้อมูลที่ถูกเลือกขึ้นมาเป็นจุดศูนย์กลางเริ่มต้นทุกตัวอยู่ห่างกันมากที่สุด โดยมีรายละเอียดของอัลกอริทึมดังนี้

1) เลือกข้อมูลขึ้นมา 1 มิติเพื่อใช้เป็นมิติหลักที่ใช้เป็นตัวแทนของข้อมูล โดยการคำนวณหาค่าสัมบูรณ์ของค่าสัมประสิทธิ์การแปรผัน (Variation Coefficient) ซึ่งค่าสัมประสิทธิ์การแปรผันสามารถคำนวณได้จากสมการที่ 2.10 โดยมิติที่มีค่าสัมบูรณ์ของค่าสัมประสิทธิ์การแปรผันสูงที่สุดจะถูกเลือกให้เป็นมิติหลัก

$$cv_j = \frac{s(x_j)}{\bar{x}_j}, \; j = 1, 2, \ldots, p \tag{2.10}$$

เมื่อ $s(x_j)$ คือค่าส่วนเบี่ยงเบนมาตรฐานของข้อมูลในมิติ j

$\bar{x}_j$ คือค่าเฉลี่ยของข้อมูลในมิติ j

2) เลือกข้อมูลมิติที่ 2 เพื่อใช้เป็นตัวแทนของข้อมูล โดยการคำนวณหาค่าสัมบูรณ์ของค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) ซึ่งค่าสัมประสิทธิ์สหสัมพันธ์สามารถคำนวณได้จากสมการที่ 2.11 โดยมิติที่มีค่าสัมบูรณ์ของค่าสัมประสิทธิ์สหสัมพันธ์น้อยที่สุดจะถูกเลือกให้เป็นมิติที่ 2

$$r_{jj'} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'})}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}\sqrt{\sum_{i=1}^{n}(x_{ij'} - \bar{x}_{j'})^2}} \tag{2.11}$$

เมื่อ n คือจำนวนข้อมูลทั้งหมด

j' มิติหลักที่ได้จากขั้นตอนที่ 1

3) หลังจากที่ได้มิติ 2 มิติที่สามารถใช้เป็นตัวแทนของข้อมูลได้แล้ว ให้คำนวณหาค่าเฉลี่ยของข้อมูล 2 มิตินี้ โดยค่าเฉลี่ยของข้อมูล 2 มิตินี้กำหนดให้เป็นจุด *m*

$$m = [\bar{x}_I \quad \bar{x}_{II}] \tag{2.12}$$

เมื่อ $\bar{x}_I$ คือค่าเฉลียของข้อมูลในมิติที่ 1

$\bar{x}_{II}$ คือค่าเฉลียของข้อมูลในมิติที่ 2

4) คำนวณหาระยะทางแบบยูคลิดระหว่างข้อมูลแต่ละตัวกับจุด *m* ที่ได้จากขั้นตอนที่ 3

$$d_{im} = d(x_i, m), \quad i = 1, 2, \ldots, n \tag{2.13}$$

เมื่อ $d(x_i, m)$ คือระยะทางแบบยูคลิดระหว่าง x_i กับ จุด *m*

n คือจำนวนข้อมูลทั้งหมด

ข้อมูลตัวที่มีระยะทางไปยังจุด *m* สูงที่สุดจะถูกเลือกให้เป็นจุดศูนย์กลางเริ่มต้นจุดที่ 1 (c_1)

5) คำนวณหาระยะทางแบบยูคลิดระหว่างข้อมูลแต่ละตัวกับจุดศูนย์กลางเริ่มต้นจุดที่ 1 ที่ได้จากขั้นตอนที่ 4

$$d_{ic1} = d(x_i, c_1), \quad i = 1, 2, \ldots, n \tag{2.14}$$

เมื่อ $d(x_i, c_1)$ คือระยะทางแบบยูคลิดระหว่าง x_i กับ จุด c_1

n คือจำนวนข้อมูลทั้งหมด

ข้อมูลตัวที่มีระยะทางไปยังจุด c_1 สูงที่สุดจะถูกเลือกให้เป็นจุดศูนย์กลางเริ่มต้นจุดที่ 2 (c_2)

6) คำนวณหาจุดศูนย์กลางเริ่มต้นตัวต่อไป (c_r) โดยการคำนวณหาผลรวมของระยะทางระหว่างข้อมูลแต่ละตัวไปยังจุดศูนย์กลางเริ่มต้นก่อนหน้าทั้งหมด

$$Sd_{ir} = d_{ic1} + d_{ic2} + \ldots + d_{ic(r-1)}, \quad i = 1, 2, \ldots, n \tag{2.15}$$

เมื่อ r คือลำดับที่ของจุดศูนย์กลางเริ่มต้นที่ต้องการคำนวณหา

n คือจำนวนข้อมูลทั้งหมด

ข้อมูลตัวที่มีค่า Sd_{ir} สูงที่สุดจะถูกเลือกให้เป็นจุดศูนย์กลางเริ่มต้นตัวต่อไป

7) ทำซ้ำขั้นตอนที่ 6 จนกระทั่งได้จุดศูนย์กลางเริ่มต้นครบตามจำนวนกลุ่ม (K)

8) จัดกลุ่มแบบเคมีนส์โดยใช้จุดศูนย์กลางเริ่มต้นที่คำนวณได้ โดยจัดกลุ่มข้อมูลเพียง 2 มิติหลัก

9) คำนวณหาจุดศูนย์กลางเริ่มต้นของทุกมิติโดยการหาค่าเฉลี่ยของข้อมูลที่อยู่ในกลุ่มเดียวกันโดยใช้ค่าความเป็นสมาชิกของกลุ่ม (Cluster Membership) ที่ได้จากขั้นตอนที่ 8

10)จัดกลุ่มโดยใช้เคมีนส์ตามปกติ

งานวิจัยนี้ได้ทดสอบประสิทธิภาพในการจัดกลุ่มโดยใช้ชุดข้อมูลจากคลังข้อมูลยูซีไอ ชุดข้อมูลที่ใช้ได้แก่ Iris, Wine, Letter Image Recognition, Ruspini และ Spambase โดยงานวิจัยนี้จะแบ่งการทดลองออกเป็น 2 การทดลอง การทดลองที่ 1 จะทดลองเพื่อเปรียบเทียบประสิทธิภาพการจัดกลุ่มแบบเคมีนส์โดยใช้วิธีการคำนวณหาจุดศูนย์กลางเริ่มต้นที่งานวิจัยนี้นำเสนอกับวิธีการกำหนดจุดศูนย์กลางเริ่มต้นแบบสุ่ม โดยเกณฑ์ที่ใช้วัดประสิทธิภาพของการจัดกลุ่มในการทดลองที่ 1 ได้แก่ Error percentage, Rand index และ Wilks' lambda test statistic ซึ่งผลการวัดประสิทธิภาพของการทดลองที่ 1 แสดงได้ดังตารางที่ 2.4

ตารางที่ 2.4 ผลการทดสอบประสิทธิภาพของการทดลองที่ 1

Dataset	Method	Error percentage	Rand index	Wilks' lambda
Iris	Improved K-means (Erisoglu et.al, 2006)	10.7	0.8797	0.0322
	Random	13.83	0.8639	0.0376
Wine	Improved K-means (Erisoglu et.al, 2006)	3.4	0.9543	0.0196
	Random	10.58	0.9018	0.0329
Letter	Improved K-means (Erisoglu et.al, 2006)	7.9046	0.8543	0.0877
	Random	9.738	0.6364	0.1071
Ruspini	Improved K-means (Erisoglu et.al, 2006)	0	1	0.0034
	Random	21.8667	0.8887	0.016
Spambase	Improved K-means (Erisoglu et.al, 2006)	36.4051	0.5369	0.4171
	Random	39.3393	0.5226	0.5912

การทดลองที่ 2 จะทดลองเพื่อเปรียบเทียบประสิทธิภาพการจัดกลุ่มแบบเคมีนส์โดยใช้วิธีการคำนวณหาจุดศูนย์กลางเริ่มต้นที่งานวิจัยนี้นำเสนอกับวิธีการที่นำเสนอโดย Khan และ Ahmad (2004) หรือมีชื่ออัลกอริธึมคือ "CCIA" และวิธีการที่นำเสนอโดย Deelers และ Auwatanamongkol (2007) โดยเกณฑ์ที่ใช้วัดประสิทธิภาพของการจัดกลุ่มในการทดลองที่ 2 คือ Error percentage ซึ่งผลการวัดประสิทธิภาพของการทดลองที่ 2 แสดงได้ดังตารางที่ 2.5

ตารางที่ 2.5 ผลการทดสอบประสิทธิภาพของการทดลองที่ 2

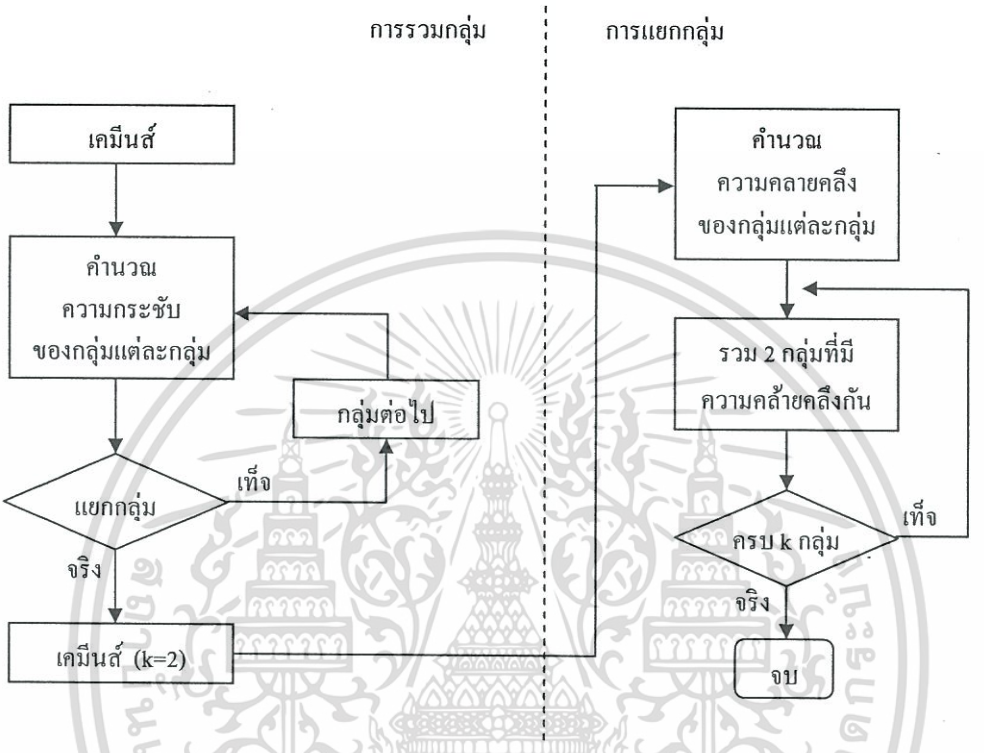
Dataset	Method	Error percentage
Iris	CCIA	10.8
	Deeler etc	10.1
	Improved K-means (Erisoglu et.al, 2006)	10
Wine	CCIA	4.15
	Deeler etc	4.9
	Improved K-means (Erisoglu et.al, 2006)	3.2
Letter	CCIA	7.7
	Deeler etc	8.05
	Improved K-means (Erisoglu et.al, 2006)	7.1
Ruspini	CCIA	3.15
	Deeler etc	3.15
	Improved K-means (Erisoglu et.al, 2006)	0

จากการทดลองทั้ง 2 การทดลองจะเห็นว่าวิธีการคำนวณหาจุดศูนย์กลางเริ่มต้นที่นำวิจัยนี้นำเสนอให้ประสิทธิภาพในการจัดกลุ่มที่ดีกว่าวิธีการกำหนดจุดศูนย์กลางเริ่มต้นแบบสุ่ม รวมถึงให้ประสิทธิภาพที่ดีกว่าวิธีการคำนวณหาจุดศูนย์กลางเริ่มต้นที่นำเสนอโดยงานวิจัยอีก 2 งานวิจัย แต่อย่างไรก็ตาม งานวิจัยนี้ก็ยังมีข้อจำกัดบางประการ คือ ในขั้นตอนของการคำนวณหาจุดศูนย์กลางเริ่มต้นตั้งแต่ตัวที่ 3 ขึ้นไปจะเลือกข้อมูลตัวที่มีค่าผลรวมของระยะทางระหว่างข้อมูลแต่ละตัวไปยังจุดศูนย์กลางเริ่มต้นก่อนหน้าทั้งหมด (Sd_{ir}) ซึ่งวิธีการนี้อาจจะทำให้จุดศูนย์กลางตัวที่ถูกเลือกขึ้นมาไม่ได้อยู่ห่างจากจุดศูนย์กลางเริ่มต้นตัวก่อนหน้าทั้งหมดมากที่สุด

2.2.5 An Improved Clustering Method Based on K-means (Lin et.al. 2012)

ปกติแล้วการจัดกลุ่มแบบเคมีนส์จะมีปัญหาในการจัดกลุ่มที่มีรูปร่างซับซ้อน หรือกลุ่มแต่ละกลุ่มมีขนาดและความหนาแน่นแตกต่างกัน ซึ่งงานวิจัยนี้นำเสนอการปรับปรุงอัลกอริทึมของเคมีนส์ให้สามารถแก้ปัญหานี้ได้ โดยใช้หลักการแยกกลุ่ม (Split) และการรวมกลุ่ม (Merge)

ในการพิจารณาว่าจะทำการแยกกลุ่มหรือไม่ จะพิจารณาจากความหนาแน่นของข้อมูล และในการพิจารณาว่าจะทำการรวมกลุ่มหรือไม่ จะพิจารณาจากค่าเฉลี่ยของระยะทางในแต่ละกลุ่ม โดยในรูปที่ 2.4 จะแสดงกระบวนการในภาพรวมทั้งในส่วนของการแยกกลุ่มและการรวมกลุ่ม



รูปที่ 2.4 กระบวนการในภาพรวมของงานวิจัยนี้

รายละเอียดของอัลกอริทึมในขั้นตอนของการแยกกลุ่มมีดังนี้

1) จัดกลุ่มโดยใช้เคมีนส์ตามปกติ

2) หลังจากที่จัดกลุ่มโดยใช้เคมีนส์ตามปกติแล้ว จะต้องมาพิจารณาว่าข้อมูลแต่ละกลุ่มจะต้องถูกแยกกลุ่มอีกครั้งหรือไม่ โดยขั้นตอนแรกจะต้องวัดระยะห่างที่มากที่สุดภายในกลุ่มนั้นๆ (D_{max})

3) สร้างจุดขึ้นมาชั่วคราวที่ระยะ 1/4 , 2/4 และ 3/4 ของช่วงระยะ D_{max}

4) คำนวณหาค่ารัศมี r จาก

$$r = a \times D_{max} / 8 \qquad (2.16)$$

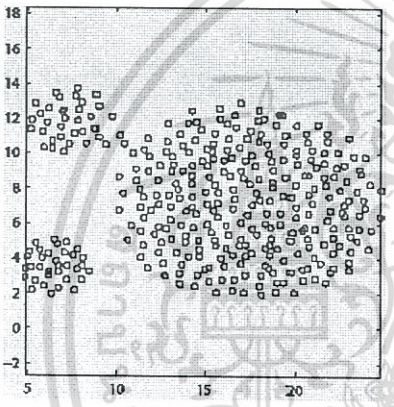
เมื่อ a คือพารามิเตอร์ที่ต้องกำหนด ซึ่งในงานวิจัยนี้กำหนดค่า a เท่ากับ 0.5

5) พิจารณาข้อมูลรอบๆ จุดที่สร้างขึ้นมาชั่วคราวทั้ง 3 จุดจากข้อ 3 ภายในรัศมี r ที่คำนวณได้จากข้อ 4 หากมีรัศมีใดๆ มีข้อมูลภายในน้อยกว่าค่าที่กำหนด (*ch*) กลุ่มนั้นจะต้องถูกแยกออกเป็น 2

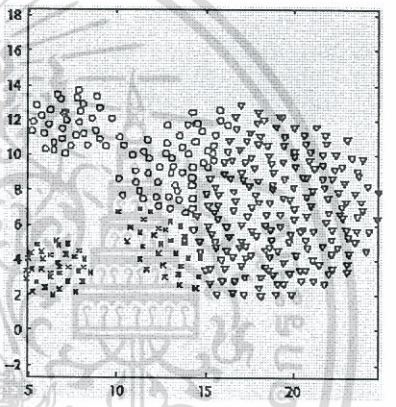
กลุ่มด้วยเคมีนส์อีกครั้งหนึ่ง กลุ่ม 2 กลุ่มที่ได้จากการแยกกลุ่มให้มาร์คค่าเท่ากับ 0 ส่วนกลุ่มที่ไม่มีการแยกกลุ่มออกอีกครั้งให้มาร์คค่าเท่ากับ 1

6) ทำซ้ำตั้งแต่ขั้นตอนที่ 2-6 จนครบทุกกลุ่ม และตรวจสอบว่ายังมีกลุ่มใดที่ยังถูกมาร์คค่าไว้เท่ากับ 0 หรือไม่ หากยังมีกลุ่มที่ถูกมาร์คค่าไว้เท่ากับ 0 ให้ทำซ้ำตั้งแต่ข้อ 2-5 จนกระทั่งทุกกลุ่มถูกมาร์คค่าไว้เท่ากับ 1

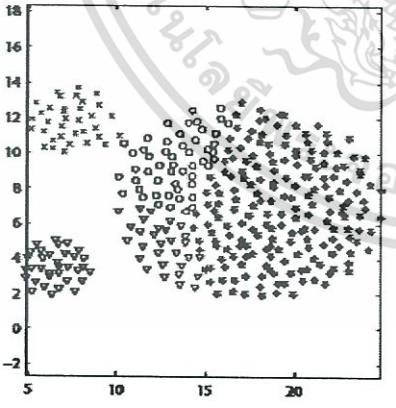
ในรูปที่ 2.5 แสดงตัวอย่างในขั้นตอนของการแยกกลุ่ม โดยข้อมูลที่ใช้จะมีทั้งหมด 3 กลุ่ม ดังแสดงในรูปที่ 2.5 (ก) ในขั้นตอนแรกคือการจัดกลุ่มข้อมูลโดยใช้การจัดกลุ่มแบบเคมีนส์ ดังแสดงในรูปที่ 2.5 (ข) หลังจากนั้นจะเข้าสู่ขั้นตอนของการแยกกลุ่ม โดยสามารถแยกกลุ่มได้ทั้งหมด 4 ครั้ง ได้ผลลัพธ์สุดท้ายทั้งหมด 8 กลุ่มดังแสดงในรูปที่ 2.5 (ค)-2.5 (ฉ)



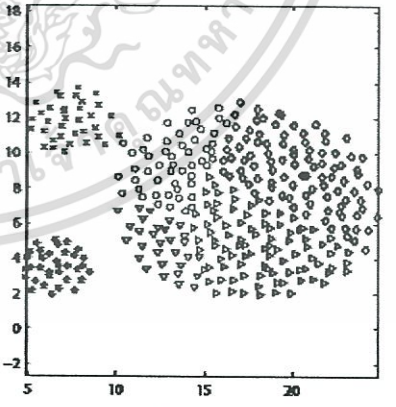
(ก) ข้อมูลที่นำมาใช้ในการจัดกลุ่ม



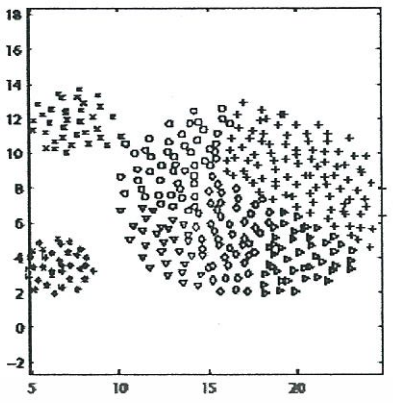
(ข) ผลลัพธ์ที่ได้จากการจัดกลุ่มแบบเคมีนส์



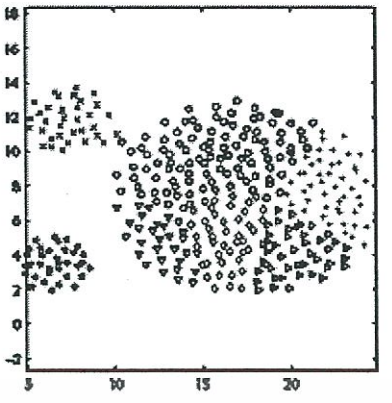
(ค) แยกกลุ่มครั้งที่ 1 ได้ทั้งหมด 4 กลุ่ม



(ง) แยกกลุ่มครั้งที่ 2 ได้ทั้งหมด 6 กลุ่ม



(จ) แยกกลุ่มครั้งที่ 3 ได้ทั้งหมด 7 กลุ่ม



(ฉ) แยกกลุ่มครั้งที่ 4 ได้ทั้งหมด 8 กลุ่ม

รูปที่ 2.5 ตัวอย่างขั้นตอนการแยกกลุ่ม (Lin et.al. 2012)

หลังจากขั้นตอนของการแยกกลุ่มแล้วก็จะเข้าสู่ขั้นตอนของการรวมกลุ่ม เพื่อรวมกลุ่มให้ได้ผลลัพธ์สุดท้ายเท่ากับ k กลุ่ม ซึ่งในที่นี้จะรวม 8 กลุ่มให้เหลือ 3 กลุ่ม ขั้นตอนการรวมกลุ่มมีรายละเอียดของอัลกอริทึมดังนี้

1) เลือกกลุ่มขึ้นมา 2 กลุ่ม กำหนดให้เป็น C_i และ C_j ($1 \leq i \leq k$ และ $1 \leq j \leq k$)

2) หาค่าระยะทางเฉลี่ยภายในกลุ่มของทั้ง 2 กลุ่ม กำหนดให้ระยะทางเฉลี่ยของทั้ง 2 กลุ่มเป็น D_1 และ D_2 โดยสามารถหาระยะทางเฉลี่ยภายในกลุ่มได้จาก

$$D = \left(\sum_{i=1}^{m-1} \left(\sum_{j=i+1}^{m} d_{ij}\right)\right) / n \qquad (2.17)$$

$$n = \sum_{s=1}^{m-1} s \qquad (2.18)$$

เมื่อ m คือจำนวนข้อมูลทั้งหมดใน 1 กลุ่ม

d_{ij} คือระยะทางระหว่างข้อมูล 1 คู่ในกลุ่มเดียวกัน

3) เปรียบเทียบค่า D_1 และ D_2 เลือกค่าที่น้อยกว่า กำหนดให้เป็น D_m

4) คำนวณค่าระยะทางเฉลี่ยระหว่าง C_i และ C_j โดยคำนวณจากสมการ 2.17 เปลี่ยนค่า d_{ij} เป็นระยะทางระหว่าง 2 กลุ่ม กำหนดให้ค่าระยะทางเฉลี่ยระหว่าง 2 กลุ่มเป็น D_3

5) หาผลต่างระหว่าง D_m และ D_3 จาก

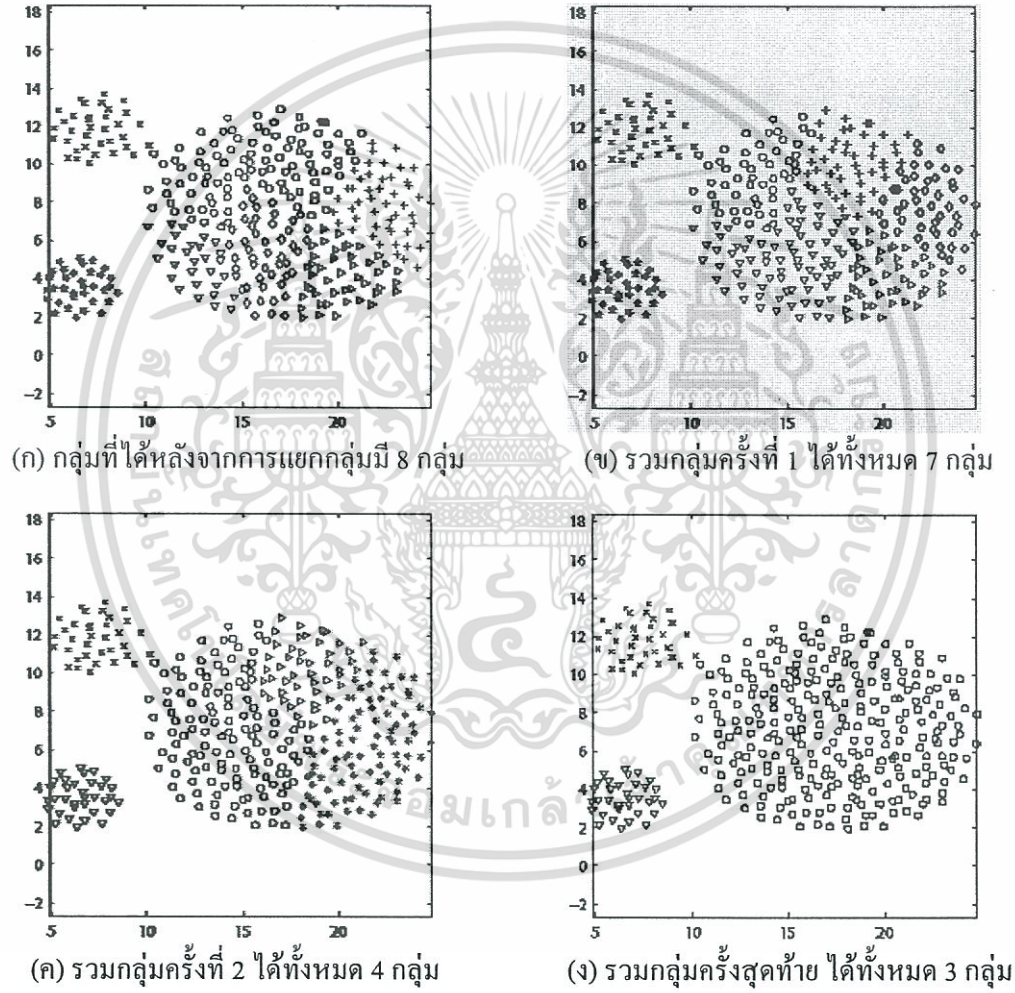
$$D_{ch} = |D_m - D_3| \qquad (2.19)$$

6) ทำซ้ำตั้งแต่ขั้นตอนที่ 1-5 จนกว่าทุกคู่ของกลุ่มจะถูกคำนวณ

7) เปรียบเทียบค่า D_{ch} ของทุกคู่ของกลุ่ม คู่ใดที่มีค่า D_{ch} น้อยที่สุด ให้รวม 2 กลุ่มนั้นเป็นกลุ่มเดียวกัน

8) ทำซ้ำตั้งแต่ขั้นตอนที่ 1-6 จนกว่าจะได้จำนวนกลุ่มเท่ากับค่า k

ในรูปที่ 2.6 แสดงตัวอย่างในขั้นตอนของการรวมกลุ่ม โดยกลุ่มที่ได้หลังจากขั้นตอนการแยกกลุ่มจะมีทั้งหมด 8 กลุ่ม จะต้องรวมกลุ่มให้ผลลัพธ์สุดท้ายเท่ากับ k กลุ่ม ซึ่งในที่นี้คือ 3 กลุ่ม



(ก) กลุ่มที่ได้หลังจากการแยกกลุ่มมี 8 กลุ่ม (ข) รวมกลุ่มครั้งที่ 1 ได้ทั้งหมด 7 กลุ่ม

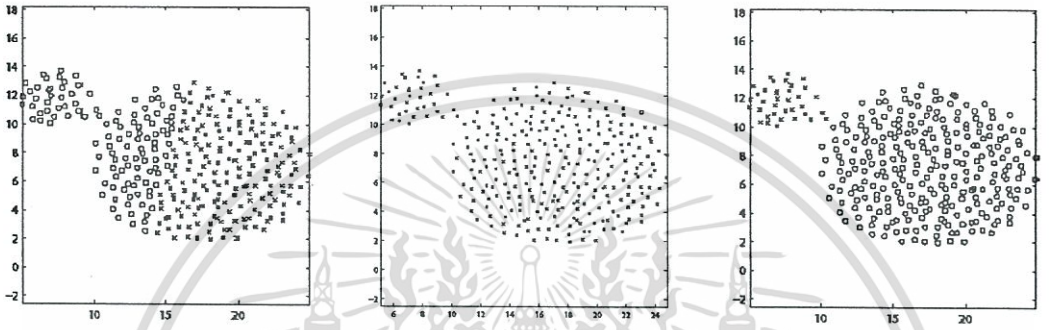
(ค) รวมกลุ่มครั้งที่ 2 ได้ทั้งหมด 4 กลุ่ม (ง) รวมกลุ่มครั้งสุดท้าย ได้ทั้งหมด 3 กลุ่ม

รูปที่ 2.6 ตัวอย่างขั้นตอนการรวมกลุ่ม (Lin et.al. 2012)

งานวิจัยนี้ได้ทดสอบประสิทธิภาพในการจัดกลุ่มเปรียบเทียบกับการจัดกลุ่มแบบเคมีนส์ดั้งเดิมและการจัดกลุ่มแบบการเชื่อมโยงแบบเดียว (Single-link clustering method) โดยในการทดลองแรกจะทดลองจัดกลุ่มจากข้อมูลที่สร้างขึ้นมา 3 ชุด

- ข้อมูลชุดที่ 1 : ข้อมูลมีทั้งหมด 2 กลุ่มที่มีขนาดแตกต่างกัน
- ข้อมูลชุดที่ 2 : ข้อมูลมีทั้งหมด 7 กลุ่มที่มีรูปร่างแตกต่างกัน
- ข้อมูลชุดที่ 3 : ข้อมูลมีทั้งหมด 6 กลุ่มที่มีความหนาแน่นแตกต่างกัน

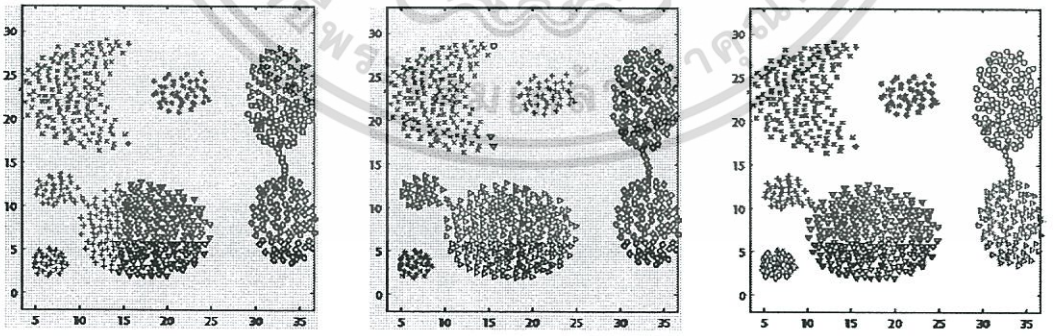
ผลการทดลองในการจัดกลุ่มของข้อมูลชุดที่ 1 ซึ่งมีขนาดของกลุ่มแตกต่างกัน โดยเปรียบเทียบอัลกอริทึมที่งานวิจัยนี้นำเสนอ การจัดกลุ่มแบบเคมีนส์ดั้งเดิม และการจัดกลุ่มแบบการเชื่อมโยงแบบเดียว แสดงได้ดังรูปที่ 2.7



(ก) เคมีนส์ดั้งเดิม (ข) การเชื่อมโยงแบบเดี่ยว (ค) อัลกอริทึมของงานวิจัยนี้

รูปที่ 2.7 เปรียบเทียบการจัดกลุ่มของข้อมูลที่ 1 ของอัลกอริทึมทั้ง 3 วิธี (Lin et.al. 2012)

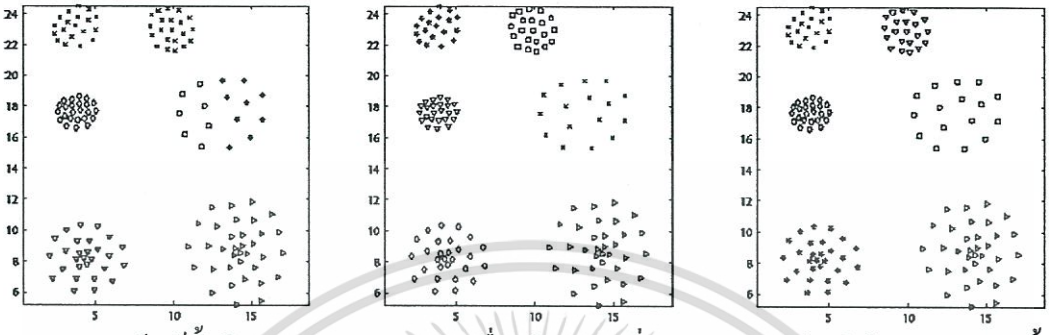
ผลการทดลองจัดกลุ่มของข้อมูลชุดที่ 2 ซึ่งมีรูปร่างของกลุ่มแตกต่างกัน โดยเปรียบเทียบอัลกอริทึมที่งานวิจัยนี้นำเสนอ การจัดกลุ่มแบบเคมีนส์ดั้งเดิม และการจัดกลุ่มแบบการเชื่อมโยงแบบเดียว แสดงได้ดังรูปที่ 2.8



(ก) เคมีนส์ดั้งเดิม (ข) การเชื่อมโยงแบบเดี่ยว (ค) อัลกอริทึมของงานวิจัยนี้

รูปที่ 2.8 เปรียบเทียบการจัดกลุ่มของข้อมูลที่ 2 ของอัลกอริทึมทั้ง 3 วิธี (Lin et.al. 2012)

ผลการทดลองจัดกลุ่มของข้อมูลชุดที่ 3 ซึ่งมีความหนาแน่นของกลุ่มแตกต่างกัน โดยเปรียบเทียบอัลกอริทึมที่งานวิจัยนี้นำเสนอ การจัดกลุ่มแบบเคมีนส์ดั้งเดิม และการจัดกลุ่มแบบการเชื่อมโยงแบบเดียว แสดงได้ดังรูปที่ 2.9



(ก) เคมีนส์ดั้งเดิม (ข) การเชื่อมโยงแบบเดี่ยว (ค) อัลกอริทึมของงานวิจัยนี้

รูปที่ 2.9 เปรียบเทียบการจัดกลุ่มของข้อมูลที่ 3 ของอัลกอริทึมทั้ง 3 วิธี (Lin et.al. 2012)

ในการทดลองที่ 2 จะใช้ชุดข้อมูลจากคลังข้อมูลยูซีไอ ชุดข้อมูลที่ใช้ได้แก่ ชุดข้อมูลดอกไอริส (Iris) และชุดข้อมูลไวน์ (Wine) โดยการวัดประสิทธิภาพในการจัดกลุ่มจะใช้การวัดความคล้ายคลึง 4 วิธี ได้แก่ Rand, Adjusted Rand, Jaccard และ FM โดยเปรียบเทียบอัลกอริทึมที่งานวิจัยนี้นำเสนอ การจัดกลุ่มแบบเคมีนส์ดั้งเดิม และการจัดกลุ่มแบบการเชื่อมโยงแบบเดียว ประสิทธิภาพในการจัดกลุ่มของชุดข้อมูลดอกไอริสแสดงดังตารางที่ 2.6 และประสิทธิภาพในการจัดกลุ่มของชุดข้อมูลไวน์แสดงดังตารางที่ 2.7

ตารางที่ 2.6 ประสิทธิภาพในการจัดกลุ่มของอัลกอริทึมทั้ง 3 วิธีของชุดข้อมูลดอกไอริส

Method	Rand	Adjusted Rand	Jaccard	FM
Improved K-means (Lin et.al. 2012)	0.9001	0.8500	0.8188	0.8992
K-means	0.8797	0.7302	0.6959	0.8208
Single-Link	0.7766	0.5638	0.5891	0.7653

ตารางที่ 2.7 ประสิทธิภาพในการจัดกลุ่มของอัลกอริทึมทั้ง 3 วิธีของชุดข้อมูลไวน์

Method	Rand	Adjusted Rand	Jaccard	FM
Proposed K-means (Lin et.al. 2012)	0.7391	0.4071	0.4286	0.7498
K-means	0.7183	0.3711	0.4120	0.7302
Single-Link	0.3628	0.0054	0.3325	0.5650

จากตารางที่ 2.6 และ 2.7 จะเห็นว่าค่าความคล้ายคลึงที่ได้จากการจัดกลุ่มโดยใช้อัลกอริทึมที่งานวิจัยนี้นำเสนอมีค่ามากกว่าการจัดกลุ่มที่ได้จากอีก 2 วิธี แสดงว่าอัลกอริทึมที่งานวิจัยนี้นำเสนอมีประสิทธิภาพในการจัดกลุ่มมากกว่า

จากการศึกษางานวิจัยนี้ สามารถวิเคราะห์จุดเด่นของอัลกอริทึมที่งานวิจัยนี้นำเสนอได้ว่า การจัดกลุ่มที่ใช้หลักการของการแยกกลุ่มและการรวมกลุ่มที่งานวิจัยนี้นำเสนอสามารถให้ประสิทธิภาพในการจัดกลุ่มที่ดีกว่าการจัดกลุ่มแบบเคมีนส์ดั้งเดิมและการจัดกลุ่มแบบการเชื่อมโยงแบบเดียว อัลกอริทึมที่งานวิจัยนี้นำเสนอสามารถจัดกลุ่มที่มีขนาด รูปร่าง และความหนาแน่นแตกต่างกันได้อย่างมีประสิทธิภาพ แต่อย่างไรก็ตาม อัลกอริทึมที่งานวิจัยนี้นำเสนอยังไม่สามารถจัดกลุ่มข้อมูลที่มีมิติมากๆ ได้อย่างมีประสิทธิภาพเท่าที่ควร ซึ่งทีมผู้วิจัยของงานวิจัยนี้จะมุ่งพัฒนาในประเด็นนี้ต่อไป

2.2.6 K-Means for Spherical Clusters with Large Variance in Sizes (Fahim et.al. 2009)

งานวิจัยนี้นำเสนอการปรับปรุงอัลกอริทึมการจัดกลุ่มแบบเคมีนส์เพื่อแก้ปัญหาในการจัดกลุ่มที่มีขนาดแตกต่างกัน โดยแนวความคิดหลักที่ใช้คือการเลื่อนจุดศูนย์กลางของกลุ่มใหญ่เข้าไปหากลุ่มเล็ก โดยรายละเอียดของอัลกอริทึมที่งานวิจัยนี้นำเสนอมีดังนี้

1) จัดกลุ่มโดยใช้เคมีนส์ตามปกติ

2) คำนวณรัศมีเฉลี่ยของกลุ่มทุกกลุ่มจาก

$$\text{radius}\,(c_i) = \frac{\sum_{p \in c_i} d^2(p, m_i)}{n_i} \qquad (2.20)$$

เมื่อ c_i คือกลุ่ม i

m_i คือจุดศูนย์กลางของกลุ่ม i

n_i คือจำนวนข้อมูลทั้งหมดในกลุ่ม i

3) เลือกกลุ่มที่มีค่ารัศมีเฉลี่ยมากที่สุดมาเปรียบเทียบกับกลุ่มอื่นๆ ที่เหลือ โดยเปรียบเทียบทีละคู่ เริ่มจากการคำนวณหาผลรวมของรัศมีทั้ง 2 กลุ่ม และหาระยะห่างระหว่าง 2 กลุ่ม

โดยผลรวมของรัศมีคำนวณได้จาก

$$\text{sumofradius} = (\text{radius(L)} + \text{radius(S)})*0.80 \quad (2.21)$$

เมื่อ L คือกลุ่มใหญ่

S คือกลุ่มเล็ก

และระยะห่างระหว่างกลุ่มคำนวณได้จาก

$$\text{meandistance} = \sqrt{\sum_{i=1}^{d} (m_{Li} - m_{Si})^2} \quad (2.22)$$

เมื่อ m_L คือค่าเฉลี่ยของกลุ่มใหญ่

m_S คือค่าเฉลี่ยของกลุ่มเล็ก

d คือมิติของข้อมูล

4) หากผลรวมของรัศมีมีค่ามากกว่าระยะห่างระหว่างกลุ่ม และอัตราส่วนของรัศมีกลุ่มเล็กต่อรัศมีกลุ่มใหญ่มีค่าน้อยกว่า 0.9 จะต้องมีการนำข้อมูลบางส่วนภายในกลุ่มเล็กไปจัดรวมเข้ากับกลุ่มใหญ่ ในการพิจารณาว่าข้อมูลตัวใดในกลุ่มเล็กจะถูกจัดเข้าไปอยู่ในกลุ่มใหญ่จะต้องเริ่มจากการหาค่าเฉลี่ยของจุดศูนย์กลางของกลุ่มทั้ง 2 กลุ่ม

$$\text{Avmean} = \frac{M_L + M_S}{2} \quad (2.23)$$

ค่าเฉลี่ยของจุดศูนย์กลางของกลุ่มทั้ง 2 กลุ่มที่คำนวณได้จะอยู่ภายในกลุ่มใหญ่และอยู่กึ่งกลางของระยะทางระหว่างจุดศูนย์กลางของกลุ่มทั้ง 2 กลุ่ม ข้อมูลในกลุ่มเล็กที่จะถูกจัดเข้าไปในกลุ่มใหญ่จะต้องอยู่ใกล้กับ *Avmean* มากกว่า m_S

หรือ

$$\text{Dis}(p_i,\text{Avmean}) \leq \text{Dis}(p_i,m_S) + \frac{\text{radius(L)}}{\text{radius(S)}*0.8} \quad (2.24)$$

หากสมการ 2.24 เป็นจริง ข้อมูล p_i จะถูกจัดเข้าไปรวมอยู่ในกลุ่มใหญ่ แต่หากสมการที่ 2.24 เป็นเท็จ ข้อมูล p_i จะยังคงอยู่ในกลุ่มเล็กเช่นเดิม

5) ทำซ้ำตั้งแต่ขั้นตอนที่ 3-4 จนครบทุกกลุ่ม

งานวิจัยนี้ได้ทดสอบประสิทธิภาพในการจัดกลุ่มโดยสร้างข้อมูล 2 มิติขึ้นมา 5 ชุด โดยข้อมูลทุกชุดสามารถแบ่งออกได้เป็น 3 กลุ่ม ซึ่งกลุ่มแต่ละกลุ่มมีรูปร่างลักษณะเป็นวงกลมและมีขนาด

ของกลุ่มแตกต่างกัน ทดสอบโดยเปรียบเทียบประสิทธิภาพในการจัดกลุ่มที่ได้จากอัลกอริทึมของงานวิจัยนี้กับการจัดกลุ่มแบบเคมีนส์ดั้งเดิม

ในการทดสอบประสิทธิภาพในการจัดกลุ่มของชุดข้อมูลที่ 1 และ 2 จะได้ผลลัพธ์การจัดกลุ่มจึงถูกต้อง 100 เปอร์เซ็นต์ เนื่องจากข้อมูลแต่ละกลุ่มแยกออกจากกันอย่างชัดเจน

ในการทดสอบประสิทธิภาพในการจัดกลุ่มของชุดข้อมูลที่ 3 จะมีความผิดพลาดเกิดขึ้นเล็กน้อยเนื่องจากข้อมูลทั้ง 3 กลุ่มแยกออกจากกันไม่ชัดเจน

ในการทดสอบประสิทธิภาพในการจัดกลุ่มของชุดข้อมูลที่ 4 จะมีความผิดพลาดเกิดขึ้นมากขึ้น เนื่องจากจุดศูนย์กลางของกลุ่มเล็ก 2 กลุ่มตกอยู่ที่ขอบของกลุ่มใหญ่

ในการทดสอบประสิทธิภาพในการจัดกลุ่มของชุดข้อมูลที่ 5 ผลลัพธ์ที่ได้จากการจัดกลุ่มโดยใช้วิธีการคำนวณหาจุดศูนย์กลางเริ่มต้นที่งานวิจัยนี้นำเสนอจะเหมือนกับผลลัพธ์ที่ได้จากการจัดกลุ่มแบบเคมีนส์ดั้งเดิม เนื่องจากความหนาแน่นของข้อมูลกลุ่มใหญ่แตกต่างจากความหนาแน่นของข้อมูลในกลุ่มเล็ก จุดศูนย์กลางของกลุ่มเล็กจะตกอยู่ในบริเวณของกลุ่มใหญ่ ทำให้ประสิทธิภาพในการจัดกลุ่มลดลง

ผลลัพธ์การทดสอบประสิทธิภาพที่ได้จากการจัดกลุ่มแบบเคมีนส์ดั้งเดิมและจากอัลกอริทึมของงานวิจัยนี้แสดงได้ดังตารางที่ 2.8

ตารางที่ 2.8 การเปรียบเทียบผลลัพธ์การจัดกลุ่มที่ได้จากอัลกอริทึมทั้ง 2 วิธี

Data sets	Exact clusters	K-means cluster	K-means error	Improved K-means clusters (Fahim et.al. 2009)	Improved K-means error (Fahim et.al. 2009)
Set 1	1815	1210	605 points	1815	0 points
	683	973		683	
	660	975		660	
Set 2	1582	1025	557 points	1582	0 points
	703	1015		703	
	642	887		642	
Set 3	1582	1043	539 points	1585	7 points
	557	816		552	
	522	802		524	

ตารางที่ 2.8 (ต่อ) การเปรียบเทียบผลลัพธ์การจัดกลุ่มที่ได้จากอัลกอริทึมทั้ง 2 วิธี

Data sets	Exact clusters	K-means cluster	K-means error	Improved K-means clusters (Fahim et.al. 2009)	Improved K-means error (Fahim et.al. 2009)
Set 4	2129	1306	823 points	2067	62 points
	505	916		534	
	510	922		543	
Set 5	2363	1417	946 points	1417	946 points

จากการศึกษางานวิจัยนี้ สามารถวิเคราะห์จุดเด่นของอัลกอริทึมที่งานวิจัยนี้นำเสนอได้ว่า อัลกอริทึมที่งานวิจัยนี้นำเสนอสามารถจัดกลุ่มที่มีขนาดแตกต่างกันได้อย่างมีประสิทธิภาพ แต่อย่างไรก็ตาม อัลกอริทึมของงานวิจัยนี้ก็ยังมีข้อจำกัดบางประการคือ กลุ่มของข้อมูลจะต้องมีรูปร่างลักษณะเป็นทรงกลมเท่านั้น และหากกลุ่มแต่ละกลุ่มแยกออกจากกันไม่ชัดเจนหรือหากข้อมูลในแต่ละกลุ่มมีความหนาแน่นไม่เท่ากัน จุดศูนย์กลางจะตกอยู่ในกลุ่มที่มีขนาดใหญ่ ทำให้ประสิทธิภาพในการจัดกลุ่มที่ได้จากอัลกอริทึมที่งานวิจัยนี้นำเสนอจะลดลง

บทที่ 3
วิธีดำเนินการวิจัย

3.1 แนวความคิดที่ใช้ในการคำนวณหาจุดศูนย์กลางเริ่มต้น

อัลกอริทึมที่ใช้ในการจัดกลุ่มมีอยู่ด้วยกันหลายวิธี แต่วิธีที่นิยมใช้คือการจัดกลุ่มแบบเคมีนส์ เนื่องจากเป็นวิธีที่เข้าใจง่าย ไม่ซับซ้อน แต่อย่างไรก็ตาม การจัดกลุ่มแบบเคมีนส์ก็ยังมีข้อจำกัดอยู่หลายประการ หนึ่งในนั้นคือประสิทธิภาพที่ได้จากการจัดกลุ่มแบบเคมีนส์จะขึ้นอยู่กับจุดศูนย์กลางเริ่มต้นซึ่งโดยปกติแล้วจะได้มาจากการสุ่ม ดังนั้นงานวิจัยนี้จึงมุ่งเน้นที่จะแก้ไขข้อจำกัดนี้โดยการคิดค้นวิธีการกำหนดจุดศูนย์กลางเริ่มต้นแทนวิธีการสุ่มข้อมูล

อัลกอริทึมในการคำนวณหาจุดศูนย์กลางเริ่มต้นที่งานวิจัยนี้นำเสนอประกอบด้วย 3 ขั้นตอนหลักๆ ในขั้นตอนแรกของอัลกอริทึมที่งานวิจัยนี้นำเสนอจะเป็นการคำนวณหาจุดศูนย์กลางเริ่มต้นจุดแรก โดยแนวความคิดที่ใช้ในการคำนวณหาจุดศูนย์กลางเริ่มต้นจุดแรกคือ บริเวณใดที่มีความหนาแน่นของข้อมูลมาก บริเวณนั้นก็น่าจะเป็นบริเวณที่ข้อมูลกระจุกรวมตัวกันเป็นกลุ่มหรือคลัสเตอร์ ดังนั้นข้อมูลบริเวณนี้จึงถูกดึงมาใช้เป็นจุดศูนย์กลางเริ่มต้นจุดแรก

ขั้นตอนต่อมาคือการหาจุดศูนย์กลางเริ่มต้นที่เหลือทั้งหมด โดยกำหนดให้จำนวนจุดศูนย์กลางเริ่มต้นทั้งหมดเป็นสองเท่าของจำนวนกลุ่ม โดยแนวความคิดที่ใช้ในการคำนวณหาจุดศูนย์กลางเริ่มต้นที่เหลือคือ ตามปกติแล้วข้อมูลที่อยู่ในกลุ่มเดียวกันจะมีระยะทางใกล้กันและข้อมูลที่อยู่ต่างกลุ่มกันจะมีระยะทางห่างกัน ดังนั้นอัลกอริทึมที่งานวิจัยนี้นำเสนอจึงพยายามจะคำนวณให้ได้จุดศูนย์กลางเริ่มต้นทั้งหมดอยู่ห่างกันมากที่สุด และเหตุผลที่กำหนดให้จำนวนจุดศูนย์กลางเริ่มต้นทั้งหมดเป็นสองเท่าของจำนวนกลุ่มคือ หากกลุ่มแต่ละกลุ่มมีขนาดที่แตกต่างกันมาก จุดศูนย์กลางเริ่มต้นที่คำนวณได้อาจจะตกอยู่ในกลุ่มเดียวกัน ดังนั้นจึงมีการเผื่อจำนวนจุดศูนย์กลางเริ่มต้นให้มากกว่าจำนวนกลุ่ม เพื่อที่จะสามารถรวมจุดศูนย์กลางเริ่มต้นที่ตกอยู่ในกลุ่มเดียวกันให้เป็นจุดศูนย์กลางเริ่มต้นจุดเดียว

ในขั้นตอนสุดท้ายคือการรวมจุดศูนย์กลางเริ่มต้นที่คาดว่าน่าจะตกอยู่ในกลุ่มเดียวกันให้เป็นจุดศูนย์กลางเริ่มต้นจุดเดียวจนกระทั่งจำนวนจุดศูนย์กลางเริ่มต้นเท่ากับจำนวนกลุ่ม (K) โดยมีแนวความคิดคือ หากข้อมูลที่อยู่ระหว่างจุดศูนย์กลางเริ่มต้นคู่ใดๆ มีการกระจายแบบสม่ำเสมอสันนิษฐานว่าจุดศูนย์กลางเริ่มต้นคู่นั้นอาจจะตกอยู่ในกลุ่มเดียวกัน ก็ให้รวมจุดศูนย์กลางคู่นั้นเป็นจุดเดียว

งานวิจัยนี้จะใช้ขั้นตอนและแนวความคิดดังกล่าวในการหาจุดศูนย์กลางเริ่มต้นเพื่อแก้ไขข้อจำกัดของการจัดกลุ่มแบบเคมีนส์ดั้งเดิมเพื่อให้ผลลัพธ์ของการจัดกลุ่มมีความเสถียรภาพและมีประสิทธิภาพมากขึ้น

3.2 ขั้นตอนในการคำนวณหาจุดศูนย์กลางเริ่มต้น

ขั้นตอนในการคำนวณหาจุดศูนย์กลางเริ่มต้นประกอบด้วย 3 ขั้นตอนหลักๆ ได้แก่

1) ขั้นตอนในการคำนวณหาจุดศูนย์กลางเริ่มต้นจุดที่ 1

2) ขั้นตอนในการคำนวณหาจุดศูนย์กลางเริ่มต้นจุดที่เหลือทั้งหมด

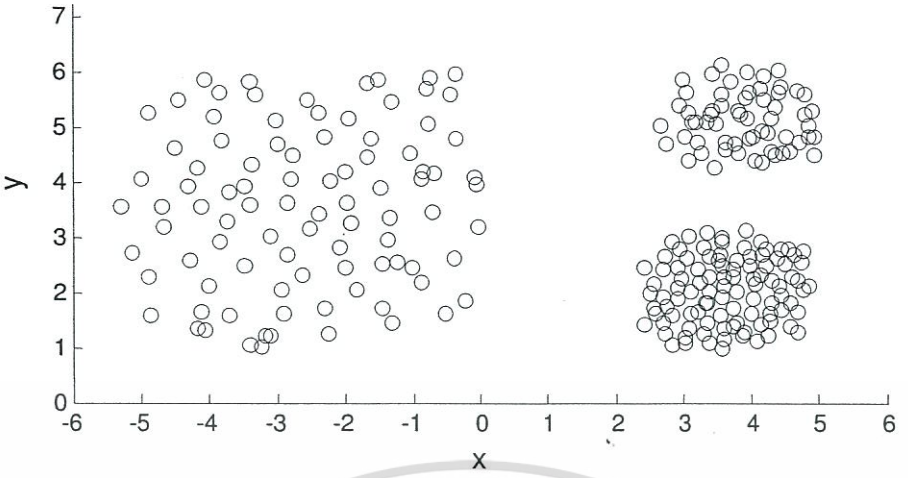
3) ขั้นตอนในการรวมจุดศูนย์กลางเริ่มต้นให้เหลือจำนวนจุดศูนย์กลางเริ่มต้นเท่ากับ K

โดยจะใช้ข้อมูลในตารางที่ 3.1 เป็นตัวอย่างประกอบการอธิบายขั้นตอนการคำนวณหาจุดศูนย์กลางเริ่มต้น ซึ่งข้อมูลที่นำมาใช้เป็นข้อมูล 2 มิติ มีจำนวนทั้งหมด 252 แถว

ตารางที่ 3.1 ตัวอย่างข้อมูลที่ใช้เป็นตัวอย่างประกอบการอธิบายขั้นตอนการหาจุดศูนย์กลางเริ่มต้น

ลำดับที่	มิติที่ 1	มิติที่ 2
1	-1.6311	4.7718
2	-1.3217	1.4521
3	-4.1700	4.2642
4	-1.8268	2.0355
5	-4.3007	3.9060
6	-3.4075	5.8048
7	-2.9820	4.6702
8	-0.3548	5.9552
9	-0.7027	3.4541
10	-4.8800	5.2553
...	-...	...
252	-2.3985	3.4185

จากข้อมูลตัวอย่างในตารางที่ 3.1 สามารถพล็อตเป็นกราฟได้ดังแสดงในรูปที่ 3.1



รูปที่ 3.1 ข้อมูลตัวอย่างที่ใช้ประกอบการอธิบายขั้นตอนการหาจุดศูนย์กลางเริ่มต้น

3.2.1 ขั้นตอนในการคำนวณหาจุดศูนย์กลางเริ่มต้นจุดที่ 1

1) คำนวณหาค่าส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation : S.D.) ของข้อมูลแต่ละมิติ โดยใช้สมการที่ 3.1

$$SD_m = \sqrt{\frac{\sum_i^n (x_{im} - \bar{x}_m)^2}{N}} \qquad (3.1)$$

เมื่อ m คือลำดับที่ของมิติของข้อมูล

x_{im} คือข้อมูลในแถวที่ i มิติที่ m

$\bar{x}$ คือค่าเฉลี่ยของข้อมูลในมิติที่ m

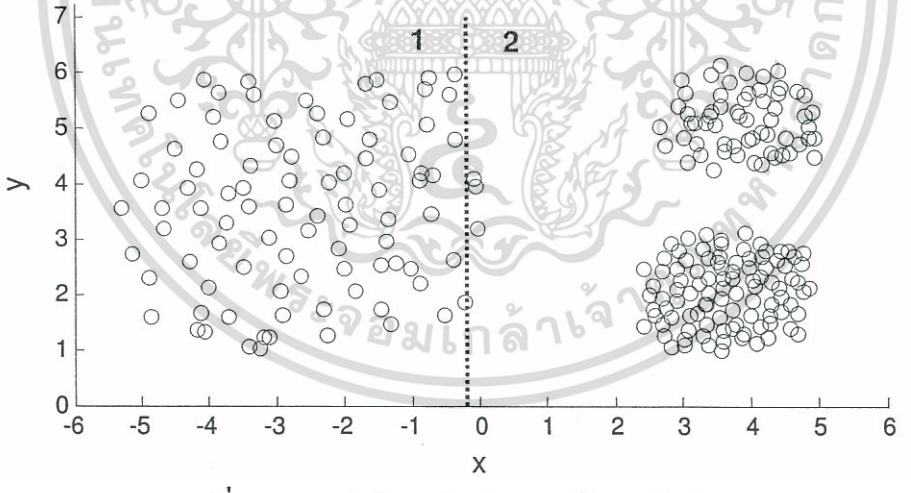
N คือจำนวนข้อมูลทั้งหมด

จากข้อมูลตัวอย่างจะสามารถคำนวณหาค่าส่วนเบี่ยงเบนมาตรฐานของข้อมูลทั้ง 2 มิติได้ดังแสดงในตารางที่ 3.2

ตารางที่ 3.2 ค่าส่วนเบี่ยงเบนมาตรฐานที่คำนวณได้ของข้อมูลตัวอย่าง

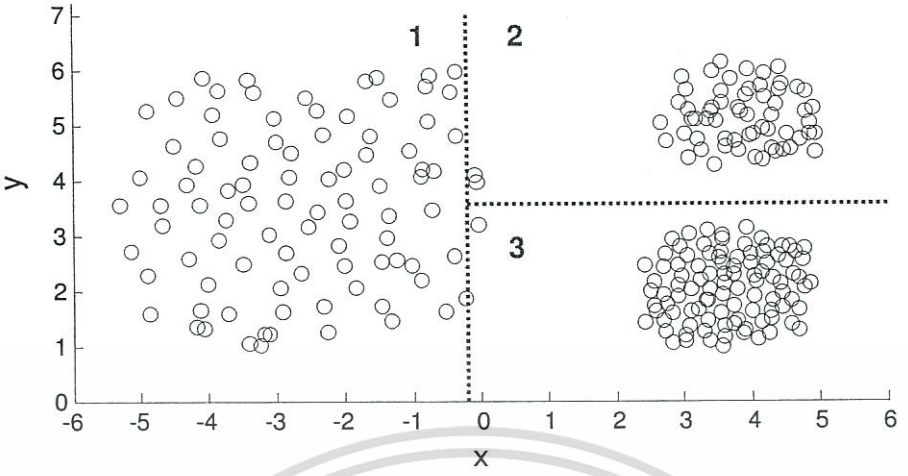
ลำดับที่	มิติที่ 1	มิติที่ 2
1	-1.6311	4.7718
2	-1.3217	1.4521
3	-4.1700	4.2642
4	-1.8268	2.0355
5	-4.3007	3.9060
...	...	...
252	-2.3985	3.4185
S.D.	3.2024	1.5316

2) แบ่งข้อมูลทั้งหมดออกเป็น 2 กลุ่มย่อยๆ โดยใช้จุดกึ่งกลางของพิสัยของมิติที่มีค่าส่วนเบี่ยงเบนมาตรฐานสูงสุดเป็นตัวแบ่ง ซึ่งจากข้อมูลตัวอย่าง มิติที่มีค่าส่วนเบี่ยงเบนมาตรฐานสูงสุดคือมิติที่ 1 ดังนั้นข้อมูลตัวอย่างจะถูกแบ่งออกเป็น 2 กลุ่มย่อยๆ โดยใช้จุดกึ่งกลางของพิสัยของมิติที่ 1 ดังแสดงในรูปที่ 3.2



รูปที่ 3.2 การแบ่งข้อมูลตัวอย่างออกเป็น 2 กลุ่มย่อยๆ

3) ทำซ้ำขั้นตอนที่ 1-2 บนกลุ่มย่อยที่มีจำนวนข้อมูลสูงสุด หรือในอีกนัยหนึ่งคือมีความหนาแน่นของข้อมูลมากที่สุด ทำซ้ำจนกระทั่งมีจำนวนกลุ่มย่อยทั้งหมดเท่ากับ K โดยกลุ่มย่อยทั้งหมดที่แบ่งได้แสดงได้ดังรูปที่ 3.3



รูปที่ 3.3 กลุ่มย่อยทั้งหมดที่แบ่งได้

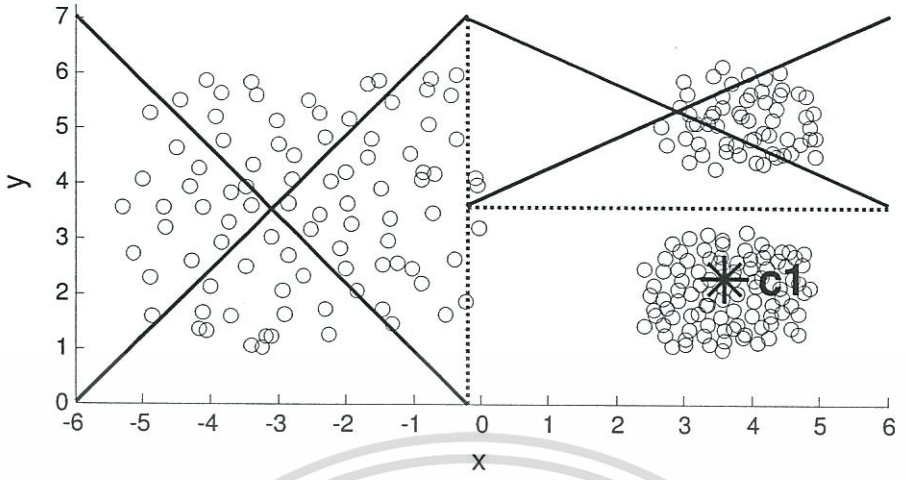
4) เปรียบเทียบจำนวนข้อมูลของกลุ่มย่อย 2 กลุ่มสุดท้าย แล้วเลือกกลุ่มย่อยที่มีจำนวนข้อมูลสูงสุดขึ้นมา โดยจากรูปที่ 3.3 กลุ่มย่อย 2 กลุ่มสุดท้ายคือกลุ่มที่ 2 และกลุ่มที่ 3 ซึ่งเมื่อเปรียบเทียบจำนวนข้อมูลของกลุ่มที่ 2 และกลุ่มที่ 3 แล้ว กลุ่มย่อยที่มีจำนวนข้อมูลสูงสุดคือกลุ่มที่ 3 ดังนั้นกลุ่มที่ 3 จะถูกเลือกขึ้นมาสำหรับกระบวนการในขั้นตอนต่อไป

5) เรียงลำดับข้อมูลภายในกลุ่มที่ได้มาจากข้อ 4 ตามมิติที่มีค่าส่วนเบี่ยงเบนมาตรฐานสูงสุด หลังจากนั้นคำนวณหาค่ามัธยฐานของข้อมูลภายในกลุ่ม

5.1) หากข้อมูลภายในกลุ่มที่ได้มาจากข้อ 4 มีจำนวนเป็นเลขคี่ ค่ามัธยฐานคือค่าที่อยู่ในตำแหน่งที่ (n+1)/2 เมื่อ n คือจำนวนข้อมูลที่อยู่ภายในกลุ่ม

5.2) หากข้อมูลภายในกลุ่มที่ได้มาจากข้อ 4 มีจำนวนเป็นเลขคู่ ให้คำนวณหาระยะทางระหว่างข้อมูลในตำแหน่งที่ (n/2) และข้อมูลในตำแหน่งที่ (n/2)-1 และคำนวณหาระยะทางระหว่างข้อมูลในตำแหน่งที่ (n/2)+1 และข้อมูลในตำแหน่งที่ (n/2)+2 หลังจากนั้นนำระยะทางที่คำนวณได้มาเปรียบเทียบกัน หากระยะทางระหว่างข้อมูลในตำแหน่งที่ (n/2) และข้อมูลในตำแหน่งที่ (n/2)-1 มีค่าน้อยกว่าระยะทางระหว่างข้อมูลในตำแหน่งที่ (n/2)+1 และข้อมูลในตำแหน่งที่ (n/2)+2 ค่ามัธยฐานคือข้อมูลในตำแหน่งที่ (n/2) แต่หากระยะทางระหว่างข้อมูลในตำแหน่งที่ (n/2) และข้อมูลในตำแหน่งที่ (n/2)-1 มีค่ามากกว่าระยะทางระหว่างข้อมูลในตำแหน่งที่ (n/2)+1 และข้อมูลในตำแหน่งที่ (n/2)+2 ค่ามัธยฐานคือข้อมูลในตำแหน่งที่ (n/2)+1

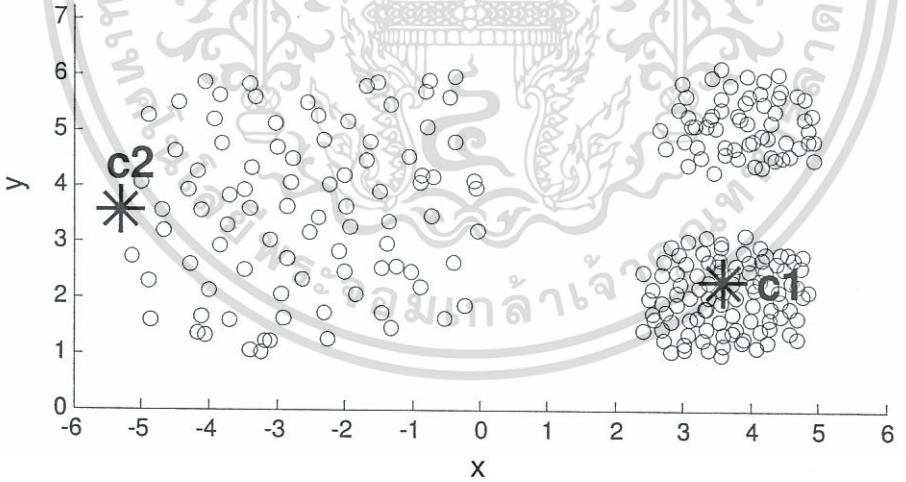
6) นำค่ามัธยฐานที่คำนวณได้จากข้อ 5 มาใช้เป็นจุดศูนย์กลางเริ่มต้นจุดแรก ซึ่งจากข้อมูลตัวอย่าง จะสามารถคำนวณหาจุดศูนย์กลางเริ่มต้นจุดแรก (*c1*) ได้ดังแสดงในรูปที่ 3.4



รูปที่ 3.4 จุดศูนย์กลางเริ่มต้นจุดที่ 1 ที่คำนวณได้

3.2.2 ขั้นตอนในการคำนวณหาจุดศูนย์กลางเริ่มต้นจุดที่เหลือ

1) คำนวณหาระยะทางแบบยูคลิดระหว่างข้อมูลที่เหลือแต่ละตัวกับจุดศูนย์กลางเริ่มต้นจุดแรกที่คำนวณได้ ข้อมูลตัวที่มีระยะทางไปยังจุดศูนย์กลางเริ่มต้นจุดแรกสูงที่สุดจะถูกนำมาใช้เป็นจุดศูนย์กลางเริ่มต้นจุดที่ 2 (*c2*) จากข้อมูลตัวอย่าง จุดศูนย์กลางเริ่มต้นจุดที่ 2 สามารถแสดงได้ดังรูปที่ 3.5



รูปที่ 3.5 จุดศูนย์กลางเริ่มต้นจุดที่ 2 ที่คำนวณได้

2) คำนวณหาระยะทางแบบยูคลิดระหว่างข้อมูลที่เหลือแต่ละตัวไปยังจุดศูนย์กลางเริ่มต้นที่คำนวณได้ก่อนหน้าแต่ละจุด แล้วเลือกระยะทางที่น้อยที่สุดของข้อมูลแต่ละตัวขึ้นมา

$$d_i^{min} = \min_{k=1}^{A}(dc_{ki}) \qquad (3.2)$$

เมื่อ dc_{ki} คือระยะทางจากข้อมูลแต่ละตัวไปยังจุดศูนย์กลางเริ่มต้นที่คำนวณได้ก่อนหน้า

A คือจำนวนของจุดศูนย์กลางเริ่มต้นที่คำนวณได้ก่อนหน้า

i คือลำดับที่ของข้อมูลที่เหลือแต่ละตัว

จากข้อมูลตัวอย่าง สามารถคำนวณหาค่าระยะทางจากข้อมูลแต่ละตัวไปยังจุดศูนย์กลางเริ่มต้นจุดที่ 1 และจุดศูนย์กลางเริ่มต้นจุดที่ 2 ได้ดังตารางที่ 3.3 และตัวเลขที่ขีดเส้นใต้ไว้คือระยะทางที่น้อยที่สุดจากข้อมูลตัวนั้นๆ ไปยังจุดศูนย์กลางเริ่มต้นแต่ละจุด

ตารางที่ 3.3 ระยะทางที่น้อยที่สุดระหว่างข้อมูลแต่ละตัวไปยังจุดศูนย์กลางเริ่มต้นก่อนหน้า

ข้อมูลที่เหลือ (i)	ระยะทางไปยังจุดศูนย์กลางเริ่มต้นจุดที่ 1 (dc_{1i})	ระยะทางไปยังจุดศูนย์กลางเริ่มต้นจุดที่ 2 (dc_{2i})
1	6.9559	<u>3.2398</u>
2	<u>4.6568</u>	5.1182
3	6.2517	<u>3.7294</u>
4	4.6772	<u>4.6519</u>
5	<u>4.4646</u>	4.6351
6	6.0755	<u>3.1154</u>
7	7.3875	<u>1.6018</u>
8	<u>5.3938</u>	5.5152
9	5.9077	<u>3.3805</u>
10	8.6491	<u>2.1222</u>
...	...	...
251	6.0968	<u>2.9141</u>

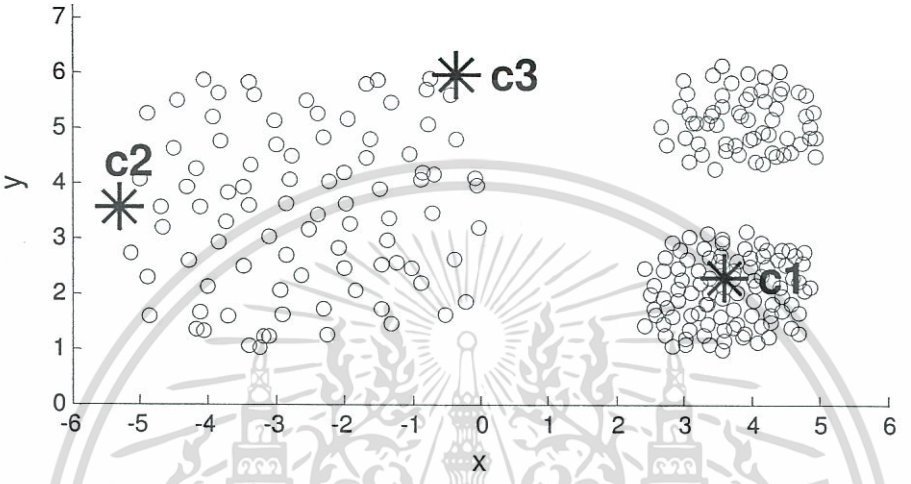
3) เลือกข้อมูลตัวที่มีค่า d_i^{min} สูงที่สุดขึ้นมาใช้เป็นจุดศูนย์กลางเริ่มต้นตัวต่อไป

$$I = \underset{i=1}{\overset{N}{\operatorname{argmax}}}\left(d_i^{min}\right) \tag{3.3}$$

เมื่อ N คือจำนวนข้อมูลที่เหลือทั้งหมด

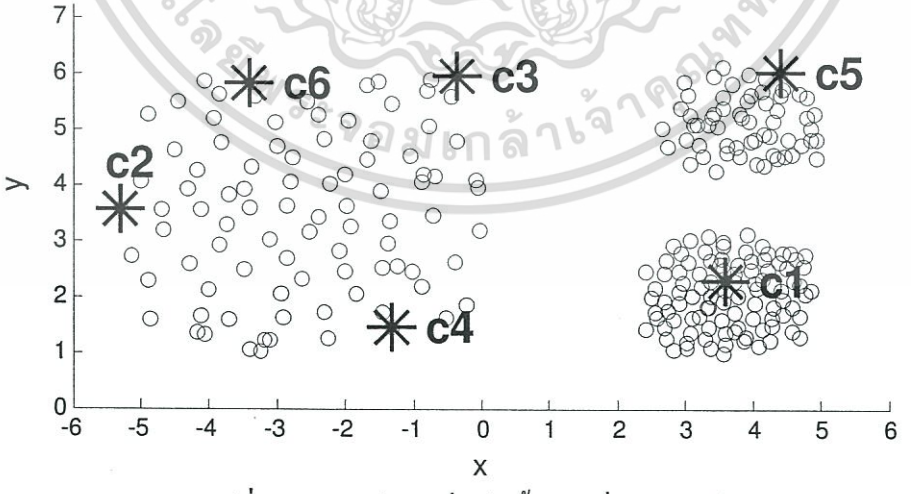
I คือลำดับที่ของข้อมูลตัวที่มีค่า d_i^{min} สูงสุด

จากข้อมูลตัวอย่างในตารางที่ 3.3 ตัวเลขที่ขีดเส้นใต้ในข้อมูลแถวที่ 8 มีค่าสูงที่สุดเมื่อเปรียบเทียบกับตัวเลขที่ขีดเส้นใต้ในข้อมูลแถวอื่น กล่าวคือ ข้อมูลแถวที่ 8 มีค่า d_i^{min} สูงที่สุด ดังนั้น ข้อมูลแถวที่ 8 จะถูกเลือกขึ้นมาเป็นจุดศูนย์กลางเริ่มต้นจุดที่ 3 (c3) โดยจุดศูนย์กลางเริ่มต้นจุดที่ 3 ที่คำนวณได้แสดงได้ดังรูปที่ 3.6



รูปที่ 3.6 จุดศูนย์กลางเริ่มต้นจุดที่ 3 ที่คำนวณได้

4) ทำซ้ำขั้นตอนที่ 2-3 จนกระทั่งได้จำนวนจุดศูนย์กลางเริ่มต้นเท่ากับ 2*K ซึ่งจากข้อมูลตัวอย่างจะต้องได้จำนวนจุดศูนย์กลางเริ่มต้นเท่ากับ 6 จุด โดยจุดศูนย์กลางเริ่มต้นที่คำนวณได้ทั้ง 6 จุดแสดงได้ดังรูปที่ 3.7



รูปที่ 3.7 จุดศูนย์กลางเริ่มต้นทั้งหมดที่คำนวณได้

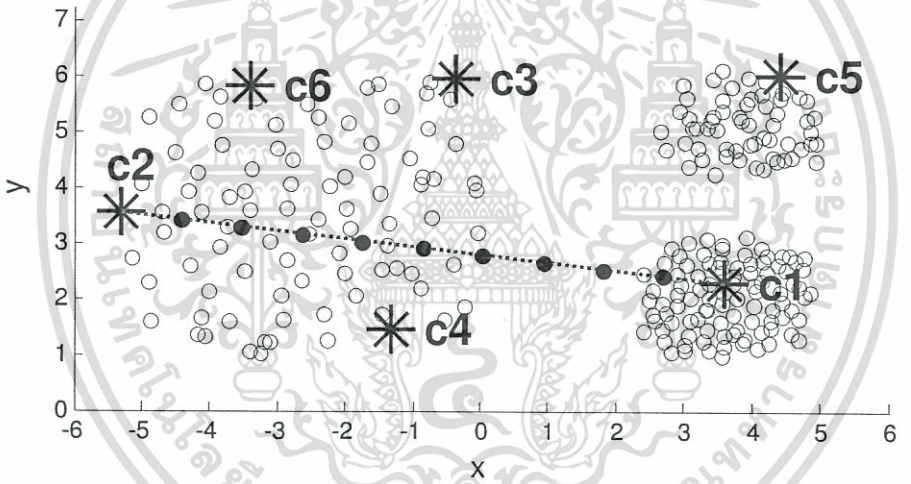
3.2.3 ขั้นตอนในการรวมจุดศูนย์กลางเริ่มต้น

1) แบ่งระยะทางระหว่างจุดศูนย์กลางแต่ละคู่ออกเป็น 10 ช่วงเท่าๆ กัน หรือในอีกนัยหนึ่งคือให้สร้างจุดขึ้นมาทั้งหมด 9 จุด ที่ระยะ 1/10, 2/10, 3/10, ..., 9/10 ของระยะทางระหว่างจุดศูนย์กลางเริ่มต้นแต่ละคู่ กล่าวคือจุดทั้ง 9 จุดของจุดศูนย์กลางเริ่มต้นคู่นั้นๆ จะอยู่ห่างกันเป็นระยะทางเท่าๆ กัน ซึ่งสามารถคำนวณได้จากสมการที่ 3.4

$$r = \frac{d_{ij}}{10} \tag{3.4}$$

เมื่อ d_{ij} คือระยะทางแบบยูคลิดระหว่างจุดศูนย์กลางเริ่มต้น i และจุดศูนย์กลางเริ่มต้น j

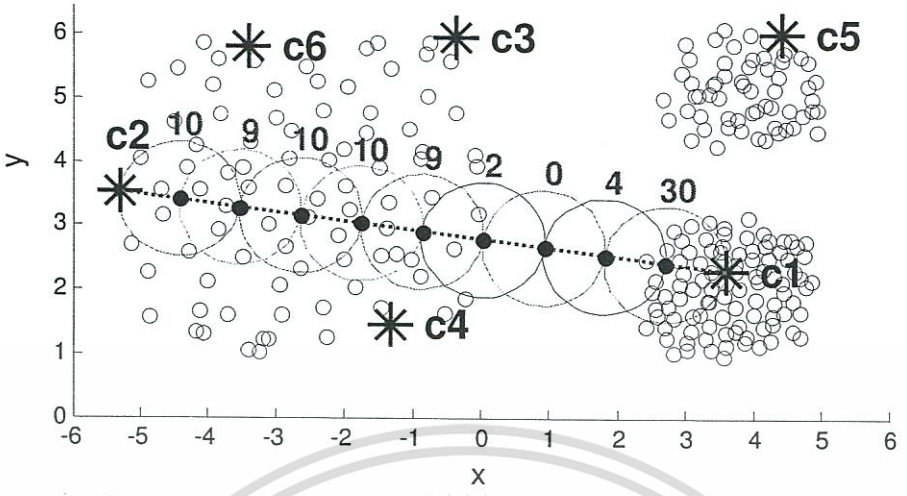
ในรูปที่ 3.8 จะแสดงตัวอย่างของการสร้างจุดขึ้นมาทั้งหมด 9 จุดบนระยะทางระหว่างจุดศูนย์กลางเริ่มต้นจุดที่ 1 และจุดศูนย์กลางเริ่มต้นจุดที่ 2



รูปที่ 3.8 ตัวอย่างการสร้างจุด 9 จุดบนระยะทางระหว่างจุดศูนย์กลางเริ่มต้นแต่ละคู่

2) นับจำนวนข้อมูลที่ตกอยู่ภายในขอบเขตรอบๆ จุดที่สร้างขึ้นมาทั้ง 9 จุด ซึ่งขอบเขตแต่ละขอบเขตจะมีระยะทางเท่ากับ r เมื่อวัดจากจุดที่สร้างขึ้นมาแต่ละจุด โดยกำหนดให้ n_{max} แสดงจำนวนข้อมูลของขอบเขตที่มีจำนวนข้อมูลมากที่สุดของจุดศูนย์กลางเริ่มต้นแต่ละคู่ และ n_{min} แสดงจำนวนข้อมูลของขอบเขตที่มีจำนวนข้อมูลน้อยที่สุดของจุดศูนย์กลางเริ่มต้นแต่ละคู่

ในรูปที่ 3.9 จะแสดงตัวอย่างการนับจำนวนข้อมูลที่ตกอยู่ภายในขอบเขตแต่ละขอบเขตระหว่างจุดศูนย์กลางเริ่มต้นจุดที่ 1 และจุดศูนย์กลางเริ่มต้นจุดที่ 2



รูปที่ 3.9 ตัวอย่างการนับจำนวนข้อมูลที่ตกอยู่ภายในแต่ขอบเขตละขอบเขต

จากข้อมูลตัวอย่างสามารถนับจำนวนข้อมูลที่ตกอยู่ภายในขอบเขตแต่ละขอบเขตระหว่างจุดศูนย์กลางเริ่มต้นแต่ละคู่ได้ดังแสดงในตารางที่ 3.4

ตารางที่ 3.4 การนับจำนวนข้อมูลที่ตกอยู่ภายในแต่ละขอบเขตระหว่างจุดศูนย์กลางเริ่มต้นแต่ละคู่

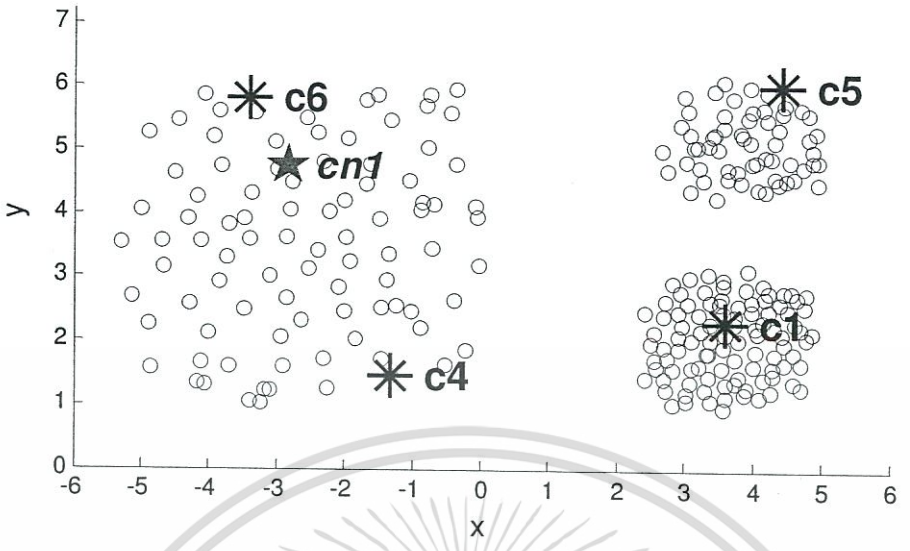
จุดศูนย์กลางเริ่มต้น 1	จุดศูนย์กลางเริ่มต้น 2	n_{max}	n_{min}
c1	c2	30	0
c1	c3	18	0
c1	c4	14	0
c1	c5	12	0
c1	c6	27	0
c2	c3	5	3
c2	c4	3	1
c2	c5	26	0
c2	c6	2	0
c3	c4	4	2
c3	c5	6	0
c3	c6	2	0
c4	c5	23	0
c4	c6	5	2
c5	c6	15	0

3) คำนวณหาค่าอัตราส่วนระหว่าง n_{max} และ n_{min} สำหรับคู่ของจุดศูนย์กลางเริ่มต้นที่มีค่า n_{min} ไม่เท่ากับ 0 ดังแสดงในตารางที่ 3.5

ตารางที่ 3.5 การคำนวณอัตราส่วนระหว่าง n_{max} และ n_{min}

จุดศูนย์กลางเริ่มต้น 1	จุดศูนย์กลางเริ่มต้น 2	n_{max}	n_{min}	n_{max} / n_{min}
c1	c2	30	0	-
c1	c3	18	0	-
c1	c4	14	0	-
c1	c5	12	0	-
c1	c6	27	0	-
c2	c3	5	3	1.6667
c2	c4	3	1	3
c2	c5	26	0	-
c2	c6	2	0	-
c3	c4	4	2	2
c3	c5	6	0	-
c3	c6	2	0	-
c4	c5	23	0	-
c4	c6	5	2	2.5
c5	c6	15	0	-

4) รวมคู่ของจุดศูนย์กลางเริ่มต้นที่มีอัตราส่วนระหว่าง n_{max} และ n_{min} น้อยที่สุดให้เป็นจุดศูนย์กลางเริ่มต้นจุดเดียว จากการคำนวณหาอัตราส่วนระหว่าง n_{max} และ n_{min} ของข้อมูลตัวอย่างในตารางที่ 3.5 จุดศูนย์กลางเริ่มต้นคู่ที่มีอัตราส่วนระหว่าง n_{max} และ n_{min} น้อยที่สุดคือคู่ของจุดศูนย์กลางเริ่มต้นจุดที่ 2 และจุดศูนย์กลางเริ่มต้นจุดที่ 3 ดังนั้น จุดศูนย์กลางเริ่มต้นจุดที่ 2 และจุดศูนย์กลางเริ่มต้นจุดที่ 3 จะถูกรวมให้เป็นจุดศูนย์กลางเริ่มต้นจุดใหม่จุดเดียว (c_{n1}) โดยใช้ค่าเฉลี่ยทางเลขคณิต (Arithmetic Mean) โดยจุดศูนย์กลางเริ่มต้นจุดที่ 2 และจุดศูนย์กลางเริ่มต้นจุดที่ 3 จะถูกลบทิ้งไป ดังแสดงได้ในรูปที่ 3.10



รูปที่ 3.10 ตัวอย่างการรวมจุดศูนย์กลางเริ่มต้นจุดที่ 2 และจุดศูนย์กลางเริ่มต้นจุดที่ 3

อย่างไรก็ตาม การรวมจุดศูนย์กลางเริ่มต้นยังต้องอาศัยเงื่อนไขเพิ่มเติมนอกเหนือจากการรวมโดยใช้อัตราส่วนระหว่าง n_{max} และ n_{min} ที่น้อยที่สุด อาทิ ในกรณีที่มีคู่ของจุดศูนย์กลางเริ่มต้นที่มีค่าอัตราส่วนระหว่าง n_{max} และ n_{min} ที่น้อยที่สุดเท่ากัน หรือในกรณีที่ค่า n_{min} ของทุกคู่มีค่าเท่ากับ 0 เป็นต้น ซึ่งเงื่อนไขเพิ่มเติมในการรวมจุดศูนย์กลางเริ่มต้นสามารถแบ่งออกได้เป็นหลายกรณี ดังนี้

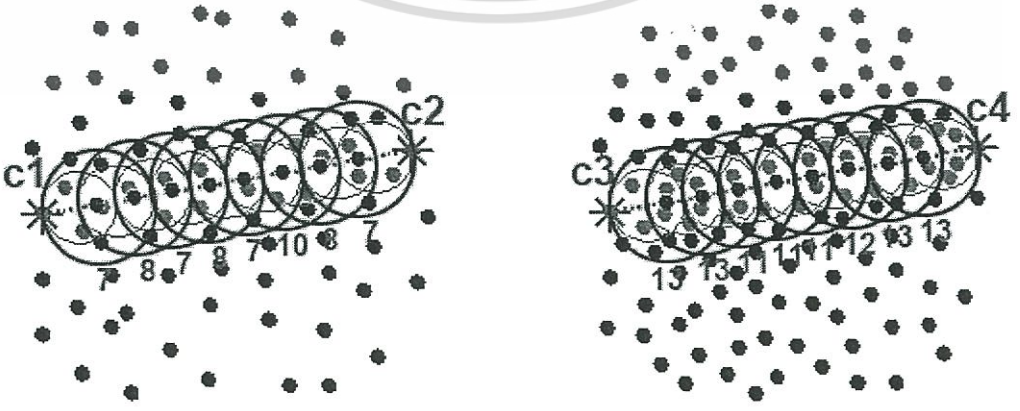
- **กรณีที่ 1** : หากมีคู่ของจุดศูนย์กลางเริ่มต้นที่มีค่าอัตราส่วนระหว่าง n_{max} และ n_{min} ที่น้อยที่สุดเท่ากัน ให้ลดจำนวนจุดที่สร้างขึ้นมา 9 จุดระหว่างคู่นั้นๆ ให้เหลือ 8 จุด โดยวิธีการลดจำนวนจุดที่สร้างขึ้นมาทำได้โดยการรวมจุดที่ 1 และจุดที่ 2 ให้เป็นขอบเขตเดียว รวมจุดที่ 2 และจุดที่ 3 ให้เป็นขอบเขตเดียว รวมจุดที่ 3 และจุดที่ 4 ให้เป็นขอบเขตเดียว รวมเช่นนี้ไปเรื่อยๆ จนครบทุกจุด ก็จะสามารถลดจำนวนจุดที่สร้างขึ้น 9 จุดให้เหลือ 8 จุดได้ ซึ่งขอบเขตแต่ละขอบเขตที่ได้จะมีขนาดใหญ่ขึ้น คือมีระยะทางเท่ากับ $d_{ij}/6.6667$ เมื่อวัดจากจุดที่สร้างขึ้นมาแต่ละจุด หลังจากที่ลดจำนวนจุดเหลือ 8 จุดแล้วให้คำนวณหาค่าอัตราส่วนระหว่าง n_{max} และ n_{min} ใหม่ โดยรูปที่ 3.11 คือตัวอย่างของข้อมูลที่มีคู่ของจุดศูนย์กลางเริ่มต้นที่มีอัตราส่วนระหว่าง n_{max} และ n_{min} น้อยที่สุดเท่ากัน 2 คู่ และตารางที่ 3.6 แสดงผลลัพธ์ของการคำนวณอัตราส่วนระหว่าง n_{max} และ n_{min} ของข้อมูลในรูปที่ 3.11 จะเห็นว่ามี 2 คู่ที่มีอัตราส่วนระหว่าง n_{max} และ n_{min} น้อยที่สุดเท่ากันคือ 1.5 คือคู่ของจุดศูนย์กลางเริ่มต้นจุดที่ 1 และจุดศูนย์กลางเริ่มต้นจุดที่ 2 และคู่ของจุดศูนย์กลางเริ่มต้นจุดที่ 3 และจุดศูนย์กลางเริ่มต้นจุดที่ 4 ดังนั้นคู่ของจุดศูนย์กลางเริ่มต้น 2 คู่นี้จะต้องถูกลดจำนวนจุดจาก 9 จุดให้เหลือ 8 จุด ดังแสดงในรูปที่ 3.12 และผลลัพธ์การคำนวณอัตราส่วนระหว่าง n_{max} และ n_{min} หลังจากลดจำนวนจุดจาก 9 จุดให้เหลือ 8 จุดแสดงได้ดังตารางที่ 3.7 ซึ่งจะเห็นว่าหลังจากที่ได้ลดจำนวนจุดจาก 9 จุดเหลือ 8 จุดแล้ว อัตราส่วนระหว่าง n_{max} และ n_{min} ของจุดศูนย์กลางเริ่มต้น 2 คู่นี้จะมีค่าไม่เท่ากัน เมื่อได้ค่าอัตราส่วนที่แตกต่างกันแล้ว ให้เลือกคู่ที่มีค่าอัตราส่วนน้อยที่สุดมารวม

เป็นจุดศูนย์กลางเริ่มต้นจุดเดียว ซึ่งจากตัวอย่างนี้ คู่ของจุดศูนย์กลางเริ่มต้นจุดที่ 3 และจุดศูนย์กลางเริ่มต้นจุดที่ 4 ได้ค่าอัตราส่วนน้อยกว่า ดังนั้นจุดศูนย์กลางเริ่มต้นจุดที่ 3 และจุดศูนย์กลางเริ่มต้นจุดที่ 4 จะถูกรวมให้เป็นจุดศูนย์กลางเริ่มต้นจุดเดียว

รูปที่ 3.11 ตัวอย่างข้อมูลที่มีคู่ของจุดศูนย์กลางที่มีอัตราส่วนระหว่าง n_{max} และ n_{min} น้อยที่สุดเท่ากัน

ตารางที่ 3.6 ผลลัพธ์การคำนวณอัตราส่วนระหว่าง n_{max} และ n_{min} ของข้อมูลในรูปที่ 3.11

จุดศูนย์กลางเริ่มต้น 1	จุดศูนย์กลางเริ่มต้น 2	n_{max}	n_{min}	n_{max} / n_{min}
c1	c2	3	2	1.5
c1	c3	4	0	-
c1	c4	8	0	-
c2	c3	0	0	-
c2	c4	7	0	-
c3	c4	6	4	1.5

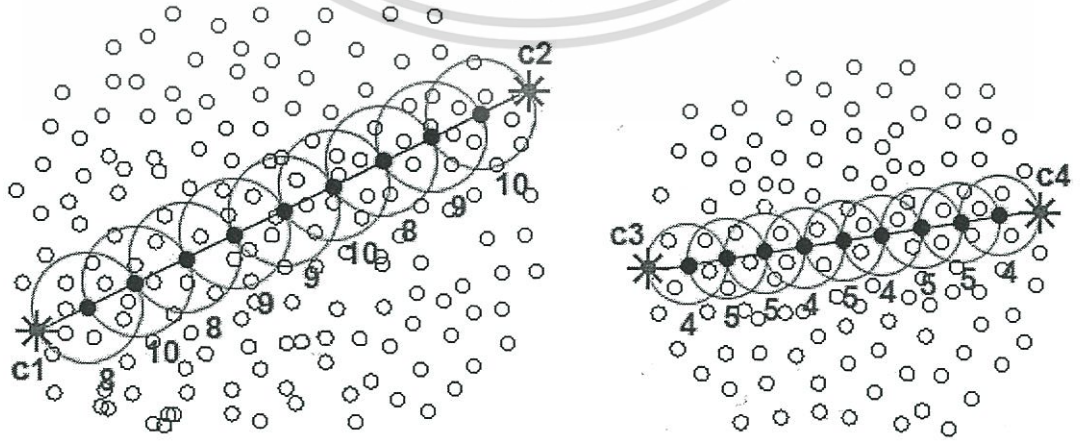


รูปที่ 3.12 ตัวอย่างการลดจำนวนจุดที่สร้างขึ้นจาก 9 จุดให้เหลือ 8 จุดของข้อมูลในรูปที่ 3.11

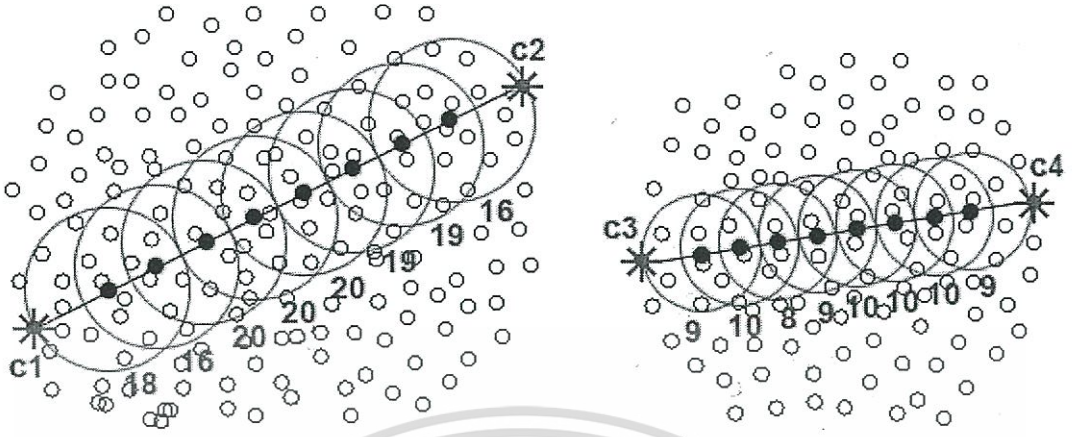
ตารางที่ 3.7 ผลลัพธ์การคำนวณอัตราส่วนระหว่าง n_{max} และ n_{min} ของข้อมูลในรูปที่ 3.11 หลังจากลดจำนวนจุดให้เหลือ 8 จุดแล้ว

จุดศูนย์กลางเริ่มต้น 1	จุดศูนย์กลางเริ่มต้น 2	n_{max}	n_{min}	n_{max} / n_{min}
c1	c2	10	7	1.4286
c1	c3	4	0	-
c1	c4	8	0	-
c2	c3	0	0	-
c2	c4	7	0	-
c3	c4	13	11	1.1818

อย่างไรก็ตาม หากทำการลดจำนวนจุดที่สร้างขึ้นจาก 9 จุดให้เหลือ 8 จุดแล้ว ค่าอัตราส่วนระหว่าง n_{max} และ n_{min} ยังคงมีค่าเท่ากันอยู่ ให้ลดจำนวนจุดที่สร้างขึ้นจาก 8 จุดให้เหลือ 7 จุด จะได้ระยะทางของขอบเขตแต่ละขอบเขตเท่ากับ $d_{ij}/5$ หากค่าอัตราส่วนระหว่าง n_{max} และ n_{min} ก็ยังคงมีค่าเท่ากันอยู่ ให้ลดจำนวนจุดที่สร้างขึ้นจาก 7 จุดให้เหลือ 6 จุด จะได้ระยะทางของขอบเขตแต่ละขอบเขตเท่ากับ $d_{ij}/4$ และสุดท้ายหากค่าอัตราส่วนระหว่าง n_{max} และ n_{min} ยังคงมีค่าเท่ากัน ให้เลือกจุดศูนย์กลางเริ่มต้นคู่ที่มีระยะทางใกล้กันมากที่สุดมารวมกันก่อน ซึ่งจะเห็นได้จากตัวอย่างในรูปที่ 3.13 เมื่อลดจำนวนจุดที่สร้างขึ้นมาจนกระทั่งเหลือ 6 จุด ค่าอัตราส่วนระหว่าง n_{max} และ n_{min} ก็ยังคงมีค่าเท่ากันอยู่ ดังนั้นจึงจะต้องเลือกคู่ของจุดศูนย์กลางที่อยู่ใกล้กันมากที่สุดมารวมให้เป็นจุดศูนย์กลางเดียวกัน ซึ่งในที่นี้ระยะทางระหว่างจุดศูนย์กลางเริ่มต้นจุดที่ 3 กับจุดศูนย์กลางเริ่มต้นจุดที่ 4 มีค่าน้อยกว่า ระยะทางระหว่างจุดศูนย์กลางเริ่มต้นจุดที่ 1 กับจุดศูนย์กลางเริ่มต้นจุดที่ 2 ดังนั้นจึงทำการรวมจุดศูนย์กลางเริ่มต้นจุดที่ 3 กับจุดศูนย์กลางเริ่มต้นจุดที่ 4 ให้เป็นจุดศูนย์กลางเดียว

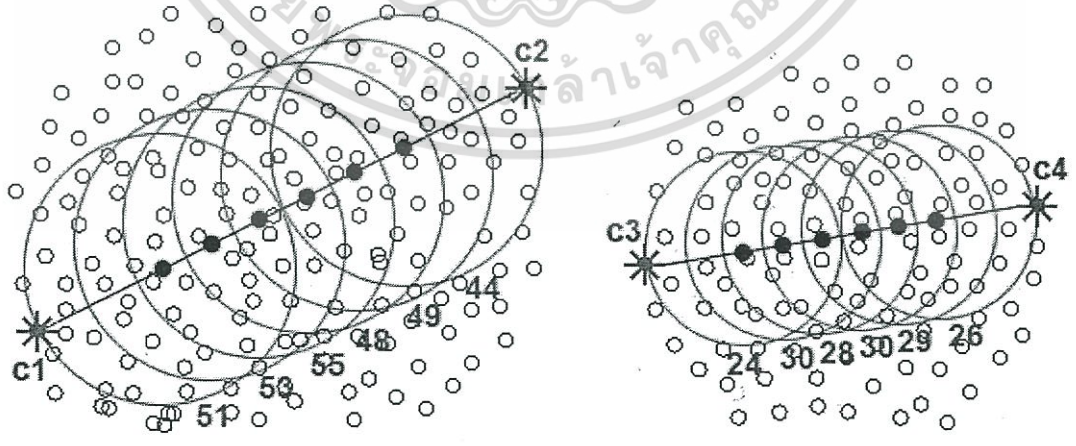


(ก) ตัวอย่างข้อมูลที่มีคู่ของจุดศูนย์กลางที่มีอัตราส่วนระหว่าง n_{max} และ n_{min} น้อยที่สุดเท่ากัน

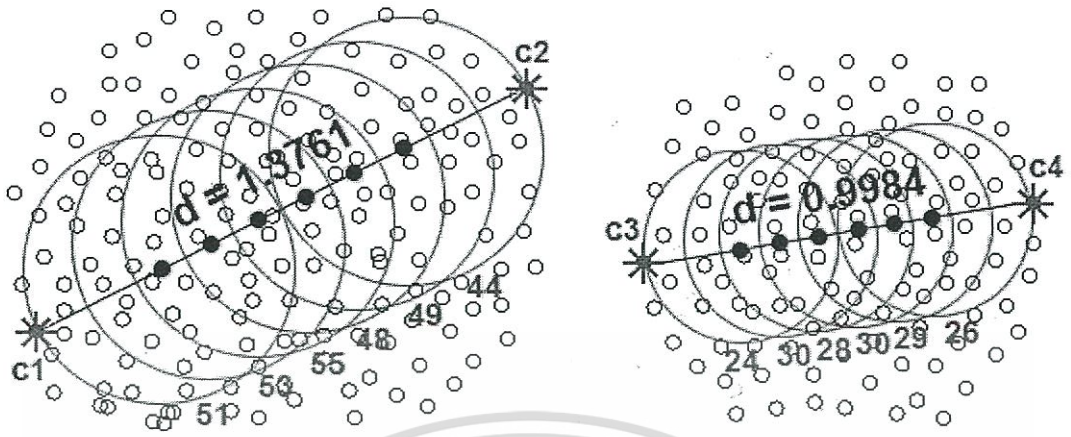


(ข) อัตราส่วนระหว่าง n_{max} และ n_{min} น้อยที่สุดเท่ากันหลังจากลดจำนวนจุดที่สร้างให้เหลือ 8 จุด

(ค) อัตราส่วนระหว่าง n_{max} และ n_{min} น้อยที่สุดเท่ากันหลังจากลดจำนวนจุดที่สร้างให้เหลือ 7 จุด



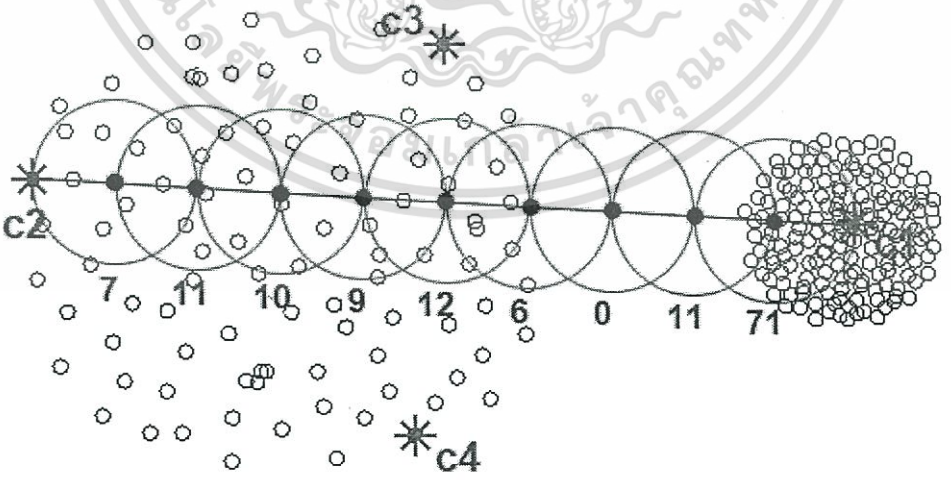
(ง) อัตราส่วนระหว่าง n_{max} และ n_{min} น้อยที่สุดเท่ากันหลังจากลดจำนวนจุดที่สร้างให้เหลือ 6 จุด



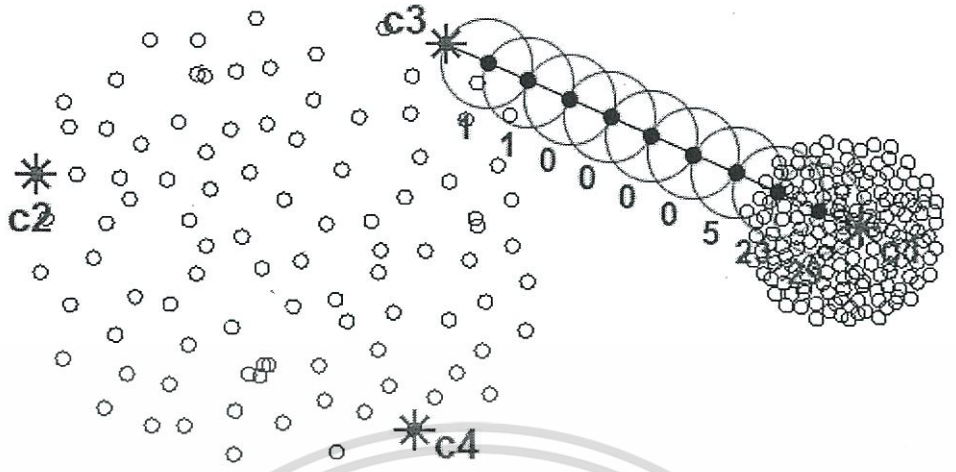
(ง) การคำนวณระยะทางระหว่างจุดศูนย์กลางคู่ที่มีอัตราส่วนระหว่าง n_{max} และ n_{min} น้อยที่สุดเท่ากัน

รูปที่ 3.13 ตัวอย่างการเลือกรวมคู่ของจุดศูนย์กลางเมื่อลดจำนวนจุดที่สร้างขึ้นจาก 9 จุดให้เหลือ 6 จุดแล้วได้อัตราส่วนระหว่าง n_{max} และ n_{min} น้อยที่สุดเท่ากัน

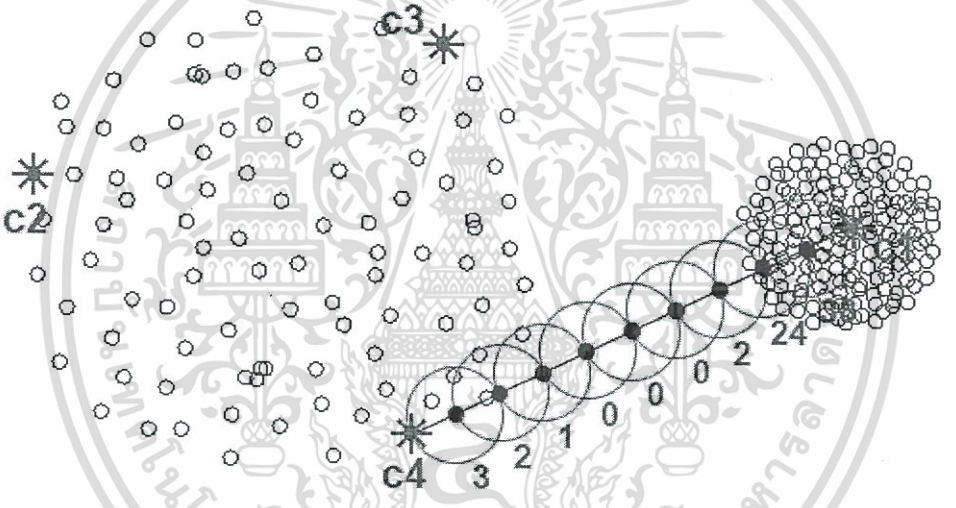
- **กรณีที่ 2** : หากคู่ของจุดศูนย์กลางเริ่มต้นทุกคู่มีค่า n_{min} เท่ากับ 0 จะไม่สามารถคำนวณหาอัตราส่วนระหว่าง n_{max} และ n_{min} ได้ ให้เลือกรวมจุดศูนย์กลางเริ่มต้นคู่ที่มีค่า n_{max} ต่ำที่สุดแทน โดยรูปที่ 3.14 คือตัวอย่างของข้อมูลที่คู่ของจุดศูนย์กลางเริ่มต้นทุกคู่มีค่า n_{min} เท่ากับ 0 และตารางที่ 3.8 แสดงผลลัพธ์ของการคำนวณหาค่า n_{max} และ n_{min} ของข้อมูลในรูปที่ 3.14 จะเห็นว่าคู่ที่มีค่า n_{max} ต่ำที่สุดคือคู่ของจุดศูนย์กลางเริ่มต้นจุดที่ 3 และจุดศูนย์กลางเริ่มต้นจุดที่ 4 ดังนั้นจึงเลือกจุดศูนย์กลางเริ่มต้นจุดที่ 3 และจุดศูนย์กลางเริ่มต้นจุดที่ 4 มารวมกันให้เป็นจุดศูนย์กลางเดียว



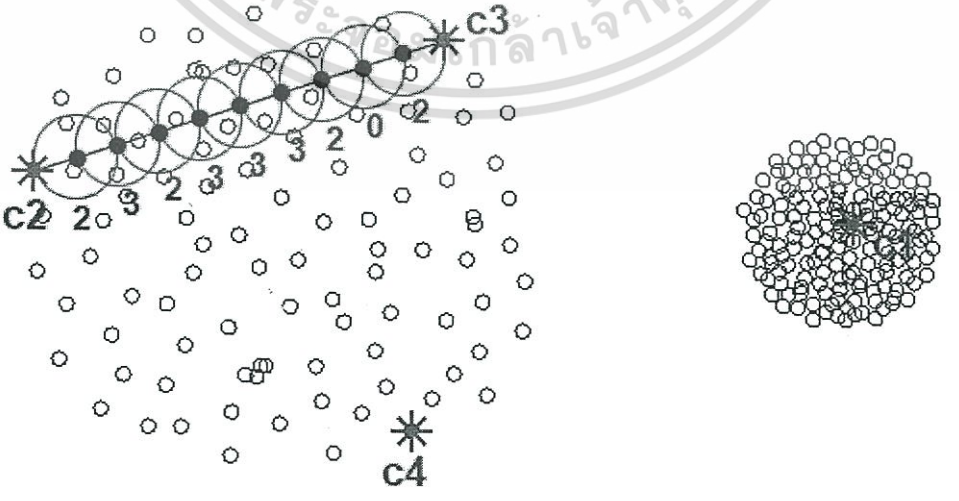
(ก) การนับจำนวนข้อมูลระหว่างจุดศูนย์กลางเริ่มต้นจุดที่ 1 และ จุดศูนย์กลางเริ่มต้นจุดที่ 2



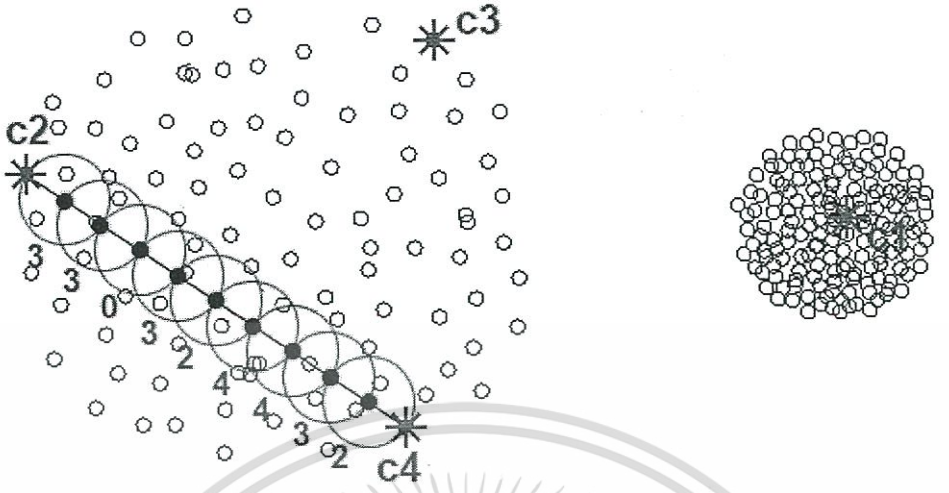
(ข) การนับจำนวนข้อมูลระหว่างจุดศูนย์กลางเริ่มต้นจุดที่ 1 และ จุดศูนย์กลางเริ่มต้นจุดที่ 3



(ค) การนับจำนวนข้อมูลระหว่างจุดศูนย์กลางเริ่มต้นจุดที่ 1 และ จุดศูนย์กลางเริ่มต้นจุดที่ 4



(ง) การนับจำนวนข้อมูลระหว่างจุดศูนย์กลางเริ่มต้นจุดที่ 2 และ จุดศูนย์กลางเริ่มต้นจุดที่ 3



(จ) การนับจำนวนข้อมูลระหว่างจุดศูนย์กลางเริ่มต้นจุดที่ 2 และ จุดศูนย์กลางเริ่มต้นจุดที่ 4

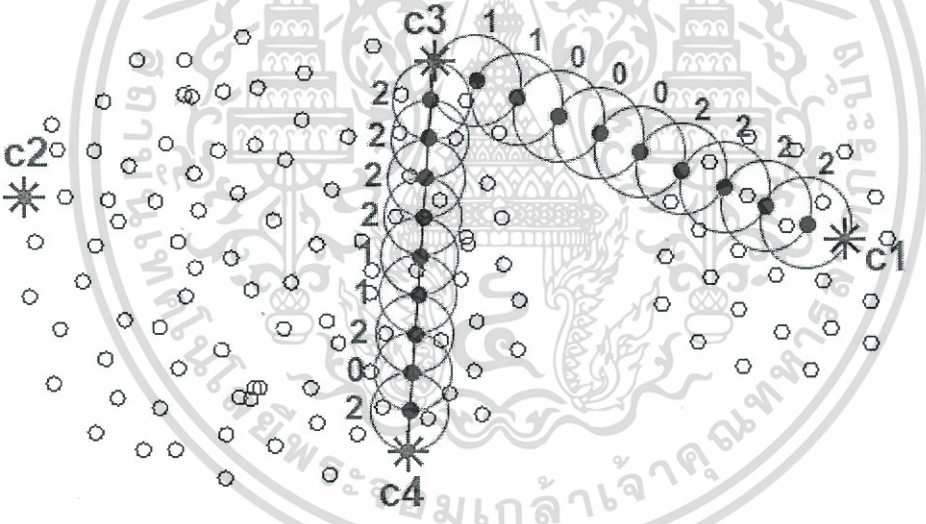
(ฉ) การนับจำนวนข้อมูลระหว่างจุดศูนย์กลางเริ่มต้นจุดที่ 3 และ จุดศูนย์กลางเริ่มต้นจุดที่ 4

รูปที่ 3.14 ตัวอย่างข้อมูลที่คู่ของจุดศูนย์กลางเริ่มต้นทุกคู่มีค่า n_{min} เท่ากับ 0

ตารางที่ 3.8 ผลลัพธ์การคำนวณหาค่า n_{max} และ n_{min} ของข้อมูลในรูปที่ 3.14

จุดศูนย์กลางเริ่มต้น 1	จุดศูนย์กลางเริ่มต้น 2	n_{max}	n_{min}
c1	c2	71	0
c1	c3	29	0
c1	c4	38	0
c2	c3	3	0
c2	c4	4	0
c3	**c4**	**2**	**0**

อย่างไรก็ตาม หากมีคู่ของจุดศูนย์กลางมากกว่า 1 คู่ที่มีค่า n_{max} ต่ำที่สุดเท่ากัน ให้นับจำนวนข้อมูลทั้งหมดที่ตกอยู่ภายในขอบเขตทั้ง 9 ขอบเขตระหว่างจุดศูนย์กลางคู่เหล่านั้น (n_{count}) แล้วเลือกจุดศูนย์กลางเริ่มต้นคู่ที่มีจำนวนข้อมูลมากที่สุดมารวมเป็นจุดศูนย์กลางเดียว โดยรูปที่ 3.15 คือตัวอย่างของข้อมูลที่มีคู่ของจุดศูนย์กลางเริ่มต้นที่มีค่า n_{max} ต่ำที่สุดเท่ากัน 2 คู่ และตารางที่ 3.9 แสดงผลลัพธ์ของการคำนวณหาค่า n_{max} และ n_{min} และผลลัพธ์การนับจำนวนข้อมูลทั้งหมดที่ตกอยู่ภายในขอบเขตทั้ง 9 ขอบเขตระหว่างคู่ของจุดศูนย์กลางที่มีค่า n_{max} ต่ำที่สุดเท่ากันของข้อมูลในรูปที่ 3.15 จากตารางที่ 3.9 จะเห็นว่าคู่ของจุดศูนย์กลางที่มีค่า n_{max} ต่ำที่สุดเท่ากันมีทั้งหมด 2 คู่ คือคู่ของจุดศูนย์กลางเริ่มต้นจุดที่ 1 และจุดศูนย์กลางเริ่มต้นจุดที่ 3 และคู่ของจุดศูนย์กลางเริ่มต้นจุดที่ 3 และจุดศูนย์กลางเริ่มต้นจุดที่ 4 เมื่อนับจำนวนข้อมูลทั้งหมดที่ตกอยู่ภายในขอบเขตทั้ง 9 ขอบเขตระหว่างจุดศูนย์กลางทั้ง 2 คู่นี้แล้ว จะเห็นว่าคู่ที่มีจำนวนข้อมูลมากกว่าคือคู่ของจุดศูนย์กลางเริ่มต้นจุดที่ 3 และจุดศูนย์กลางเริ่มต้นจุดที่ 4 ดังนั้นจึงทำการรวมจุดศูนย์กลางเริ่มต้นจุดที่ 3 และจุดศูนย์กลางเริ่มต้นจุดที่ 4 ให้เป็นจุดศูนย์กลางเดียว



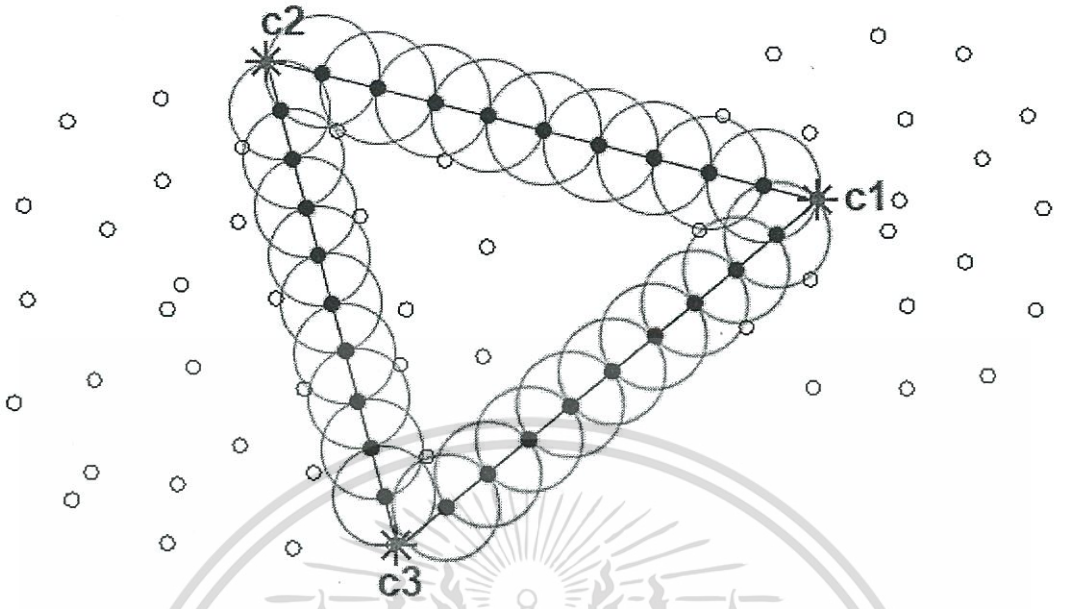
รูปที่ 3.15 ตัวอย่างข้อมูลที่มีคู่ของจุดศูนย์กลางเริ่มต้นที่มีค่า n_{max} ต่ำสุดเท่ากัน

ตารางที่ 3.9 ผลลัพธ์การคำนวณหาค่า n_{max}, n_{min} และจำนวนข้อมูลทั้งหมดที่ตกอยู่ภายในขอบเขตทั้ง 9 ขอบเขตระหว่างคู่ของจุดศูนย์กลางที่มีค่า n_{max} ต่ำที่สุด

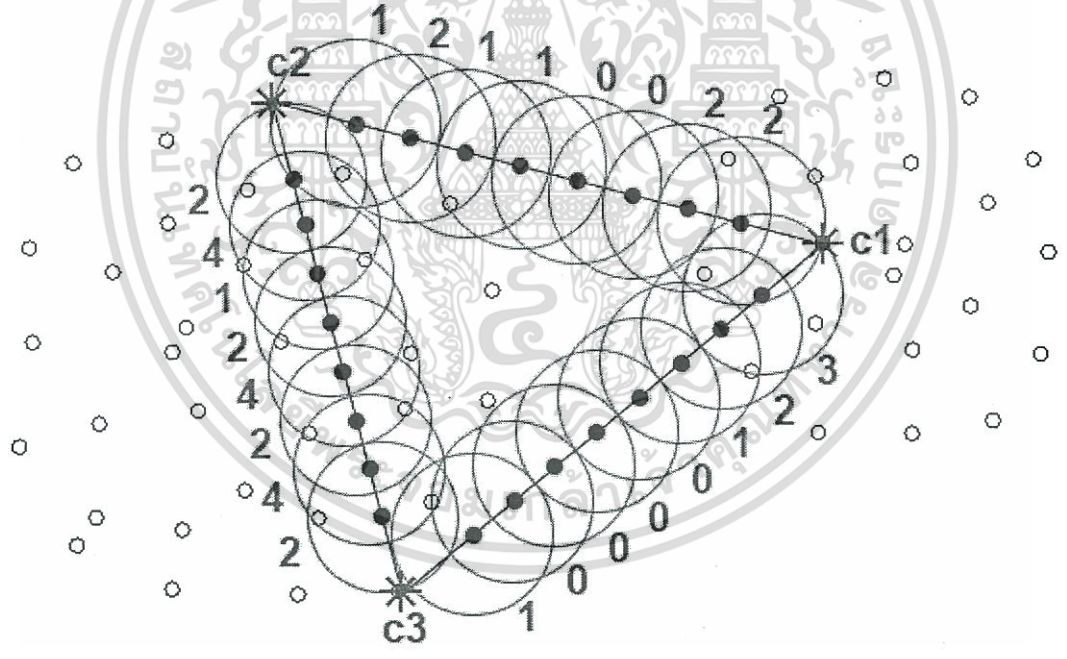
จุดศูนย์กลางเริ่มต้น 1	จุดศูนย์กลางเริ่มต้น 2	n_{max}	n_{min}	n_{count}
c1	c2	12	0	-
c1	c3	2	0	7
c1	c4	4	0	-
c2	c3	3	0	-
c2	c4	4	0	-
c3	c4	2	0	12

ทั้งนี้ หากจุดศูนย์กลางคู่ที่มีค่า n_{max} ต่ำสุดเท่ากันมีจำนวนข้อมูลทั้งหมดที่ตกอยู่ภายในขอบเขตทั้ง 9 ขอบเขตเท่ากัน ให้ลดจำนวนจุดที่สร้างขึ้นมาโดยใช้วิธีเช่นเดียวกันกับในกรณีที่ 1 แล้วคำนวณหาค่า n_{max} และ n_{min} ของคู่เหล่านั้นใหม่

- **กรณีที่ 3 :** หากคู่ของจุดศูนย์กลางเริ่มต้นทุกคู่มีค่า n_{max} และ n_{min} เท่ากับ 0 ให้ลดจำนวนจุดที่สร้างขึ้นมาโดยใช้วิธีเช่นเดียวกันกับในกรณีที่ 1 แล้วคำนวณหาค่า n_{max} และ n_{min} ใหม่ โดยรูปที่ 3.16 แสดงตัวอย่างของข้อมูลที่คู่ของจุดศูนย์กลางเริ่มต้นทุกคู่มีค่า n_{max} และ n_{min} เท่ากับ 0 รูปที่ 3.17 แสดงผลลัพธ์ที่ได้หลังจากลดจำนวนจุดที่สร้างขึ้นมาจาก 9 จุดให้เหลือ 8 จุดของตัวอย่างในรูปที่ 3.16 และหลังจากที่ลดจำนวนจุดให้เหลือ 8 จุดแล้ว จะสามารถคำนวณค่า n_{max} และ n_{min} ใหม่ได้ดังแสดงในตารางที่ 3.10 จากตารางที่ 3.10 จะเห็นว่ามีเพียงคู่เดียวที่สามารถคำนวณหาค่าอัตราส่วนระหว่าง n_{max} และ n_{min} ได้ (ค่า n_{min} ไม่เท่ากับ 0) คือคู่ของจุดศูนย์กลางเริ่มต้นจุดที่ 2 และจุดศูนย์กลางเริ่มต้นจุดที่ 3 ดังนั้นจึงเลือกรวมจุดศูนย์กลางเริ่มต้นจุดที่ 2 และจุดศูนย์กลางเริ่มต้นจุดที่ 3 ให้เป็นจุดศูนย์กลางเดียว



รูปที่ 3.16 ตัวอย่างข้อมูลที่คู่ของจุดศูนย์กลางเริ่มต้นทุกคู่มีค่า n_{max} และ n_{min} เท่ากับ 0

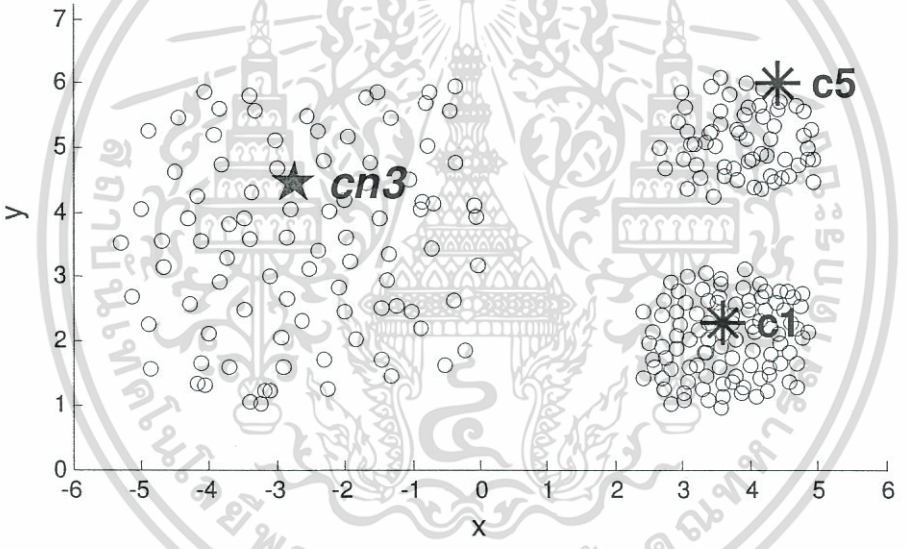


รูปที่ 3.17 ตัวอย่างการลดจำนวนจุดที่สร้างขึ้นจาก 9 จุดให้เหลือ 8 จุดของข้อมูลในรูปที่ 3.16

ตารางที่ 3.10 ผลลัพธ์การคำนวณอัตราส่วนระหว่าง n_{max} และ n_{min} ของข้อมูลในรูปที่ 3.16 หลังจากลดจำนวนจุดให้เหลือ 8 จุดแล้ว

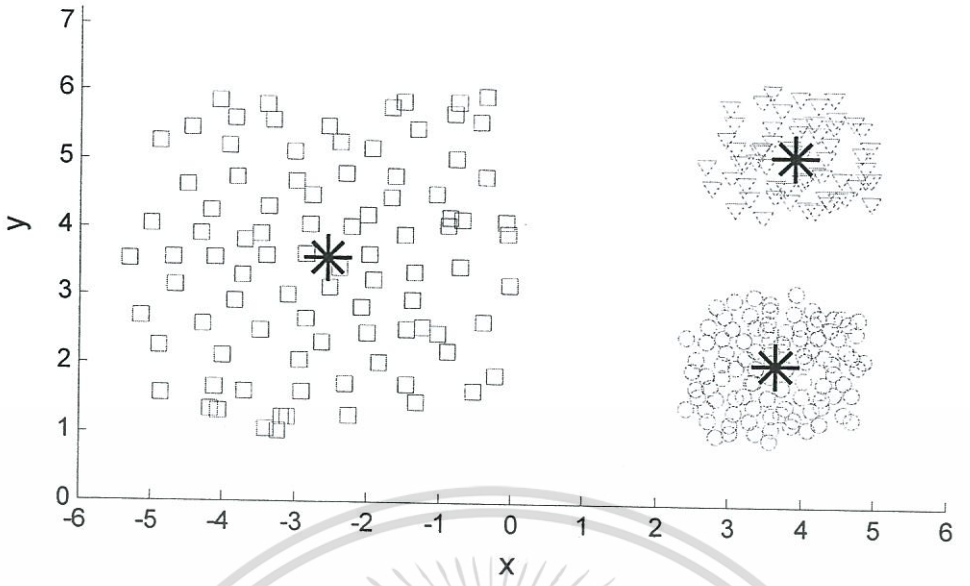
จุดศูนย์กลางเริ่มต้น 1	จุดศูนย์กลางเริ่มต้น 2	n_{max}	n_{min}	n_{max} / n_{min}
c1	c2	2	0	-
c1	c3	3	0	-
c2	**c3**	**4**	**1**	**4**

5) ทำซ้ำขั้นตอนที่ 1-4 จนกระทั่งเหลือจำนวนจุดศูนย์กลางเริ่มต้นเท่ากับเท่ากับ K โดยผลลัพธ์สุดท้ายของการคำนวณหาจุดศูนย์กลางเริ่มต้นของข้อมูลตัวอย่างในรูปที่ 3.1 แสดงได้ดังรูปที่ 3.18



รูปที่ 3.18 จุดศูนย์กลางเริ่มต้นทั้งหมดที่คำนวณได้

6) นำจุดศูนย์กลางเริ่มต้นที่คำนวณได้ไปใช้ในการจัดกลุ่มแบบเคมีนส์ตามปกติ โดยผลลัพธ์ที่ได้การจัดกลุ่มแบบเคมีนส์ของข้อมูลตัวอย่างในรูปที่ 3.1 โดยใช้จุดศูนย์กลางเริ่มต้นที่คำนวณได้จากข้อ 5 แสดงได้ดังรูปที่ 3.19



รูปที่ 3.19 ผลลัพธ์ที่ได้การจัดกลุ่มแบบเคมีนส์โดยใช้จุดศูนย์กลางเริ่มต้นที่คำนวณได้

อัลกอริทึมในขั้นตอนการคำนวณหาจุดศูนย์กลางเริ่มต้นจุดที่ 1 แสดงได้ดังรูปที่ 3.20 อัลกอริทึมในในขั้นตอนการคำนวณหาจุดศูนย์กลางเริ่มต้นจุดที่เหลือทั้งหมดแสดงได้ดังรูปที่ 3.21 และอัลกอริทึมในขั้นตอนการรวมจุดศูนย์กลางเริ่มต้นทีละคู่จนมีจำนวนเหลือเท่ากับจำนวนกลุ่ม แสดงได้ดังรูปที่ 3.22

ALGORITHM 1: Finding the first initial centroid

BEGIN
  Calculate the standard deviation *SD* of each variable
  Repeat
    For the sub-groups with the highest density of data points
      Divide data points into two sub-groups
    End for
  Until the number of cluster is equal to *2*K*
  Compare the number of data pointd of the last two sub-groups
  Calculate the median value for the sub-group with higher density of data points
  Assign the median value as the first initial centroid C_1
END

รูปที่ 3.20 อัลกอริทึมในขั้นตอนการคำนวณหาจุดศูนย์กลางเริ่มต้นจุดที่ 1

ALGORITHM 2: Finding the remaining initial centroids

BEGIN
  For each data points
    Calculate the Euclidean distance between it and the first initial centroid dc_1
  End for
  Select the data point with the highest dc_1 as the second initial centroid C_2
  Repeat
    For each remaining data points
      Calculate the minimum of the distances to the previously centroids d_i^{min}
    End for
    Assign the data point with the maximum d_i^{min} as the next initial centroid
  Until the number of initial centroids is equal to $2*K$
END

รูปที่ 3.21 อัลกอริทึมในขั้นตอนการคำนวณหาจุดศูนย์กลางเริ่มต้นจุดที่เหลือ

ALGORITHM 3: Merging pairs of initial centroids

```
BEGIN
  Repeat
    For each pair of initial centroids
      Create points at 1/10, 2/10, …, and 9/10 of the distance between each pair
      For each of the nine created points
        Determine the number of data points located within r distance (Eq.3.4)
      End for
      n_max ← The maximum number of data points
      n_min ← The maximum number of data points
      If n_min ≠ 0 Then
        ratio ← n_max / n_min
        If the number of pair with the lowest ratio = 1 Then
          Merge the pair with the lowest ratio into one centroid
        Else if the number of pair with the lowest ratio > 1 & the number created
        points > 6
          For each pair with the lowest ratio
            Reduce the number created points
            Recalculate n_max and n_min
          End for
        Else if the number of pair with the lowest ratio > 1 & the number created
        points = 6
          For each pair with the lowest ratio
            Merge the pair with the lowest distance between each other
          End for
        End if
      Else if n_max ≠ 0 & n_min = 0 Then
        If the number of pair with the lowest n_max = 1 Then
          Merge the pair with the lowest n_max into one centroid
        Else
          For each pair with the lowest n_max
            Determine the number of data points located between each pair
            Merge the pair with the highest number of data points
          End for
        End if
      Else if n_max = 0 & n_min = 0 Then
        If the number created points > 6 Then
          Reduce the number created points
          Recalculate n_max and n_min
        Else if number created points = 6
          Merge the pair with the lowest distance between each other
        End if
      End if
    End for
  Until the number of initial centroids is equal to K
END
```

รูปที่ 3.22 อัลกอริทึมในขั้นตอนการรวมจุดศูนย์กลางเริ่มต้น

บทที่ 4
ผลการทดลอง

ในบทนี้จะกล่าวถึงการทดสอบประสิทธิภาพที่ได้จากการจัดกลุ่มแบบเคมีนส์โดยใช้วิธีการคำนวณหาจุดศูนย์กลางเริ่มต้นที่งานวิจัยนี้นำเสนอ เปรียบเทียบกับประสิทธิภาพที่ได้จากวิธีการกำหนดจุดศูนย์กลางเริ่มต้นแบบอื่นๆ โดยเนื้อหาในบทนี้สามารถแบ่งออกได้เป็น 3 ส่วนหลักๆ ได้แก่ ข้อมูลที่นำมาใช้ในการทดลอง เกณฑ์ที่ใช้ในการวัดประสิทธิภาพของการจัดกลุ่ม และผลการทดลอง

4.1 ข้อมูลที่นำมาใช้ในการทดลอง

ในงานวิจัยนี้จะทดสอบประสิทธิภาพของการจัดกลุ่มโดยใช้ชุดข้อมูลทั้งหมด 11 ชุด โดยข้อมูล 9 ชุดจะมาจากคลังข้อมูลยูซีไอ (Bache and Lichman. 2013) ได้แก่ Glass Identification, Iris, Wine, Abalone, Soybean (Small), Wall-Following Robot Navigation, Statlog (Landsat Satellite), Ecoli, และ User Knowledge Modeling ส่วนข้อมูลอีก 2 ชุดจะเป็นชุดข้อมูลที่นำเสนอโดย Yeung, Haynor และ Ruzzo (2006) ได้แก่ Yeast Cell Cycle (subset 1) และ Yeast Cell Cycle (subset 2)

4.1.1 Glass Identification

เป็นข้อมูลที่ประกอบด้วยตัวอย่างจำนวน 214 ตัวอย่าง แอตทริบิวท์ 9 แอตทริบิวท์ โดยสามารถจำแนกคลาสออกได้ 6 คลาส คือ 1, 2, 3, 5, 6 และ 7 คลาสแต่ละคลาสประกอบด้วยตัวอย่างจำนวน 70, 76, 17, 13, 9 และ 29 ตัวอย่าง ตามลำดับ

4.1.2 Iris

เป็นข้อมูลที่ประกอบด้วยตัวอย่างจำนวน 150 ตัวอย่าง แอตทริบิวท์ 4 แอตทริบิวท์ โดยสามารถจำแนกคลาสออกได้ 3 คลาส คือ Iris-setosa, Iris-versicolor และ Iris-virginica คลาสแต่ละคลาสประกอบด้วยตัวอย่างจำนวน 50 ตัวอย่างเท่าๆ กัน

4.1.3 Wine

เป็นข้อมูลที่ประกอบด้วยตัวอย่างจำนวน 178 ตัวอย่าง แอตทริบิวท์ 13 แอตทริบิวท์ โดยสามารถจำแนกคลาสออกได้ 3 คลาส คือ 1, 2 และ 3 คลาสแต่ละคลาสประกอบด้วยตัวอย่างจำนวน 59, 71 และ 48 ตัวอย่าง ตามลำดับ

4.1.4 Abalone

เป็นข้อมูลที่ประกอบด้วยตัวอย่างจำนวน 4177 ตัวอย่าง แอตทริบิวท์ 8 แอตทริบิวท์ โดยสามารถจำแนกคลาสออกได้ 3 คลาส คือ F, I และ M คลาสแต่ละคลาสประกอบด้วยตัวอย่างจำนวน 1307, 1342 และ 1528 ตัวอย่าง ตามลำดับ

4.1.5 Soybean (Small)

เป็นข้อมูลที่ประกอบด้วยตัวอย่างจำนวน 47 ตัวอย่าง แอตทริบิวท์ 35 แอตทริบิวท์ โดยสามารถจำแนกคลาสออกได้ 4 คลาส คือ 1, 2, 3, และ 4 คลาสแต่ละคลาสประกอบด้วยตัวอย่างจำนวน 10, 10, 10 และ 17 ตัวอย่าง ตามลำดับ

4.1.6 Wall-Following Robot Navigation

เป็นข้อมูลที่ประกอบด้วยตัวอย่างจำนวน 5456 ตัวอย่าง แอตทริบิวท์ 24 แอตทริบิวท์ โดยสามารถจำแนกคลาสออกได้ 4 คลาส คือ Move-Forward, Sharp-Right-Turn, Slight-Left-Turn และ Slight-Right-Turn คลาสแต่ละคลาสประกอบด้วยตัวอย่างจำนวน 2205, 2097, 328 และ 826 ตัวอย่าง ตามลำดับ

4.1.7 Statlog (Landsat Satellite)

เป็นข้อมูลที่ประกอบด้วยตัวอย่างจำนวน 6435 ตัวอย่าง แอตทริบิวท์ 36 แอตทริบิวท์ โดยสามารถจำแนกคลาสออกได้ 6 คลาส คือ 1, 2, 3, 4, 5 และ 7 คลาสแต่ละคลาสประกอบด้วยตัวอย่างจำนวน 1533, 703, 1358, 626, 707 และ 1508 ตัวอย่าง ตามลำดับ

4.1.8 Ecoli

เป็นข้อมูลที่ประกอบด้วยตัวอย่างจำนวน 336 ตัวอย่าง แอตทริบิวท์ 7 แอตทริบิวท์ โดยสามารถจำแนกคลาสออกได้ 8 คลาส คือ cp, im, imL, imS, imU, om, omL และ pp คลาสแต่ละคลาสประกอบด้วยตัวอย่างจำนวน 143, 77, 2, 2, 35, 20, 5 และ 52 ตัวอย่าง ตามลำดับ

4.1.9 User Knowledge Modeling

เป็นข้อมูลที่ประกอบด้วยตัวอย่างจำนวน 403 ตัวอย่าง แอตทริบิวท์ 5 แอตทริบิวท์ โดยสามารถจำแนกคลาสออกได้ 4 คลาส คือ High, Middle, Low และ Very_low คลาสแต่ละคลาสประกอบด้วยตัวอย่างจำนวน 102, 122, 129 และ 50 ตัวอย่าง ตามลำดับ

4.1.10 Yeast Cell Cycle (subset 1)

เป็นข้อมูลที่ประกอบด้วยตัวอย่างจำนวน 384 ตัวอย่าง แอตทริบิวท์ 17 แอตทริบิวท์ โดยสามารถจำแนกคลาสออกได้ 5 คลาส คือ 1, 2, 3, 4 และ 5 คลาสแต่ละคลาสประกอบด้วยตัวอย่างจำนวน 67, 135, 75, 52 และ 55 ตัวอย่าง ตามลำดับ

4.1.11 Yeast Cell Cycle (subset 2)

เป็นข้อมูลที่ประกอบด้วยตัวอย่างจำนวน 237 ตัวอย่าง แอตทริบิวท์ 17 แอตทริบิวท์ โดยสามารถจำแนกคลาสออกได้ 4 คลาส คือ 1, 2, 3 และ 4 คลาสแต่ละคลาสประกอบด้วยตัวอย่างจำนวน 49, 31, 18 และ 139 ตัวอย่าง ตามลำดับ

4.2 เกณฑ์ที่ใช้ในการวัดประสิทธิภาพในการจัดกลุ่ม

ในงานวิจัยนี้จะทดสอบประสิทธิภาพในการจัดกลุ่มเพื่อเปรียบเทียบประสิทธิภาพที่ได้จากการจัดกลุ่มแบบเคมีนส์โดยใช้จุดศูนย์กลางเริ่มต้นที่คำนวณได้จากวิธีการที่งานวิจัยนี้นำเสนอเปรียบเทียบกับการใช้วิธีการคำนวณจุดศูนย์กลางเริ่มต้นแบบอื่นๆ โดยงานวิจัยนี้จะใช้ค่าเปอร์เซ็นต์ความผิดพลาด (Error Percentage) มาเป็นเกณฑ์ในการวัดประสิทธิภาพในการจัดกลุ่ม โดยค่าเปอร์เซ็นต์ความผิดพลาดสามารถคำนวณได้จากสมการที่ 4.1

$$\text{Error Percentage} = \frac{\varepsilon}{N} \tag{4.1}$$

เมื่อ ε คือจำนวนข้อมูลที่ถูกจัดกลุ่มผิด

N คือจำนวนข้อมูลทั้งหมด

วิธีใดที่คำนวณได้ค่าเปอร์เซ็นต์ความผิดพลาดต่ำ หมายความว่าวิธีการกำหนดศูนย์กลางเริ่มต้นวิธีนั้นมีประสิทธิภาพ ให้ผลลัพธ์ในการจัดกลุ่มที่มีความถูกต้องสูง ส่วนวิธีใดที่คำนวณได้ค่าเปอร์เซ็นต์ความผิดพลาดสูง หมายความว่าวิธีการกำหนดศูนย์กลางเริ่มต้นวิธีนั้นมีประสิทธิภาพต่ำ ยังไม่สามารถให้ผลลัพธ์ในการจัดกลุ่มที่มีความถูกต้องมากนัก

4.3 ผลการทดลอง

ในงานวิจัยนี้จะมีการทดลองเพื่อเปรียบเทียบประสิทธิภาพในการจัดกลุ่มแบบเคมีนส์โดยใช้วิธีการกำหนดจุดศูนย์กลางเริ่มต้นตามที่งานวิจัยนี้นำเสนอ เปรียบเทียบกับวิธีการกำหนดจุดศูนย์กลางเริ่มต้นแบบสุ่ม และเปรียบเทียบกับวิธีการกำหนดจุดศูนย์กลางเริ่มต้นที่นำเสนอโดย Erisoglu, Calis และ Sakallioglu (2011) ซึ่งสำหรับวิธีการกำหนดจุดศูนย์กลางเริ่มต้นแบบสุ่มจะทดลองโดยการสุ่มจุดศูนย์กลางเริ่มต้นขึ้นมาทั้งหมด 10 ครั้งของข้อมูลแต่ละชุด แล้วคำนวณคำนวณหาค่าเปอร์เซ็นต์ความผิดพลาดของการสุ่มจุดศูนย์กลางเริ่มต้นแต่ละครั้ง

โดยเครื่องคอมพิวเตอร์ที่ใช้ในการทดลองมีคุณสมบัติดังนี้

- หน่วยประมวลผลเอเอ็มดี (Advance Micro Devices: AMD) C-50 (1.00 GHz, 1MB L2Cache)

- แรม (Random Access Memory: RAM) 2GB DDR3
- ฮาร์ดดิสก์ (Hard Disk) 320 GB 5400 RPM

และใช้โปรแกรมแมทแลป (MATLAB) เวอร์ชั่น R2010a ในการเขียนโค้ดเพื่อทำการทดลอง

โดยผลการทดลองเพื่อเปรียบเทียบประสิทธิภาพในการจัดกลุ่มแบบเคมีนส์โดยใช้วิธีการกำหนดจุดศูนย์กลางเริ่มต้นแบบต่างๆ ซึ่งใช้ค่าเปอร์เซ็นต์ความผิดพลาดเป็นเกณฑ์ในการวัดประสิทธิภาพ แสดงได้ดังตารางที่ 4.1

ตารางที่ 4.1 ผลการทดสอบประสิทธิภาพในการจัดกลุ่มโดยใช้วิธีการกำหนดจุดศูนย์กลางเริ่มต้นแบบต่างๆ

Datasets	Error Percentage			
	Original K-means		Proposed method	Erisoglu et al. [8]
Glass Identification	best	41.1215	45.7944	46.2617
	average	46.1215		
	worst	51.4019		
Iris	best	10.6667	10.6667	10.6667
	average	19.7333		
	worst	33.3333		
Wine	best	29.7753	32.5843	29.7753
	average	30.0562		
	worst	32.5843		
Yeast Cell Cycle (subset 1)	best	27.0833	33.0729	26.0417
	average	33.2552		
	worst	39.3229		
Yeast Cell Cycle (subset 2)	best	25.7384	25.7384	25.7384
	average	26.7089		
	worst	30.8017		
Abalone	best	48.5037	48.5037	48.5037
	average	49.2531		
	worst	51.0175		

ตารางที่ 4.1 (ต่อ) ผลการทดสอบประสิทธิภาพในการจัดกลุ่มโดยใช้วิธีการกำหนดจุดศูนย์กลางเริ่มต้นแบบต่างๆ

Datasets	Error Percentage			
	Original K-means		Proposed method	Improved K-means (Erisoglu et.al, 2011)
Soybean (Small)	best	0	21.2766	21.2766
	average	24.2553		
	worst	42.5532		
Wall-Following Robot Navigation	best	49.7617	50.4949	50.4949
	average	50.7331		
	worst	51.1364		
Statlog (Landsat Satellite)	best	26.216	38.073	46.1228
	average	38.5734		
	worst	46.589		
Ecoli	best	15.7738	19.0476	20.8333
	average	19.3155		
	worst	24.7024		
User Knowledge Modeling	best	36.7246	37.2208	38.7097
	average	41.34		
	worst	50.6204		

จากตารางที่ 4.1 เมื่อเปรียบเทียบประสิทธิภาพในการจัดกลุ่มที่ได้จากวิธีการกำหนดจุดศูนย์กลางเริ่มต้นที่งานวิจัยนี้นำเสนอกับประสิทธิภาพในการจัดกลุ่มที่ได้จากวิธีการกำหนดจุดศูนย์กลางแบบสุ่ม โดยเปรียบเทียบกับค่าเฉลี่ย (Average) ของเปอร์เซ็นต์ความผิดพลาดจากการสุ่มจุดศูนย์กลางเริ่มต้นทั้งหมด 10 ครั้ง จะเห็นว่า วิธีการกำหนดจุดศูนย์กลางเริ่มต้นที่งานวิจัยนี้นำเสนอให้ค่าเปอร์เซ็นต์ความผิดพลาดที่น้อยกว่าวิธีการกำหนดจุดศูนย์กลางเริ่มต้นแบบสุ่มทั้งสิ้น 10 ชุดข้อมูล มีเพียง 1 ชุดข้อมูลที่ให้ค่าเปอร์เซ็นต์ความผิดพลาดที่มากกว่า คือ ชุดข้อมูล Wine

ในการเปรียบเทียบประสิทธิภาพในการจัดกลุ่มที่ได้จากวิธีการกำหนดจุดศูนย์กลางเริ่มต้นที่งานวิจัยนี้นำเสนอกับประสิทธิภาพที่ได้จากวิธีการกำหนดจุดศูนย์กลางเริ่มต้นที่นำเสนอโดย Erisoglu, Calis และ Sakallioglu จะเห็นว่า วิธีการกำหนดจุดศูนย์กลางเริ่มต้นที่งานวิจัยนี้นำเสนอให้ค่าเปอร์เซ็นต์ความผิดพลาดที่น้อยกว่าวิธีการกำหนดจุดศูนย์กลางเริ่มต้นที่นำเสนอโดย Erisoglu,

Calis และ Sakallioglu ทั้งสิ้น 4 ชุดข้อมูล ได้แก่ Glass Identification, Statlog (Landsat Satellite), Ecoli และ User Knowledge Modeling ส่วนชุดข้อมูลที่ให้ค่าเปอร์เซ็นต์ความผิดพลาดเท่ากันมีทั้งสิ้น 5 ชุดข้อมูล ได้แก่ Iris, Yeast Cell Cycle (subset 2), Abalone, Soybean (Small) และ Wall-Following Robot Navigation และมีเพียง 2 ชุดข้อมูลที่ให้ค่าเปอร์เซ็นต์ความผิดพลาดที่มากกว่า ได้แก่ Wine และ Yeast Cell Cycle (subset 1)

ตารางที่ 4.2 แสดงจำนวนครั้งที่อัลกอริทึมที่นำเสนอให้ค่าเปอร์เซ็นต์ความผิดพลาดที่น้อยกว่าเท่ากับ และมากกว่าค่าเปอร์เซ็นต์ความผิดพลาดที่ได้จากการสุ่มจุดศูนย์กลางเริ่มต้นแบบเคมีนส์ดั้งเดิมทั้งหมด 10 ครั้ง ของข้อมูลทั้งหมด 11 ชุด

ตารางที่ 4.2 การเปรียบเทียบเปอร์เซ็นต์ความผิดพลาดที่ได้จากอัลกอริทึมที่นำเสนอกับเปอร์เซ็นต์ความผิดพลาดที่ได้จากการสุ่มจุดศูนย์กลางเริ่มทั้งหมด 10 ครั้ง

Datasets	จำนวนครั้งที่เปอร์เซ็นต์ความผิดพลาดน้อยกว่าเคมีนส์ดั้งเดิม	จำนวนครั้งที่เปอร์เซ็นต์ความผิดพลาดเท่ากับเคมีนส์ดั้งเดิม	จำนวนครั้งที่เปอร์เซ็นต์ความผิดพลาดมากกว่าเคมีนส์ดั้งเดิม
Glass Identification	3	-	7
Iris	4	6	-
Wine	-	2	8
Yeast Cell Cycle (subset 1)	6	2	2
Yeast Cell Cycle (subset 2)	3	7	-
Abalone	3	7	-
Soybean (Small)	3	6	1
Wall-Following Robot Navigation	8	-	2
Statlog (Landsat Satellite)	5	5	-
Ecoli	5	5	-
User Knowledge Modeling	7	-	3

และผลการทดลองเพื่อเปรียบเทียบเวลาที่ใช้ในการประมวลผลโดยใช้วิธีการกำหนดจุดศูนย์กลางเริ่มต้นแบบต่างๆแสดงได้ดังตารางที่ 4.3

ตารางที่ 4.3 ผลการทดสอบเวลาที่ใช้ในการประมวลผลโดยใช้วิธีการกำหนดจุดศูนย์กลางเริ่มต้นแบบต่างๆ

Datasets	Time (sec)		
	Original K-means	Proposed method	Improved K-means (Erisoglu et.al, 2011)
Glass Identification	24.560316	10.346430	7.181547
Iris	24.324212	6.929355	5.930217
Wine	23.904464	6.774086	5.907031
Yeast Cell Cycle (subset 1)	6.008016	16.722695	6.596300
Yeast Cell Cycle (subset 2)	5.851093	10.087512	6.363757
Abalone	7.841718	42.484876	8.222498
Soybean (Small)	5.761039	8.021462	6.507600
Wall-Following Robot Navigation	14.376440	206.099392	21.163142
Statlog (Landsat Satellite)	25.730006	231.288009	39.058802
Ecoli	6.375262	43.973114	5.989103
User Knowledge Modeling	7.146451	9.768558	6.510594
Average	13.80718336	53.86322627	10.85732645

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

งานวิจัยนี้เป็นงานวิจัยที่มุ่งเน้นข้อจำกัดบางประการของการจัดกลุ่มแบบเคมีนส์ กล่าวคือ ปกติแล้วการจัดกลุ่มแบบเคมีนส์จะต้องมีการสุ่มข้อมูลขึ้นมาเพื่อใช้เป็นจุดศูนย์กลางเริ่มต้น ซึ่งผลลัพธ์ที่ได้จากการจัดกลุ่มแบบเคมีนส์จะขึ้นอยู่กับจุดศูนย์กลางเริ่มต้นที่สุ่มขึ้นมาได้ หากสุ่มได้จุดศูนย์กลางเริ่มต้นที่ไม่เหมาะสม ก็อาจทำให้ผลลัพธ์ที่ได้จากการจัดกลุ่มแบบเคมีนส์ไม่มีประสิทธิภาพ ดังนั้น งานวิจัยนี้จึงนำเสนอวิธีการคำนวณหาจุดศูนย์กลางเริ่มต้นแทนการใช้วิธีการสุ่มแบบเดิม

โดยแนวความคิดที่งานวิจัยนี้นำมาใช้ในการหาจุดศูนย์กลางเริ่มต้น ได้แก่

1) บริเวณใดที่มีความหนาแน่นของข้อมูลมาก บริเวณนั้นก็น่าจะเป็นบริเวณที่ข้อมูลกระจุกรวมตัวกันเป็นกลุ่มหรือคลัสเตอร์

2) ข้อมูลที่อยู่ในกลุ่มเดียวกันจะมีระยะทางใกล้กันและข้อมูลที่อยู่ต่างกลุ่มกันจะมีระยะทางห่างกัน

3) หากข้อมูลที่อยู่ระหว่างจุดศูนย์กลางเริ่มต้นคู่ใดๆ มีการกระจายแบบสม่ำเสมอ สันนิษฐานว่าจุดศูนย์กลางเริ่มต้นคู่นั้นอาจจะตกอยู่ในกลุ่มเดียวกัน

งานวิจัยนี้ได้ทดสอบประสิทธิภาพในการจัดกลุ่มโดยใช้ชุดข้อมูลมาตรฐานทั้งหมด 11 ชุด และใช้ค่าเปอร์เซ็นต์ความผิดพลาดมาเป็นเกณฑ์ในการวัดประสิทธิภาพ ซึ่งจากผลการทดลองในบทที่ 4 แสดงให้เห็นว่า วิธีการคำนวณหาจุดศูนย์กลางเริ่มต้นที่งานวิจัยนี้นำเสนอให้ผลลัพธ์ที่มีประสิทธิภาพเมื่อเปรียบเทียบกับวิธีการกำหนดจุดศูนย์กลางเริ่มต้นแบบสุ่มและวิธีการกำหนดจุดศูนย์กลางเริ่มต้นที่นำเสนอโดยงานวิจัยอื่นที่งานวิจัยนี้นำมาเปรียบเทียบ

5.2 ข้อดีของงานวิจัย

1) การจัดกลุ่มแบบเคมีนส์โดยใช้วิธีการกำหนดจุดศูนย์กลางเริ่มต้นที่งานวิจัยนี้นำเสนอให้ผลลัพธ์การจัดกลุ่มที่มีความถูกต้องแม่นยำมากกว่าการกำหนดจุดศูนย์กลางเริ่มต้นโดยใช้วิธีการสุ่มข้อมูล

2) การจัดกลุ่มแบบเคมีนส์โดยใช้วิธีการกำหนดจุดศูนย์กลางเริ่มต้นที่งานวิจัยนี้นำเสนอให้ผลลัพธ์ที่มีความเสถียรภาพ ซึ่งแตกต่างจากการจัดกลุ่มแบบเคมีนส์ดั้งเดิม ที่ผลลัพธ์การจัดกลุ่มจะขึ้นอยู่กับจุดศูนย์กลางเริ่มต้นที่สุ่มขึ้นมาได้

3) สามารถหาจุดศูนย์กลางเริ่มต้นที่เหมาะสมกับข้อมูลที่กลุ่มแต่ละกลุ่มมีขนาดหรือความหนาแน่นของข้อมูลแตกต่างกันได้

5.3 ปัญหาที่พบในงานวิจัย

1) ในขั้นตอนการคำนวณหาจุดศูนย์กลางเริ่มต้นทั้งหมด งานวิจัยนี้ได้กำหนดให้หาจุดศูนย์กลางเริ่มต้นให้เป็นจำนวนสองเท่าของจำนวนกลุ่ม (แล้วค่อยมายุบรวมจุดศูนย์กลางทีละคู่จนเหลือจำนวนจุดศูนย์กลางเริ่มต้นเท่ากับจำนวนกลุ่ม) ซึ่งในความเป็นจริงแล้วจะไม่สามารถทราบได้แน่ชัดว่าจำนวนจุดศูนย์กลางเริ่มต้นที่จะต้องคำนวณหาควรจะเป็นจำนวนเท่าไร

2) ขั้นตอนของการรวมจุดศูนย์กลางเริ่มต้นให้เหลือจำนวนจุดศูนย์กลางเริ่มต้นเท่ากับจำนวนกลุ่ม จะต้องมีการสร้างขอบเขตขึ้นมาระหว่างจุดศูนย์กลางเริ่มต้นแต่ละคู่ ซึ่งงานวิจัยนี้กำหนดให้สร้างขอบเขตเริ่มต้นขึ้นมาทั้งหมด 9 ขอบเขต หรือแบ่งระยะทางระหว่างจุดศูนย์กลางแต่ละคู่ออกเป็น 10 ช่วง ซึ่งในความเป็นจริงแล้วจะไม่สามารถทราบว่าจำนวนขอบเขตที่เหมาะสมควรจะเป็นเท่าไร

3) วิธีการคำนวณหาจุดศูนย์กลางเริ่มต้นที่งานวิจัยนี้นำเสนอประกอบด้วยขั้นตอนหลายขั้นตอน ทำให้มีความซับซ้อนค่อนข้างสูง

5.4 แนวทางในการพัฒนาต่อ

1) คิดค้นวิธีการกำหนดจำนวนจุดศูนย์กลางเริ่มต้นทั้งหมด

2) คิดค้นวิธีการกำหนดจำนวนขอบเขตระหว่างจุดศูนย์กลางเริ่มต้นแต่ละคู่ในขั้นตอนของการรวมจุดศูนย์กลางเริ่มต้นให้เหลือจำนวนจุดศูนย์กลางเริ่มต้นเท่ากับจำนวนกลุ่ม

3) ปรับปรุงหรือลดกระบวนการบางขั้นตอนของวิธีการคำนวณหาจุดศูนย์กลางเริ่มต้นที่งานวิจัยนี้นำเสนอเพื่อลดความซับซ้อนและเวลาที่ใช้ในการประมวลผล

4) นำวิธีการจัดกลุ่มแบบเคมีนส์โดยใช้วิธีการกำหนดจุดศูนย์กลางเริ่มต้นที่งานวิจัยนี้นำเสนอไปประยุกต์ใช้กับระบบสารสนเทศต่างๆ

บรรณานุกรม

Bache, K. and Lichman, M. 2013. **UCI machine learning repository**. [Online]. Available : http://archive.ics.uci.edu/ml

Deelers, S. and Auwatanamongkol, S. 2007. "Enhancing k-means algorithm with initial cluster centers derived from data partitioning along the data axis with the highest variance." **Internat. J. Comput. Sci.** 2 : 247–252.

Erisoglu, M., Calis, N. and Sakallioglu, S. 2011. "A new algorithm for initial cluster centers in K-means algorithm." **Pattern Recognition Letters**. 32 : 1701-1705.

Fahim, A.M. et.al. 2006. "An Efficient Enhanced K-means Clustering Algorithm." **Journal of Zhejiang University**. 10(7) : 1626-1633.

Fahim, A.M. et.al. 2009. "K-Means for Spherical Clusters with Large Variance in Sizes." **International Journal of Electrical and Computer Engineering.** 4(3) : 145-150.

Khan, S.S. and Ahmad, A. 2004. "Cluster center initialization algorithm for k-means algorithm." **Pattern Recognition Letters.** 25 : 1293–1302.

Li, Y. and Wu, H. 2012. "A Clustering Method Based on K-means Algorithm." 1104-1109. in **International Conference on Solid State Devices and Materials Science.**

Lin, Y. et.al. 2012. "A Improved Clustering Method Based on K-means." 734-737. in **9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2012).**

Nazeer, K.A., Kumar, S.D. and Sebastian, M.P. 2011. "Enhancing the K-means Clustering Algorithm by Using a O(n logn) Heuristic Method for Finding Better Initial Centroids." 261-264. in **Second International Conference on Emerging Applications of Information Technology.**

Tajunisha, N. and Saravanan V. 2011. "An efficient method to improve the clustering performance for high dimensional data by Principal Component Analysis and modified K-means." **International Journal of Database Management Systems (IJDMS).** 3(1) : 196-205.

Yeung, K.Y., Haynor, D.R. and Ruzzo, W.L. 2006. **Validating Clustering for Gene Expression Data.** Available : http://www.cs.washington.edu/homes/kayee/cluster

ภาคผนวก

ผลงานวิจัยที่ได้รับการตีพิมพ์เผยแพร่

Arit Thammano and Pathcharnee Chattanes. "Efficient cluster center initialization method for K-means clustering.". **The Nineteenth International Symposium on Artificial Life and Robotics 2014 (AROB 19th 2014)**, B-Con Plaza, Beppu, Japan, 2014

Program

THE NINETEENTH INTERNATIONAL SYMPOSIUM ON ARTIFICIAL LIFE AND ROBOTICS

(AROB 19th 2014)

January 22– 24, 2014

B-Con Plaza, Beppu, Oita, JAPAN

International Society of Artificial Life and Robotics

The Nineteenth International Symposium on Artificial Life and Robotics 2014 (AROB 19th 2014),
B-Con Plaza, Beppu, Japan, January 22-24, 2014

Efficient cluster center initialization method for K-means clustering

Arit Thammano[1] and Pathcharnee Chattanes[2]

Computational Intelligence Laboratory
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok, 10520 Thailand
(Tel: 662-723-4964, Fax: 662-723-4910)

[1]arit@it.kmitl.ac.th and [2]pepathchar@gmail.com

Abstract: Clustering is a technique to divide a set of objects into several clusters. Among the clustering methods, K-means algorithm is one of the most well-known methods. However, the clustering results of the K-means algorithm depend heavily on the randomly chosen initial centroids. Therefore, this paper proposes a method for finding the appropriate initial centroids for K-means algorithm. The experimental results show that the proposed initialization method produces more accurate clusters than the original K-means algorithm for most of the datasets.

Keywords: Clustering, Initialization method, K-means algorithm

1 INTRODUCTION

Clustering is a technique to divide a set of objects into several groups or clusters. The objects in the same cluster are more similar to one another than to the objects in other clusters. In contrast to classification, clustering does not rely on class-labeled training examples. Therefore, clustering is a form of learning by observation, rather than learning by examples [1]. Clustering has been widely used in numerous applications, such as medicine, business, biology, information retrieval, pattern recognition, and image processing [2]. Currently, there are many clustering algorithms available. One of the most well-known clustering algorithms is K-means algorithm because of its simplicity, speed, and effectiveness. However, the performance of K-means depends heavily on the initial centroids which are selected randomly. If the initial centroids are not properly chosen, a bad clustering result will be achieved [2]. Therefore, several methods for determining the initial centroids have been proposed to overcome this shortcoming.

Redmond and Heneghan [3] proposed a method to determine the initial centroids for K-means algorithm. The proposed method uses a kd-tree to estimate the density of the data at various locations. Then it sequentially selects K initial centroids by using the distance and the density information to aid each selection.

Li and Wu [4] proposed an improved K-means clustering algorithm which is a combination of the largest minimum distance algorithm and the original K-means algorithm. The largest minimum distance algorithm is used to initialize the cluster centers. The experimental results show that the proposed algorithm outperforms the original K-means in terms of the convergence speed, cluster precision, and stability.

Tajunisha and Saravanan [5] proposed a method to improve the performance of K-means algorithm. In the proposed algorithm, Principal Component Analysis (PCA) is used to find the initial centroids and to reduce the dimension of the data. Moreover, K-means algorithm is modified to reduce the computational complexity of the process of assigning the data points to clusters.

Pavan et al. [6] proposed a new method, called Single Pass Seed Selection (SPSS), to determine the initial centroids. This proposed method is an extension of K-means++, which is proposed by Arthur and Vassilvitskii [7]. The advantage of the SPSS is that it produces a single optimal solution which is insensitive to outliers.

Erisoglu, Calis, and Sakallioglu [8] proposed an algorithm to compute the initial centroids for K-means algorithm. The proposed algorithm chooses two variables that best describe the change in the dataset. First, the variable which has the maximum value of the coefficient of variation is selected as the main axis. After determining the main axis, the variable which has the minimum absolute correlation coefficient between the main axis and itself is selected as the second axis. After that, the proposed algorithm chooses the initial centroids that are far away from one another.

Nazeer, Kumar, and Sebastian [9] proposed an algorithm, based on the concepts of sorting and partitioning the input data, to determine the initial centroids for K-means algorithm. The proposed algorithm sorts the input

The Nineteenth International Symposium on Artificial Life and Robotics 2014 (AROB 19th 2014),
B-Con Plaza, Beppu, Japan, January 22-24, 2014

data set and partitions the sorted data set into K number of sets. Then the mean value of each set is used as the initial centroid. The experimental results show that the proposed algorithm produces better clustering results than the original K-means algorithm. Moreover, it also uses less computation time.

This paper proposes a new initialization method for K-means clustering algorithm with the aim to improve the accuracy of the clustering results.

The rest of the paper is organized as follows. Section 2 describes the original K-means algorithm. The proposed initialization method is presented in section 3. The experimental results are discussed in section 4. Finally, section 5 is the conclusions.

2 K-MEANS CLUSTERING ALGORITHM

K-means algorithm is one of the most well-known clustering algorithms. K-means algorithm comprises of two main phases. In the first phase, K data points are randomly selected from the data set and used as the initial centroids. In the second phase, the rest of the data points are assigned to the closest centroid by calculating the Euclidean distances between each data point and the initial centroids. After all data points are assigned to one of the clusters, the arithmetic mean of each cluster is calculated and used as a new centroid. This process is repeated until no change in the cluster centers occurs. The pseudo code of K-means algorithm is shown as follows:

```
Define the number of clusters (K).
Randomly select K data points as the initial centroids.
Repeat
    Assign the data points to their closest centroids.
    Update each cluster center by computing the
    arithmetic mean of all data points in the cluster.
Until all centroids are unchanged.
```

Although K-means algorithm is simple, fast, and very effective, it has two limitations. First, K-means algorithm requires the user to predetermine the number of clusters, which is denoted as K. It is practically very difficult to decide on the suitable cluster number. Second, the performance of K-means depends heavily on the initial selection of the cluster centers. If the initial centroids are not properly chosen, a bad clustering result will be achieved.

3 PROPOSED METHOD

To improve the performance of the K-means clustering algorithm, this paper proposes a method for determining the appropriate initial centroids. The proposed method tries to choose the initial centroids that are far away from one another. The proposed method consists of three main phases. In the first phase, the first initial centroid is determined. In the second phase, the remaining initial centroids are determined. The total number of initial centroids must be twice the value of the predefined number of clusters K. In the last phase, a pair of initial centroids is merged together until the number of centroids is equal to K. A pair of initial centroids is merged together if there exist many data points uniformly located along the area between the two centroids. When all initialization phases are complete, the original K-means algorithm is performed. The detailed steps of each initialization phase are described as follows:

The first phase:

1. Calculate the standard deviation of each variable.
2. Divide all data points into two sub-groups by using the middle of the range of the variable with the highest standard deviation as a cut-off point.
3. Repeat steps 1 – 2 on the sub-group with the highest density of data points until the total number of sub-groups is equal to K.
4. Compare the last two sub-groups. The one with higher density of data points is selected.
5. Based on the variable with the highest standard deviation, perform the followings: i) Sort the data points in the selected sub-group, and ii) Calculate the median value of the data points in the selected sub-group.
   - 5.1 If the number of data points in the selected sub-group is odd, the median value will be at the position (n+1)/2.
   - 5.2 If the number of data points in the selected sub-group is even, the median value will be determined according to the following process. First, the Euclidean distance between the data point at position (n/2) and the data point at position (n/2)-1 is calculated. Second, the Euclidean distance between the data point at position (n/2)+1 and the data point at position (n/2)+2 is calculated. Third, the above two distances are compared. If the former is smaller than the latter, the median value will be at the

The Nineteenth International Symposium on Artificial Life and Robotics 2014 (AROB 19th 2014),
B-Con Plaza, Beppu, Japan, January 22-24, 2014

position (n/2). Otherwise, the median value will be at the position (n/2)+1.

6. Assign the median value obtained in step 5 as the first initial centroid.

The second phase:

1. Calculate the Euclidean distance between each data point and the first initial centroid. The Euclidean distance between each data point and the first initial centroid is denoted as dc_{1i}.
2. Assign the data point with the highest dc_{1i} as the second initial centroid.
3. Calculate the minimum of the Euclidean distances between each remaining data point and the previously assigned initial centroids.

$$d_i^{min} = \min_{k=1}^{A}(dc_{ki}) \quad (1)$$

where k = 1, 2, ..., A. A is the number of previously assigned initial centroids.

4. Assign the data point with the maximum d_i^{min} as the next initial centroid.

$$I = \arg\max_{i=1}^{N}\left(d_i^{min}\right) \quad (2)$$

where N is the number of remaining data points. I is the index of the data point with the maximum d_i^{min}.

5. Repeat steps 3 – 5 until the number of initial centroids reaches twice the value of K.

The third phase:

1. Create points at 1/10, 2/10, 3/10, ..., and 9/10 of the distance between each pair of initial centroids. Let d_{ij} denote the distance between the i and j initial centroids. Therefore, the distance between any pair of adjacent points (r) is defined as:

$$r = \frac{d_{ij}}{10} \quad (3)$$

2. For each pair of initial centroids, determine the number of data points located within r distance from each of the nine created points. The maximum number of data points located within r distance from each of the nine created points is denoted as n_{max} while the minimum number of data points located within r distance from each of the nine created points is denoted as n_{min}.
3. Pick a pair of the initial centroids and merge them together. The merging process is performed as follows:

Case 1:

If there is at least one pair of initial centroids whose both n_{max} and n_{min} are not equal to 0, calculate the ratio between n_{max} and n_{min}. After the ratios of all pairs of initial centroids are calculated, they are sorted from lowest to highest ratio. The pair with the lowest ratio is the first to be merged, and the pair with the highest ratio is the last to be merged. The new centroid is the arithmetic mean of these two initial centroids.

Case 2:

If n_{max} is not equal to 0 but n_{min} is equal to 0, the pair with the lowest n_{diff} (= n_{max} – n_{min}) will be the first to be merged and the pair with the highest n_{diff} will be the last to be merged. However, if there is more than one pair with the same lowest n_{diff}, determine the total number of data points located between each of these pairs. The pair with the highest number of data points is the first to be merged and the pair with the lowest number of data points is the last to be merged.

Case 3:

If both n_{max} and n_{min} are equal to 0, the number of points which are created in step 1 will be reduced by 1. This causes r to increase. For example, if the number of created points is reduced from nine to eight, r will increase from $d_{ij}/10$ to $d_{ij}/6.6667$. After reducing the number of created points, redetermine n_{max} and n_{min}. According to the conditions in Case 1 and 2, perform the merging of initial centroids. However, if both n_{max} and n_{min} are still equal to 0, the number of created points will be further reduced by 1 until the number of created points is equal to six. In the case that the number of created points is equal to six, and both n_{max} and n_{min} are still equal to 0, the pair with the lowest d_{ij} will be the first to be merged and the pair with the highest d_{ij} will be the last.

4. The merging is repeated until the number of centroids is equal to K.

4 EXPERIMENTAL RESULTS

To evaluate the proposed algorithm, the performance of the proposed algorithm is compared with that of the original K-means algorithm and the initialization method proposed

The Nineteenth International Symposium on Artificial Life and Robotics 2014 (AROB 19th 2014),
B-Con Plaza, Beppu, Japan, January 22-24, 2014

by Erisoglu, Calis and Sakallioglu [8]. The experiments have been conducted using 11 benchmark datasets. Nine out of eleven datasets are obtained from UCI machine learning repository [10]. The other two, Yeast 1 and Yeast 2, are retrieved from Yeung, Haynor, and Ruzzo [11]. The details of the datasets are given in Table 1. Since the performance of K-means algorithm is sensitive to the initial selection of centroids, for each datasets, ten experimental repetitions are performed with different initial centroids each time.

Table 1. Details of the benchmark datasets

Datasets	The number of instances	The number of attributes	K
Glass	214	9	6
Iris	150	4	3
Wine	178	13	3
Yeast 1	384	17	5
Yeast 2	237	17	4
Abalone	4177	8	3
Soybean (Small)	47	35	4
Wall-Following Robot Navigation	5456	24	4
Landsat	6435	36	6
Ecoli	336	7	8
User Knowledge Modeling	403	5	4

In this study, the performance is measured in terms of the error rate. The error rate is a ratio of the number of misclassified data points to the total number of data points.

$$\text{Error} = \frac{\varepsilon}{N} \tag{4}$$

where ε is the number of misclassified data points.
N is the total number of data points.

Table 2 shows the experimental results of the proposed algorithm in comparison to the original K-means algorithm and the initialization method proposed by Erisoglu, Calis, and Sakallioglu [8]. The clustering results in Table 2 demonstrate that the proposed method achieves better accuracy (lower error rate) than the average accuracy of the original k-means in 10 out of 11 dataset. In comparison to the initialization method proposed by Erisoglu, Calis and Sakallioglu [8], the proposed method obtains better accuracy in 4 datasets, obtains the same accuracy in 5 datasets, and obtains lower accuracy in 2 datasets.

Table 2. The experimental results

Datasets	Error rate (%)			
	Original K-means		Proposed method	Erisoglu et al. [8]
Glass	best	41.1215	45.7944	46.2617
	average	46.1215		
	worst	51.4019		
Iris	best	10.6667	10.6667	10.6667
	average	19.7333		
	worst	33.3333		
Wine	best	29.7753	32.5843	29.7753
	average	30.0562		
	worst	32.5843		
Yeast 1	best	27.0833	33.0729	26.0417
	average	33.2552		
	worst	39.3229		
Yeast 2	best	25.7384	25.7384	25.7384
	average	26.7089		
	worst	30.8017		
Abalone	best	48.5037	48.5037	48.5037
	average	49.2531		
	worst	51.0175		
Soybean (Small)	best	0	21.2766	21.2766
	average	24.2553		
	worst	42.5532		
Wall-Following Robot Navigation	best	49.7617	50.4949	50.4949
	average	50.7331		
	worst	51.1364		
Landsat	best	26.216	38.073	46.1228
	average	38.5734		
	worst	46.589		
Ecoli	best	15.7738	19.0476	20.8333
	average	19.3155		
	worst	24.7024		
User Knowledge Modeling	best	36.7246	37.2208	38.7097
	average	41.34		
	worst	50.6204		

5 CONCLUSIONS

K-means algorithm is one of the most well-known clustering algorithms. However, its performance depends heavily on the initial centroids which are selected randomly. Therefore, this paper proposes a new initialization method

The Nineteenth International Symposium on Artificial Life and Robotics 2014 (AROB 19th 2014),
B-Con Plaza, Beppu, Japan, January 22-24, 2014

for K-means algorithm with the aim to improve the accuracy of the clustering results. The main idea of the proposed algorithm is to choose the initial centroids that are far away from one another. First, the process of creating the initial centroids is performed. Then, a pair of initial centroids is merged together if there exist many data points uniformly located along the area between the two centroids. The experimental results show that the proposed method produces more accurate clusters than the original K-means algorithm for most of the datasets.

REFERENCES

[1] Han J and Kamber M (2006), Data mining: Concepts and techniques. Morgan Kaufmann, San Francisco

[2] Tan PN, Steinbach M, and Kumar V (2006), Introduction to data mining. Addison-Wesley, Boston

[3] Redmond SJ and Heneghan C (2007), A method for initialising the K-means clustering algorithm using kd-trees, Pattern Recognition Letters, 28, pp. 965-973

[4] Li Y and Wu H (2012), A clustering method based on K-means algorithm, Physics Procedia, 25, pp. 1104-1109

[5] Tajunisha N and Saravanan V (2011), An efficient method to improve the clustering performance for high dimensional data by Principal Component Analysis and modified K-means, International Journal of Database Management Systems (IJDMS), 3(1), pp. 196-205

[6] Pavan KK, Rao AA, Rao AVD, and Sridhar GR (2011), Robust seed selection algorithm for K-means type algorithms, International Journal of Computer Science & Information Technology (IJCSIT), 3(5), pp. 147-163

[7] Arthur D and Vassilvitskii S (2007), K-means++: The advantages of careful seeding, Proceedings of the eighteenth annual ACM-SIAM symposium of Discrete algorithm, pp. 1027-1035

[8] Erisoglu M, Calis N, and Sakallioglu S (2011), A new algorithm for initial cluster centers in K-means algorithm, Pattern Recognition Letters, 32, pp. 1701-1705

[9] Nazeer KAA, Kumar SDM, and Sebastian MP (2011), Enhancing the K-means clustering algorithm by using a O(n logn) heuristic method for finding better initial centroids, Proceedings of the Second International Conference on Emerging Applications of Information Technology, pp. 261-254

[10] Bache K and Lichman M (2013), UCI machine learning repository [http://archive.ics.uci.edu/ml], University of California, School of Information and Computer Science

[11] Yeung KY, Haynor DR, and Ruzzo WL (2006), Validating Clustering for Gene Expression Data [http://www.cs.washington.edu/homes/kayee/cluster]

ประวัติผู้เขียน

ชื่อ-นามสกุล	นางสาวพัชนี ฉัตรธเนศ
วัน เดือน ปีเกิด	8 กรกฎาคม 2531
ที่อยู่	41 หมู่บ้านนครินทร์การ์เด้น ถ.ร่มเกล้า แขวงคลองสามประเวศ เขตลาดกระบัง กรุงเทพมหานคร 10520
ประวัติการศึกษา	2553 วิศวกรรมศาสตรบัณฑิต สาขาวิชาวิศวกรรมสารสนเทศ (เกียรตินิยมอันดับ 1) สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ประสบการณ์การทำงาน	-