ขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานที่มีประสิทธิภาพบนระบบคลัสเตอร์

AN EFFICIENT PARALLEL K-MEANS CLUSTERING ALGORITHM ON A CLUSTER SYSTEM

นเรศ ผ่องสวัสดิ์กุล

NARED PONGSAWATKUL

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต

สาขาวิชาวิทยาการคอมพิวเตอร์

คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าธนบุรี มหาวิทยาลัยแห่งชาติ

พ.ศ. 2553

KMITL-2010-SC-M-002-024

ขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานที่มีประสิทธิภาพบนระบบคลัสเตอร์

AN EFFICIENT PARALLEL K-MEANS CLUSTERING ALGORITHM ON A CLUSTER SYSTEM

นเรศ ผ่องสวัสดิ์กุล

NARED PONGSAWATKUL

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาวิทยาการคอมพิวเตอร์

คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2553

KMITL-2010-SC-M-002-024

AN EFFICIENT PARALLEL K-MEANS CLUSTERING ALGORITHM ON A CLUSTER SYSTEM

NARED PONGSAWATKUL

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF SCIENCE IN COMPUTER SCIENCE

FACULTY OF SCIENCE

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2010

KMITL-2010-SC-M-002-024

คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ใบรับรองวิทยานิพนธ์

หัวข้อวิทยานิพนธ์ ขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานที่มีประสิทธิภาพบนระบบคลัสเตอร์

An Efficient Parallel K-means Clustering Algorithm on a Cluster Systems

นักศึกษา นายนเรศ ผ่องสวัสดิ์กุล

รหัสประจำตัว 50067505

ปริญญา วิทยาศาสตรมหาบัณฑิต

สาขาวิชา วิทยาการคอมพิวเตอร์

อาจารย์ที่ปรึกษาวิทยานิพนธ์ ผศ.ดร.จีรพร วีระพันธุ์

คณะกรรมการสอบวิทยานิพนธ์		ลายมือชื่อ
ผศ.ดร.จีรพร	วีระพันธุ์	
ผศ.ดร.นวลสวาท	หิรัญสกลวงศ์	
ผศ.ดร.ศรัณย์	อินทโกสุม	
ดร.เฉลิมศักดิ์	เลิศวงศ์เสถียร	

วัน / เดือน/ ปี ที่สอบ 10 พฤษภาคม พ.ศ. 2553 เวลา 13.00 น.

สถานที่สอบ ณ อาคารจุฬาภรณ 1 คณะวิทยาศาสตร์ ห้อง 219

คณะวิทยาศาสตร์รับรองแล้ว

(รองศาสตราจารย์ ดร.ดุษณี ธนะบริพัฒน์)

คณบดีคณะวิทยาศาสตร์

วันที่ 24 **เดือน** พ.ค. **พ.ศ.** 53

สำนักทะเบียนและประมวลผล สจล.
วันที่ส่งเล่มวิทยานิพนธ์ฉบับสมบูรณ์
วันที่ 26 เดือน พ.ค. พ.ศ. 53
ลงชื่อ.....

หัวข้อวิทยานิพนธ์	ขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานที่มีประสิทธิภาพบนระบบคลัสเตอร์
นักศึกษา	นายนเรศ ผ่องสวัสดิ์กุล
รหัสประจำตัว	50067505
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	วิทยาการคอมพิวเตอร์
พ.ศ.	2553
อาจารย์ที่ปรึกษา	ผศ.ดร.จีรพร วีระพันธุ์

บทคัดย่อ

เคมีนคลัสเตอร์ริงเป็นขั้นตอนวิธีสำหรับจัดกลุ่มวัตถุต่างๆ ตามจำนวนกลุ่มที่ต้องการ โดยพิจารณาจากคุณลักษณะของวัตถุเหล่านั้น และถูกนำไปพัฒนาแอพพลิเคชันจำนวนมากในหลากหลายสาขา แต่การนำไปประยุกต์ใช้ต่างก็ประสบปัญหาเดียวกันคือ เวลาและหน่วยความจำที่ต้องใช้ในการประมวลผลมากเกินไปเนื่องจากข้อมูลมีขนาดใหญ่ งานวิจัยนี้เรานำเสนอวิธีการเพิ่มประสิทธิภาพของขั้นตอนวิธีเคมีนคลัสเตอร์ริงโดยนำเสนอขั้นตอนวิธีสำหรับเลือกชุดของจุดศูนย์กลางเริ่มต้นแบบใหม่ และนำวิธีการประมวลผลแบบขนานบนระบบคลัสเตอร์มาประยุกต์ใช้ ซึ่งผลการทดลองแสดงให้เห็นว่าเวลาที่ต้องใช้สำหรับการประมวลผลลดลง และสามารถรองรับปัญหาขนาดใหญ่ได้มากกว่าเดิม

คำสำคัญ : ขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนาน

Thesis Title	An Efficient Parallel K-means Clustering Algorithm on a Cluster System
Student	Mr. Nared Pongsawatkul
Student ID	50067505
Degree	Master of Science
Program	Computer Science
Year	2010
Thesis Advisor	Asst. Prof. Dr. Jeeraporn Werapun

ABSTRACT

K-means clustering algorithm is applied to classify or to group objects into K groups by considering their attributes or features and can be used for developing applications in various fields. However, the problem of those applications are time consuming and out of memory because the data are too large. In this research we present the efficient K-means clustering algorithm by introducing the new method for selecting initial centers and applying a parallel method to solve the problem. To evaluate the performance of the proposed algorithm, a number of experiments were performed on a cluster system. The investigated results showed that the response time was improved and the system can support more data.

Keywords : Parallel K-means Clustering Algorithm

กิตติกรรมประกาศ

วิทยานิพนธ์นี้มิอาจจะสำเร็จลุล่วงไปได้ด้วยดี หากมิได้รับคำแนะนำ คำชี้แนะ ความรู้ และความเอาใจใส่จาก ผศ.ดร.จีรพร วีระพันธุ์ ผู้เป็นอาจารย์ที่ปรึกษา ซึ่งท่านได้สละเวลาให้กับข้าพเจ้าอย่างเต็มที่ จึงใคร่ขอขอบพระคุณเป็นอย่างสูง

ขอขอบพระคุณ ผศ.ดร.นวลสวาท หิรัญสกลวงศ์ ผศ.ดร.ศรัณย์ อินทโกสุม และดร.เฉลิม ศักดิ์ เลิศวงศ์เสถียร คณะกรรมการสอบหัวข้อ และโครงร่างวิทยานิพนธ์ ที่กรุณาให้คำแนะนำตลอดจนข้อชี้แนะ จนในที่สุดทำให้วิทยานิพนธ์ฉบับนี้สำเร็จลงได้

ขอขอบพระคุณบิดา มารดา และครอบครัว ที่สนับสนุนให้ได้เรียนในระดับที่ได้ตั้งใจ อีกทั้งยังได้ดูแลเรื่องค่าใช้จ่ายต่างๆ ระหว่างศึกษาเป็นอย่างดีอีกด้วย

ขอขอบคุณ นางสาวสาวิณี กิจพ่อค้า และครอบครัว ที่คอยเป็นกำลังใจ คอยกระตุ้นให้ทำงานและมอบสิ่งที่ดีให้กันตลอดมา

สำหรับคุณงามความดี และประโยชน์อันใดที่เกิดขึ้นจากวิทยานิพนธ์ฉบับนี้ ข้าพเจ้าขอมอบให้กับบิดา มารดา อาจารย์ทุกท่านซึ่งเป็นที่เคารพรักยิ่ง ตลอดจนญาติพี่น้อง และเพื่อนๆทุกคน

นเรศ ผ่องสวัสดิ์กุล

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

สารบัญรูป

สารบัญรูป (ต่อ)

รูปที่ หน้า

สารบัญรูป (ต่อ)

รูปที่ หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ขั้นตอนวิธีเคมีนคลัสเตอร์ริง (K-means Clustering Algorithm) เป็นวิธีการจัดกลุ่มข้อมูลรูปแบบหนึ่ง ซึ่งนิยมนำไปใช้ในการทำเหมืองข้อมูล (Data mining) การเรียนรู้ของเครื่อง (Machine learning) แบบไม่มีผู้สอน งานวิจัย ฯลฯ โดยจะแบ่งข้อมูล (เป็นเวกเตอร์) ออกเป็นกลุ่มตามจำนวนที่ต้องการ (K) ซึ่งนำข้อมูลที่มีคุณลักษณะเหมือนกัน หรือคล้ายกันจัดไว้ในกลุ่มเดียวกัน ขั้นตอนวิธีที่ใช้ในการแบ่งกลุ่มจะอาศัยความเหมือน (Similarity) หรือความใกล้ชิด (Proximity) โดยคำนวณจากการวัดระยะระหว่างข้อมูลกับจุดศูนย์กลางของกลุ่ม (Mean/Centroid) โดยใช้การวัดระยะแบบต่างๆ เช่น การวัดระยะแบบยูคลิด (Euclidean distance) การวัดระยะแบบแมนฮัตตัน (Manhattan distance) และการวัดระยะแบบเชบิเชฟ (Chebychev distance) เป็นต้น

แนวคิดของขั้นตอนวิธีนี้ริเริ่มโดย Hugo Steinhaus ในปี 1956 ซึ่งขั้นตอนวิธีแบบมาตรฐานถูกนำเสนอโดย Stuart Lloyd ในปี 1957 และพบว่าขั้นตอนวิธีนี้ถูกนำไปใช้ในงานวิจัยเป็นครั้งแรกโดย Jame MacQueen ในปี 1967 จนถึงปัจจุบันขั้นตอนวิธีนี้ก็ยังเป็นที่นิยมเนื่องจากเป็นขั้นตอนวิธีที่ง่าย และให้ผลลัพธ์ที่ยอมรับได้ ถึงแม้ว่าขั้นตอนวิธีนี้จะมีหลักการทำงานไม่ซับซ้อนด้วยเวลาเท่ากับ O(RKN) เมื่อใช้วิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบสุ่ม โดย R เป็นจำนวนรอบของการทำงานซ้ำ ส่วน K เป็นจำนวนกลุ่มที่ต้องการแบ่งข้อมูล และ N เป็นจำนวนของข้อมูลทั้งหมด แต่เมื่อใช้ในการแก้ปัญหาเวกเตอร์ขนาดใหญ่ก็อาจใช้เวลา และหน่วยความจำในการทำงานมากเกินไป เหล่านี้เป็นสาเหตุสำคัญที่ทำให้ขั้นตอนวิธีนี้ได้รับการพัฒนาประสิทธิภาพให้ดียิ่งขึ้นอย่างต่อเนื่อง ทั้งทางด้านความเร็ว ความแม่นยำและวิธีจัดการกับปัญหาที่มีขนาดใหญ่ เพื่อตอบสนองกับภาระงาน ซึ่งมีแนวโน้มที่เพิ่มขึ้นในปัจจุบันและอนาคต

ในอดีตที่ผ่านมางานวิจัยเกี่ยวกับการเพิ่มประสิทธิภาพขั้นตอนวิธีนี้แบ่งออกเป็น 2 แนวทาง คือ 1) วิธีเลือกชุดจุดศูนย์กลางเริ่มต้น [1,5,10] ซึ่งเป็นส่วนสำคัญที่ทำให้การแบ่งข้อมูลมีความรวดเร็ว และแม่นยำยิ่งขึ้น แต่วิธีเลือกเหล่านี้ไม่ได้ช่วยแก้ปัญหาหน่วยความจำที่ไม่เพียงพอสำหรับการประมวลผล และวิธีเลือกที่มีอยู่ก็ยังไม่ได้ให้ผลลัพธ์ที่ดีกับทุกชุดข้อมูล และจำนวนกลุ่มที่แตกต่างกัน 2) ขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบเครือข่าย เช่น ระบบเครือข่ายแบบคลัสเตอร์ของเวิร์คสเตชั่น (Cluster of Workstation) [3,4,6,8] เพื่อการแบ่งข้อมูลที่รวดเร็ว และสามารถรองรับปัญหาขนาดใหญ่ขึ้นได้ เนื่องจากใช้คอมพิวเตอร์หลายเครื่องในการทำงานร่วมกัน ซึ่งเครื่องคอมพิวเตอร์เหล่านั้นมีทรัพยากรเป็นของตนเอง เช่น หน่วยประมวลผล หน่วยความจำ

เป็นต้น จึงง่ายต่อการขยายเครือข่าย เพื่อเพิ่มประสิทธิภาพให้กับการแบ่งข้อมูล อย่างไรก็ตาม ขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานส่วนใหญ่ยังไม่ได้นำวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบต่างๆ มาประยุกต์ใช้ เนื่องจากวิธีการเลือกที่มีอยู่นั้นยังใช้เวลามากในการทำงาน และอาจให้ชุดจุดศูนย์กลางเริ่มต้นที่ไม่เหมาะสม ดังนั้นวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบสุ่ม จึงเป็นที่นิยมมากที่สุดในการนำมาใช้ร่วมกับขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนเครือข่ายในขณะนี้

งานวิจัยนี้ได้นำเสนอขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์ โดยนำวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่มาประยุกต์ใช้กับขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานที่มีอยู่ เพื่อเพิ่มประสิทธิภาพด้านความเร็ว (Response Time) ความแม่นยำ (Accuracy) ในการแบ่งข้อมูล และยังสามารถรองรับปัญหาที่มีขนาดใหญ่มากกว่าคอมพิวเตอร์เพียงเครื่องเดียวจะสามารถทำงานได้

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

วิทยานิพนธ์นี้มีวัตถุประสงค์เพื่อศึกษาและเพิ่มประสิทธิภาพให้กับขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบเครือข่าย ดังนี้

1) ออกแบบ และปรับวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่ ให้สามารถเลือกชุดจุดศูนย์กลางเริ่มต้นสำหรับการแบ่งข้อมูลได้อย่างเหมาะสม

2) ออกแบบขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์ โดยนำวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่มาประยุกต์ใช้

1.3 สมมติฐานของการศึกษา

ขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์สามารถแบ่งข้อมูลได้อย่างรวดเร็ว และแม่นยำ เมื่อนำวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่มาประยุกต์ใช้ และยังสามารถรองรับปัญหาที่มีขนาดใหญ่ได้

1.4 ขอบเขตการวิจัย

วิทยานิพนธ์นี้มีขอบเขตการวิจัย เพื่อศึกษาวิธีการเพิ่มประสิทธิภาพด้านความเร็ว และความแม่นยำให้กับขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์

1.5 ขั้นตอนการศึกษาและดำเนินงานวิจัย

วิทยานิพนธ์นี้มีขั้นตอนการศึกษาและการดำเนินงานวิจัย ดังนี้

1) ศึกษาการแบ่งข้อมูลด้วยขั้นตอนวิธีเคมีนคลัสเตอร์ริงทั้ง 2 แบบ คือ แบบอนุกรม (Sequential) และแบบขนาน (Parallel) รวมถึงวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบต่างๆ เช่น แบบสุ่ม (Random) แบบคัดสรร (Seeding) และแบบให้ค่าน้ำหนักดีกำลังสอง (D^2 Weighting)

2) ทำการตั้งสมมติฐานโดยคาดว่าขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์จะสามารถแบ่งข้อมูลได้อย่างรวดเร็ว และแม่นยำ เมื่อนำวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่มาประยุกต์ใช้ และยังสามารถรองรับปัญหาที่มีขนาดใหญ่ได้

3) ออกแบบวิธีเลือกชุดจุดศูนย์กลางเริ่มต้น และขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์ พร้อมกับวิเคราะห์ความซับซ้อนด้านเวลา และด้านหน่วยความจำ

4) ทดลอง วัดผลการทดลองของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้น และขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์ สรุปผลการทดลอง พร้อมเสนอแนวทางการพัฒนางานวิจัย

5) เขียนวิทยานิพนธ์

1.6 ประโยชน์ที่คาดว่าจะได้รับ

วิทยานิพนธ์นี้มีประโยชน์ที่คาดว่าจะได้รับ ดังนี้

1) ขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์สามารถแบ่งข้อมูลได้อย่างรวดเร็ว และแม่นยำ

2) ขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์สามารถรองรับปัญหาที่มีขนาดใหญ่มากกว่าคอมพิวเตอร์เพียงเครื่องเดียวจะสามารถทำงานได้

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

เนื้อหาในบทนี้จะกล่าวถึงทฤษฎีที่เกี่ยวข้องกับการแบ่งข้อมูลด้วยขั้นตอนวิธีเคมีนแบบอนุกรม (Sequential K-means Algorithm) และระบบคลัสเตอร์ (Cluster System) จากนั้นจะกล่าวถึงงานวิจัยที่เกี่ยวข้องกับวิธีเลือกชุดจุดศูนย์กลางเริ่มต้น และขั้นตอนวิธีเคมีนแบบขนานบนเครือข่าย (Parallel K-means Algorithm) ส่วนสุดท้ายคือ การวัดสมรรถนะ (Performance Measurement)

2.1 ขั้นตอนวิธีเคมีนแบบอนุกรม

2.1.1 ขั้นตอนวิธีเคมีนคลัสเตอร์ริง

ขั้นตอนวิธีเคมีนคลัสเตอร์ริง (K-means Clustering Algorithm) คือขั้นตอนวิธีสำหรับแบ่งข้อมูลออกเป็นกลุ่มตามจำนวนที่ต้องการ (K) ซึ่งพิจารณาจากคุณลักษณะของข้อมูลเหล่านั้น โดยนำข้อมูลที่มีคุณลักษณะ (Attributes) เหมือนกัน หรือคล้ายกันจัดไว้ในกลุ่มเดียวกัน ดังนั้นข้อมูลที่อยู่ต่างกลุ่มกันจึงมีคุณลักษณะแตกต่างกัน ซึ่งคุณลักษณะเหล่านั้นต้องเป็นข้อมูลเชิงปริมาณเท่านั้น ขั้นตอนวิธีนี้ต้องอาศัยความเหมือน (Similarity) หรือความใกล้ชิด (Proximity) ระหว่างข้อมูลกับจุดศูนย์กลางของกลุ่ม (Mean/ Centroid) เป็นเกณฑ์ในการแบ่งข้อมูล ซึ่งสามารถคำนวณได้โดยใช้การวัดระยะทางแบบต่างๆ (อธิบายในหัวข้อ 2.1.2) ขั้นตอนวิธีนี้มีหลักการทำงานดังนี้

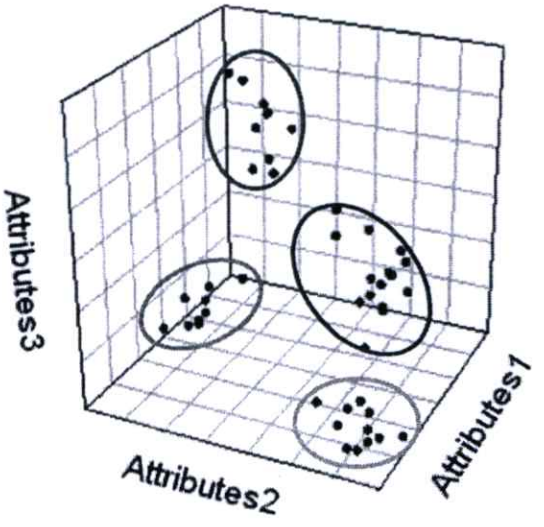
K-means Clustering Algorithm: O(RKN) // if take any random objects as initial means.

- Calculate initial means.
- Assign objects into clusters by using initial means.
- Do while objects move to another clusters.
  - Recalculate means of clusters by using objects belonging to them.
  - Assign objects into clusters by using calculated means.
- End of while.

รูปที่ 2.1 ขั้นตอนวิธีเคมีนแบบอนุกรม

การแบ่งข้อมูลด้วยขั้นตอนวิธีนี้ต้องระบุจำนวนกลุ่มที่ต้องการก่อนทุกครั้ง ซึ่งใช้ตัวอักษร K เขียนแทนจำนวนกลุ่มที่ต้องการ โดยจำนวนกลุ่มที่ต้องการเป็นจำนวนเต็มบวก และมีค่ามากกว่า 1 เสมอ ส่วนผลลัพธ์ที่ได้จากการแบ่งข้อมูลคือ กลุ่มข้อมูลจำนวน K กลุ่ม (แสดงดังรูปที่ 2.2)

ขั้นตอนวิธีนี้สามารถแบ่งการทำงานออกเป็น 2 ส่วนหลัก คือ 1) การเลือกชุดจุดศูนย์กลางเริ่มต้น ซึ่งเป็นตัวแปรหนึ่งที่เป็นตัวกำหนดความเร็วในการแบ่งข้อมูล และความแม่นยำของผลลัพธ์ ชุดจุดศูนย์กลางเริ่มต้นสามารถคำนวณได้จากวิธีการเลือกชุดจุดศูนย์กลางเริ่มต้นแบบต่างๆ เช่น วิธีการสุ่ม (Random) เป็นต้น (อธิบายในหัวข้อ 2.5) ซึ่งชุดจุดศูนย์กลางเหล่านั้นประกอบด้วยจุดศูนย์กลางทั้งหมด K จุด โดยจุดศูนย์กลางแต่ละจุดเป็นตัวแทนของกลุ่มต่างๆ 2) การแบ่งข้อมูล ซึ่งต้องอาศัยกระบวนการทำงานซ้ำเพื่อค้นหาผลลัพธ์ตามเงื่อนไขที่กำหนด เช่น ข้อมูลไม่มีการย้ายกลุ่ม เป็นต้น โดยมีความซับซ้อนด้านเวลาเท่ากับ O(RKN) เมื่อ R คือจำนวนรอบของการทำงานซ้ำ และ N คือจำนวนข้อมูลทั้งหมด กระบวนการย่อยภายในส่วนนี้ คือการคำนวณชุดจุดศูนย์กลาง ซึ่งสามารถคำนวณได้จากสมการมัชฌิมเลขคณิต หรือค่าเฉลี่ยของสมาชิกในกลุ่ม (อธิบายในหัวข้อถัดไป) และการกำหนดข้อมูลเข้าสู่กลุ่ม หรือการจัดกลุ่มข้อมูลโดยอาศัยชุดจุดศูนย์กลางที่ถูกคำนวณขึ้นมาใหม่ จึงต้องคำนวณระยะทางระหว่างข้อมูลกับจุดศูนย์กลางของแต่ละกลุ่ม เพื่อค้นหาระยะทางที่สั้นที่สุดระหว่างข้อมูลกับจุดศูนย์กลาง ซึ่งระยะทางที่สั้นที่สุดบอกให้รู้ว่าข้อมูลเหล่านั้นต้องเป็นสมาชิกของกลุ่มใด เนื่องจากค่าของระยะทางแสดงถึงความเหมือนกันระหว่างข้อมูลกับจุดศูนย์กลางของกลุ่ม โดยกระบวนการทำซ้ำที่ใช้ในการแบ่งข้อมูลนั้นไม่สามารถคาดเดาได้เลยว่าการทำงานจะสิ้นสุดภายในกี่รอบ ซึ่งการทำงานแต่ละรอบต้องคำนวณชุดจุดศูนย์กลาง และกำหนดข้อมูลเข้าสู่กลุ่มทุกครั้งจนกว่าการแบ่งข้อมูลจะเสร็จสิ้นตามเงื่อนไขที่กำหนด



รูปที่ 2.2 ผลลัพธ์ของขั้นตอนวิธีเคมีน

จากรูปที่ 2.2 แสดงตัวอย่างผลลัพธ์ที่ได้จากการแบ่งข้อมูล โดยการแบ่งข้อมูลครั้งนี้มีจำนวนกลุ่มที่ต้องการเท่ากับ 4 กลุ่ม และพิจารณาคุณลักษณะทั้งหมด 3 คุณลักษณะ ซึ่งใช้รูปทรงวงรีแสดงขอบเขตสมาชิกของแต่ละกลุ่ม

2.1.2 ชุดจุดศูนย์กลาง

ชุดจุดศูนย์กลาง (Means/ Centroids) คือตัวแทนของกลุ่มต่างๆ เปรียบเสมือนหลักยึดที่ใช้จัดข้อมูลทั้งหมดเข้าสู่กลุ่มที่มีอยู่ ซึ่งข้อมูลทั้งหมดต้องมีกลุ่มอยู่ และเป็นสมาชิกของกลุ่มใดกลุ่มหนึ่งเท่านั้น ในที่นี้จะใช้ตัวอักษร C แทนชุดจุดศูนย์กลาง โดยชุดจุดศูนย์กลางประกอบด้วยจุดศูนย์กลางจำนวน K จุด ตามจำนวนกลุ่มที่ต้องการ ดังนั้น $C = \{c_1, c_2, \ldots, c_K\}$ ซึ่งนำมาใช้ในขั้นตอนวิธีเคมีนตลอดการทำงาน และสามารถคำนวณได้จากสมการมัชฌิมเลขคณิตดังนี้

นิยาม 2.1

มัชฌิมเลขคณิต คือผลบวกของสมาชิกทุกจำนวน หารด้วยจำนวนสมาชิกที่อยู่ภายในกลุ่มนั้น โดยทั่วไปมักเรียกกันว่า ค่าเฉลี่ย หรือมัชฌิม ซึ่งสามารถเขียนแทนด้วย $\bar{x}$ อ่านว่า *เอกซ์บาร์*

กำหนดชุดข้อมูล $x = \{x_1, x_2, \ldots, x_n\}$ ซึ่ง n เป็นจำนวนของข้อมูลทั้งหมด มัชฌิมเลขคณิตของชุดข้อมูลนี้สามารถคำนวณได้ดังสมการที่ 2.1

$$\bar{x} = \frac{1}{n}(x_1 + \ldots + x_n) = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \qquad (2.1)$$

2.1.3 การวัดระยะทางระหว่างข้อมูลกับจุดศูนย์กลางของกลุ่ม

การวัดระยะทางระหว่างข้อมูลกับจุดศูนย์กลางของกลุ่มสามารถทำได้หลายวิธี ตัวอย่างเช่น การวัดระยะแบบยูคลิด (Euclidean distance) การวัดระยะแบบเชบิเชฟ (Chebychev distance) และการวัดระยะแบบแมนฮัตตัน (Manhattan distance) เป็นต้น ซึ่งการวัดระยะทางดังกล่าวเป็นการวัดข้อมูลเชิงปริมาณ และสามารถนำมาใช้ในการแบ่งข้อมูลด้วยขั้นตอนวิธีเคมีนได้ทั้งสิ้น โดยทั่วไปนิยมใช้การวัดระยะทางแบบยูคลิดเนื่องจากการวัดระยะทางแบบนี้สามารถเข้าใจ และใช้งานได้ง่าย ซึ่งถูกนำมาใช้ในงานวิจัยนี้ด้วย

นิยาม 2.2

ระยะทางแบบยูคลิด (Euclidean distance) คือระยะทางระหว่างข้อมูล หรือจุดสองจุดในแนวเส้นตรง แนวคิดมาจากทฤษฎีของพีทาโกรัส ด้วยหลักการทำงานของขั้นตอนวิธีเคมีนระยะทาง คือค่าของความเหมือนระหว่างข้อมูลที่ต้องการเปรียบเทียบ ซึ่งการแบ่งกลุ่มด้วยวิธีนี้ต้องใช้ระยะทางที่สั้นที่สุดระหว่างข้อมูลกับจุดศูนย์กลางเป็นเกณฑ์ในการจัดข้อมูลเข้ากลุ่มต่างๆ

กำหนดข้อมูล x และ y ซึ่ง x = $(x_1, x_2, \ldots, x_n)$ และ y = $(y_1, y_2, \ldots, y_n)$ เป็นจุดสองจุดบนปริภูมิยูคลิด n มิติ ส่วน n คือจำนวนคุณลักษณะ หรือจำนวนมิติของข้อมูล ดังนั้นระยะทางระหว่างจุด x กับ y สามารถคำนวณได้ดังสมการที่ 2.2

$$D(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (2.2)$$

นิยาม 2.3

ค่านอร์มแบบยูคลิด (Euclidean norm) คือระยะทางจากจุดหนึ่งจุดไปยังจุดกำเนิดบนปริภูมิยูคลิด โดยจุดกำเนิด คือจุดที่เส้นแกนของระบบตัดกัน เขียนแทนด้วยตัวอักษร O ซึ่งจุดกำเนิดต้องมีค่าเป็นศูนย์เสมอตามจำนวนมิติของข้อมูล ดังนั้น O = (0, 0, …, 0) ซึ่งสามารถนำมาใช้บอกตำแหน่งของจุดต่างๆ ได้โดยใช้ระยะอ้างอิงจากจุดกำเนิด และเป็นตัวแบ่งเส้นแกนออกเป็น 2 ส่วน คือด้านบวกและด้านลบ ซึ่งค่านอร์มแบบยูคลิดถูกนำมาใช้ในงานวิจัยนี้ด้วย

กำหนดข้อมูล x ซึ่ง x = $(x_1, x_2, \ldots, x_n)$ เป็นจุดหนึ่งบนปริภูมิยูคลิด n มิติ ส่วน n คือจำนวนคุณลักษณะของข้อมูล ดังนั้นค่านอร์มแบบยูคลิดของจุด x สามารถคำนวณได้ดังสมการที่ 2.3

$$\| x \| = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2} \qquad (2.3)$$

2.1.4 การวัดความแม่นยำของการจัดกลุ่มข้อมูล

ความแม่นยำของการจัดกลุ่มข้อมูล (Accuracy) คือ ผลลัพธ์ของการจัดกลุ่มที่มีค่าใกล้เคียงกับค่าจริง ซึ่งสามารถวัดได้จากค่าความคลาดเคลื่อน (Error function: E) โดยค่าความคลาดเคลื่อน คือ ผลรวมของผลต่างระหว่างจุดศูนย์กลางกับข้อมูลในกลุ่มยกกำลังสอง และถ้าความคลาดเคลื่อนมีค่าน้อยจะแสดงถึงความแม่นยำของการจัดกลุ่มข้อมูลครั้งนั้นมีค่ามาก ดังนั้นการจัดกลุ่มข้อมูลแต่ละครั้งต้องพิจารณาค่าความคลาดเคลื่อนให้น้อยที่สุดเสมอ เพื่อให้ผลลัพธ์ที่ได้มีความแม่นยำสูงที่สุด สามารถคำนวณได้ดังสมการที่ 2.4

$$E = \sum_{i=1}^{K}\sum_{x \in c_i} \| x - c_i \|^2 \qquad (2.4)$$

เมื่อ C คือ ชุดจุดศูนย์กลาง ซึ่ง C = $\{c_1, c_2, \ldots, c_K\}$

x คือ ข้อมูลใดๆ ที่เป็นสมาชิกของกลุ่ม c_i ซึ่ง $1 \le i \le K$

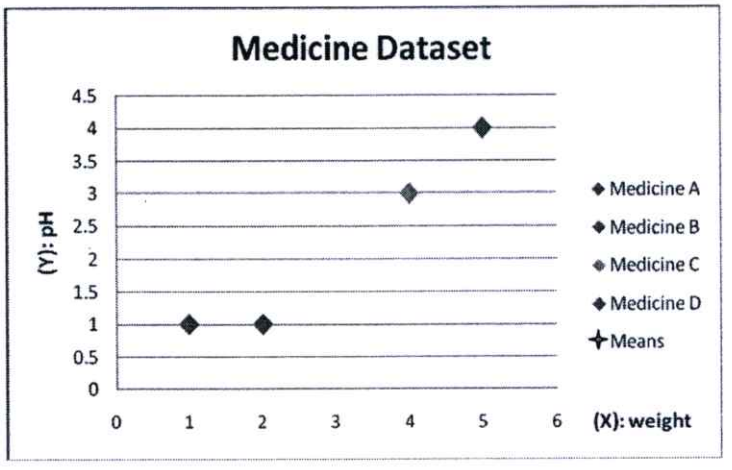
2.1.5 ตัวอย่างการแบ่งข้อมูลด้วยขั้นตอนวิธีเคมีน

ในหัวข้อนี้เป็นการอธิบายการทำงานของขั้นตอนวิธีเคมีนกับชุดข้อมูลตัวอย่างขนาดเล็กคือ ข้อมูลยาประเภทต่างๆ มีทั้งหมด 4 ประเภท ซึ่งยาแต่ละประเภทมี 2 คุณลักษณะ คือน้ำหนัก และค่า pH ที่แสดงความเป็นกรดเป็นเบสของสารเคมี เป้าหมาย คือการแบ่งข้อมูลทั้งหมดออกเป็น 2 กลุ่ม โดยพิจารณาน้ำหนักและค่า pH ซึ่งแสดงรายละเอียดได้ดังนี้

ตารางที่ 2.1 ข้อมูลยาประเภทต่างๆ เมื่อจำนวนกลุ่มที่ต้องการเท่ากับ 2 กลุ่ม

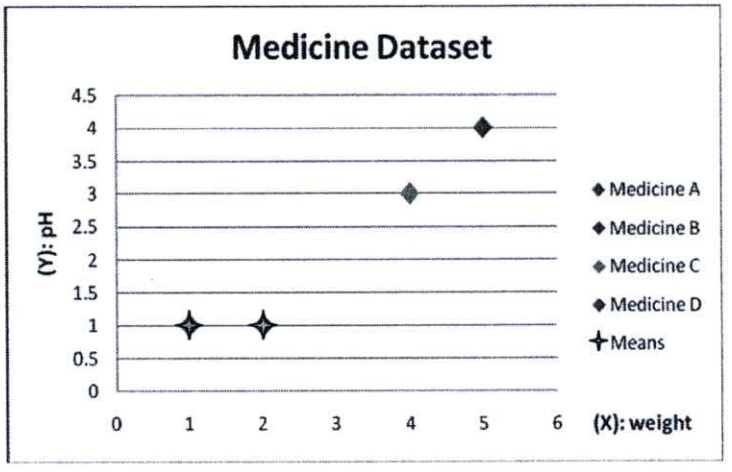
Objects	Attribute1: weight	Attribute2: pH	Clusters: 2
Medicine A	1	1	-
Medicine B	2	1	-
Medicine C	4	3	-
Medicine D	5	4	-

ขั้นตอนที่ 1 คือ การอ่านข้อมูลจากชุดข้อมูล คือยาประเภทต่างๆ ส่วนจุดศูนย์กลางทั้งสองจุดยังคงว่างเปล่า ซึ่งจะทำการเลือกในขั้นตอนถัดไป และแต่ละกลุ่มก็ยังไม่มีสมาชิกเป็นของตนเอง



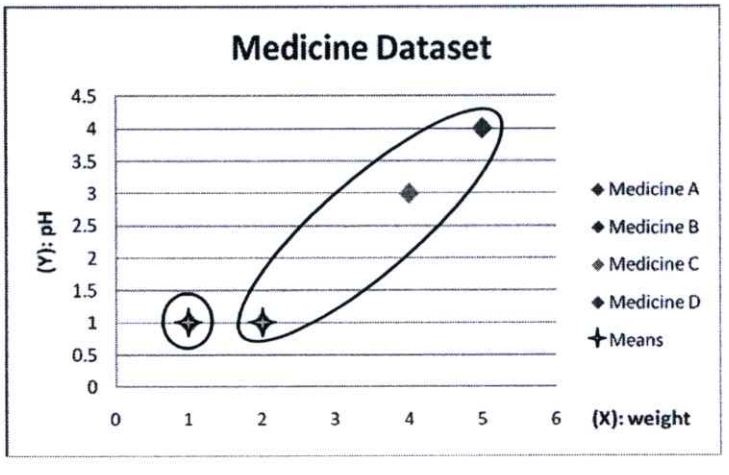
รูปที่ 2.3 ขั้นตอนแรกของการแบ่งข้อมูล

ขั้นตอนที่ 2 คือ การกำหนดชุดจุดศูนย์กลางเริ่มต้น สมมุติว่าขั้นตอนนี้ใช้วิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบสุ่ม (Random Method) โดยให้ยาประเภท A และ B เป็นชุดจุดศูนย์กลางเริ่มต้นเพื่อใช้ในการแบ่งข้อมูลในขั้นตอนถัดไป ดังนั้นชุดจุดศูนย์กลางเริ่มต้น C = {(1, 1), (1, 2)}



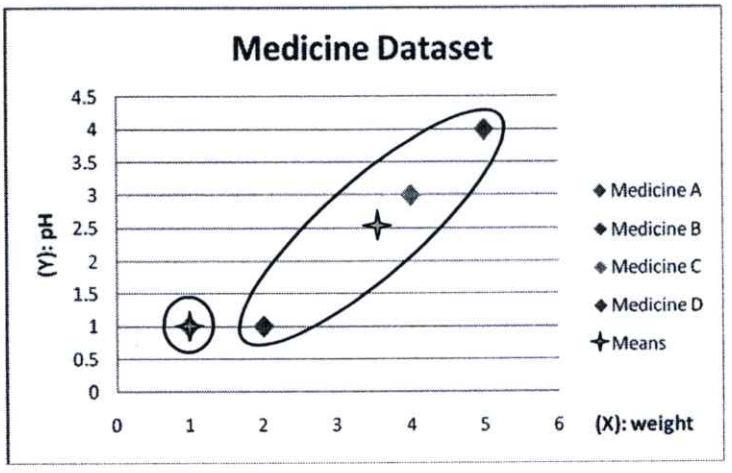
รูปที่ 2.4 ขั้นตอนที่สองของการแบ่งข้อมูล

ขั้นตอนที่ 3 คือ การจัดกลุ่มข้อมูล โดยพิจารณาจากระยะทางที่สั้นที่สุดระหว่างข้อมูลกับจุดศูนย์กลาง คือข้อมูลไหนอยู่ใกล้กับจุดศูนย์กลางใดก็จัดให้อยู่กลุ่มนั้น ซึ่งข้อมูลทั้งหมดจะต้องมีกลุ่มอยู่ และเป็นสมาชิกของกลุ่มใดกลุ่มหนึ่งเท่านั้น จากรูปที่ 2.5 ยาประเภท A ถูกจัดอยู่ในกลุ่มที่ 1 ส่วนยาประเภทอื่นถูกจัดไว้ในกลุ่มที่ 2 ซึ่งให้รูปทรงวงรีแสดงขอบเขตสมาชิกของแต่ละกลุ่ม



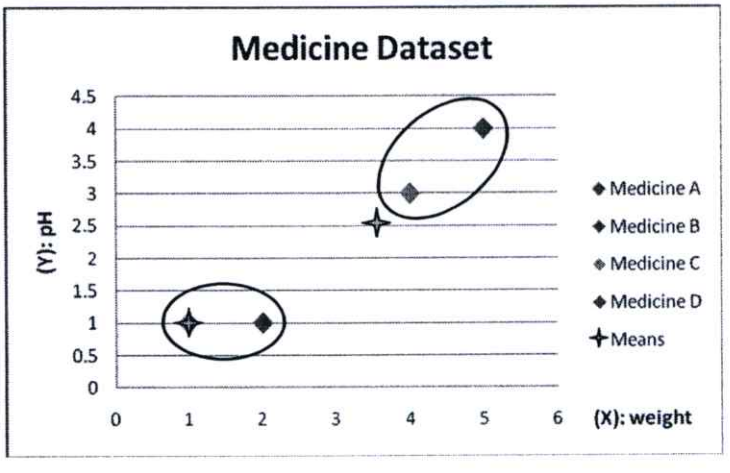
รูปที่ 2.5 ขั้นตอนที่สามของการแบ่งข้อมูล

ขั้นตอนที่ 4 คือ การกำหนดชุดจุดศูนย์กลาง โดยคำนวณจากค่าเฉลี่ยของสมาชิกภายในกลุ่ม ซึ่งกลุ่มที่ 1 มีสมาชิกเพียงตัวเดียวเท่านั้นคือ ยาประเภท A ส่วนกลุ่มที่ 2 มีสมาชิกทั้งหมด 3 ตัวคือ ยาประเภท B, C และ D ค่าเฉลี่ยของกลุ่มที่ 2 สามารถคำนวณได้ดังนี้ c_2 = {(2+4+5)/3, (1+3+4)/3} ดังนั้นจุดศูนย์กลาง C = {(1, 1), (3.67, 2.67)}



รูปที่ 2.6 ขั้นตอนที่สี่ของการแบ่งข้อมูล

ขั้นตอนที่ 5 คือ การจัดกลุ่มข้อมูลอีกครั้ง ด้วยชุดจุดศูนย์กลางชุดใหม่คือ (1, 1) และ (3.67, 2.67) ตามลำดับ จากรูปที่ 2.7 ยาประเภท A และ B ถูกจัดให้อยู่ในกลุ่มที่ 1 ส่วนยาประเภทอื่นๆ ถูกจัดไว้ในกลุ่มที่ 2 ทั้งหมด

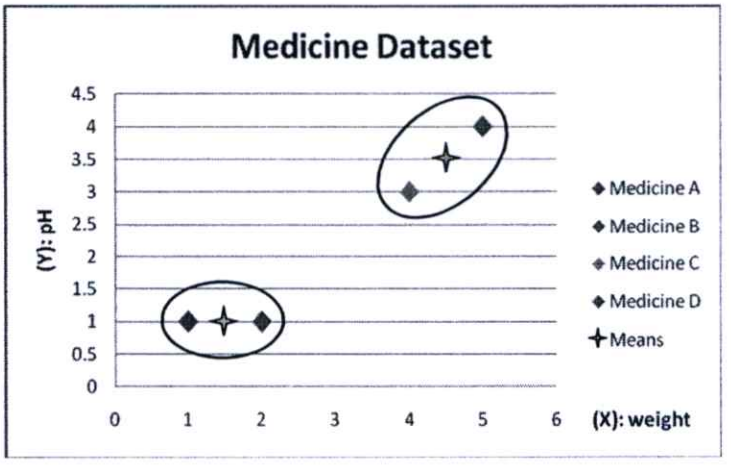


รูปที่ 2.7 ขั้นตอนที่ห้าของการแบ่งข้อมูล

ขั้นตอนที่ 6 คือ การกำหนดชุดจุดศูนย์กลางชุดใหม่อีกครั้ง โดยกลุ่มที่ 1 มีสมาชิกทั้งหมด 2 ตัวคือ ยาประเภท A และ B ค่าเฉลี่ยของกลุ่มที่ 1 สามารถคำนวณได้ดังนี้ c_1 = {(1+2)/2, (1+1)/2} ส่วนค่าเฉลี่ยของกลุ่มที่ 2 คือ c_2 = {(4+5)/2, (3+4)/2} ดังนั้นจุดศูนย์กลาง C = {(1.5, 1), (4.5, 3.5)}

รูปที่ 2.8 ขั้นตอนที่หกของการแบ่งข้อมูล

ขั้นตอนที่ 7 คือ การจัดกลุ่มข้อมูลอีกครั้ง ด้วยชุดจุดศูนย์กลางชุดใหม่คือ (1.5, 1) และ (4.5, 3.5) ตามลำดับ จากรูปที่ 2.9 ยาประเภท A และ B ถูกจัดให้อยู่ในกลุ่มที่ 1 ส่วนยาประเภทอื่นๆ ถูกจัดไว้ในกลุ่มที่ 2 ทั้งหมด เช่นเดียวกับขั้นตอนที่ห้าแสดงว่า ข้อมูลทั้งหมดไม่มีการเปลี่ยนกลุ่มแล้ว

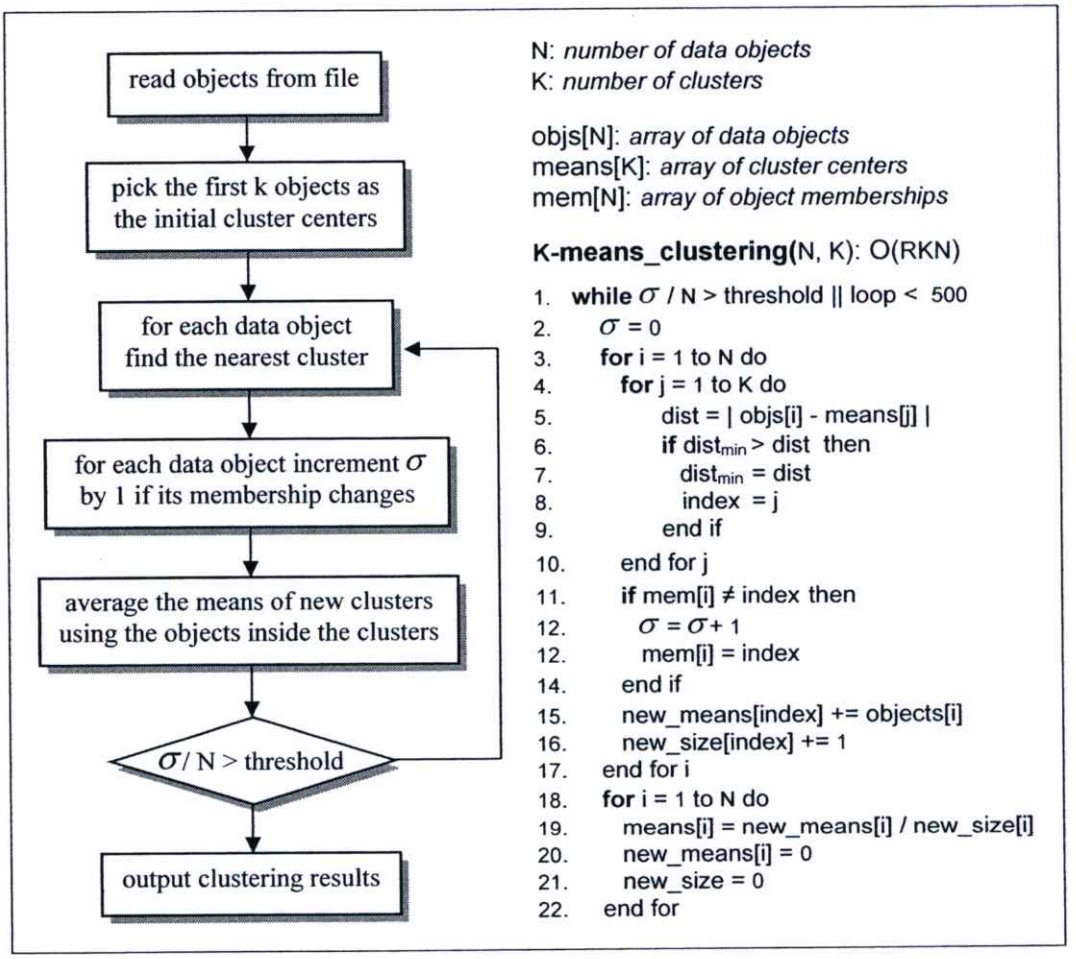


รูปที่ 2.9 ขั้นตอนที่เจ็ดของการแบ่งข้อมูล

ขั้นตอนสุดท้าย คือแสดงผลลัพธ์ที่ได้จากการแบ่งข้อมูล ซึ่งรายละเอียดทั้งหมดเหมือนกับรูปที่ 2.9 เนื่องจากข้อมูลทั้งหมดไม่มีการเปลี่ยนกลุ่มตามเงื่อนไขของกระบวนการทำงานซ้ำ ดังนั้นการแบ่งข้อมูลครั้งนี้จึงสิ้นสุดลง ผลลัพธ์ที่ได้คือกลุ่มข้อมูล 2 กลุ่ม ซึ่งจำนวนรอบของการทำงานซ้ำ (R) เท่ากับ 2 รอบเท่านั้นตามหลักการทำงานของขั้นตอนวิธีเคมีนในรูปที่ 2.1

2.1.6 ตัวอย่างโปรแกรมเคมีนแบบอนุกรม

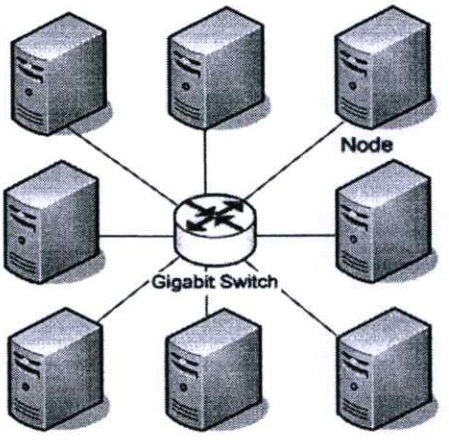
ขั้นตอนวิธีเคมีนคลัสเตอร์ริงถูกพัฒนาโปรแกรมขึ้นมาหลายรูปแบบ ซึ่งแต่ละรูปแบบยังคงยึดหลักการทำงานของขั้นตอนวิธีเดิมอยู่ โดยโปรแกรมเคมีนส่วนใหญ่ที่ถูกนำเสนอจะแตกต่างกันตรงที่เงื่อนไขของการจัดกลุ่มข้อมูลที่ใช้หยุดกระบวนการทำงานซ้ำ จึงทำให้กระบวนการภายในแตกต่างกันตามไปด้วย โดยในหัวข้อนี้จะนำเสนอโปรแกรมเคมีนแบบอนุกรมของ J. MacQueen ซึ่งนำมาใช้กับงานวิจัยนี้ รายละเอียดทั้งหมดแสดงได้ดังนี้



รูปที่ 2.10 โปรแกรมเคมีนแบบอนุกรมของแมคควีน

2.2 ระบบคลัสเตอร์

ระบบคลัสเตอร์ (Cluster System) คือกลุ่มของเครื่องคอมพิวเตอร์ที่มีการเชื่อมต่อกันผ่านเครือข่ายความเร็วสูง ซึ่งสามารถนำมาใช้การประมวลผลแบบขนานได้ โดยเครื่องคอมพิวเตอร์แต่ละเครื่องมีทรัพยากรเป็นของตนเองคือ หน่วยประมวลผล หน่วยความจำ ระบบปฏิบัติการ ส่วนต่อเชื่อมกับเครือข่าย อุปกรณ์อินพุตและเอาต์พุต ถือว่าเป็นเครื่องคอมพิวเตอร์ที่สมบูรณ์ ซึ่งแต่ละเครื่องอาจมีหน่วยประมวลผลมากกว่า 1 หน่วยก็ได้ ทำให้ระบบคลัสเตอร์ง่ายต่อการขยาย ราคาถูก และมีประสิทธิภาพสูงโดยอาจเทียบเท่าเครื่องซุปเปอร์คอมพิวเตอร์หรือสูงกว่าสำหรับการประมวลผลงานที่มีความซับซ้อน



รูปที่ 2.11 ระบบพีซีคลัสเตอร์

คอมพิวเตอร์แต่ละเครื่องในระบบคลัสเตอร์จะถูกเรียกว่า "โหนด (Node)" ซึ่งอยู่ใกล้กันหรืออยู่ในพื้นที่ใกล้เคียงกัน โหนดทั้งหมดเชื่อมต่อกันด้วยเครือข่าย เช่น เครือข่ายอีเทอร์เน็ต (Gigabit Ethernet) มายรีเน็ต (Myrinet) หรือเครือข่ายรูปแบบอื่นๆ ซึ่งต้องเป็นเทคโนโลยีเครือข่ายที่มีความเร็วสูง และค่า Latency Time ต่ำ เพื่อให้การติดต่อสื่อสารระหว่างกันสามารถเป็นไปได้อย่างรวดเร็ว และสูญเสียเวลาในเครือข่ายน้อยที่สุดเสมือนว่าหน่วยประมวลผลอยู่ใกล้กันมาก หรืออยู่บนแผงวงจรรวม (Main Board) เดียวกัน การแลกเปลี่ยนข้อมูลระหว่างเครื่องคอมพิวเตอร์ทำได้โดยการส่งผ่านข้อความ (Message Passing) เท่านั้น ซึ่งมาตรฐานที่นิยมใช้คือ ระบบการส่งผ่านข้อมูลหรือ Msessage Passing Interface (MPI) ส่วนการทำให้กลุ่มของคอมพิวเตอร์สามารถทำงานร่วมกันได้นั้นจะต้องมีซอฟแวร์เป็นตัวกลางในการเชื่อมการทำงานระหว่างเครื่องคอมพิวเตอร์เข้าด้วยกัน ตัวอย่างระบบพีซีคลัสเตอร์สามารถแสดงได้ดังรูปที่ 2.11

2.2.1 การจำแนกคอมพิวเตอร์แบบคลัสเตอร์

ระบบคลัสเตอร์สามารถจำแนกประเภทได้หลายรูปแบบตามเกณฑ์ต่างๆคือ รูปแบบการใช้งาน และลักษณะของเครื่องคอมพิวเตอร์ที่อยู่ในระบบคลัสเตอร์ ซึ่งมีรายละเอียดดังนี้

2.2.1.1 จำแนกตามรูปแบบการใช้งานระบบคลัสเตอร์

ระบบคลัสเตอร์สามารถประยุกต์ใช้ได้กับงานที่หลากหลาย ไม่ว่าจะเป็นงานทางด้านการคำนวณ หรืองานทางด้านการเป็นเครื่องแม่ข่ายให้บริการงานต่างๆ โดยสามารถแบ่งย่อยได้เป็นสองประเภทดังนี้

1) ระบบคลัสเตอร์แบบประสิทธิภาพสูง (High Performance Clusters: HPC) ระบบคลัสเตอร์แบบนี้มักถูกนำไปประยุกต์ใช้ในการคำนวณทางด้านวิทยาศาสตร์และคณิตศาสตร์ ซึ่งถูกสร้างขึ้นมาเพื่อให้มีความรวดเร็วในการคำนวณมากที่สุด ดังนั้นประสิทธิภาพของหน่วยประมวลผลแต่ละหน่วยจะต้องสูงเพียงพอ อีกทั้งเครือข่ายที่ใช้ในการเชื่อมต่อต้องมีคุณภาพดีมาก ประสิทธิภาพในการคำนวณจึงจะสูงตามไปด้วย

2) ระบบคลัสเตอร์แบบเสถียรภาพสูง (HAC: High Availability Clusters) ระบบคลัสเตอร์แบบนี้จะเน้นไปทางด้านเครื่องแม่ข่ายที่ให้บริการงานต่างๆ เช่น การให้บริการเป็นเว็บเซอร์เวอร์ (Web Server) การให้บริการพื้นที่เก็บข้อมูลบนเครือข่าย (Storage Server) การให้บริการฐานข้อมูล ไม่ว่าจะเป็น Oracle 10 g หรือ MySQL Cluster เพื่อทำให้มั่นใจได้ว่า ผู้ใช้งานทั่วๆ ไปจะสามารถเข้าถึงทรัพยากร หรือบริการต่างๆ ได้ตลอดเวลา

2.2.1.2 จำแนกตามลักษณะของเครื่องคอมพิวเตอร์ในระบบคลัสเตอร์

ระบบคลัสเตอร์เกิดจากการเชื่อมต่อเครื่องคอมพิวเตอร์หลายๆ เครื่องเข้าด้วยกัน โดยเครื่องคอมพิวเตอร์นี้อาจจะมีลักษณะเหมือนกันทั้งหมด หรือไม่เหมือนกันเลยก็ได้ ซึ่งถ้าหากจำแนกระบบคลัสเตอร์ตามลักษณะของเครื่องคอมพิวเตอร์ สามารถจำแนกออกได้สองประเภทดังนี้

1) ระบบคลัสเตอร์แบบเนื้อเดียว (Homogeneous Cluster) ระบบคลัสเตอร์แบบนี้เป็นระบบที่มีองค์ประกอบทางด้านฮาร์ดแวร์และซอร์ฟแวร์ที่ใช้ในแต่ละเครื่องเหมือนกันทั้งหมดได้แก่ หน่วยประมวลผลกลาง ชนิดและขนาดของหน่วยความจำ ชนิดและขนาดของฮาร์ดดิสก์ และชนิดของระบบปฏิบัติการ เป็นต้น โดยระบบคลัสเตอร์นี้เป็นแบบที่นิยมสร้าง เนื่องจากการบริหารจัดการระบบสามารถทำได้อย่างสะดวก นอกจากนั้นการเขียนโปรแกรมเพื่อทำการประมวลผลบนระบบคลัสเตอร์สามารถเขียนในครั้งเดียวแล้วทำงานได้กับทุกๆ เครื่องในระบบ

2) ระบบคลัสเตอร์แบบเนื้อผสม (Heterogeneous Cluster) ระบบคลัสเตอร์แบบนี้เป็นระบบที่มีความยืดหยุ่นสูง โดยสามารถสร้างจากเครื่องแบบใดก็ได้ที่สนับสนุนการประมวลผลแบบขนานแต่ปัญหาของระบบนี้คือการสร้างโปรแกรมสำหรับประมวลผลแบบขนานจะมีความยุ่งยากและซับซ้อนมากขึ้นเนื่องจากต้องทำการสร้างโปรแกรมที่สามารถประมวลผลเฉพาะของแต่ละเครื่องแต่ละระบบปฏิบัติการ เช่น เครื่องที่ใช้ระบบปฏิบัติการลีนุกส์ (Linux) และเครื่องที่ใช้ไมโครซอฟส์วินโดว์ (Microsoft Windows) เป็นระบบปฏิบัติการ โปรแกรมที่สร้างขึ้นมาอาจไม่สามารถใช้ด้วยกันได้ วิธีแก้ปัญหาอย่างหนึ่งคือสร้างโปรแกรมให้สนับสนุนมาตรฐานเช่น ANSI เป็นต้น หรืออาจสร้างโปรแกรมที่สามารถประมวลผลได้บนทุกระบบโดยที่ใช้รหัสต้นฉบับ (Source Code) ตัวเดียวกัน เช่น ภาษาซี (C) หรือภาษาจาวา (Java)

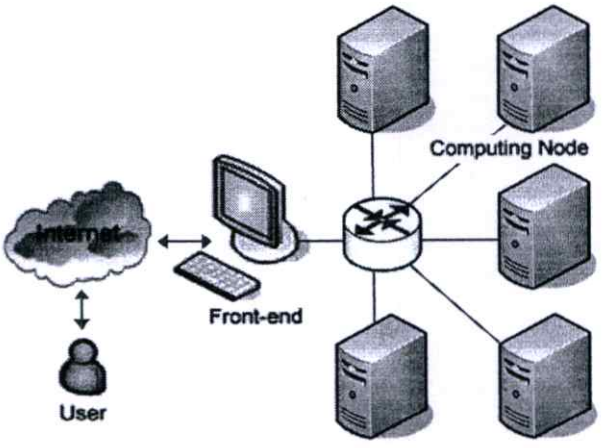
สิ่งหนึ่งที่ต้องให้ความสำคัญมากในระบบคลัสเตอร์แบบเนื้อผสมนี้ คือ การสมดุลงานในระบบ หรือ Load balancing เนื่องจากคอมพิวเตอร์ที่เชื่อมต่ออยู่กับระบบอาจมีองค์ประกอบภายในที่แตกต่างกัน ทำให้คอมพิวเตอร์แต่ละโหนดมีความสามารถและประสิทธิภาพในการประมวลผลได้รวดเร็วแตกต่างกัน ดังนั้นเครื่องที่มีประสิทธิภาพมากจึงต้องรับงานไปคำนวณมากกว่าเครื่องที่มีประสิทธิภาพต่ำตามความเหมาะสม ถ้าหากทำการสมดุลงานได้ดี ประสิทธิภาพของระบบก็จะดี แต่ถ้าสมดุลงานไม่ดี ประสิทธิภาพโดยรวมก็จะต่ำลงตามไปด้วย

2.2.2 รูปแบบการเชื่อมต่อกับเครือข่ายภายนอก

สำหรับรูปแบบวิธีการเชื่อมต่อเครือข่าย (Network Topology) ของระบบคลัสเตอร์นั้นมีหลายรูปแบบ ซึ่งรูปแบบของเครือข่ายมีเพียงแค่สองรูปแบบที่เป็นไปได้ในทางปฏิบัติและไม่ต้องลงทุนเกี่ยวกับเครือข่ายมากนัก ดังนี้

2.2.2.1 การเชื่อมต่อกับเครือข่ายภายนอกแบบเครื่องเดียว

การเชื่อมต่อระบบคลัสเตอร์ด้วยวิธีนี้ต้องมีคอมพิวเตอร์เครื่องหนึ่งที่ทำหน้าที่เป็นทางเข้า-ทางออก หรือเกตเวย์ (Gate way) ให้แก่ระบบทั้งหมด เมื่อผู้ใช้ติดต่อเข้ามาในระบบจะต้องทำการติดต่อกับเครื่องนี้ เรียกว่า “Front-end Node” ซึ่งทำหน้าที่ควบคุมการทำงานของโหนดอื่น ๆ ในระบบ ส่วนเครื่องอื่นๆ จะทำงานอยู่เบื้องหลังเท่านั้น หรือทำหน้าที่ประมวลผล เรียกว่า “Compute Node” ทำให้วิธีนี้มีประโยชน์ในเรื่องของการรักษาความปลอดภัยของระบบ เพราะมีเพียงเครื่องเดียวเท่านั้นที่เชื่อมต่อกับภายนอก แต่การเชื่อมต่อด้วยวิธีนี้ก็รับประกันไม่ได้เสมอไป การเชื่อมต่อสามารถแสดงได้ดังนี้

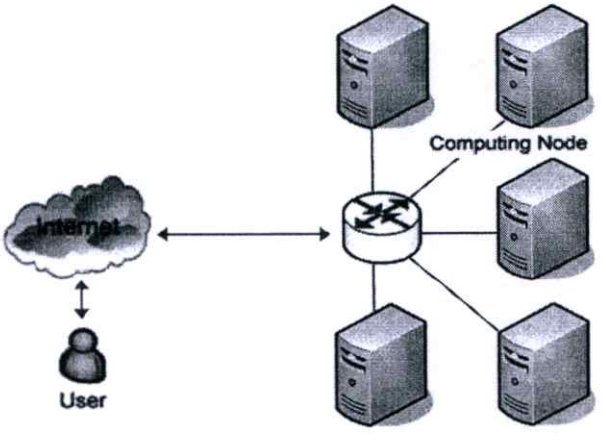


รูปที่ 2.12 การเชื่อมต่อกับเครือข่ายภายนอกแบบเครื่องเดียว

การเชื่อมต่อแบบนี้มีข้อดีในส่วนของการดูแลเรื่องความปลอดภัย แต่ไม่ค่อยเหมาะนักกับงานที่เป็นด้านการให้บริการ หรือคลัสเตอร์ที่ต้องการความคงทนสูง (High Availability Cluster) อย่างเช่น เครื่องแม่ข่ายให้บริการเว็บ หรืออีเมล์ ดังนั้นจึงต้องใช้รูปแบบการเชื่อมต่อตามวิธีการในหัวข้อถัดไป

2.2.2.2 การเชื่อมต่อกับเครื่องข่ายภายนอกแบบทุกเครื่อง

การเชื่อมต่อเครือข่ายแบบนี้ ทุกเครื่องจะทำการเชื่อมต่อกับเครือข่ายภายนอก ทำให้เครื่องจากภายนอกสามารถเข้าถึงทรัพยากรของแต่ละเครื่องได้โดยตรง ซึ่งมีประโยชน์ในกรณีที่เครื่องคลัสเตอร์เหล่านี้ทำหน้าที่ให้บริการงานต่างๆ เช่น เครื่องแม่ข่ายของงานเว็บไซต์ การเชื่อมต่อแบบนี้มักจะนำเอาไฟร์วอล (Firewall) มาวางไว้ด้านหน้าของระบบอีกชั้น เพื่อเป็นการป้องกันการบุกรุกจากภายนอกได้ในระดับหนึ่ง



รูปที่ 2.13 การเชื่อมต่อกับเครือข่ายภายนอกแบบทุกเครื่อง

2.3 งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้องกับงานวิจัยนี้แบ่งออกเป็น 2 ประเภท คือ วิธีเลือกชุดจุดศูนย์กลางเริ่มต้นและขั้นตอนวิธีเคมีนแบบขนานบนเครือข่าย ซึ่งงานวิจัยเหล่านั้นมีเป้าหมายเช่นเดียวกัน คือต้องการเพิ่มประสิทธิภาพให้ขั้นตอนวิธีเคมีน โดยมีรายละเอียดทั้งหมดดังนี้

2.3.1 วิธีเลือกชุดจุดศูนย์กลางเริ่มต้น

วิธีเลือกชุดจุดศูนย์กลางเริ่มต้น (Select Initial Means Method) คือกระบวนการเลือกข้อมูลหรือตัวแทนตามจำนวนที่ต้องการ (K) จากข้อมูลทั้งหมด (N) โดยตัวแทนเหล่านั้นจะถูกกำหนดให้เป็นชุดจุดศูนย์กลางเริ่มต้น ซึ่งชุดจุดศูนย์กลางเริ่มต้นที่มีความเหมาะสมกับชุดข้อมูลจะส่งผลให้การแบ่งข้อมูลด้วยขั้นตอนวิธีเคมีนสำเร็จได้อย่างรวดเร็ว และให้ผลลัพธ์ที่แม่นยำ โดยการจัดกลุ่มข้อมูลด้วยชุดจุดศูนย์กลางเริ่มต้นที่แตกต่างกันก็อาจได้ผลลัพธ์ของการจัดกลุ่มแตกต่างกันตามไปด้วย ส่วนเหตุผลที่ต้องเลือกชุดจุดศูนย์กลางเริ่มต้นจากข้อมูลบางส่วนของข้อมูลทั้งหมดนั้น เพื่อทำให้มั่นใจได้ว่าแต่ละกลุ่มจะมีสมาชิกเข้าร่วมหลังจากผ่านขั้นตอนการจัดกลุ่มข้อมูลครั้งแรกไปแล้วตามหลักการทำงานของขั้นตอนวิธีเคมีน (อธิบายในหัวข้อ 2.1) ซึ่งในหัวข้อนี้มีงานวิจัยที่เกี่ยวข้องทั้งหมด 3 งานวิจัย ดังนี้

1) วิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบสุ่ม (Random Method)
2) วิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบคัดสรร (Seeding Method)
3) วิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบให้ค่าน้ำหนักดีกำลังสอง (D^2 Weighting Method)

2.3.1.1 วิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบสุ่ม

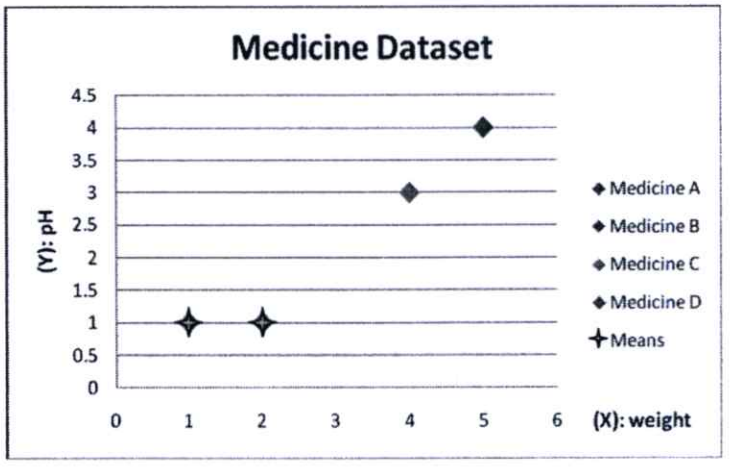
วิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบสุ่ม (Random Method) เป็นวิธีเลือกที่ได้รับความนิยมมากที่สุดเนื่องจากเป็นวิธีที่ง่าย และไม่เสียเวลาในการเลือกชุดจุดศูนย์กลางเริ่มต้น โดยวิธีเลือกนี้อ้างอิงมาจากงานวิจัย [8] ในปี ค.ศ. 2005 ซึ่งอาจเรียกว่า การสุ่มตามลำดับของข้อมูล คือเลือกข้อมูลลำดับแรกจนถึงลำดับที่ K แล้วกำหนดให้เป็นชุดจุดศูนย์กลางเริ่มต้น โดยมีหลักการทำงานดังนี้

Random Method: O(K)

- Choose a center $c_i = x_i$, that $X = \{x_1, x_2, \ldots, x_n\}$ and $1 \leq i \leq K$.
- Repeat Step 1 until we have chosen a total of k centers.

รูปที่ 2.14 วิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบสุ่มตามลำดับของข้อมูล

ตัวอย่าง อ้างอิงจากข้อมูลยาประเภทต่างๆ (อธิบายในหัวข้อ 2.1.5) เพื่อนำมาใช้ในการเลือกชุดจุดศูนย์กลางเริ่มต้นด้วยวิธีการนี้ ซึ่งจำนวนกลุ่มที่ต้องการเท่ากับ 2 กลุ่ม ดังนี้



รูปที่ 2.15 การเลือกชุดจุดศูนย์กลางเริ่มต้นแบบสุ่มตามลำดับของข้อมูล

จากรูปที่ 2.15 ชุดจุดศูนย์กลางเริ่มต้นที่เลือกได้ด้วยวิธีการนี้คือ C = {(1, 1), (2, 1)} เนื่องจาก ยาประเภท A และ B เป็นข้อมูลลำดับแรก และลำดับที่สองของชุดข้อมูล โดยจำนวนกลุ่มที่ต้องการเท่ากับ 2 กลุ่ม

2.3.1.2 วิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบคัดสรร

วิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบคัดสรร (Seeding Method) เป็นวิธีเลือกที่อ้างอิงมาจากงานวิจัย [10] โดย R. Ostrovsky และคณะในปี ค.ศ. 2006 ซึ่งมีหลักการทำงานดังนี้

Seeding Method: $O(N^2)$

- Choose two centers c_1, c_2 from $x, y \in X$ with probability $\| x-y \|^2 / \sum_{x,y \in X} \| x-y \|^2$.
- Choose the next center c_i, selecting $c_i = x \in X$ with probability $\min_{j \in \{1,\ldots,i\}} \| x-c_j \|^2 / \sum_{x \in X} \min_{j \in \{1,\ldots,i\}} \| x-c_j \|^2$ that $1 \le i \le K$.
- Repeat Step 2 until we have chosen a total of k centers.

รูปที่ 2.16 วิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบคัดสรร

ตัวอย่าง อ้างอิงจากข้อมูลยาประเภทต่างๆ (อธิบายในหัวข้อ 2.1.5) เพื่อนำมาใช้ในการเลือกชุดจุดศูนย์กลางเริ่มต้นด้วยวิธีการนี้ แต่กำหนดให้กลุ่มที่ต้องการเท่ากับ 3 กลุ่ม เพื่อแสดงการเลือกชุดจุดศูนย์กลางเริ่มต้นได้ในทุกกรณี รายละเอียดทั้งหมดแสดงได้ดังนี้

ตารางที่ 2.2 ข้อมูลยาประเภทต่างๆ เมื่อจำนวนกลุ่มที่ต้องการเท่ากับ 3 กลุ่ม

Objects	Attribute1: weight	Attribute2: pH	Clusters: 3
Medicine A	1	1	-
Medicine B	2	1	-
Medicine C	4	3	-
Medicine D	5	4	-

ขั้นตอนที่ 1 คือ เลือกจุดศูนย์กลางเริ่มต้นจำนวน 2 จุดแรก โดยเลือกข้อมูลหนึ่งคู่จากข้อมูลทั้งหมดตามค่าความน่าจะเป็น (Probability: P) ที่กำหนดไว้ ซึ่งสามารถคำนวณได้ดังสมการที่ 2.5

$$P = \frac{\| x-y \|^2}{\sum_{x,y \in X} \| x-y \|^2} \qquad (2.5)$$

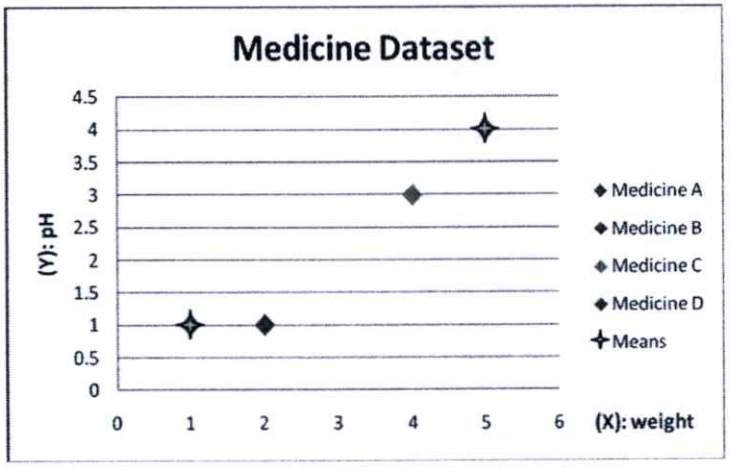
เมื่อ X คือ ชุดข้อมูลที่ต้องการจัดกลุ่มด้วยขั้นตอนวิธีเคมีน

x, y คือ ข้อมูลใดๆ ที่อยู่ภายในชุดข้อมูล

ตารางที่ 2.3 การเลือกจุดศูนย์กลางเริ่มต้น 2 จุดแรก ด้วยวิธีเลือกแบบคัดสรร

คู่ของข้อมูล	ระยะทาง: $D(x, y)^2$	ระยะทางสะสม	ค่าความน่าจะเป็น: (P)
A และ B	$\{(1-2)^2 + (1-1)^2\} = 1$	1	(1 / 67) = 0.01
A และ C	$\{(1-4)^2 + (1-3)^2\} = 13$	14	(13/ 67) = 0.19
A และ D	$\{(1-5)^2 + (1-4)^2\} = 25$	39	(25/ 67) = 0.37
B และ C	$\{(2-4)^2 + (1-3)^2\} = 8$	47	(8 / 67) = 0.12
B และ D	$\{(2-5)^2 + (1-4)^2\} = 18$	65	(18/ 67) = 0.27
C และ D	$\{(4-5)^2 + (3-4)^2\} = 2$	67	(2 / 67) = 0.03

จากตารางที่ 2.3 แสดงการคำนวณค่าความน่าจะเป็นของขั้นตอนที่ 1 ของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบคัดสรร สรุปว่าคู่ของยาประเภท A และ D ถูกกำหนดให้เป็นจุดศูนย์กลาง 2 จุดแรก เนื่องจากค่าความน่าจะเป็นของยาประเภท A และ D มากที่สุด ผลลัพธ์แสดงได้ดังรูปที่ 2.17



รูปที่ 2.17 การเลือกจุดศูนย์กลางเริ่มต้น 2 จุดแรก ด้วยวิธีเลือกแบบคัดสรร

ขั้นตอนที่ 2 คือ เลือกจุดศูนย์กลางเริ่มต้นจุดถัดไป คือจุดศูนย์กลางจุดที่สาม โดยเลือกจากข้อมูลตามค่าความน่าจะเป็น (Probability: P) ที่กำหนดไว้ ซึ่งสามารถคำนวณได้ดังสมการที่ 2.6

$$P = \frac{\min_{j\in\{1,\ldots,i\}} \| x - c_j \|^2}{\sum_{x\in X} \min_{j\in\{1,\ldots,i\}} \| x - c_j \|^2} \qquad (2.6)$$

เมื่อ X คือ ชุดข้อมูลที่ต้องการจัดกลุ่มด้วยขั้นตอนวิธีเคมีน

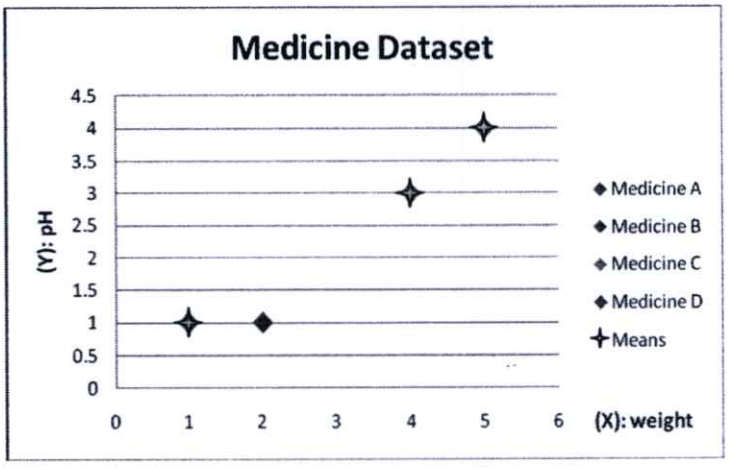
x คือ ข้อมูลใดๆ ที่อยู่ภายในชุดข้อมูล

i คือ จำนวนจุดศูนย์กลางที่ถูกเลือกไว้แล้ว ซึ่ง $1 \le i \le K$

ตารางที่ 2.4 การเลือกจุดศูนย์กลางเริ่มต้นจุดที่สาม ด้วยวิธีเลือกแบบคัดสรร

ข้อมูล	ระยะทางที่สั้นที่สุด	ระยะทางสะสม	ค่าความน่าจะเป็น: (P)
A	$\{(1-1)^2 + (1-1)^2\} = 0$	0	(0 / 3) = 0
B	$\{(2-1)^2 + (1-1)^2\} = 1$	1	(1 / 3) = 0.33
C	$\{(4-5)^2 + (3-4)^2\} = 2$	3	(2 / 3) = 0.67
D	$\{(5-5)^2 + (5-4)^2\} = 0$	3	(0 / 3) = 0

จากตารางที่ 2.4 แสดงการคำนวณค่าความน่าจะเป็นของขั้นตอนที่ 2 ของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบคัดสรร สรุปว่าข้อมูลของยาประเภท C ถูกกำหนดให้เป็นจุดศูนย์กลางจุดที่สาม เนื่องจากค่าความน่าจะเป็นของยาประเภท C มากที่สุด ดังนั้นชุดจุดศูนย์กลางเริ่มต้นที่เลือกได้ด้วยวิธีการนี้คือ C = {(1, 1), (5, 4), (4, 3)} ผลลัพธ์แสดงได้ดังรูปที่ 2.18



รูปที่ 2.18 การเลือกจุดศูนย์กลางเริ่มต้นจุดที่สาม ด้วยวิธีเลือกแบบคัดสรร

2.3.1.3 วิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบให้ค่าน้ำหนักดีกำลังสอง

วิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบให้ค่าน้ำหนักดีกำลังสอง (D^2 Weighting Method) เป็นวิธีเลือกที่อ้างอิงมาจากงานวิจัย [1] ในปี ค.ศ. 2007 โดยมีหลักการทำงานดังนี้

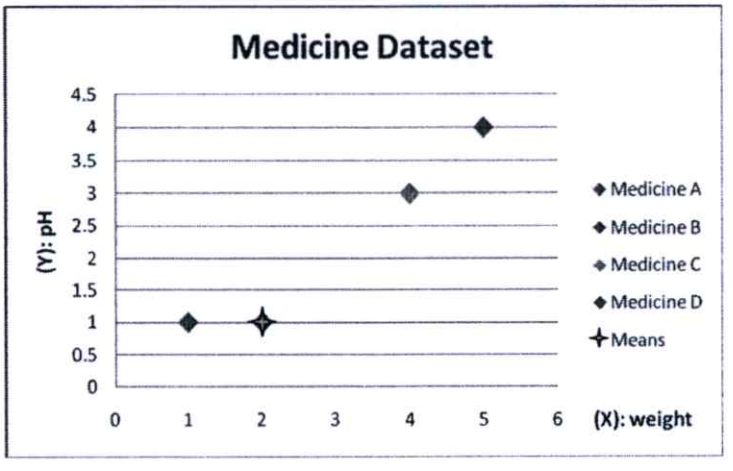
D^2 Weighting Method: O(KN)

- Choose an initial center c_1 uniformly at random from X.
- Choose the next center c_i, selecting $c_i = x' \in X$ with probability $D(x')^2 / \sum_{x \in X} D(x)^2$.
- Repeat Step 2 until we have chosen a total of k centers.

รูปที่ 2.19 วิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบให้ค่าน้ำหนักดีกำลังสอง

ตัวอย่าง อ้างอิงจากข้อมูลยาประเภทต่างๆ (อธิบายในหัวข้อ 2.1.5) เพื่อนำมาใช้ในการเลือกจุดศูนย์กลางเริ่มต้นด้วยวิธีการนี้ ซึ่งจำนวนกลุ่มที่ต้องการเท่ากับ 2 กลุ่ม ดังนี้

ขั้นตอนที่ 1 คือ เลือกจุดศูนย์กลางเริ่มต้นจุดแรก โดยเลือกข้อมูลจากชุดข้อมูลแบบสุ่ม ซึ่งสมมุติให้ยาประเภท B = (2, 1) เป็นจุดศูนย์กลางเริ่มต้นจุดแรก ผลลัพธ์แสดงได้ดังรูปที่ 2.20



รูปที่ 2.20 การเลือกจุดศูนย์กลางเริ่มต้นจุดแรก ด้วยวิธีเลือกแบบให้ค่าน้ำหนักดีกำลังสอง

ขั้นตอนที่ 2 คือ เลือกจุดศูนย์กลางเริ่มต้นจุดถัดไป คือจุดศูนย์กลางจุดที่สอง โดยเลือกข้อมูลตามค่าความน่าจะเป็น (Probability: P) ที่กำหนดไว้ ซึ่งสามารถคำนวณได้ดังสมการที่ 2.7

$$P = \frac{D(x')^2}{\sum_{x \in X} D(x)^2} \tag{2.7}$$

เมื่อ X คือ ชุดข้อมูลที่ต้องการจัดกลุ่มด้วยขั้นตอนวิธีเคมีน

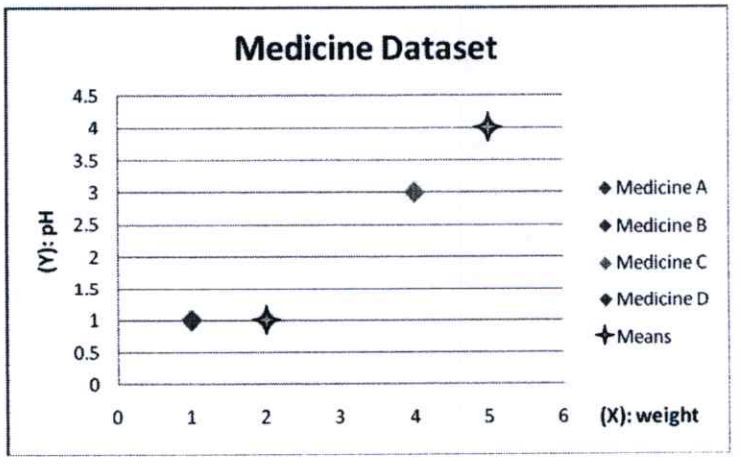
x คือ ข้อมูลใดๆ ที่อยู่ภายในชุดข้อมูล

D(x') คือ ระยะทางระหว่างข้อมูลที่สนใจกับจุดศูนย์กลางของข้อมูล

ตารางที่ 2.5 การเลือกจุดศูนย์กลางเริ่มต้นจุดที่สองด้วยวิธีเลือกแบบให้ค่าน้ำหนักดีกำลังสอง

ข้อมูล	ระยะทาง: $D(x')^2$	ระยะทางสะสม	ค่าความน่าจะเป็น: (P)
A	$\{(1-2)^2 + (1-1)^2\} = 1$	1	(1 / 27) = 0.04
B	$\{(2-2)^2 + (1-1)^2\} = 0$	1	(0 / 27) = 0
C	$\{(4-2)^2 + (3-1)^2\} = 8$	9	(8 / 27) = 0.30
D	$\{(5-2)^2 + (4-1)^2\} = 18$	27	(18/ 27) = 0.67

จากตารางที่ 2.5 แสดงการคำนวณค่าความน่าจะเป็นของขั้นตอนที่สองของวิธีเลือกจุดศูนย์กลางเริ่มต้นแบบให้ค่าน้ำหนักดีกำลังสอง สรุปว่าข้อมูลของยาประเภท D ถูกกำหนดให้เป็นจุดศูนย์กลางจุดที่สอง เนื่องจากค่าความน่าจะเป็นของยาประเภท D มากที่สุด ดังนั้นชุดจุดศูนย์กลางเริ่มต้นที่เลือกได้ด้วยวิธีการนี้คือ C = {(2, 1), (4, 3)} ผลลัพธ์แสดงได้ดังรูปที่ 2.21



รูปที่ 2.21 การเลือกจุดศูนย์กลางเริ่มต้นจุดที่สอง ด้วยวิธีเลือกแบบให้ค่าน้ำหนักดีกำลังสอง

2.3.2 ขั้นตอนวิธีเคมีนแบบขนาน

การแบ่งข้อมูลด้วยขั้นตอนวิธีเคมีนแบบขนาน (Parallel K-means Algorithm: PK-means) ถูกนำเสนอในปี 2005 โดย W. Liao ตามงานวิจัย [8] ซึ่งนำวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบสุ่มตามลำดับ และอ้างอิงหลักการทำงานของขั้นตอนวิธีเคมีนแบบอนุกรมของ J. MacQueen (อธิบายในหัวข้อ 2.1.6) มาใช้ในออกแบบขั้นตอนวิธีเคมีนแบบขนาน โดยมีขั้นตอนการทำงานดังนี้

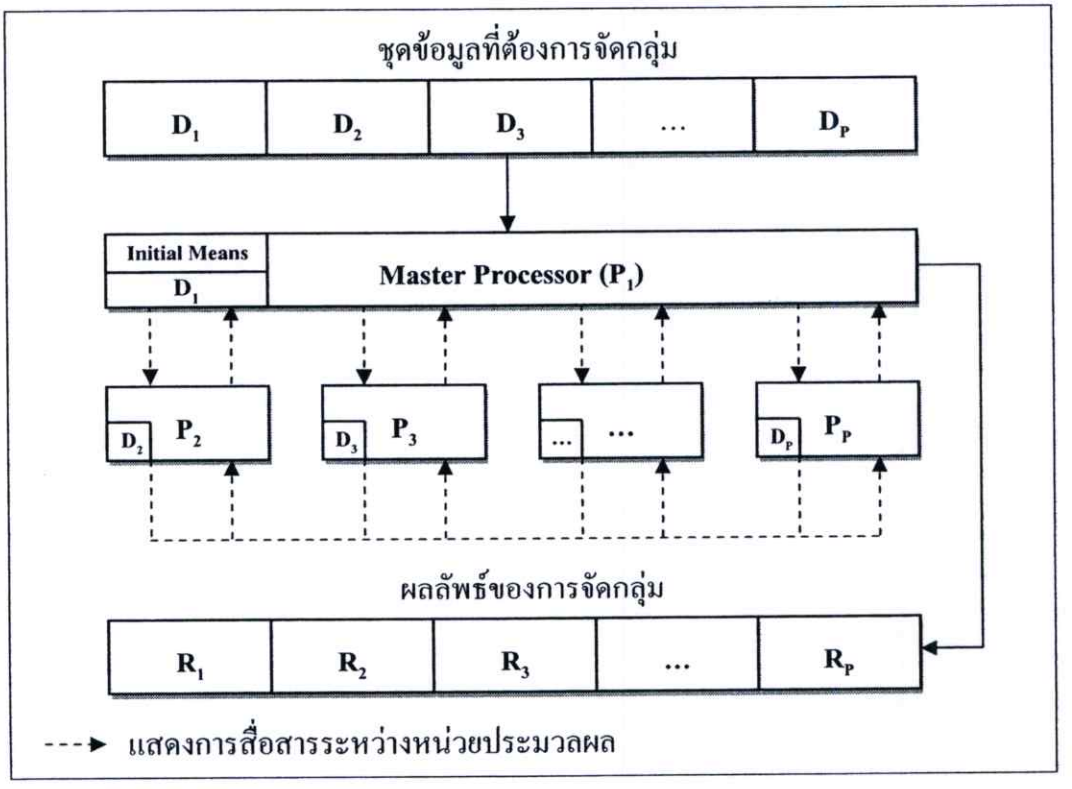
Master Process

1. Read objects from file.
2. Randomly form equal subsets follow about number of processors.
3. Select initial centers.
4. Send each subset to each slave process.
5. Broadcast initial centers to all slave processes.
6. Receive result sets from all slave processes.

Slave Process

1. Receive a subset P and set C from master process.
2. Assign internal objects into internal clusters and count of objects moved.
3. Calculate sum and count of objects into internal cluster.
4. Broadcast sum, count of objects and count of objects moved to all slave processes.
5. Calculate new centers by using sum and counts.
6. Repeat step 2 if objects moved.
7. Send result set to master process.

รูปที่ 2.22 ขั้นตอนวิธีเคมีนแบบขนานของเลียโอ



รูปที่ 2.23 การทำงานร่วมกันระหว่างหน่วยประมวลผลตามขั้นตอนวิธีเคมีนแบบขนาน

หลักการประมวลผลแบบขนาน คือการแตกงานใหญ่ออกเป็นงานย่อย แล้วให้หน่วยประมวลผลทั้งหมดช่วยกันทำงาน สุดท้ายก็รวบรวมผลลัพธ์ที่ได้จากงานย่อย ซึ่งการพัฒนาโปรแกรมแบบขนานบนระบบคลัสเตอร์เป็นวิธีหนึ่งที่ได้รับความนิยม เนื่องจากให้ความเร็วในการประมวลผลสูง และสามารถรองรับข้อมูลในระดับที่เครื่องคอมพิวเตอร์เครื่องใดเครื่องหนึ่งไม่สามารถรองรับได้ ทำให้แก้ปัญหาขนาดใหญ่ได้โดยใช้เวลาลดลง (ซึ่งปกติต้องใช้เวลานาน)

จากรูปที่ 2.22 และ 2.23 แสดงการออกแบบ และการทำงานของการขั้นตอนวิธีแบบขนาน โดยแบ่งหน่วยประมวลผลออกเป็น 2 ประเภท คือหน่วยประมวลผลหลัก (Master Processor) จำนวน 1 หน่วยประมวลผล และหน่วยประมวลผลรอง (Slave Processor) จำนวน P-1 หน่วยประมวลผล ซึ่งออกแบบให้แต่ละหน่วยประมวลผลจัดกลุ่มข้อมูลของทุกกลุ่ม และจัดกลุ่มข้อมูลไปพร้อมๆ กัน แต่หน่วยประมวลผลหลักต้องรับผิดชอบหน้าที่มากกว่าหน่วยประมวลผลรองอยู่ 3 กระบวนการ คือกระบวนการแบ่งชุดข้อมูลที่ต้องการจัดกลุ่ม เลือกชุดจุดศูนย์กลางเริ่มต้นด้วยวิธีการสุ่มตามลำดับของชุดข้อมูล และกระบวนการรวบรวมผลลัพธ์จากการจัดกลุ่มเพื่อบันทึกลงในไฟล์ข้อมูล ส่วนหน่วยประมวลผลรองต้องรอรับชุดข้อมูลย่อยที่มีขนาดเท่ากัน หรือใกล้เคียงกัน และชุดจุดศูนย์กลางเริ่มต้นจากหน่วยประมวลผลหลักก่อนจึงจะเริ่มทำงานได้ ซึ่งความซับซ้อนด้านเวลา และหน่วยความจำของขั้นตอนวิธีนี้เท่ากับ O(KR(N/P)) และ O(N) ตามลำดับ โดยการออกแบบของขั้นตอนวิธีนี้เป็นไปตามแบบจำลองการเขียนโปรแกรมเพียงโปรแกรมเดียว และใช้โปรแกรมนี้ทำงานบนทุกหน่วยประมวลผลกับชุดข้อมูลย่อยที่แตกต่างกัน (SPMD) ซึ่งขนาดชุดข้อมูลย่อยของแต่ละหน่วยประมวลผลสามารถคำนวณได้ดังสมการที่ 2.8

โดยที่

$$N = \sum_{i=1}^{P} D_i \tag{2.8}$$

$$D_i \approx \frac{N}{P}$$

เมื่อ N คือ ขนาดของชุดข้อมูลที่ต้องการจัดกลุ่มด้วยขั้นตอนวิธีเคมีนแบบขนาน

P คือ จำนวนหน่วยประมวลผลที่ใช้ในการทำงาน

D_i คือ ชุดข้อมูลย่อยของแต่ละหน่วยประมวลผลที่ i ซึ่ง $1 \le i \le P$

จากรูปที่ 2.23 ชุดข้อมูลที่ผ่านกระบวนการแบ่งข้อมูลแล้วจะได้ชุดข้อมูลย่อยจำนวน P ชุด ซึ่งชุดข้อมูลย่อยทั้งหมดจะมีลำดับดังนี้ D_1, D_2, D_3, ..., D_P โดยชุดข้อมูลย่อยแต่ละชุดจะถูกส่งไปเก็บไว้ยังหน่วยความจำของแต่ละหน่วยประมวลผลตามลำดับนี้ P_1, P_2, P_3, ..., P_P และชุดข้อมูลย่อยที่ผ่านการจัดกลุ่มเรียบร้อยแล้วก็จะให้ผลลัพธ์ตามลำดับดังนี้ R_1, R_2, R_3, ..., R_P จากนั้นหน่วยประมวลผลหลักก็ทำหน้าที่รวบรวมผลลัพธ์ และบันทึกลงไฟล์ข้อมูล

2.4 การวัดสมรรถนะ

การวัดสมรรถนะ (Performance Measurement) ของขั้นตอนวิธีแบบอนุกรมและแบบขนาน (Sequential/ Parallel algorithm) โดยส่วนใหญ่จะประเมินผลจากเวลาที่ใช้ในการประมวลผล (Response Time) เป็นหลัก ซึ่งในการประมวลผลจริงบนระบบคลัสเตอร์นั้นเวลาที่ใช้ในการประมวลผลงานหนึ่งสามารถวัดได้โดยจับเวลาตั้งแต่เริ่มต้นจนสิ้นสุดการประมวลผล ซึ่งวัดได้ทั้งแบบอนุกรม (Sequential Programming) และแบบขนาน (Parallel Programming) โดยเวลาที่ใช้ในการประมวลผลเหล่านั้นไม่รวมเวลาที่ใช้ในการย้ายข้อมูลมาเก็บไว้ในหน่วยความจำสำรอง (Hard disk) และการตั้งค่าระบบ (Initialization) ของ MPI ซึ่งกระบวนการทั้งสองจะถูกทำเพียงครั้งเดียว

การวัดสมรรถนะของขั้นตอนวิธีแบบขนานยังสามารถคำนวณค่าอัตราการเพิ่มขึ้นของความเร็ว (Speedup) และค่าประสิทธิภาพ (Efficiency) ได้อีกด้วย เพื่อเปรียบเทียบประสิทธิภาพของขั้นตอนวิธีแบบขนานที่เพิ่มขึ้นเมื่อเพิ่มจำนวนหน่วยประมวลผล โดยค่าในอุดมคติของทุกสมการจะสูงสุดได้เมื่อเวลาที่ใช้ในการสื่อสารระหว่างหน่วยประมวลผลเป็นศูนย์เท่านั้น ซึ่งในกรณีนี้บนระบบคลัสเตอร์เกิดขึ้นได้ยาก เนื่องจากแต่ละหน่วยประมวลผลเชื่อมต่อกันผ่านเครือข่าย และปกติเมื่อเพิ่มจำนวนหน่วยประมวลผลเข้าไปก็จะทำให้เวลาที่ใช้ในการประมวลผลลดลง แต่เวลาที่ใช้ในการสื่อสารระหว่างหน่วยประมวลผลจะเพิ่มขึ้นแทน สมการคำนวณค่าเหล่านี้สามารถแสดงได้ดังนี้

2.4.1 เวลาที่ใช้ในการประมวลผล

การวัดเวลาที่ใช้ในการประมวลผล (Response Time) ของการประมวลผลแบบขนาน (Parallel Computation) แตกต่างจากการประมวลผลแบบอนุกรม เนื่องจากการประมวลผลแบบขนานต้องการใช้หน่วยประมวลผลมากกว่าหนึ่งหน่วยในการทำงาน ซึ่งหน่วยประมวลผลแต่ละหน่วยอาจมีการสื่อสาร และแลกเปลี่ยนข้อมูลกัน ดังนั้นเวลาที่ใช้ในการประมวลผลของโปรแกรมแบบขนานจึงสามารถคำนวณได้ดังสมการที่ 2.9

$$T_P = T_{comm} + T_{comp} \leq \frac{T_S}{P} \tag{2.9}$$

เมื่อ T_P คือ เวลาทั้งหมดที่ใช้ในการประมวลผลแบบขนานด้วย P หน่วยประมวลผล ซึ่งมีค่าในอุดมคติสูงสุดเท่ากับ T_S/P

T_{comm} คือ เวลาที่ใช้ในการประมวลผลแบบขนานด้วย P หน่วยประมวลผล

T_{comp} คือ เวลาที่ใช้ในการสื่อสารระหว่างหน่วยประมวลผล

T_S คือ เวลาที่ใช้ในการประมวลผลแบบอนุกรม หรือเวลาที่ใช้ในการประมวลผลแบบขนานด้วย 1 หน่วยประมวลผล

P คือ จำนวนหน่วยประมวลผลที่ใช้ในการทำงาน

2.4.2 ค่าอัตราการเพิ่มของความเร็ว

การวัดอัตราการเพิ่มของความเร็ว (Speedup) ของการประมวลผลแบบขนาน คือการเปรียบเทียบเวลาที่ใช้ในการประมวลผลแบบอนุกรมกับเวลาทั้งหมดที่ใช้ในการประมวลผลแบบขนานด้วย P หน่วยประมวลผล เพื่อบอกว่าโปรแกรมแบบขนานที่พัฒนาขึ้นมีความเร็วของการประมวลผลเพิ่มขึ้นเท่าไหร่ ซึ่งสามารถคำนวณได้ดังสมการที่ 2.10

$$S_P = \frac{T_S}{T_P} \leq P \tag{2.10}$$

เมื่อ S_P คือ อัตราการเพิ่มของความเร็วที่ใช้ในการประมวลผลแบบขนานด้วย P หน่วยประมวลผล ซึ่งมีค่าในอุดมคติสูงสุดเท่ากับ P

T_S คือ เวลาที่ใช้ในการประมวลผลแบบอนุกรม หรือเวลาที่ใช้ในการประมวลผลแบบขนานด้วย 1 หน่วยประมวลผล

T_P คือ เวลาทั้งหมดที่ใช้ในการประมวลผลแบบขนานด้วย P หน่วยประมวลผล

2.4.3 ค่าประสิทธิภาพ

การวัดประสิทธิภาพ (Efficiency) ของการประมวลผลแบบขนาน คือการเปรียบเทียบค่าอัตราการเพิ่มของความเร็วกับจำนวนหน่วยประมวลผลที่ใช้ในการทำงาน เพื่อบอกว่าโปรแกรมแบบขนานที่พัฒนาขึ้นได้รับประสิทธิภาพมากน้อยเพียงใด ซึ่งสามารถคำนวณได้ดังสมการที่ 2.11

$$E_P = \frac{S_P}{P} \leq 1 \tag{2.11}$$

เมื่อ E_P คือ ประสิทธิภาพของการประมวลผลแบบขนาน ซึ่งมีค่าในอุดมคติสูงสุดเท่ากับ 1

S_P คือ อัตราการเพิ่มของความเร็วที่คำนวณได้จากสมาการที่ 2.10

P คือ จำนวนหน่วยประมวลผลที่ใช้ในการทำงาน

บทที่ 3
ขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์

งานวิจัยนี้ทำการศึกษาและเพิ่มประสิทธิภาพให้กับขั้นตอนวิธีเคมีนแบบขนานที่มีผู้เสนอไว้แล้ว [8] ด้วยวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่ที่นำเสนอในงานวิจัยนี้เพื่อลดเวลา และเพิ่มความแม่นยำให้กับการจัดกลุ่มข้อมูลด้วยขั้นตอนวิธีเคมีนแบบขนานดังกล่าว ซึ่งขั้นตอนวิธีเคมีนแบบขนานบนระบบคลัสเตอร์ที่งานวิจัยนี้นำเสนอได้นำวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่มาประยุกต์ใช้ร่วมกัน โดยแบ่งการทำงานทั้งหมดออกเป็น 2 กระบวนการ คือ กระบวนการหลัก และกระบวนการรอง (Master-Slave Process) ขั้นตอนวิธีเคมีนแบบขนานในงานวิจัยนี้มี 2 วิธี ดังนี้

1) ขั้นตอนวิธีเคมีนแบบขนานที่กำหนดให้กระบวนการหลักเลือกชุดจุดศูนย์กลางเริ่มต้น (Parallel K-means Algorithm by Master Process Select Initial Means: MIM PK-means)

2) ขั้นตอนวิธีเคมีนแบบขนานที่กำหนดให้กระบวนการรองเลือกชุดจุดศูนย์กลางเริ่มต้น (Parallel K-means Algorithm by Slave Process Select Initial Means: SIM PK-means)

โดยวิธีแรก คือขั้นตอนวิธีเคมีนแบบขนานที่กำหนดให้กระบวนการหลักเลือกชุดจุดศูนย์กลางเริ่มต้น (MIM PK-means) ออกแบบการทำงานทั้งหมดตามงานวิจัย [8] และนำวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่ไปประยุกต์ใช้ ซึ่งการเลือกชุดจุดศูนย์กลางเริ่มต้นถูกกำหนดไว้ในกระบวนการหลัก (Master Process) โดยให้หน่วยประมวลผลหลัก (Master Processor) รับผิดชอบงานของกระบวนการนี้ หน่วยประมวลผลหลักจึงต้องเลือกชุดจุดศูนย์กลางเริ่มต้นเพียงฝ่ายเดียวเท่านั้น ส่วนวิธีที่สอง คือขั้นตอนวิธีเคมีนแบบขนานที่กำหนดให้กระบวนการรองเลือกชุดจุดศูนย์กลางเริ่มต้น (SIM PK-means) ซึ่งปรับการออกแบบจากวิธีแรก โดยนำวิธีเลือกจุดศูนย์กลางเริ่มต้นแบบใหม่มากำหนดไว้ในกระบวนการรอง (Slave Process) และกำหนดให้ทุกหน่วยประมวลผล (Master-Slave Processor) รับผิดชอบงานของกระบวนการรอง ดังนั้นทุกหน่วยประมวลผลจึงต้องทำหน้าที่เลือกชุดจุดศูนย์กลางร่วมกัน โดยขั้นตอนวิธีเคมีนแบบขนานบนระบบคลัสเตอร์ทุกวิธีที่กล่าวมานั้นถูกพัฒนาขึ้นด้วยภาษาซีตามมาตราฐานเอ็มพีไอบนระบบคลัสเตอร์ของสำนักวิจัยและบริการคอมพิวเตอร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เนื้อหาในบทที่ 3 นี้แบ่งออกเป็น 3 ส่วนดังต่อไปนี้ ส่วนที่ 1 วิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่ ส่วนที่ 2 ขั้นตอนวิธีเคมีนแบบขนานบนระบบคลัสเตอร์ และส่วนสุดท้ายการวิเคราะห์ความซับซ้อนด้านเวลา และหน่วยความจำของขั้นตอนวิธีเคมีนแบบขนานบนระบบคลัสเตอร์

3.1 วิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่

ในหัวข้อนี้นำเสนอวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่ ซึ่งนำแนวคิดของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบคัดสรรจากงานวิจัย [10] (อธิบายในหัวข้อ 2.3.1.2) มาศึกษา เพื่อพัฒนาเป็นวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่ โดยอาศัยค่านอร์มยูคลิด (Euclidean norm) ในการค้นหาจุดศูนย์กลางเริ่มต้นจุดแรก และคำนวณค่าเฉลี่ยของสมาชิกแต่ละกลุ่มในขั้นตอนสุดท้าย ซึ่งการทำงานของวิธีเลือกแบบใหม่นี้จะเลือกจุดศูนย์กลางเริ่มต้นทีละจุดสลับกับการจัดกลุ่มข้อมูลส่วนผลลัพธ์ที่ได้จากวิธีเลือกแบบใหม่นี้คือ ชุดจุดศูนย์กลางเริ่มต้น และกลุ่มข้อมูลจำนวน K กลุ่ม ขั้นตอนการทำงานทั้งหมดมีดังนี้

วิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่: O(KN)

1. เลือกวัตถุที่มีระยะทางน้อยที่สุดจากพิกัดศูนย์
2. กำหนดจุดศูนย์กลางจากวัตถุที่ถูกเลือกไว้
3. จัดกลุ่มวัตถุทั้งหมดตามระยะทางที่น้อยที่สุดระหว่างข้อมูลกับจุดศูนย์กลางที่ถูกกำหนดไว้
4. เลือกวัตถุที่มีระยะทางมากที่สุดจากทุกกลุ่ม
5. ทำงานซ้ำขั้นตอนที่ 2 จนกว่าจุดศูนย์กลางจะครบ K จุดศูนย์กลาง
6. คำนวณค่าเฉลี่ยของแต่ละกลุ่มเพื่อกำหนดเป็นชุดของจุดศูนย์กลางเริ่มต้น

รูปที่ 3.1 วิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่

ตัวอย่าง อ้างอิงจากข้อมูลยาประเภทต่างๆ (อธิบายในหัวข้อ 2.1.5) เพื่ออธิบายการทำงานของวิธีเลือกจุดศูนย์กลางเริ่มต้นแบบใหม่ โดยยาประเภทต่างๆ มีทั้งหมด 4 ประเภท ซึ่งยาแต่ละประเภทมี 2 คุณลักษณะ คือน้ำหนัก และค่า pH ที่แสดงความเป็นกรดเป็นเบสของสารเคมี โดยจำนวนกลุ่มที่ต้องการแบ่งข้อมูลเท่ากับ 2 กลุ่ม รายละเอียดทั้งหมดแสดงได้ดังนี้

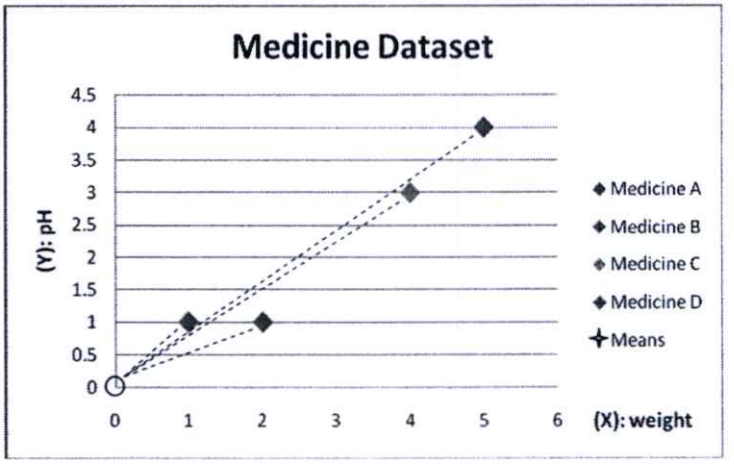
ตารางที่ 3.1 ข้อมูลยาประเภทต่างๆ เมื่อจำนวนกลุ่มที่ต้องการเท่ากับ 2 กลุ่ม

Objects	Attribute1: weight	Attribute2: pH	Clusters: 2
Medicine A	1	1	-
Medicine B	2	1	-
Medicine C	4	3	-
Medicine D	5	4	-

<u>รอบที่ 1</u> ขั้นตอนที่ 1 คือเลือกข้อมูลที่มีระยะทางน้อยที่สุดจากพิกัดศูนย์ โดยอาศัยค่านอร์มแบบยูคลิด (Euclidean norm) ในการหาระยะทางระหว่างข้อมูลกับพิกัดศูนย์ หรือจุดตัดของแนวแกน (อธิบายในหัวข้อ 2.1.3) รายละเอียดทั้งหมดแสดงได้ดังนี้

ตารางที่ 3.2 ระยะทางระหว่างข้อมูลกับพิกัดศูนย์

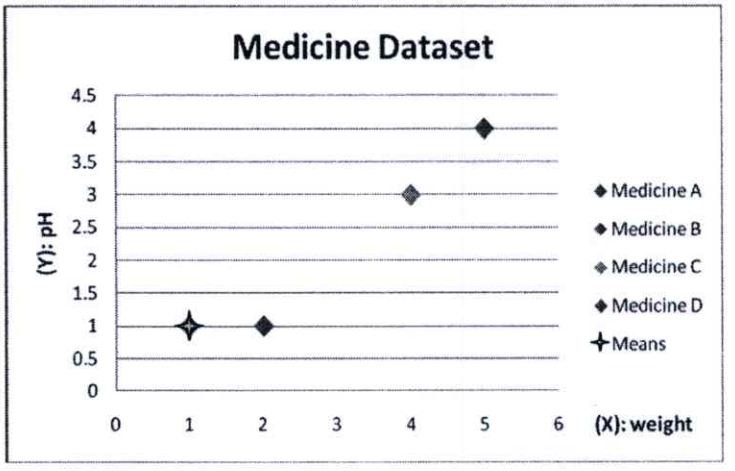
ข้อมูล	กลุ่ม	ค่านอร์มแบบยูคลิด
A	-	$\{(1)^2 + (1)^2\} = 2$
B	-	$\{(2)^2 + (1)^2\} = 5$
C	-	$\{(4)^2 + (3)^2\} = 25$
D	-	$\{(5)^2 + (4)^2\} = 41$



รูปที่ 3.2 รอบที่ 1 ขั้นตอนที่ 1 ของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่

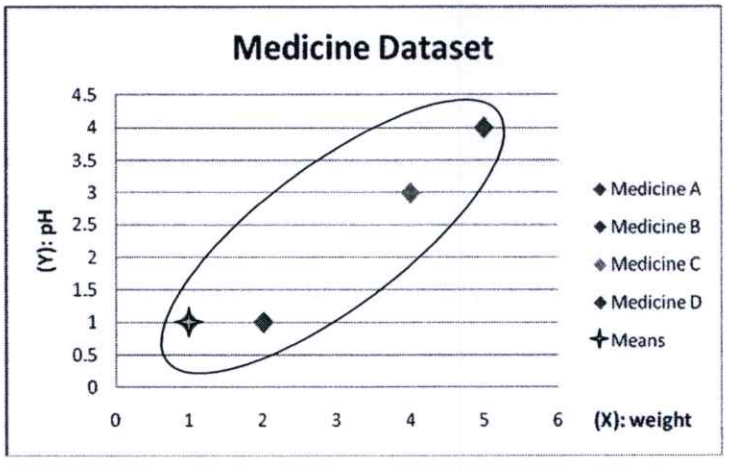
จากตารางที่ 3.2 แสดงให้เห็นการคำนวณระยะทางระหว่างข้อมูลกับพิกัดศูนย์ โดยอาศัยค่านอร์มแบบยูคลิด ซึ่งระยะทางที่วัดได้นั้นมีลักษณะเป็นเส้นตรง และผลลัพธ์ของขั้นตอนนี้คือ ยาประเภท A เพราะมีระยะทางน้อยที่สุดจึงถูกเลือกให้เป็นจุดศูนย์กลางเริ่มต้นจุดแรก ซึ่งในขั้นตอนนี้ยาแต่ละประเภทยังไม่ได้เป็นสมาชิกของกลุ่มใดเลย

<u>รอบที่ 1</u> ขั้นตอนที่ 2 คือกำหนดจุดศูนย์กลางจากข้อมูลที่ถูกเลือกไว้ จากขั้นตอนที่ 1 ยาประเภท A ถูกเลือกไว้ เนื่องจากยาประเภทนี้มีระยะทางจากพิกัดศูนย์น้อยที่สุดเมื่อเปรียบเทียบกับยาประเภทอื่นๆ ดังนั้นจุดศูนย์กลางเริ่มต้นจุดแรกคือ $c_1 = (1, 1)$



รูปที่ 3.3 รอบที่ 1 ขั้นตอนที่ 2 ของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่

<u>รอบที่ 1</u> ขั้นตอนที่ 3 คือจัดกลุ่มข้อมูลทั้งหมดตามระยะทางที่น้อยที่สุดระหว่างข้อมูลกับจุดศูนย์กลางที่ถูกกำหนดไว้ ในตอนนี้จุดศูนย์กลางเริ่มต้นถูกกำหนดไว้เพียงจุดเดียวเท่านั้นคือ $c_1 = (1, 1)$ ดังนั้นยาทุกประเภทจึงถูกจัดไว้ในกลุ่มที่ 1 ทั้งหมด



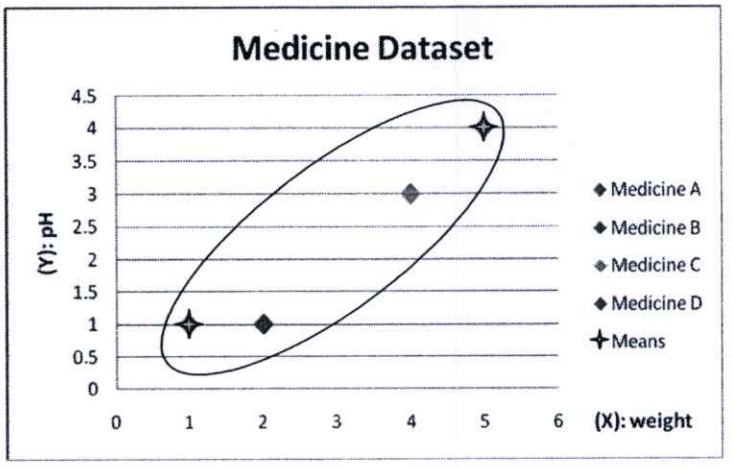
รูปที่ 3.4 รอบที่ 1 ขั้นตอนที่ 3 ของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่

<u>รอบที่ 1</u> ขั้นตอนที่ 4 คือเลือกวัตถุที่มีระยะทางมากที่สุดจากทุกกลุ่ม ในขณะนี้ยาทุกประเภทถูกจัดไว้ในกลุ่มที่ 1 ทั้งหมด ดังนั้นระยะทางระหว่างข้อมูลกับจุดศูนย์กลางของข้อมูลสามารถแสดงได้ดังตารางที่ 3.3 ซึ่งพบว่ายาประเภท D มีระยะทางมากที่สุด จึงถูกเลือกเป็นจุดศูนย์กลางจุดถัดไป

ตารางที่ 3.3 ระยะทางระหว่างข้อมูลกับจุดศูนย์กลางของข้อมูล (รอบที่ 1)

ข้อมูล	กลุ่ม	ระยะทางระหว่างข้อมูลกับจุดศูนย์กลางของข้อมูล
A	1	$\{(1-1)^2 + (1-1)^2\} = 0$
B	1	$\{(2-1)^2 + (1-1)^2\} = 1$
C	1	$\{(4-1)^2 + (3-1)^2\} = 13$
D	1	$\{(5-1)^2 + (4-1)^2\} = 25$

<u>รอบที่ 2</u> ขั้นตอนที่ 2 คือกำหนดจุดศูนย์กลางจากข้อมูลที่ถูกเลือกไว้ จากขั้นตอนที่ 4 ยาประเภท D ถูกเลือกไว้ เนื่องจากยาประเภทนี้มีระยะทางมากที่สุดจากทุกกลุ่ม ซึ่งตอนนี้มีเพียงกลุ่มที่ 1 เท่านั้น ดังนั้นจุดศูนย์กลางเริ่มต้นจุดที่สองคือ $c_2 = (5, 4)$



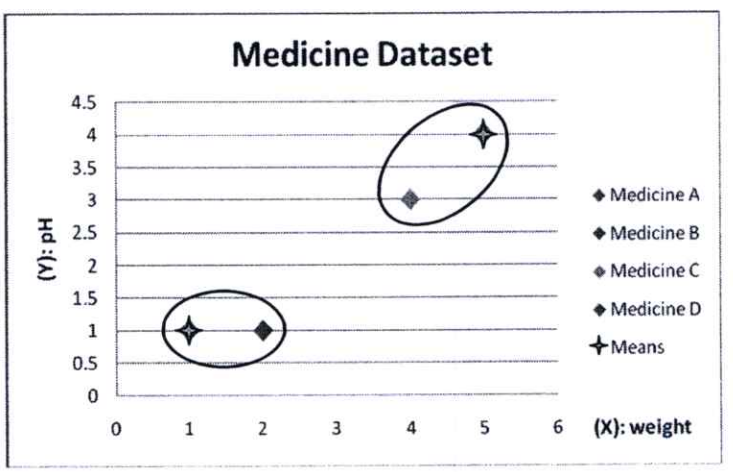
รูปที่ 3.5 รอบที่ 2 ขั้นตอนที่ 2 ของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่

จากรูปที่ 3.5 ใช้รูปทรงวงรีแสดงขอบเขตสมาชิกของแต่ละกลุ่ม ซึ่งตอนนี้ยาทุกประเภทถูกจัดไว้ในกลุ่มที่ 1 เท่านั้น ส่วนกลุ่มที่ 2 ยังไม่มีสมาชิกอยู่เลย เนื่องจากจุดศูนย์กลางเริ่มต้นของกลุ่มที่ 2 พึ่งถูกกำหนดในขั้นตอนนี้ และยังไม่ได้ผ่านขั้นตอนการจัดกลุ่มข้อมูล ซึ่งจะเกิดขึ้นอีกครั้งในขั้นตอนถัดไป

<u>รอบที่ 2</u> ขั้นตอนที่ 3 คือจัดกลุ่มข้อมูลทั้งหมดตามระยะทางที่น้อยที่สุดระหว่างข้อมูลกับจุดศูนย์กลางที่ถูกกำหนดไว้ ถึงขั้นตอนนี้จุดศูนย์กลางเริ่มต้นถูกกำหนดไว้สองจุดแล้ว ยาทุกประเภทจึงต้องคำนวณระยะทางระหว่างข้อมูลกับจุดศูนย์กลางของข้อมูลอีกครั้ง เพื่อค้นหากลุ่มที่เหมาะสมกับข้อมูลนั้น รายละเอียดทั้งหมดแสดงได้ดังนี้

ตารางที่ 3.4 ระยะทางระหว่างข้อมูลกับจุดศูนย์กลางของข้อมูล (รอบที่ 2)

ข้อมูล	กลุ่ม	ระยะทางระหว่างข้อมูลกับจุดศูนย์กลางของข้อมูล
A	1	$\{(1-1)^2 + (1-1)^2\} = 0$
B	1	$\{(2-1)^2 + (1-1)^2\} = 1$
C	2	$\{(4-5)^2 + (3-4)^2\} = 2$
D	2	$\{(5-5)^2 + (4-4)^2\} = 0$



รูปที่ 3.6 รอบที่ 2 ขั้นตอนที่ 3 ของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่

จากตารางที่ 3.4 แสดงให้เห็นการคำนวณระยะทางระหว่างข้อมูลกับจุดศูนย์กลางของข้อมูลเพื่อใช้ในการจัดข้อมูลเข้าสู่กลุ่มที่มีความเหมาะสม โดยพิจารณาจากระยะทางที่น้อยที่สุด ซึ่งพบว่ายาประเภท A และ B ถูกจัดให้อยู่กลุ่มที่ 1 ส่วนยาประเภท C และ D ถูกจัดให้อยู่กลุ่มที่ 2 ผลลัพธ์การจัดกลุ่มข้อมูลครั้งนี้สามารถแสดงได้ดังรูปที่ 3.6

3.2 ขั้นตอนวิธีเคมีนแบบขนานบนระบบคลัสเตอร์

3.2.1 ขั้นตอนวิธี MIM PK-means

ขั้นตอนวิธีเคมีนแบบขนานที่กำหนดให้กระบวนการหลักเลือกชุดจุดศูนย์กลางเริ่มต้น (MIM PK-means) ได้พัฒนาขึ้นตามขั้นตอนวิธีเคมีนแบบขนานจากงานวิจัย [8] (อธิบายในหัวข้อ 2.3.2) และนำวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่มาประยุกต์ใช้ร่วมกัน โดยแบ่งการทำงานทั้งหมดออกเป็น 2 กระบวนการ คือ กระบวนการหลัก และกระบวนการรอง (Master-Slave Process) ส่วนหน่วยประมวลผลที่ใช้ในการทำงานก็แบ่งเป็น 2 ประเภทเช่นเดียวกัน คือ หน่วยประมวลผลหลัก (Master Processor) จำนวน 1 หน่วย และหน่วยประมวลรอง (Slave Processor) จำนวน P-1 หน่วย ซึ่ง P คือจำนวนหน่วยประมวลผลที่ใช้ในการทำงาน โดยหน่วยประมวลผลหลักรับผิดชอบงานของกระบวนการหลักและรอง ส่วนหน่วยประมวลผลรองรับผิดชอบงานของกระบวนการรองเท่านั้น ขั้นตอนการทำงานทั้งหมดมีดังนี้

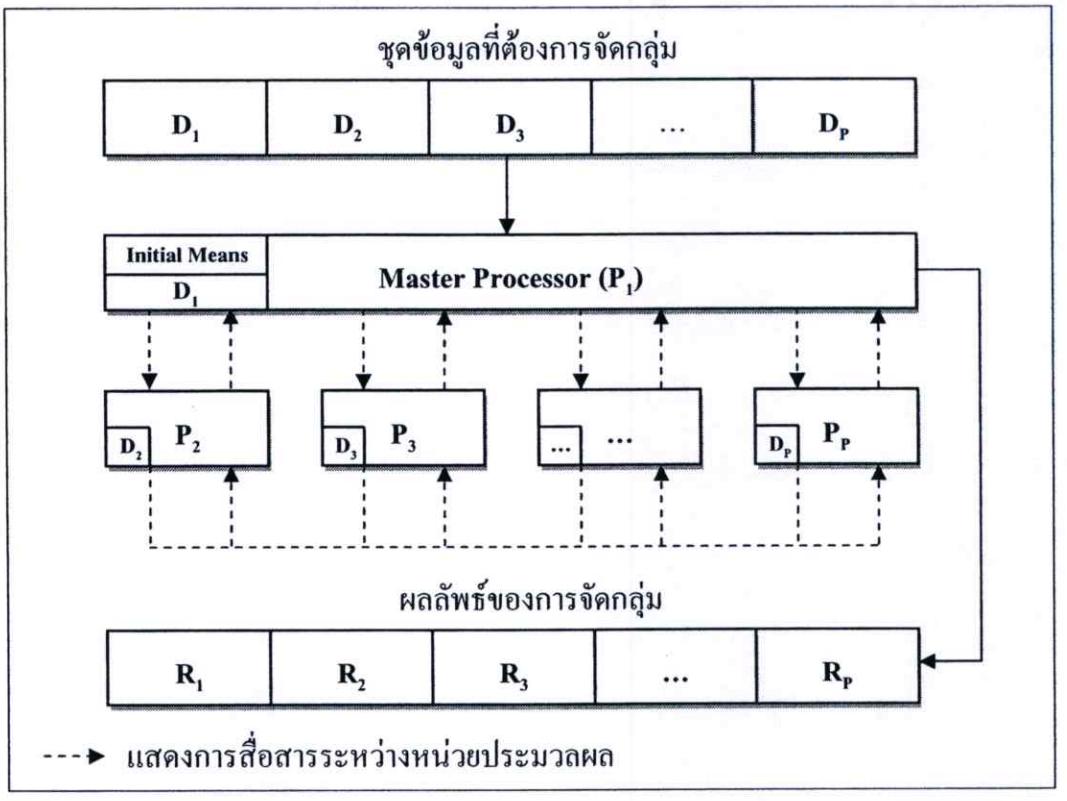
กระบวนการหลัก (Master Process)

ขั้นตอนที่ 1 คืออ่านข้อมูลจากชุดข้อมูล
ขั้นตอนที่ 2 คือแบ่งข้อมูลตามจำนวนหน่วยประมวลผล
ขั้นตอนที่ 3 คือเลือกชุดจุดศูนย์กลางเริ่มต้น
ขั้นตอนที่ 4 คือส่งชุดข้อมูลย่อยไปยังแต่ละกระบวนการรอง
ขั้นตอนที่ 5 คือกระจายชุดจุดศูนย์กลางเริ่มต้นไปยังกระบวนการรองทั้งหมด
ขั้นตอนที่ 6 คือส่งผลลัพธ์การจัดกลุ่มย่อยไปยังแต่ละกระบวนการรอง
ขั้นตอนที่ 7 คือรับผลลัพธ์การจัดกลุ่มจากกระบวนการรองทั้งหมด

กระบวนการรอง (Slave Process)

ขั้นตอนที่ 1 คือรับชุดข้อมูลย่อย และชุดจุดศูนย์กลางเริ่มต้น
ขั้นตอนที่ 2 คือจัดกลุ่มข้อมูล และนับจำนวนข้อมูลที่มีการเปลี่ยนกลุ่ม
ขั้นตอนที่ 3 คือคำนวณผลรวม และนับจำนวนสมาชิกของแต่ละกลุ่ม
ขั้นตอนที่ 4 คือกระจายผลรวม จำนวนสมาชิกของแต่ละกลุ่ม และจำนวนข้อมูลที่มีการเปลี่ยนกลุ่มไปยังกระบวนการรองทั้งหมด
ขั้นตอนที่ 5 คือคำนวณหาค่าเฉลี่ยแล้วกำหนดให้เป็นชุดจุดศูนย์กลางชุดใหม่
ขั้นตอนที่ 6 คือส่งผลลัพธ์การจัดกลุ่มไปยังกระบวนการหลัก

รูปที่ 3.8 ขั้นตอนวิธี MIM PK-means



รูปที่ 3.9 การทำงานร่วมกันระหว่างหน่วยประมวลผลของขั้นตอนวิธี MIM PK-means

จากรูปที่ 3.8 และ 3.9 แสดงการออกแบบ และการทำงานร่วมกันระหว่างหน่วยประมวลผลของขั้นตอนวิธี MIM PK-means โดยหน่วยประมวลผลหลัก (Master Processor) รับผิดชอบงานของทั้งสองกระบวนการ เริ่มต้นด้วยการอ่านข้อมูลจากไฟล์ข้อมูล และแบ่งข้อมูลทั้งหมดออกเป็นชุดข้อมูลย่อยจำนวน P ชุด ซึ่ง P คือ จำนวนหน่วยประมวลผลที่ใช้ในการทำงาน โดยชุดข้อมูลย่อยประกอบด้วยแถวข้อมูลที่อยู่ลำดับติดกัน ซึ่งมีขนาดเท่ากับ N/P (อธิบายในสมการ 2.7) โดยเรียกว่าการแบ่งข้อมูลแบบ Row Block และชุดข้อมูลย่อยทั้งหมดจะมีลำดับดังนี้ D_1, D_2, D_3, ..., D_P จากนั้นเลือกชุดจุดศูนย์กลางเริ่มต้น และส่งชุดข้อมูลย่อยไปเก็บไว้ยังหน่วยประมวลผลตามลำดับเดียวกัน P_1, P_2, P_3, ..., P_P ตามด้วยชุดจุดศูนย์กลางเริ่มต้น และผลลัพธ์การจัดกลุ่มเบื้องต้น G_1, G_2, G_3, …, G_P ที่ได้จากวิธีเลือกจุดศูนย์กลางแบบใหม่ ซึ่งไม่ได้แสดงไว้ในรูปที่ 3.9 หลังจากนั้นหน่วยประมวลผลทุกหน่วยก็เริ่มทำงานภายในกระบวนการรองคือ การจัดกลุ่มข้อมูล สุดท้ายผลลัพธ์การจัดกลุ่มของทุกหน่วยประมวลผลจะถูกส่งไปให้กับหน่วยประมวลผลหลัก (P_1) ตามลำดับดังนี้ R_1, R_2, R_3, ..., R_P เพื่อรวบรวมผลลัพธ์ และบันทึกลงไฟล์ข้อมูล การออกแบบขั้นตอนวิธีดังกล่าว หน่วยประมวลผลหลักต้องทำการเลือกชุดจุดศูนย์กลางเริ่มต้นเพียงฝ่ายเดียวกับชุดข้อมูลที่มีขนาดเท่ากับ N ซึ่งความซับซ้อนด้านเวลาเท่ากับ O(KN) ดังนั้นเมื่อข้อมูลมีขนาดใหญ่ขึ้นเวลาที่ใช้ในส่วนนี้ก็จะมากขึ้นตามไปด้วย แม้ว่าจะเพิ่มจำนวนหน่วยประมวลผลให้มากขึ้นเพียงใดก็ตามเวลาที่ใช้ในส่วนนี้ก็ยังคงเท่าเดิม เหล่านี้เป็นเหตุผลที่ทำให้เกิดการพัฒนาขั้นตอนวิธี SIM PK-means ขึ้น

MIM PK-means Algorithm: O(KN + RK(N/P))

```
Start MPI
  if rank = 0 then  /* rank = 0 is Master processor. */
    // Read data from dataset and Create subdata size N/P.
    // Generate initial means.
    for i = 1 to P-1 do
      MPI_Send (sub_data[i], num_data[i] * dimension, MPI_FLOAT, i, tag, MPI_COMM_WORLD)
      MPI_Send (sub_mem[i], num_data[i], MPI_INT, i, tag, MPI_COMM_WORLD)
    end for
  else if  /* rank ≠ 0 is Slave processor. */
    MPI_Recv (sub_data, num_data* dimension, MPI_FLOAT, 0, tag, MPI_COMM_WORLD, status)
    MPI_Recv (sub_mem, num_data, MPI_FLOAT, 0, tag, MPI_COMM_WORLD, status)
  end if
  MPI_Bcast (means, K* dimension, MPI_FLOAT, 0, MPI_COMM_WORLD)
  MPI_Allreduce (num_data, total_data, 1, MPI_INT, MPI_SUM, MPI_COMM_WORLD)
  while delta > threshold  || loop < 500  /* delta is number of data moved clusters.*/
    delta = 0
    // Assign data into internal clusters.
    // Count number of data moved clusters.
    // Calculate sum and Count number of data into internal clusters.
    MPI_Allreduce (new_means, recv_means, K* dimension, MPI_FLOAT, MPI_SUM, MPI_COMM_WORLD)
    MPI_Allreduce (new_size, recv_size, K, MPI_FLOAT, MPI_SUM, MPI_COMM_WORLD)
    MPI_Allreduce (delta, recv_delta, 1, MPI_FLOAT, MPI_SUM, MPI_COMM_WORLD)
    delta = recv_delta/ total_data
  end  while
  if rank = 0 then
    for i = 1 to P-1 do
      MPI_Recv (mem[i], num_data[i], MPI_FLOAT, i, tag, MPI_COMM_WORLD, status)
    end for
  else if
      MPI_Send (sub_mem, num_data, MPI_INT, 0, tag, MPI_COMM_WORLD)
  end if
End MPI
```

รูปที่ 3.10 ตัวอย่างรหัสเหมือนเอ็มพีไอของขั้นตอนวิธี MIM PK-means

3.2.2 ขั้นตอนวิธี SIM PK-means

ขั้นตอนวิธีเคมีนแบบขนานที่กำหนดให้กระบวนการรองเลือกชุดจุดศูนย์กลางเริ่มต้น (SIM PK-means) มีการออกแบบคล้ายกับขั้นตอนวิธี MIM PK-means โดยแบ่งการทำงานทั้งหมดออกเป็น 2 กระบวนการ คือ กระบวนการหลัก และกระบวนการรอง (Master-Slave Process) ส่วนหน่วยประมวลผลที่ใช้ในการทำงานก็แบ่งเป็น 2 ประเภทเช่นเดียวกัน คือ หน่วยประมวลผลหลัก (Master Processor) จำนวน 1 หน่วย และหน่วยประมวลรอง (Slave Processor) จำนวน P-1 หน่วย ซึ่ง P คือจำนวนหน่วยประมวลผลที่ใช้ในการทำงาน และหน่วยประมวลผลหลักรับผิดชอบงานของทั้งสองกระบวนการ ส่วนหน่วยประมวลผลรองรับผิดชอบงานของกระบวนการรองเพียงอย่างเดียว โดยการเลือกชุดจุดศูนย์กลางเริ่มต้นด้วยวิธีเลือกแบบใหม่ถูกกำหนดไว้ในกระบวนการรอง ดังนั้นทุกหน่วยประมวลผลต้องทำหน้าที่เลือกชุดจุดศูนย์กลางเริ่มต้นร่วมกัน การออกแบบของขั้นตอนวิธีนี้ช่วยลดภาระงานให้กับหน่วยประมวลผลหลักที่เกิดขึ้นในขั้นตอนวิธี MIM PK-means เพื่อทำให้เวลาที่ใช้ในการเลือกชุดจุดศูนย์กลางเริ่มต้นน้อยลงกว่าเดิม ขั้นตอนการทำงานทั้งหมดมีดังนี้

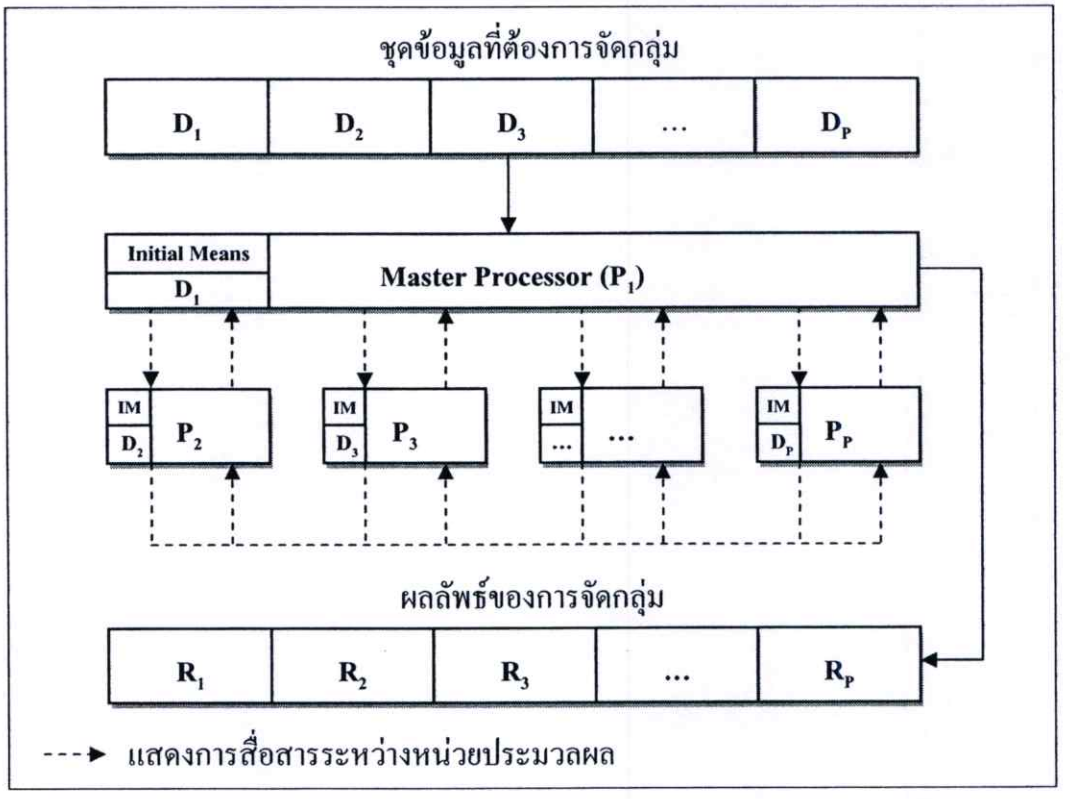
กระบวนการหลัก (Master Process)

ขั้นตอนที่ 1 คืออ่านข้อมูลจากชุดข้อมูล
ขั้นตอนที่ 2 คือแบ่งข้อมูลตามจำนวนหน่วยประมวลผล
ขั้นตอนที่ 3 คือส่งชุดข้อมูลย่อยไปยังแต่ละกระบวนการรอง
ขั้นตอนที่ 4 คือรับผลลัพธ์การจัดกลุ่มจากกระบวนการรองทั้งหมด

กระบวนการรอง (Slave Process)

ขั้นตอนที่ 1 คือรับชุดข้อมูลย่อย
ขั้นตอนที่ 2 คือเลือกชุดจุดศูนย์กลางเริ่มต้น
ขั้นตอนที่ 3 คือจัดกลุ่มข้อมูล และนับจำนวนข้อมูลที่มีการเปลี่ยนกลุ่ม
ขั้นตอนที่ 4 คือคำนวณผลรวม และนับจำนวนสมาชิกของแต่ละกลุ่ม
ขั้นตอนที่ 5 คือกระจายผลรวม จำนวนสมาชิกของแต่ละกลุ่ม และจำนวนข้อมูลที่มีการเปลี่ยนกลุ่มไปยังกระบวนการรองทั้งหมด
ขั้นตอนที่ 6 คือคำนวณหาค่าเฉลี่ยแล้วกำหนดให้เป็นชุดจุดศูนย์กลางชุดใหม่
ขั้นตอนที่ 7 คือส่งผลลัพธ์การจัดกลุ่มไปยังกระบวนการหลัก

รูปที่ 3.11 ขั้นตอนวิธี SIM PK-means



รูปที่ 3.12 การทำงานร่วมกันระหว่างหน่วยประมวลผลของขั้นตอนวิธี SIM PK-means

จากรูปที่ 3.11 และ 3.12 แสดงการออกแบบ และการทำงานร่วมกันระหว่างหน่วยประมวลผลของขั้นตอนวิธี SIM PK-means โดยกำหนดให้หน่วยประมวลผลหลัก (Master Processor: P_1) อ่านข้อมูลจากไฟล์ข้อมูล และแบ่งข้อมูลทั้งหมดออกเป็นชุดข้อมูลย่อยจำนวน P ชุด ซึ่ง P คือจำนวนหน่วยประมวลผลที่ใช้ในการทำงาน โดยชุดข้อมูลย่อยประกอบด้วยแถวข้อมูลที่อยู่ลำดับติดกัน ซึ่งมีขนาดเท่ากับ N/P (อธิบายในสมการ 2.7) โดยเรียกว่า การแบ่งข้อมูลแบบ Row Block และชุดข้อมูลย่อยทั้งหมดจะมีลำดับดังนี้ D_1, D_2, D_3, ..., D_P จากนั้นส่งชุดข้อมูลย่อยไปเก็บไว้ยังหน่วยประมวลผลตามลำดับเดียวกันดังนี้ P_1, P_2, P_3, ..., P_P หลังจากนั้นหน่วยประมวลผลทุกหน่วยจึงเริ่มทำงานภายในกระบวนการรองคือ เลือกชุดจุดศูนย์กลางเริ่มต้นด้วยวิธีเลือกแบบใหม่ ซึ่งจากรูปที่ 3.12 ใช้ตัวอักษร IM แทนชุดจุดศูนย์กลางเริ่มต้น (Initial Means) และทำการจัดกลุ่มข้อมูล จากนั้นผลลัพธ์ที่ได้จากการจัดกลุ่มของทุกหน่วยประมวลผลจะถูกส่งไปให้กับหน่วยประมวลผลหลักตามลำดับดังนี้ R_1, R_2, R_3,..., R_P เพื่อรวบรวม และบันทึกผลลัพธ์ลงไปยังไฟล์ข้อมูล โดยการออกแบบขั้นตอนวิธีดังกล่าว หน่วยประมวลผลทุกหน่วยสามารถทำงานได้อย่างเต็มประสิทธิภาพมากกว่าขั้นตอนวิธี MIM PK-means เนื่องจากทุกหน่วยประมวลผลต้องช่วยกันเลือกชุดจุดศูนย์กลางเริ่มต้นกับชุดข้อมูลย่อยที่มีขนาดเท่ากับ N/P ทำให้ความซับซ้อนด้านเวลาเท่ากับ O(K(N/P)) ซึ่งขนาดของชุดข้อมูลย่อยจะลดลงเมื่อจำนวนหน่วยประมวลผลเพิ่มขึ้น ส่งผลให้เวลาที่ใช้ในส่วนนี้ลดลง และทำให้เวลาทั้งหมดที่ใช้ในการประมวลผลแบบขนานลดลงตามไปด้วย

SIM PK-means Algorithm: O(RK(N/P))

```
Start MPI
  if rank = 0 then  /* rank = 0 is Master processor. */
    // Read data from dataset and Create subdata size N/P.
    for i = 1 to P-1 do
      MPI_Send (sub_data[i], num_data[i] * dimension, MPI_FLOAT, i, tag, MPI_COMM_WORLD)
    end for
  else if /* rank ≠ 0 is Slave processor. */
    MPI_Recv (sub_data, num_data* dimension, MPI_FLOAT, 0, tag, MPI_COMM_WORLD, status)
  end if
  for i = 0 to K-1 then
    // Generate initial mean c_i.
    MPI_Bcast (means[i], dimension, MPI_FLOAT, rank, MPI_COMM_WORLD)
    // Assign data into new cluster.
  end for
  // Calculate initial means.
  MPI_Allreduce (num_data, total_data, 1, MPI_INT, MPI_SUM, MPI_COMM_WORLD)
  while delta > threshold || loop < 500  /* delta is number of data moved clusters.*/
    delta = 0
    //  Assign data into internal clusters
    // Count number of data moved clusters.
    // Calculate sum and Count number of data into internal clusters.
    MPI_Allreduce (new_means, recv_means, K* dimension, MPI_FLOAT, MPI_SUM, MPI_COMM_WORLD)
    MPI_Allreduce (new_size, recv_size, K, MPI_FLOAT, MPI_SUM, MPI_COMM_WORLD)
    MPI_Allreduce (delta, recv_delta, 1, MPI_FLOAT, MPI_SUM, MPI_COMM_WORLD)
    delta = recv_delta/ total_data
  end  while
  if rank = 0 then
    for i = 1 to P-1 do
      MPI_Recv (mem[i], num_data[i], MPI_INT, i, tag, MPI_COMM_WORLD, status)
    end for
  else if
      MPI_Send (sub_mem, num_data, MPI_INT, 0, tag, MPI_COMM_WORLD)
  end if
End MPI
```

รูปที่ 3.13 ตัวอย่างรหัสเหมือนเอ็มพีไอของขั้นตอนวิธี SIM PK-means

3.3 การวิเคราะห์ความซับซ้อนด้านเวลา

การวิเคราะห์ความซับซ้อนด้านเวลา (Time Complexity Analysis) ของขั้นตอนวิธีแบบขนานบนระบบคลัสเตอร์นั้น เวลาทั้งหมดที่ใช้ในการประมวลผลแบบขนาน (Parallel Execution Time) จะแบ่งออกเป็นเวลา 2 ส่วน คือเวลาที่ใช้ในการประมวลผลแบบขนานด้วย P หน่วยประมวลผล หรือการคำนวณ (Computation Time) และเวลาที่ใช้ในการติดต่อสื่อสารระหว่างหน่วยประมวลผล (Communication Time) ซึ่งสามารถคำนวณเวลาเหล่านี้ได้ตามสมการดังนี้

$$T_P = T_{comm} + T_{comp}$$

เมื่อ T_P คือ เวลาทั้งหมดที่ใช้ในการประมวลผลแบบขนานด้วย P หน่วยประมวลผล
T_{comm} คือ เวลาที่ใช้ในการสื่อสารระหว่างหน่วยประมวลผล
T_{comp} คือ เวลาที่ใช้ในการคำนวณ

ซึ่งสมการทั่วไปของเวลาที่ใช้ในการสื่อสารระหว่างหน่วยประมวลผลมีรูปแบบดังนี้

$$T_{comm} = T_{startup} + (N/P)T_{data}$$

เมื่อ $T_{startup}$ คือ ค่าคงที่ของเวลาที่ใช้ในการจัดการกับข้อมูล และเริ่มทำการส่งข้อมูล
T_{data} คือ ค่าคงที่ของเวลาที่ใช้ในการส่งข้อมูล 1 หน่วย
N คือ จำนวนข้อมูลทั้งหมด

ในงานวิจัยนี้ได้นำเสนอขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์ทั้งหมด 2 วิธี คือ ขั้นตอนวิธี MIM PK-means และขั้นตอนวิธี SIM PK-means ซึ่งสามารถทำการวิเคราะห์ความซับซ้อนด้านเวลาของขั้นตอนวิธีแบบขนานดังกล่าวได้ในหัวข้อถัดไป

3.3.1 ความซับซ้อนด้านเวลาของขั้นตอนวิธี MIM PK-means

ขั้นตอนวิธีเคมีนแบบขนานที่กำหนดให้กระบวนการหลักเลือกชุดจุดศูนย์กลางเริ่มต้น (MIM PK-means) มีเวลาที่ใช้ในการคำนวณ และเวลาที่ใช้ในการสื่อสารระหว่างหน่วยประมวลผลทั้งหมด 11 ขั้นตอน โดยไม่รวมเวลาที่ใช้ในการแบ่งชุดข้อมูลย่อย การย้ายข้อมูลมาเก็บไว้ในหน่วยความจำสำรอง (Hard disk) การตั้งค่าระบบ (Initialization) ของ MPI และการบันทึกผลลัพธ์การจัดกลุ่มลงไฟล์ข้อมูล ซึ่งกระบวนการดังกล่าวถูกทำเพียงครั้งเดียวเท่านั้น ขั้นตอนทั้งหมดมีดังนี้

ขั้นตอนที่ 1 คือ หน่วยประมวลผลหลัก (Master Processor) ทำการเลือกชุดจุดศูนย์กลางเริ่มต้นตามจำนวนกลุ่มที่ต้องการ (K) ด้วยวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่ที่นำเสนอในงานวิจัยนี้ โดยชุดข้อมูลที่ต้องการจัดกลุ่มมีขนาดเท่ากับ nxD ซึ่ง n คือจำนวนข้อมูลทั้งหมด และ D คือจำนวนคุณลักษณะ หรือมิติของข้อมูลเหล่านั้น ซึ่งชุดข้อมูลสามารถเขียนแทนด้วย N

$$T_{comp1} = KN$$

ขั้นตอนที่ 2 คือ หน่วยประมวลผลหลักส่งชุดข้อมูลย่อยขนาด N/P ไปยังแต่ละหน่วยประมวลผลรอง หรือหน่วยประมวลผลทำงาน ซึ่ง P คือจำนวนหน่วยประมวลผลทั้งหมด โดยทำการส่งทั้งหมด P-1 ครั้ง เนื่องจากต้องคงเหลือข้อมูลจำนวนหนึ่งให้หน่วยประมวลผลหลักทำงานด้วย

$$T_{comm1} = P(T_{startup} + (N/P)T_{data})$$

ขั้นตอนที่ 3 คือ หน่วยประมวลผลหลักทำการกระจายชุดจุดศูนย์กลางเริ่มต้นที่เลือกไว้ไปยังหน่วยประมวลผลรองทั้งหมด ซึ่งชุดจุดศูนย์กลางเริ่มต้นมีขนาดเท่ากับ KxD และเขียนแทนด้วย C

$$T_{comm2} = P(T_{startup} + CT_{data})$$

ขั้นตอนที่ 4 คือ หน่วยประมวลผลหลักทำการส่งผลลัพธ์การจัดกลุ่มย่อยขนาด G/P ที่ได้จากการเลือกชุดจุดศูนย์กลางเริ่มต้นไปยังแต่ละหน่วยประมวลผลรอง ซึ่ง G คือผลลัพธ์การจัดกลุ่มทั้งหมดและขนาดของผลลัพธ์เท่ากับ n โดย n คือจำนวนข้อมูลทั้งหมดในชุดข้อมูล

$$T_{comm3} = P(T_{startup} + (G/P)T_{data})$$

ขั้นตอนที่ 5 คือ แต่ละหน่วยประมวลผลนำชุดข้อมูลย่อยและชุดจุดศูนย์กลางเริ่มต้นมาจัดกลุ่ม และนับจำนวนข้อมูลที่มีการเปลี่ยนกลุ่ม

$$T_{comp2} = K(N/P)$$

ขั้นตอนที่ 6 คือ แต่ละหน่วยประมวลผลคำนวณผลรวม และนับจำนวนสมาชิกของแต่ละกลุ่ม

$$T_{comp3} = N/P$$

ขั้นตอนที่ 7 คือ แต่ละหน่วยประมวลผลทำการกระจายผลรวมของสมาชิกในแต่ละกลุ่ม ซึ่งมีขนาดเท่ากับชุดจุดศูนย์กลางเริ่มต้น (C) ไปยังหน่วยประมวลผลทั้งหมด

$$T_{comm4} = P^2(T_{startup} + CT_{data})$$

ขั้นตอนที่ 8 คือ แต่ละหน่วยประมวลผลทำการกระจายจำนวนสมาชิกของแต่ละกลุ่ม ซึ่งมีขนาดเท่ากับจำนวนกลุ่มที่ต้องการ (K) ไปยังหน่วยประมวลผลทั้งหมด

$$T_{comm5} = P^2(T_{startup} + KT_{data})$$

ขั้นตอนที่ 9 คือ แต่ละหน่วยประมวลผลทำการกระจายจำนวนข้อมูลที่มีการเปลี่ยนกลุ่มไปยังหน่วยประมวลผลทั้งหมด

$$T_{comm6} = P^2(T_{startup} + T_{data})$$

ขั้นตอนที่ 10 คือ แต่ละหน่วยประมวลผลนำผลรวม และจำนวนสมาชิกของแต่ละกลุ่มมาคำนวณหาค่าเฉลี่ยแล้วกำหนดให้เป็นชุดจุดศูนย์กลางชุดใหม่

$$T_{comp4} = C$$

ขั้นตอนที่ 11 คือ แต่ละหน่วยประมวลผลส่งผลลัพธ์การจัดกลุ่มกลับไปยังหน่วยประมวลผลหลัก ซึ่งหน่วยประมวลผลหลักต้องรอรับผลลัพธ์ของการจัดกลุ่มจากหน่วยประมวลผลรองทั้งหมด P-1 ครั้ง เนื่องจากผลลัพธ์จำนวนหนึ่งได้จากหน่วยประมวลผลหลัก

$$T_{comm7} = P(T_{startup} + (G/P)T_{data})$$

จากขั้นตอนวิธี MIM PK-means (อธิบายในหัวข้อ 3.2.1) พบว่ากระบวนการรองมีการทำงานซ้ำตั้งแต่ขั้นตอนที่ 2 จนถึงขั้นตอนที่ 5 ซึ่งขั้นตอนที่ 2 คือการจัดกลุ่มข้อมูล และนับจำนวนข้อมูลที่มีการเปลี่ยนกลุ่ม ส่วนขั้นตอนที่ 5 คือการคำนวณหาค่าเฉลี่ยแล้วกำหนดให้เป็นชุดจุดศูนย์กลางชุดใหม่เพื่อใช้จัดกลุ่มข้อมูลในรอบถัดไป เวลาทั้งหมดที่ใช้ในการประมวลผลแบบขนานด้วย P หน่วยประมวลผล สามารถคำนวณได้ดังนี้

เวลาที่ใช้ในการสื่อสารระหว่างหน่วยประมวลผล (Communication Time: T_{comm})

$$\begin{aligned}
\mathbf{T_{comm}} &= T_{comm1} + T_{comm2} + T_{comm3} + T_{comm4} + T_{comm5} + T_{comm6} + T_{comm7} \\
&= P(T_{startup} + (N/P)T_{data}) + P(T_{startup} + CT_{data}) + P(T_{startup} + (G/P)T_{data}) + \\
&\quad P^2R(T_{startup} + CT_{data}) + P^2R(T_{startup} + KT_{data}) + P^2R(T_{startup} + T_{data}) + \\
&\quad P(T_{startup} + (G/P)T_{data}) \\
&= (4P + 3P^2R)T_{startup} + (N + PC + 2G + P^2R(C + K + 1))T_{data} \\
&= O(N + P^2RC)
\end{aligned}$$

เวลาที่ใช้ในการคำนวณ (Computation Time: T_{comp})

$$\begin{aligned}
\mathbf{T_{comp}} &= T_{comp1} + T_{comp2} + T_{comp3} + T_{comp4} \\
&= KN + RK(N/P) + R(N/P) + RC \\
&= O(KN + RK(N/P))
\end{aligned}$$

ดังนั้น เวลาทั้งหมดที่ใช้ในการประมวลผลแบบขนาน(Parallel Execution Time: T_P) สำหรับขั้นตอนวิธี MIM PK-means คือ

$$\begin{aligned}
\mathbf{T_P} &= T_{comm} + T_{comp} \\
&= O(N + P^2RC) + O(KN + RK(N/P)) \\
&= O(KN + RK(N/P))
\end{aligned}$$

จากเวลาทั้งหมดที่ใช้ในการประมวลผลแบบขนานด้วย P หน่วยประมวลผล พบว่าเวลาทั้งหมดขึ้นอยู่กับที่เวลาที่ใช้ในการคำนวณ โดยเวลาที่ใช้ในการคำนวณส่วนใหญ่เกิดขึ้นจากการเลือกจุดศูนย์กลางเริ่มต้นด้วยวิธีเลือกจุดศูนย์กลางเริ่มต้นแบบใหม่ ซึ่งหน่วยประมวลหลักต้องเลือกชุดจุดศูนย์กลางเริ่มต้นกับชุดข้อมูลขนาดเท่ากับ N และการจัดกลุ่มข้อมูล ซึ่งทุกหน่วยประมวลผลร่วมกันทำงานกับชุดข้อมูลย่อยขนาดเท่ากับ N/P โดย R คือจำนวนรอบของการทำงานซ้ำ ซึ่งไม่สามารถคาดเดาได้ว่าการจัดกลุ่มข้อมูลจะหยุดทำงานตามเงื่อนไขที่กำหนดไว้เมื่อใด

3.3.2 ความซับซ้อนด้านเวลาของขั้นตอนวิธี SIM PK-means

ขั้นตอนวิธีเคมีนแบบขนานที่กำหนดให้กระบวนการรองเลือกชุดจุดศูนย์กลางเริ่มต้น (SIM PK-means) มีเวลาที่ใช้ในการคำนวณ และเวลาที่ใช้ในการสื่อสารระหว่างหน่วยประมวลผลทั้งหมด 9 ขั้นตอน โดยไม่รวมเวลาที่ใช้ในการแบ่งชุดข้อมูลย่อย การย้ายข้อมูลมาเก็บไว้ในหน่วยความจำสำรอง (Hard disk) การตั้งค่าระบบ (Initialization) ของ MPI และการบันทึกผลลัพธ์การจัดกลุ่มลงไฟล์ข้อมูล ซึ่งกระบวนการดังกล่าวถูกทำเพียงครั้งเดียวเท่านั้น ขั้นตอนทั้งหมดมีดังนี้

ขั้นตอนที่ 1 คือ หน่วยประมวลผลหลัก (Master Processor) ทำการส่งชุดข้อมูลย่อยขนาด N/P ไปยังแต่ละหน่วยประมวลผลรอง โดยชุดข้อมูล (N) มีขนาดเท่ากับ nxD ซึ่ง n คือจำนวนข้อมูลทั้งหมด D คือจำนวนคุณลักษณะ และ P คือจำนวนหน่วยประมวลผลทั้งหมด โดยทำการส่งทั้งหมด P-1 ครั้ง เนื่องจากต้องคงเหลือข้อมูลจำนวนหนึ่งให้กับหน่วยประมวลผลหลักทำงานด้วย

$$T_{comm1} = P(T_{startup} + (N/P)T_{data})$$

ขั้นตอนที่ 2 คือ แต่ละหน่วยประมวลผลเลือกชุดจุดศูนย์กลางเริ่มต้นตามจำนวนกลุ่มที่ต้องการ (K) ด้วยวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่ ซึ่งวิธีเลือกแบบใหม่นี้ทำการเลือกจุดศูนย์กลางเริ่มต้นทีละจุดสลับกับการจัดกลุ่มข้อมูล ดังนั้นระหว่างการเลือกชุดจุดศูนย์กลางเริ่มต้นแต่ละจุดต้องมีการสื่อสารระหว่างหน่วยประมวลผลเกิดขึ้นทุกครั้งจนกว่าจุดศูนย์กลางเริ่มต้นจะครบ K จุด

$$T_{comp1} = K(N/P)$$

และ

$$T_{comm2} = P^2K(T_{startup} + T_{data})$$

$$T_{comm3} = PK(T_{startup} + (C/K)T_{data})$$

$$T_{comm4} = P^2(T_{startup} + CT_{data})$$

$$T_{comm5} = P^2(T_{startup} + KT_{data})$$

ขั้นตอนที่ 3 คือ แต่ละหน่วยประมวลผลจัดกลุ่มข้อมูลตามจำนวนกลุ่มที่ต้องการ และนับจำนวนข้อมูลที่มีการเปลี่ยนกลุ่ม

$$T_{comp2} = K(N/P)$$

ขั้นตอนที่ 4 คือ แต่ละหน่วยประมวลผลคำนวณผลรวม และนับจำนวนสมาชิกของแต่ละกลุ่ม

$$T_{comp3} = N/P$$

ขั้นตอนที่ 5 คือ แต่ละหน่วยประมวลผลทำการกระจายผลรวมของสมาชิกในแต่ละกลุ่ม ซึ่งมีขนาดเท่ากับชุดจุดศูนย์กลางเริ่มต้น (C) ไปยังหน่วยประมวลผลทั้งหมด ซึ่งชุดจุดศูนย์กลางเริ่มต้นมีขนาดเท่ากับ KxD

$$T_{comm6} = P^2(T_{startup} + CT_{data})$$

ขั้นตอนที่ 6 คือ แต่ละหน่วยประมวลผลทำการกระจายจำนวนสมาชิกของแต่ละกลุ่ม ซึ่งมีขนาดเท่ากับจำนวนกลุ่มที่ต้องการ (K) ไปยังหน่วยประมวลผลทั้งหมด

$$T_{comm7} = P^2(T_{startup} + KT_{data})$$

ขั้นตอนที่ 7 คือ แต่ละหน่วยประมวลผลทำการกระจายจำนวนข้อมูลที่มีการเปลี่ยนกลุ่มไปยังหน่วยประมวลผลทั้งหมด

$$T_{comm8} = P^2(T_{startup} + T_{data})$$

ขั้นตอนที่ 8 คือ แต่ละหน่วยประมวลผลนำผลรวม และจำนวนสมาชิกของแต่ละกลุ่มมาคำนวณหาค่าเฉลี่ยแล้วกำหนดให้เป็นชุดจุดศูนย์กลางชุดใหม่

$$T_{comp4} = C$$

ขั้นตอนที่ 9 คือ แต่ละหน่วยประมวลผลส่งผลลัพธ์การจัดกลุ่มกลับไปยังหน่วยประมวลผลหลัก ซึ่งหน่วยประมวลหลักต้องรอรับผลลัพธ์ของการจัดกลุ่มจากหน่วยประมวลผลรองทั้งหมด P-1 ครั้ง เนื่องจากผลลัพธ์จำนวนหนึ่งได้จากหน่วยประมวลหลัก โดยผลลัพธ์การจัดกลุ่มมีขนาดเท่ากับ n ซึ่ง n คือจำนวนข้อมูลทั้งหมดในชุดข้อมูล และผลลัพธ์การจัดกลุ่มเขียนแทนด้วย G

$$T_{comm9} = P(T_{startup} + (G/P)T_{data})$$

จากขั้นตอนวิธี SIM PK-means (อธิบายในหัวข้อ 3.2.2) พบว่ากระบวนการรองมีการทำงานซ้ำตั้งแต่ขั้นตอนที่ 3 จนถึงขั้นตอนที่ 6 ซึ่งขั้นตอนที่ 3 คือการจัดกลุ่มข้อมูล และนับจำนวนข้อมูลที่มีการเปลี่ยนกลุ่ม ส่วนขั้นตอนที่ 6 คือการคำนวณหาค่าเฉลี่ยแล้วกำหนดให้เป็นชุดจุดศูนย์กลางชุดใหม่เพื่อใช้จัดกลุ่มข้อมูลในรอบถัดไป เวลาทั้งหมดที่ใช้ในการประมวลผลแบบขนานด้วย P หน่วยประมวลผล สามารถคำนวณได้ดังนี้

เวลาที่ใช้ในการสื่อสารระหว่างหน่วยประมวลผล (Communication Time: T_{comm})

$$\begin{aligned}
\mathbf{T_{comm}} &= T_{comm1} + T_{comm2} + T_{comm3} + T_{comm4} + T_{comm5} + T_{comm6} + T_{comm7} + T_{comm8} + T_{comm9} \\
&= P(T_{startup} + (N/P)T_{data}) + P^2K(T_{startup} + T_{data}) + PK(T_{startup} + (C/K)T_{data}) + P^2(T_{startup} + CT_{data}) + P^2(T_{startup} + KT_{data}) + P^2R(T_{startup} + CT_{data}) + P^2R(T_{startup} + KT_{data}) + P^2R(T_{startup} + T_{data}) + P(T_{startup} + (G/P)T_{data}) \\
&= (2P + PK + P^2(3R + K + 2)T_{startup} + (N + PC + G + P^2(RC + RK + 2K + C + R)T_{data} \\
&= O(N + P^2RC)
\end{aligned}$$

เวลาที่ใช้ในการคำนวณ (Computation Time: T_{comp})

$$\begin{aligned}
\mathbf{T_{comp}} &= T_{comp1} + T_{comp2} + T_{comp3} + T_{comp4} \\
&= K(N/P) + RK(N/P) + R(N/P) + RC \\
&= O(RK(N/P))
\end{aligned}$$

ดังนั้น เวลาทั้งหมดที่ใช้ในการประมวลผลแบบขนาน(Parallel Execution Time: T_P) สำหรับขั้นตอนวิธี SIM PK-means คือ

$$\begin{aligned}
\mathbf{T_P} &= T_{comm} + T_{comp} \\
&= O(N + P^2RC) + O(RK(N/P)) \\
&= O(RK(N/P))
\end{aligned}$$

จากเวลาทั้งหมดที่ใช้ในการประมวลผลแบบขนานด้วย P หน่วยประมวลผล พบว่าเวลาทั้งหมดขึ้นอยู่กับที่เวลาที่ใช้ในการคำนวณ โดยเวลาที่ใช้ในการคำนวณส่วนใหญ่เกิดขึ้นจาก การจัดกลุ่มข้อมูลเพียงอย่างเดียว ซึ่งแตกต่างจากขั้นตอนวิธี MIM PK-means เนื่องจากขั้นตอนวิธีนี้ (SIM PK-means) เลือกชุดจุดศูนย์กลางเริ่มต้นโดยใช้ทุกหน่วยประมวลผลร่วมกัน ซึ่งเลือกจากชุดข้อมูลย่อยขนาด N/P เท่านั้น แต่ก็ต้องเพิ่มเวลาที่ใช้ในการสื่อสารระหว่างการเลือกจุดศูนย์กลางแต่ละจุดด้วยเช่นกันดังขั้นตอนที่ 2 เพื่อเลือกชุดจุดศูนย์กลางเริ่มต้นตามหลักการของวิธีเลือกแบบใหม่

ตารางที่ 3.6 เวลาทั้งหมดที่ใช้ในการประมวลผลแบบขนานของขั้นตอนวิธีที่นำเสนอ

การวิเคราะห์ประสิทธิภาพ	ขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนาน	
	MIM PK-means	**SIM PK-means**
เวลาที่ใช้ในการสื่อสารตอนเลือกชุดจุดศูนย์กลางเริ่มต้น	O(PKD)	$O(P^2KD)$
เวลาที่ใช้ในการคำนวณตอนเลือกชุดจุดศูนย์กลางเริ่มต้น	O(KN)	O(K(N/P))
เวลาทั้งหมดที่ใช้ในการสื่อสาร	$O(N + P^2RKD)$	$O(N + P^2RKD)$
เวลาทั้งหมดที่ใช้ในการคำนวณ	O(KN + RK(N/P))	O(RK(N/P))

จากตารางที่ 3.6 แสดงเวลาทั้งหมดที่ใช้ในการประมวลผลแบบขนาน พบว่าความซับซ้อนด้านเวลาทั้งหมดที่ใช้ในการสื่อสาร และการคำนวณของขั้นตอนวิธีทั้งสองใกล้เคียงกัน มีเพียงเวลาที่ใช้ในการสื่อสาร และการคำนวณของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นเท่านั้นที่แตกต่างกันอย่างชัดเจน เนื่องจากขั้นตอนวิธี MIM PK-means ออกแบบให้หน่วยประมวลผลหลักเลือกชุดจุดศูนย์กลางเริ่มต้นจากชุดข้อมูลเพียงฝ่ายเดียว ทำให้เวลาที่ใช้ในการคำนวณมากกว่าขั้นตอนวิธี SIM PK-means คือ O(KN) และ O(K(N/P)) ตามลำดับ แต่ขั้นตอนวิธีนี้ใช้เวลาในการสื่อสารตอนเลือกชุดจุดศูนย์กลางเริ่มต้นน้อยกว่าขั้นตอนวิธี SIM PK-means คือ O(PKD) และ$O(P^2KD)$ ตามลำดับ ซึ่ง K คือจำนวนกลุ่มที่ต้องการ และ D คือจำนวนคุณลักษณะของข้อมูล ดังนั้น KD ก็คือขนาดของชุดจุดศูนย์กลาง (C) นั้นเอง เนื่องจากหน่วยประมวลผลหลักทำการกระจายชุดจุดศูนย์กลางเริ่มต้นไปยังหน่วยประมวลผลทั้งหมดเพียงอย่างเดียวเท่านั้น แต่ขั้นตอนวิธี SIM PK-means ต้องทำการสื่อสารกันทุกครั้งขณะเลือกจุดศูนย์กลางเริ่มต้นแต่ละจุด ด้วยเหตุนี้ทำให้ขั้นตอนวิธี SIM PK-means มีเวลาที่ใช้ในการสื่อสารระหว่างหน่วยประมวลผลในส่วนนี้มากกว่าขั้นตอนวิธี MIM PK-means

3.4 วิเคราะห์ความซับซ้อนด้านหน่วยความจำ

การวิเคราะห์ความซับซ้อนด้านหน่วยความจำ (Space Complexity Analysis) ทั้งหมดที่จำเป็นต้องใช้ในการประมวลผลของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์ทั้ง 2 วิธีที่นำเสนอในงานวิจัยนี้ คือ ขั้นตอนวิธี MIM PK-means และขั้นตอนวิธี SIM PK-means เท่ากับ O(N) ซึ่ง N คือ ขนาดของชุดข้อมูล เนื่องจากขั้นตอนวิธีทั้งสองแบ่งหน่วยประมวลผลออกเป็น 2 ประเภท โดยหน่วยประมวลผลหลักทำการอ่านชุดข้อมูลและแบ่งข้อมูลทั้งหมดออกเป็น P ชุด ซึ่ง P คือ จำนวนหน่วยประมวลผลที่ใช้ในการทำงาน โดยข้อมูลแต่ละชุดเรียกว่า ชุดข้อมูลย่อย ซึ่งมีขนาดเท่ากับ N/P และจัดส่งชุดข้อมูลย่อยไปเก็บไว้ในหน่วยความจำของแต่ละหน่วยประมวลผลรองจำนวน P-1 ครั้ง เพราะชุดข้อมูลย่อยจำนวนหนึ่งชุดจะถูกเก็บไว้ในหน่วยประมวลผลหลัก ดังนั้นหน่วยความจำทั้งหมดที่จำเป็นต้องใช้ในการประมวลผลจึงเท่ากับ O(N)

บทที่ 4

การทดลองและผลการทดลอง

งานวิจัยนี้ได้นำเสนอขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์ ซึ่งได้นำเสนอขั้นตอนวิธีนี้ไว้ในบทที่ 3 สำหรับเนื้อหาในบทนี้จะกล่าวถึงการทดลองและผลการทดลองของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์ ซึ่งแบ่งการทดลองออกเป็น 2 ส่วน คือ การทดลองของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้น และการทดลองของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์

4.1 เครื่องมือที่ใช้ในการทดลอง

เครื่องมือที่ใช้ในการทดลอง คือระบบคลัสเตอร์ของสำนักวิจัยและบริการคอมพิวเตอร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ซึ่งเป็นระบบคลัสเตอร์แบบเนื้อเดียว (Homogenous Cluster) ที่มีองค์ประกอบทางด้านฮาร์ดแวร์ และซอฟต์แวร์ที่ใช้ในแต่ละเครื่องเหมือนกันทั้งหมด และเชื่อมต่อกับเครือข่ายภายนอกผ่านเครื่อง Front-end คือเครื่องคอมพิวเตอร์เครื่องหนึ่งที่เป็นทางเข้า-ทางออกให้แก่ระบบ ซึ่งระบบคลัสเตอร์ดังกล่าวมีคุณลักษณะดังนี้

4.1.1 ฮาร์ดแวร์

จำนวนเครื่องคอมพิวเตอร์	:	5 เครื่อง
หน่วยประมวลผลกลาง (CPU)	:	Intel(R) Xeon(TM) Processor 2.80 GHz
จำนวนหน่วยประมวลผลกลาง	:	20 หน่วย
หน่วยความจำหลัก (RAM)	:	4 กิกะไบต์
หน่วยความจำสำรอง (Hard Disk)	:	80 กิกะไบต์
เครือข่ายอีเทอร์เน็ต (Ethernet)	:	1000 เมกกะบิตต่อวินาที

4.1.2 ซอฟต์แวร์

ระบบปฏิบัติการ (OS)	:	Rocks Clusters 4.2.1
โปรแกรมอรรถประโยชน์	:	EditPlus 2
ชุดคำสั่งและคอมไพเลอร์	:	MPICH 1.2

4.2 ชุดข้อมูล

ชุดข้อมูลทั้งหมดที่ใช้ในการทดลองนำมาจากงานวิจัย [2] ซึ่งสามารถดาวน์โหลดชุดข้อมูลเหล่านั้นได้จากเว็บไซต์นี้ http://archive.ics.uci.edu/ml/ เพื่อนำมาใช้ในการวัดประสิทธิภาพของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้น และขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์ รายละเอียดของชุดข้อมูลทั้งหมดมีดังนี้

ตารางที่ 4.1 ชุดข้อมูลที่ใช้ในการทดลอง

ชื่อชุดข้อมูล	**ขนาดของชุดข้อมูล** (ไบต์)	**ลักษณะของชุดข้อมูล** (จำนวน x คุณลักษณะ)
Cloud.txt	108,461	1024 x 10
Spam.txt	734,043	4601 x 58
Intrusion.txt	66,052,485	494019 x 35

4.3 การทดลอง

การทดลองของงานวิจัยนี้แบ่งออกเป็น 2 ส่วน คือ การทดลองของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้น และการทดลองของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์ โดยพัฒนาโปรแกรมเหล่านี้ด้วยภาษาซีตามมาตรฐานเอ็มพีไอ และทำการทดสอบโปรแกรมกับชุดข้อมูลที่กล่าวมาแล้วข้างต้นบนระบบคลัสเตอร์ของสำนักวิจัยและบริการคอมพิวเตอร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง โดยนำหลักการออกแบบโปรแกรมเคมีนแบบอนุกรมของ J. MacQueen (อธิบายในหัวข้อ 2.1.6) มาใช้ในการพัฒนาโปรแกรมทั้งหมด

การกำหนดค่าในการทดลองครั้งนี้ เริ่มจากการกำหนดเงื่อนไขที่ใช้สำหรับหยุดการจัดกลุ่มข้อมูล โดยให้มีค่า Threshold เท่ากับ 0.01 ซึ่งค่านี้เป็นค่าสมมุติที่ตั้งไว้เพื่อเป็นจุดตรวจสอบเท่านั้น และจำนวนรอบของการทำงานซ้ำทั้งหมดต้องไม่เกิน 500 รอบ โดยอ้างอิงค่าเหล่านี้จากงานวิจัย [8] ส่วนสุดท้ายคือ จำนวนกลุ่มที่ต้องการ (K) ซึ่งเป็นข้อมูลนำเข้าของขั้นตอนวิธีเคมีน โดยกำหนดให้จำนวนกลุ่มที่ต้องการเท่ากับ 10, 20, 30, 40, 50 และ 100 ตามงานวิจัย [1] เพื่อใช้แบ่งข้อมูลให้กับชุดข้อมูลทั้งหมดในหัวข้อ 4.2 ซึ่งรายละเอียดของการทดลองในแต่ละส่วนมีดังนี้

4.3.1 การทดลองของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้น

การทดลองในส่วนนี้ต้องเขียนโปรแกรมเคมีนคลัสเตอร์ริงแบบขนานโดยใช้วิธีเลือกชุดจุดศูนย์กลางเริ่มต้นที่แตกต่างกัน (อธิบายในหัวข้อ 2.3.1 และ 3.1) และใช้หน่วยประมวลผลจำนวน 1 หน่วยในการประมวลผล เพื่อเปรียบเทียบประสิทธิภาพของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นดังกล่าว โดยพิจารณาจาก เวลาที่ใช้ในการประมวลผล ความคลาดเคลื่อน และจำนวนรอบของการจัดกลุ่มข้อมูล ซึ่งวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นทั้งหมดมีดังนี้

1) วิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบสุ่ม (Random method)
2) วิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบคัดสรร (Seeding method)
3) วิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบให้ค่าน้ำหนักดีกำลังสอง (D^2 Weighting method)
4) วิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่ที่นำเสนอในงานวิจัยนี้

จากวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นทั้งหมด พบว่าวิธีเลือกแบบที่ 3 อาจให้ชุดจุดศูนย์กลางเริ่มต้นที่แตกต่างกันในการเลือกแต่ละครั้งเมื่อจำนวนกลุ่มที่ต้องการ (K) มีค่าเท่าเดิม เนื่องจากวิธีเลือกแบบนี้ทำการเลือกจุดศูนย์กลางเริ่มต้นจุดแรกด้วยวิธีการสุ่มทำให้ชุดจุดศูนย์กลางเริ่มต้นที่ได้รับแตกต่างกัน ดังนั้นการทดลองกับวิธีเลือกแบบนี้จึงต้องกำหนดจำนวนชุดจุดศูนย์กลางเริ่มต้นในการทดลอง เพื่อให้ได้ผลลัพธ์ที่คลอบคลุมในทุกกรณีและมีความน่าเชื่อถือมากที่สุด โดยสุ่มเลือกจุดศูนย์กลางเริ่มต้นจุดแรกที่ไม่ซ้ำกัน ส่งผลให้เวลาทั้งหมดที่ใช้ในการประมวลผล จำนวนรอบ และความคลาดเคลื่อนของการจัดกลุ่มข้อมูลเป็นค่าเฉลี่ยทั้งหมด โดยกำหนดให้จำนวนชุดจุดศูนย์กลางเริ่มต้นของชุดข้อมูล Cloud.txt, Spam.txt และ Intrusion.txt เท่ากับ 200 ชุด ซึ่งอ้างอิงมาจากงานวิจัย [1]

4.3.2 การทดลองของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์

การทดลองในส่วนนี้จำเป็นต้องเขียนโปรแกรมเคมีนคลัสเตอร์ริงแบบขนาน โดยใช้หน่วยประมวลผลจำนวน 2 และ 4 หน่วยในการประมวลผลเท่านั้น ตามขั้นตอนวิธีเคมีนแบบขนานที่มีการออกแบบแตกต่างกัน เพื่อเปรียบเทียบประสิทธิภาพของขั้นตอนวิธีเคมีนแบบขนานดังกล่าว โดยพิจารณาจาก เวลาทั้งหมดที่ใช้ในการประมวลผล ความคลาดเคลื่อน และจำนวนรอบของการจัดกลุ่มข้อมูล ซึ่งขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานทั้งหมดมีดังนี้

1) ขั้นตอนวิธีเคมีนแบบขนานที่เลือกชุดจุดศูนย์กลางเริ่มต้นแบบสุ่มตามลำดับข้อมูล
2) ขั้นตอนวิธีเคมีนแบบขนานที่กำหนดให้กระบวนการหลักเลือกชุดจุดศูนย์กลางเริ่มต้น
3) ขั้นตอนวิธีเคมีนแบบขนานที่กำหนดให้กระบวนการรองเลือกชุดจุดศูนย์กลางเริ่มต้น

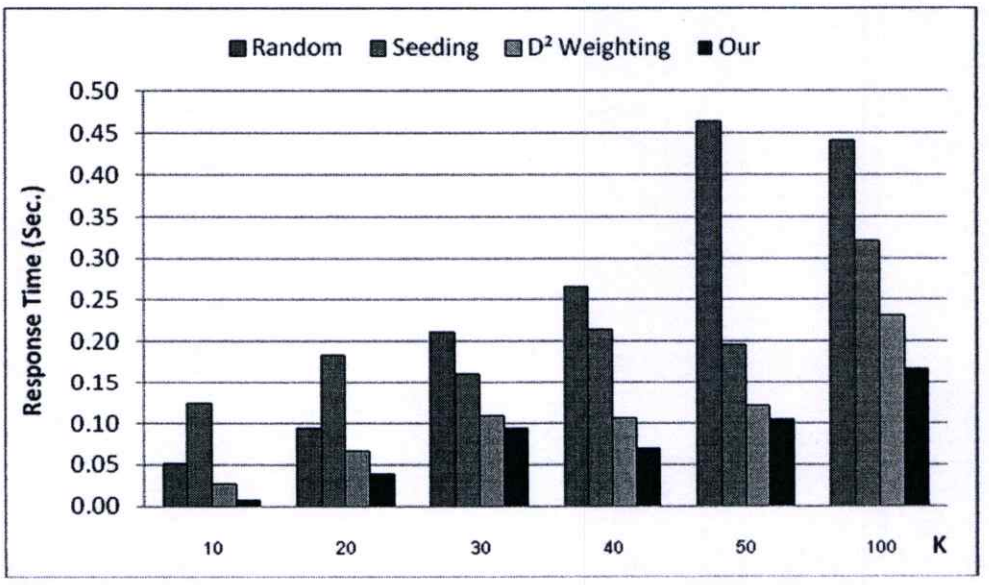
4.4 ผลการทดลอง

4.4.1 ผลการทดลองของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้น

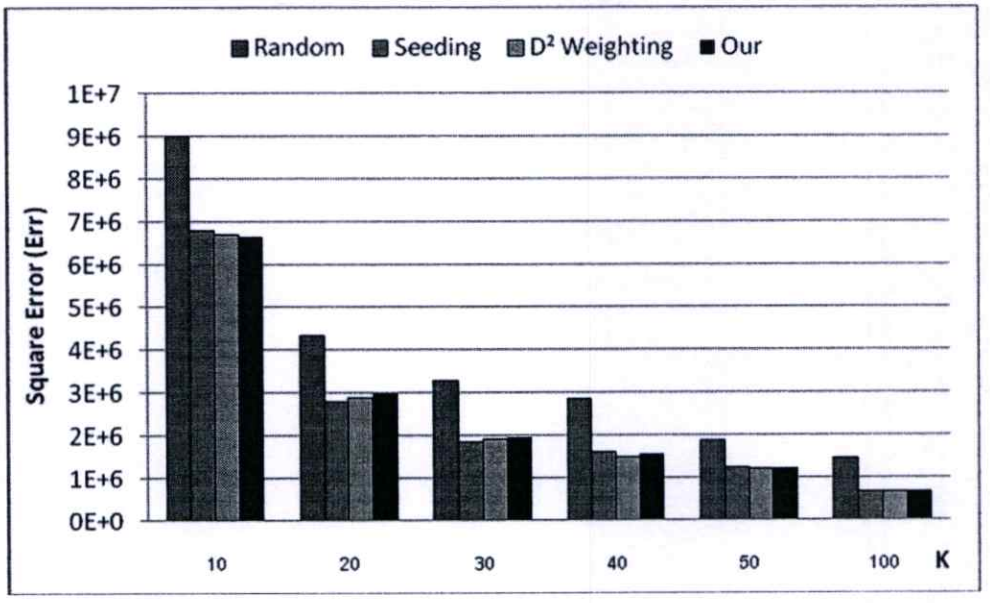
ผลการทดลองของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบต่างๆ กับชุดข้อมูลทั้งหมดที่ใช้ในการทดลองครั้งนี้มีดังนี้ ตารางที่ 4.2, 4.3 และ 4.4 แสดงผลการเปรียบเทียบประสิทธิภาพของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นทั้งหมดกับชุดข้อมูล Cloud.txt, Spam.txt และ Intrusion.txt ตามลำดับ โดยแสดงเวลาที่ใช้ในการประมวลผล ความคลาดเคลื่อน (Err) และจำนวนรอบ (R) ของการจัดกลุ่มข้อมูลตามจำนวนกลุ่มที่ต้องการ (K) ส่วนรูปที่ 4.1 และ 4.3 คือแผนภูมิแท่งแสดงเวลาที่ใช้ในการประมวลผลของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นกับชุดข้อมูล Cloud.txt, Spam.txt ตามลำดับ และรูปที่ 4.2 และ 4.4 คือแผนภูมิแท่งแสดงความคลาดเคลื่อนของการจัดกลุ่มข้อมูลของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นกับชุดข้อมูล Cloud.txt และ Spam.txt รายละเอียดทั้งหมดมีดังนี้

ตารางที่ 4.2 เปรียบเทียบประสิทธิภาพของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นกับชุดข้อมูล Cloud.txt

Cloud Dataset : 1024 x 10												
K	Random			Seeding			D^2 Weighting			Our		
	R	Err	Sec	R	Err	Sec	R	Err	Sec	R	Err	Sec
10	33	9.01E+06	0.051	32	6.80E+06	0.125	18	6.70E+06	0.027	3	6.65E+06	0.008
20	32	4.32E+06	0.094	37	2.80E+06	0.184	23	2.89E+06	0.066	12	2.99E+06	0.039
30	49	3.27E+06	0.211	20	1.86E+06	0.160	25	1.91E+06	0.109	21	1.95E+06	0.094
40	47	2.86E+06	0.266	24	1.61E+06	0.215	18	1.50E+06	0.106	12	1.57E+06	0.070
50	66	1.88E+06	0.465	17	1.26E+06	0.195	16	1.22E+06	0.121	14	1.24E+06	0.106
100	31	1.45E+06	0.441	17	6.69E+05	0.320	15	6.83E+05	0.231	11	6.96E+05	0.168



รูปที่ 4.1 เวลาที่ใช้ในการประมวลผลของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นกับชุดข้อมูล Cloud.txt

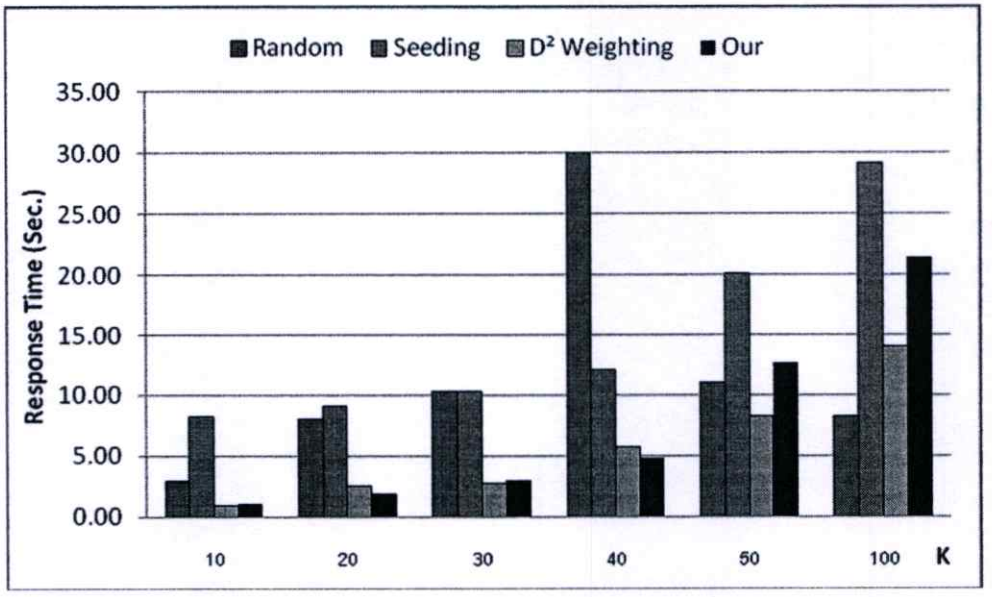


รูปที่ 4.2 ความคลาดเคลื่อนของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นกับชุดข้อมูล Cloud.txt

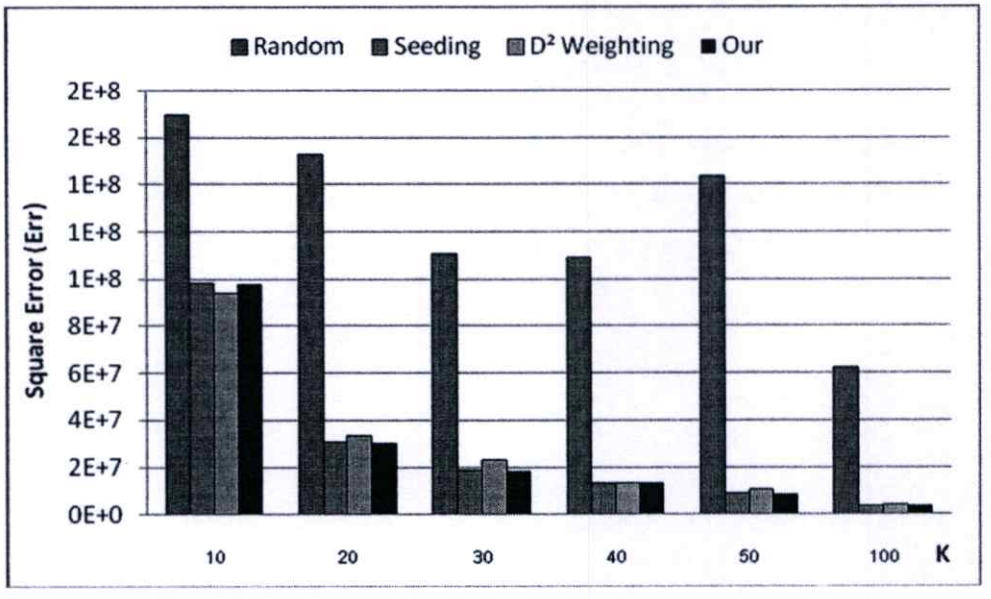
จากตารางที่ 4.2 แสดงผลการเปรียบเทียบประสิทธิภาพของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นกับชุดข้อมูล Cloud.txt พบว่าเวลาทั้งหมดที่ใช้ในการประมวลผลของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่มีค่าน้อยที่สุดในทุกจำนวนกลุ่มที่ต้องการ (K) คือ 0.008, 0.039, 0.094, 0.070, 0.106 และ 0.168 วินาที ซึ่งสอดคล้องกับจำนวนรอบของการจัดกลุ่มข้อมูล ส่วนความคลาดเคลื่อนของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบสุ่ม (Random) มีค่ามากที่สุด แสดงว่าความแม่นยำของการจัดกลุ่มข้อมูลด้วยวิธีนี้มีค่าน้อยที่สุด และวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นส่วนที่เหลือมีค่าความคลาดเคลื่อนใกล้เคียงกันทั้งหมด โดยทั่วไปค่าความคลาดเคลื่อนของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นจะมีแนวโน้มที่ลดลงเมื่อจำนวนกลุ่มที่ต้องการเพิ่มมากขึ้น เนื่องจากจำนวนกลุ่มที่เพิ่มขึ้นนั้นทำให้ข้อมูลบางส่วนเกิดการเปลี่ยนกลุ่มไปจากเดิม ดังนั้นระยะทางระหว่างข้อมูลกับจุดศูนย์กลางของกลุ่มจึงลดลง และส่งผลให้ค่าความคลาดเคลื่อนลดลงตามไปด้วย

ตารางที่ 4.3 เปรียบเทียบประสิทธิภาพของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นกับชุดข้อมูล Spam.txt

Spam Dataset : 4601 x 58												
K	Random			Seeding			D^2 Weighting			Our		
	R	Err	Sec	R	Err	Sec	R	Err	Sec	R	Err	Sec
10	85	1.70E+08	2.934	31	9.82E+07	8.313	25	9.41E+07	0.895	30	9.82E+07	1.055
20	123	1.53E+08	8.102	29	3.05E+07	9.160	37	3.33E+07	2.482	28	3.05E+07	1.891
30	106	1.11E+08	10.30	31	1.85E+07	10.29	27	2.32E+07	2.699	30	1.85E+07	2.988
40	235	1.09E+08	30.06	37	1.32E+07	12.08	44	1.34E+07	5.764	36	1.32E+07	4.766
50	69	1.44E+08	11.02	80	8.83E+06	20.10	51	1.03E+07	8.281	79	8.83E+06	12.69
100	26	6.25E+07	8.231	68	3.49E+06	29.17	43	3.83E+06	13.98	67	3.49E+06	21.50



รูปที่ 4.3 เวลาที่ใช้ในการประมวลผลของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นกับชุดข้อมูล Spam.txt



รูปที่ 4.4 ความคลาดเคลื่อนของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นกับชุดข้อมูล Spam.txt

จากตารางที่ 4.3 แสดงผลการเปรียบเทียบประสิทธิภาพของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นกับชุดข้อมูล Spam.txt พบว่าเวลาที่ใช้ในการประมวลผลของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่ใช้เวลาในการประมวลผลน้อยที่สุดเมื่อจำนวนกลุ่มที่ต้องการ (K) เท่ากับ 20 และ 40 และเมื่อจำนวนกลุ่มที่ต้องการเท่ากับ 50 และ 100 เวลาที่ใช้ในการประมวลผลจะมากกว่าวิธีเลือกแบบสุ่ม (Random) และวิธีเลือกแบบให้ค่าน้ำหนักดีกำลังสอง (D^2 Weighting) เนื่องจากวิธีเลือกดังกล่าวใช้วิธีสุ่มในการเลือกจุดศูนย์กลางเริ่มต้น จึงมีโอกาสที่จะเลือกชุดจุดศูนย์กลางเริ่มต้นที่ทำให้จัดกลุ่มข้อมูลได้เร็วกว่า (จำนวนรอบในการจัดกลุ่มข้อมูล (R) น้อยกว่า) วิธีเลือกแบบใหม่ได้ ส่วนค่าความ

คาดเคลื่อนของวิธีเลือกนี้ก็มีค่าน้อยที่สุดเช่นเดียวกับวิธีเลือกแบบคัดสรร (Seeding) ยกเว้นเมื่อจำนวนกลุ่มที่ต้องการเท่ากับ 10 เท่านั้นที่วิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบให้ค่าน้ำหนักดีกำลังสองมีค่าน้อยที่สุด และวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบสุ่มยังคงมีค่าความคลาดเคลื่อนมากที่สุด เช่นเดียวกับผลการเปรียบเทียบของชุดข้อมูล Cloud.txt ที่กล่าวมาแล้วข้างต้น

ตารางที่ 4.4 เปรียบเทียบประสิทธิภาพของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นกับชุดข้อมูล Intrusion.txt

Intrusion Dataset : 494019 x 35												
K	Random			Seeding			D^2 Weighting			Our		
	R	Err	Min	R	Err	Min	R	Err	Min	R	Err	Min
10	28	1.61E+14	1.089	2	1.75E+13	868.35	2	1.75E+13	0.110	1	1.75E+13	0.080
20	67	1.56E+14	4.924	2	1.15E+13	868.45	2	1.22E+13	0.216	1	1.23E+13	0.150
30	166	1.52E+14	18.01	3	2.08E+12	868.67	7	2.84E+12	0.857	2	3.98E+12	0.327
40	185	1.52E+14	26.44	11	1.63E+12	869.95	11	1.82E+12	1.687	11	1.60E+12	1.709
50	223	1.52E+14	39.63	18	9.06E+11	871.61	14	8.90E+11	2.649	13	8.88E+11	2.476
100	180	1.52E+14	63.25	54	2.89E+11	887.68	50	3.03E+11	17.88	47	3.09E+11	16.80

จากตารางที่ 4.4 แสดงผลการทดลองเปรียบเทียบประสิทธิภาพของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นกับชุดข้อมูล Intrusion.txt พบว่าเวลาที่ใช้ในการประมวลผลของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่มีค่าน้อยที่สุดเมื่อกลุ่มที่ต้องการมีจำนวนเท่ากับ 10, 20, 30, 50 และ 100 สอดคล้องกับจำนวนรอบของการจัดกลุ่มข้อมูล ส่วนวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบคัดสรรใช้เวลามากที่สุดในการทดลองครั้งนี้ โดยค่าความคลาดเคลื่อนของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบสุ่มมีค่ามากที่สุดเช่นเดียวกับชุดข้อมูล Cloud.txt และ Spam.txt ที่กล่าวมาแล้ว ส่วนวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบอื่นๆ มีค่าใกล้เคียงกัน

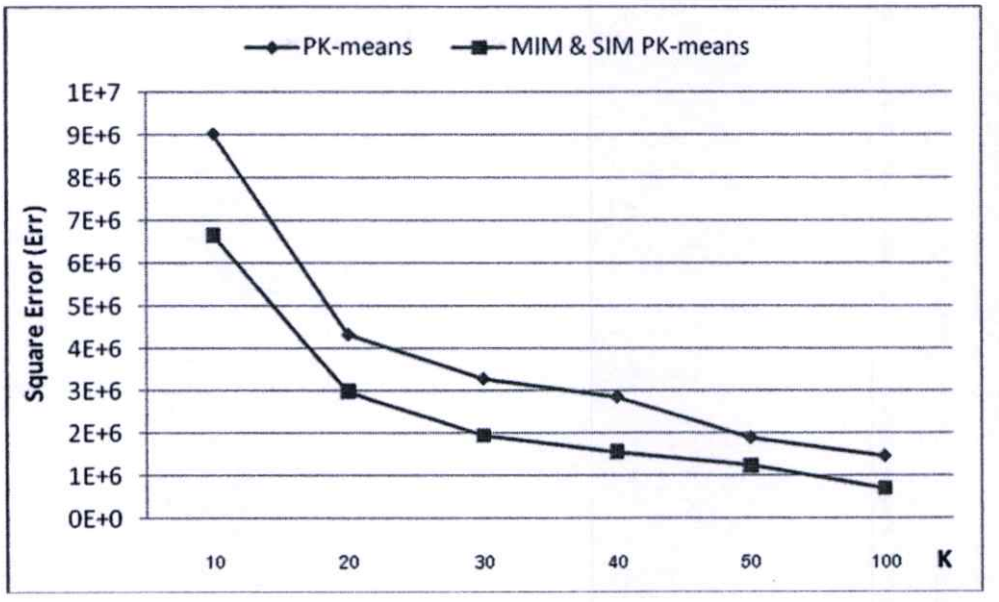
4.4.2 ผลการทดลองของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์

ผลการทดลองของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์ทั้ง 3 วิธี คือ 1) ขั้นตอนวิธีเคมีนแบบขนานที่เลือกชุดจุดศูนย์กลางเริ่มต้นแบบสุ่มตามลำดับข้อมูล (Parallel K-means Algorithm: PK-means) ซึ่งอ้างอิงมาจากงานวิจัย [8] (อธิบายในหัวข้อ 2.3.2) 2) ขั้นตอนวิธีเคมีนแบบขนานที่กำหนดให้กระบวนการหลักเลือกชุดจุดศูนย์กลางเริ่มต้น (Parallel K-means Algorithm by Master Process Select Initial Means: MIM PK-means) และ 3) ขั้นตอนวิธีเคมีนแบบขนานที่กำหนดให้กระบวนการรองเลือกชุดจุดศูนย์กลางเริ่มต้น (Parallel K-means Algorithm by Slave Process Select Initial Means: SIM PK-means) ซึ่งขั้นตอนวิธีทั้งสองถูกนำเสนอในงานวิจัยนี้ (อธิบายในหัวข้อ 3.2.1 และ 3.2.2) โดยนำวิธีเลือกชุดจุดศูนย์กลางแบบใหม่ (อธิบายในหัวข้อ 3.1) มาประยุกต์ใช้ร่วมกัน เพื่อเพิ่มประสิทธิภาพให้กับขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์ โดยใช้หน่วยประมวลผลจำนวน 2 และ 4 หน่วยเท่านั้นในการประมวลผลครั้งนี้

ผลการทดลองของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์เป็นการเปรียบเทียบเวลาทั้งหมดที่ใช้ในการประมวลผล (Response Time) จำนวนรอบ (R) และความคลาดเคลื่อน (Err) ของการจัดกลุ่มข้อมูลด้วยขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานกับชุดข้อมูลที่กล่าวมาแล้วข้างต้นคือ Cloud.txt, Spam.txt และ Intrusion.txt โดยกำหนดให้หน่วยประมวลผลที่ใช้เท่ากับ 2 และ 4 หน่วยประมวลผล ดังตารางที่ 4.5, 4.6 และ 4.7 ตามลำดับ ส่วนรูปที่ 4.5 ถึง 4.16 เป็นแผนภูมิที่แสดงความคลาดเคลื่อน (Err) ของการจัดกลุ่มข้อมูล และแสดงเวลาทั้งหมดที่ใช้ในการประมวลผลซึ่งสอดคล้องกับข้อมูลในตารางดังกล่าว จากตารางทั้งหมดในหัวข้อนี้พบว่าจำนวนรอบ (R) และความคลาดเคลื่อนของการจัดกลุ่มข้อมูลมีค่าเช่นเดียวกับผลการทดลองเปรียบเทียบประสิทธิภาพของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นที่กล่าวมาแล้วข้างต้น (อธิบายในหัวข้อ 4.4.1) เนื่องจากค่าเหล่านี้ขึ้นอยู่กับวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นเพียงอย่างเดียวเท่านั้น เมื่อใช้เงื่อนไขในการหยุดเดียวกัน ส่งผลให้ขั้นตอนวิธี MIM PK-means และขั้นตอนวิธี SIM PK-means มีจำนวนรอบ และความคลาดเคลื่อนของการจัดกลุ่มข้อมูลเท่ากันในทุกกลุ่มที่ต้องการ (K) เนื่องจากขั้นตอนวิธีทั้งสองนำวิธีเลือกแบบใหม่มาประยุกต์ใช้เหมือนกัน โดยมีเพียงเวลาทั้งหมดที่ใช้ในการประมวลผลเท่านั้นที่มีการเปลี่ยนแปลง ทำให้ทราบว่าจำนวนหน่วยประมวลผลที่ใช้ในการทำงานไม่ส่งผลกระทบต่อจำนวนรอบ และความคลาดเคลื่อนของการจัดกลุ่มข้อมูลเลย ดังนั้นไม่ว่าจะเพิ่มหรือลดจำนวนหน่วยประมวลผลเพียงใดก็ตาม จำนวนรอบ และความคลาดเคลื่อนของการจัดกลุ่มข้อมูลก็ยังคงมีค่าเท่าเดิมเสมอ

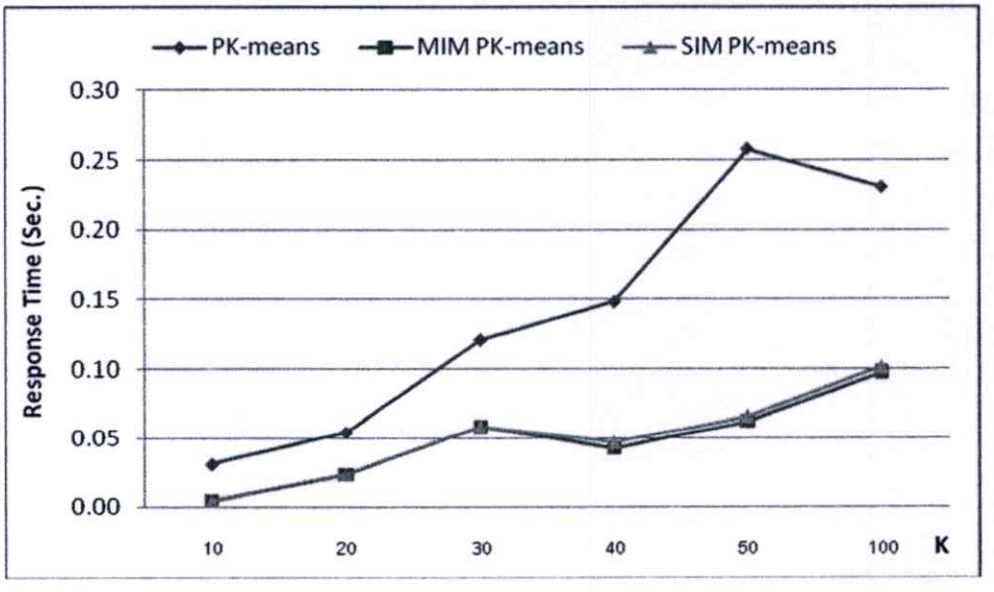
ตารางที่ 4.5 เปรียบเทียบเวลาทั้งหมดที่ใช้ในการประมวลผล จำนวนรอบ และความคลาดเคลื่อนของการจัดกลุ่มข้อมูลของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์ทั้ง 3 วิธีกับชุดข้อมูล Cloud.txt

Cloud Dataset : 1024 x 10												
K	PK-means				MIM PK-means				SIM PK-means			
	R	Err	Time: Sec		R	Err	Time: Sec		R	Err	Time: Sec	
			2 PEs	4 PEs			2 PEs	4 PEs			2 PEs	4 PEs
10	33	9.01E+06	0.031	0.031	3	6.65E+06	0.004	0.004	3	6.65E+06	0.004	0.008
20	32	4.32E+06	0.055	0.043	12	2.99E+06	0.023	0.020	12	2.99E+06	0.023	0.059
30	49	3.27E+06	0.121	0.082	21	1.95E+06	0.059	0.039	21	1.95E+06	0.059	0.078
40	47	2.86E+06	0.148	0.102	12	1.57E+06	0.043	0.031	12	1.57E+06	0.047	0.074
50	66	1.88E+06	0.258	0.160	14	1.24E+06	0.063	0.043	14	1.24E+06	0.066	0.082
100	31	1.45E+06	0.231	0.137	11	6.96E+05	0.098	0.063	11	6.96E+05	0.102	0.152

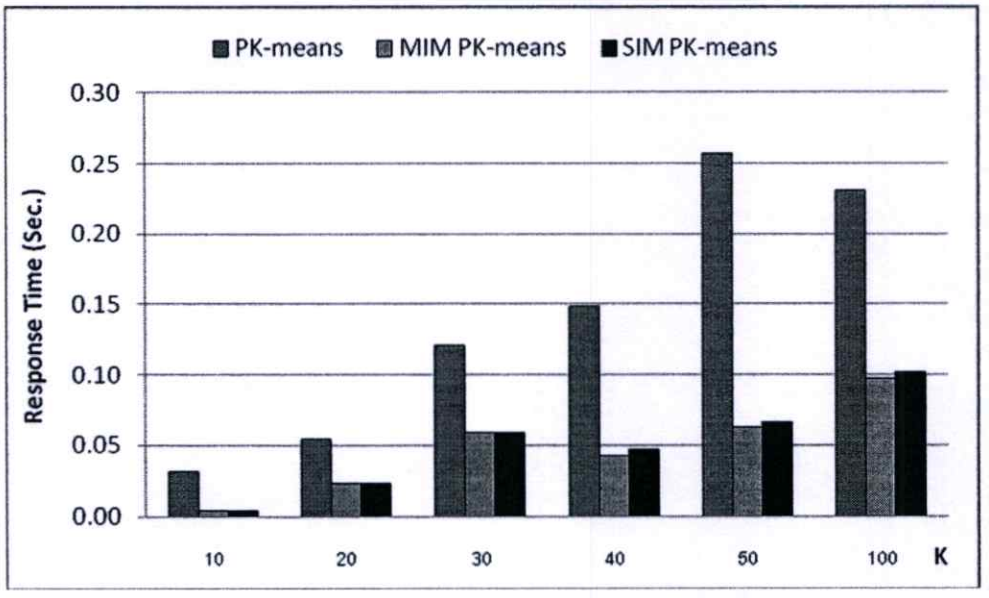


รูปที่ 4.5 เปรียบเทียบความคลาดเคลื่อนของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์กับชุดข้อมูล Cloud.txt

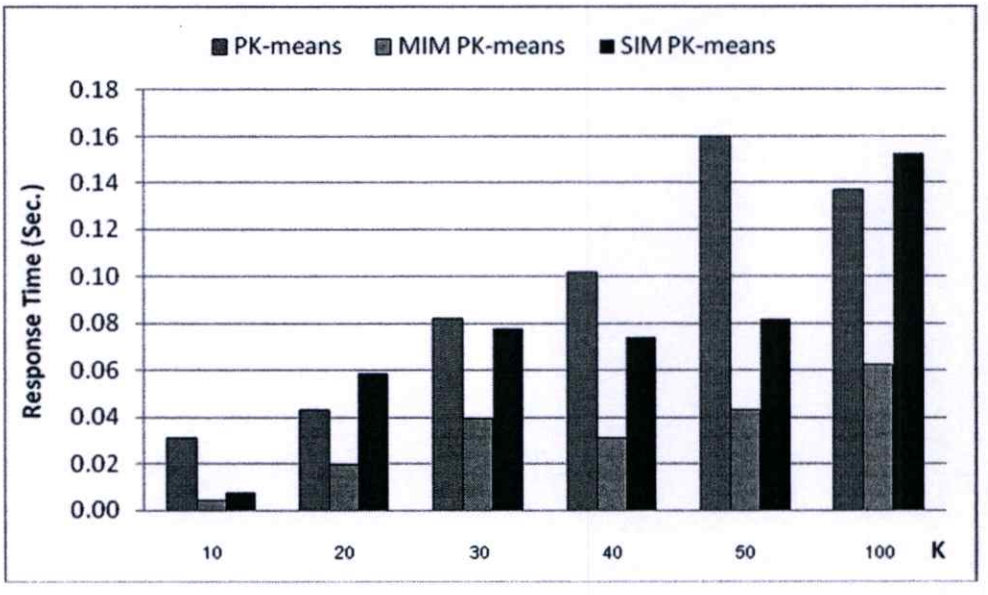
จากรูปที่ 4.5 แสดงผลการเปรียบเทียบความคลาดเคลื่อนของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์กับชุดข้อมูล Cloud.txt พบว่าขั้นตอนวิธี MIM PK-means และขั้นตอนวิธี SIM PK-means ที่นำเสนอมีค่าความคลาดเคลื่อนของการจัดกลุ่มข้อมูลน้อยกว่าขั้นตอนวิธี PK-means ในทุกกลุ่มที่ต้องการ แสดงว่าขั้นตอนวิธีทั้งสองสามารถจัดกลุ่มข้อมูลได้แม่นยำกว่าขั้นตอนวิธี PK-means



รูปที่ 4.6 แนวโน้มของเวลาทั้งหมดที่ใช้ในการประมวลผลของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์กับชุดข้อมูล Cloud.txt เมื่อใช้หน่วยประมวลผลเท่ากับ 2 หน่วย



รูปที่ 4.7 เปรียบเทียบเวลาทั้งหมดที่ใช้ในการประมวลผลของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์กับชุดข้อมูล Cloud.txt เมื่อใช้หน่วยประมวลผลเท่ากับ 2 หน่วย



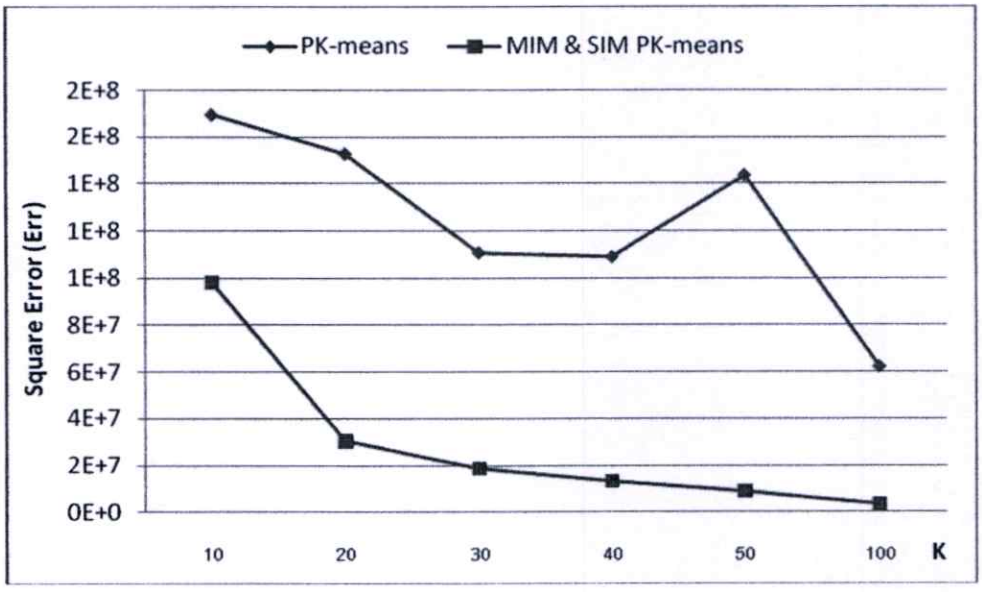
รูปที่ 4.8 เปรียบเทียบเวลาทั้งหมดที่ใช้ในการประมวลผลของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์กับชุดข้อมูล Cloud.txt เมื่อใช้หน่วยประมวลผลเท่ากับ 4 หน่วย

จากรูปที่ 4.7 และ 4.8 แสดงผลการเปรียบเทียบเวลาทั้งหมดที่ใช้ในการประมวลผลของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานกับชุดข้อมูล Cloud.txt พบว่าขั้นตอนวิธี MIM PK-means ใช้เวลาทั้งหมดในการประมวลผลน้อยที่สุด เมื่อจำนวนหน่วยประมวลผลเท่ากับ 2 และ 4 หน่วย ส่วนขั้นตอนวิธี SIM PK-means ใช้เวลาทั้งหมดในการประมวลผลรองลงมา ยกเว้นเมื่อจำนวนหน่วยประมวลผลเท่ากับ 4 หน่วย และจำนวนกลุ่มที่ต้องการเท่ากับ 20 และ 100 กลุ่มเท่านั้น เนื่องจากช่วงที่เลือกชุดจุดศูนย์กลางเริ่มต้นต้องใช้เวลามากในการสื่อสารระหว่างหน่วยประมวลผล

ทำให้เวลาที่ใช้ในเลือกชุดจุดศูนย์กลางเริ่มต้นมากเกินไป และส่งผลให้เวลาทั้งหมดที่ใช้ในการประมวลผลมากที่สุดตามไปด้วย

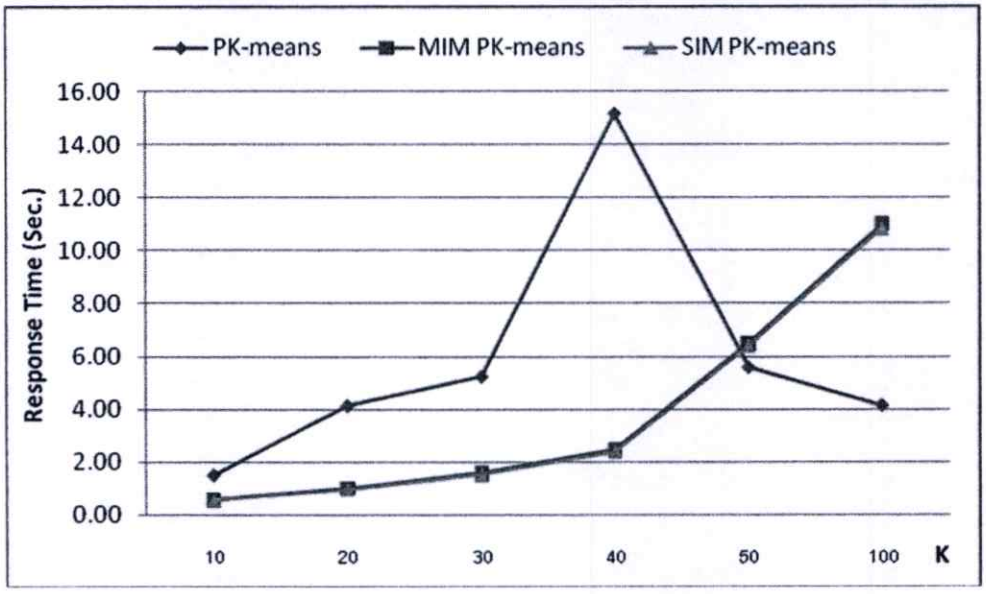
ตารางที่ 4.6 เปรียบเทียบเวลาทั้งหมดที่ใช้ในการประมวลผล จำนวนรอบ และความคลาดเคลื่อนของขั้นตอนวิธีเคมีนคลัสเตอร์ริงคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์ทั้ง 3 วิธี กับชุดข้อมูล Spam.txt

Spam Dataset : 4601 x 58												
K	PK-means				MIM PK-means				SIM PK-means			
	R	Err	Time: Sec		R	Err	Time: Sec		R	Err	Time: Sec	
			2 PEs	4 PEs			2 PEs	4 PEs			2 PEs	4 PEs
10	85	1.70E+08	1.504	0.805	30	9.82E+07	0.559	0.313	30	9.82E+07	0.543	0.289
20	123	1.53E+08	4.121	2.133	28	3.05E+07	0.992	0.551	28	3.05E+07	0.973	0.508
30	106	1.11E+08	5.227	2.688	30	1.85E+07	1.563	0.852	30	1.85E+07	1.520	0.785
40	235	1.09E+08	15.18	7.816	36	1.32E+07	2.449	1.316	36	1.32E+07	2.391	1.231
50	69	1.44E+08	5.578	2.844	79	8.83E+06	6.488	3.391	79	8.83E+06	6.410	3.301
100	26	6.25E+07	4.141	2.109	67	3.49E+06	11.02	5.723	67	3.49E+06	10.86	5.504



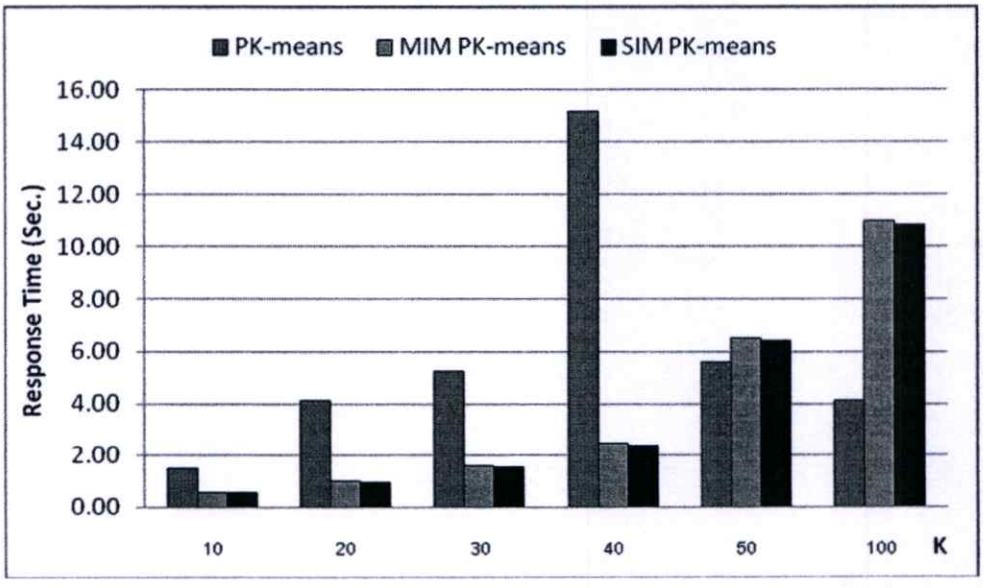
รูปที่ 4.9 เปรียบเทียบความคลาดเคลื่อนของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์กับชุดข้อมูล Spam.txt

จากรูปที่ 4.9 แสดงผลการเปรียบเทียบความคลาดเคลื่อนของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์กับชุดข้อมูล Spam.txt พบว่าขั้นตอนวิธี MIM PK-means และขั้นตอนวิธี SIM PK-means ที่นำเสนอมีค่าความคลาดเคลื่อนของการจัดกลุ่มข้อมูลน้อยกว่าขั้นตอนวิธี PK-means ในทุกกลุ่มที่ต้องการ แสดงว่าขั้นตอนวิธีทั้งสองสามารถจัดกลุ่มข้อมูลได้แม่นยำกว่าขั้นตอนวิธี PK-means เช่นเดียวกับชุดข้อมูล Cloud.txt ดังที่กล่าวมาแล้วข้างต้น

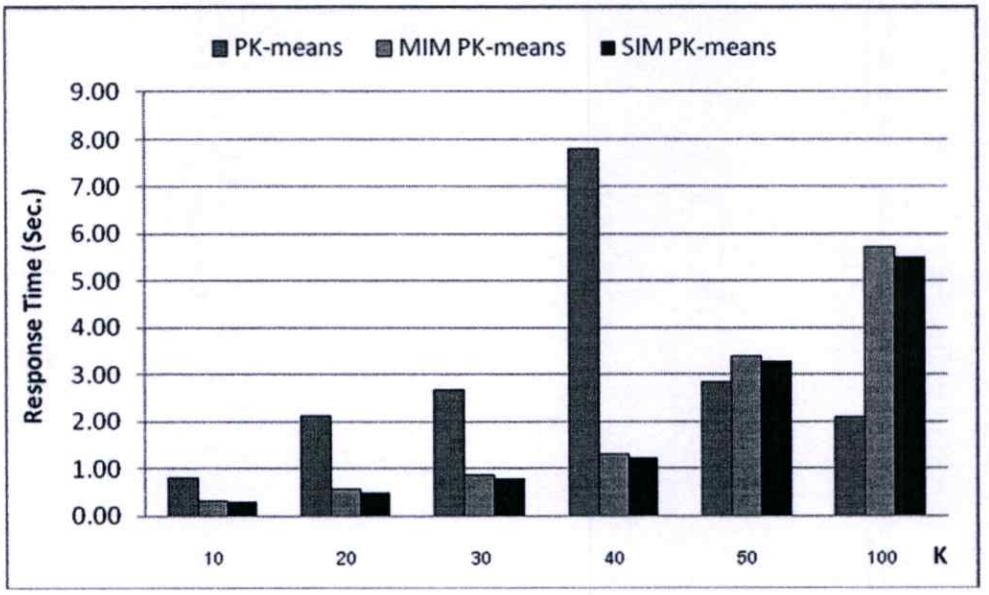


รูปที่ 4.10 แนวโน้มของเวลาทั้งหมดที่ใช้ในการประมวลผลของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์กับชุดข้อมูล Spam.txt เมื่อใช้หน่วยประมวลผลเท่ากับ 2 หน่วย

จากรูปที่ 4.10 แสดงแนวโน้มของเวลาทั้งหมดที่ใช้ในการประมวลผลของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์กับชุดข้อมูล Spam.txt เมื่อใช้หน่วยประมวลผลเท่ากับ 2 หน่วย พบว่าขั้นตอนวิธีทั้งสองที่นำเสนอใช้เวลามากกว่าขั้นตอนวิธี PK-means เมื่อจำนวนกลุ่มที่ต้องการเท่ากับ 50 และ 100 เนื่องจากจำนวนรอบในการจัดกลุ่มข้อมูล (R) ของขั้นตอนวิธีทั้งสองมากกว่าขั้นตอนวิธี PK-means ส่งผลให้เวลาที่ใช้ในช่วงนี้มากกว่าตามไปด้วย



รูปที่ 4.11 เปรียบเทียบเวลาทั้งหมดที่ใช้ในการประมวลผลของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์กับชุดข้อมูล Spam.txt เมื่อใช้หน่วยประมวลผลเท่ากับ 2 หน่วย

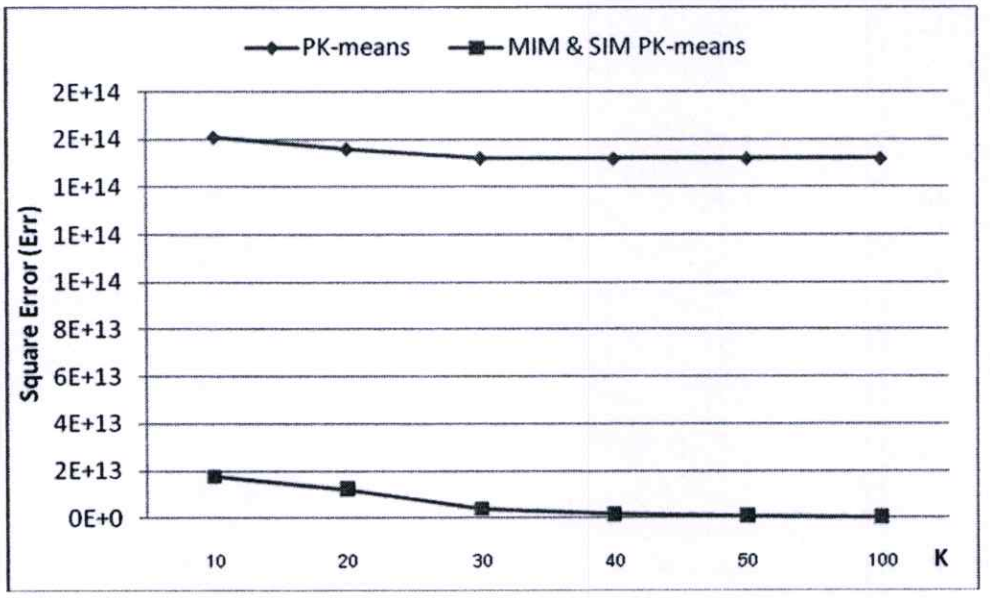


รูปที่ 4.12 เปรียบเทียบเวลาทั้งหมดที่ใช้ในการประมวลผลของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์กับชุดข้อมูล Spam.txt เมื่อใช้หน่วยประมวลผลเท่ากับ 4 หน่วย

จากรูปที่ 4.11 และ 4.12 แสดงผลการเปรียบเทียบเวลาทั้งหมดที่ใช้ในการประมวลผลของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานกับชุดข้อมูล Spam.txt พบว่าขั้นตอนวิธี SIM PK-means ใช้เวลาทั้งหมดในการประมวลผลน้อยที่สุด ส่วนขั้นตอนวิธี MIM PK-means ที่ใช้เวลาทั้งหมดรองลงมา เมื่อจำนวนหน่วยประมวลผลเท่ากับ 2 และ 4 หน่วย และจำนวนกลุ่มที่ต้องการเท่ากับ 10 ถึง 40 กลุ่ม ยกเว้นเมื่อจำนวนกลุ่มที่ต้องการเท่ากับ 50 และ 100 ขั้นตอนวิธี PK-means จะใช้เวลาทั้งหมดในการประมวลผลน้อยที่สุด เพราะวิธีเลือกชุดศูนย์กลางเริ่มต้นแบบสุ่มสามารถเลือกชุดจุดศูนย์กลางเริ่มต้นที่ทำให้จัดกลุ่มข้อมูลได้เร็วกว่า (จำนวนรอบน้อยกว่า) วิธีเลือกแบบใหม่ที่ขั้นตอนวิธีทั้งสองนำไปประยุกต์ใช้ทำให้เวลาทั้งหมดที่ใช้ในการประมวลผลน้อยตามไปด้วย

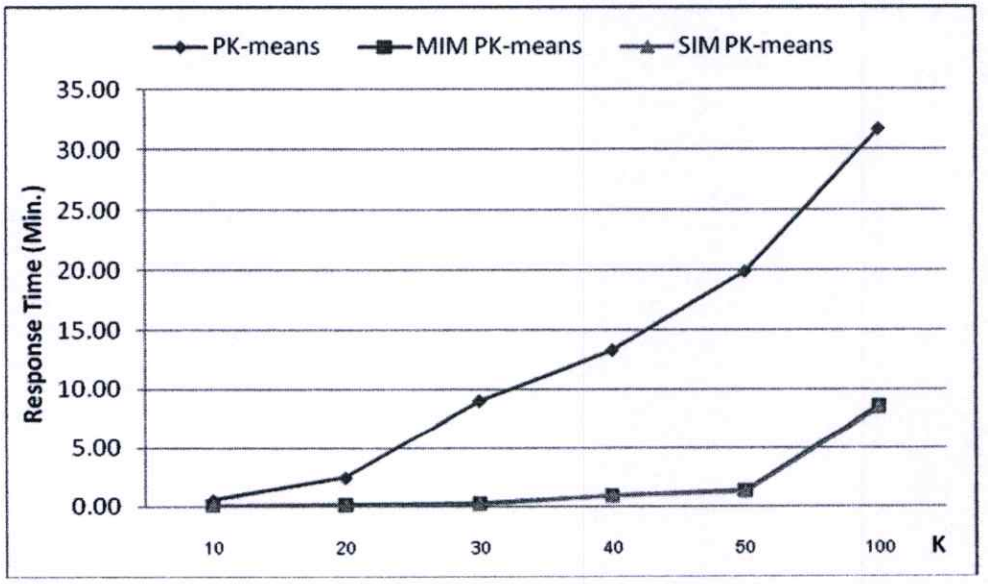
ตารางที่ 4.7 เปรียบเทียบเวลาทั้งหมดที่ใช้ในการประมวลผล จำนวนรอบ และความคลาดเคลื่อนของขั้นตอนวิธีเคมีนคลัสเตอร์ริงคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์ทั้ง 3 วิธีกับชุดข้อมูล Intrusion.txt

Intrusion Dataset : 494019 x 35												
K	PK-means				MIM PK-means				SIM PK-means			
	R	Err	Time: Min		R	Err	Time: Min		R	Err	Time: Min	
			2 PEs	4 PEs			2 PEs	4 PEs			2 PEs	4 PEs
10	28	1.61E+14	0.548	0.275	1	1.75E+13	0.060	0.050	1	1.75E+13	0.040	0.020
20	67	1.56E+14	2.465	1.238	1	1.23E+13	0.113	0.094	1	1.23E+13	0.075	0.038
30	166	1.52E+14	9.021	4.528	2	3.98E+12	0.220	0.166	2	3.98E+12	0.164	0.082
40	185	1.52E+14	13.30	6.675	11	1.60E+12	0.930	0.540	11	1.60E+12	0.858	0.428
50	223	1.52E+14	19.87	9.950	13	8.88E+11	1.335	0.761	13	8.88E+11	1.240	0.621
100	180	1.52E+14	31.69	15.88	47	3.09E+11	8.622	4.502	47	3.09E+11	8.413	4.217

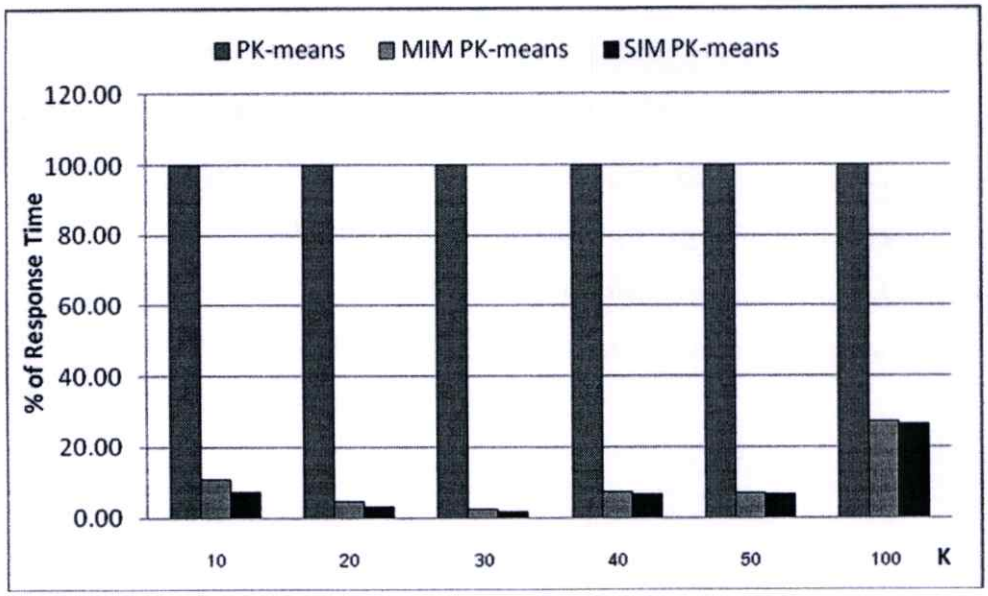


รูปที่ 4.13 เปรียบเทียบความคลาดเคลื่อนของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์กับชุดข้อมูล Intrusion.txt

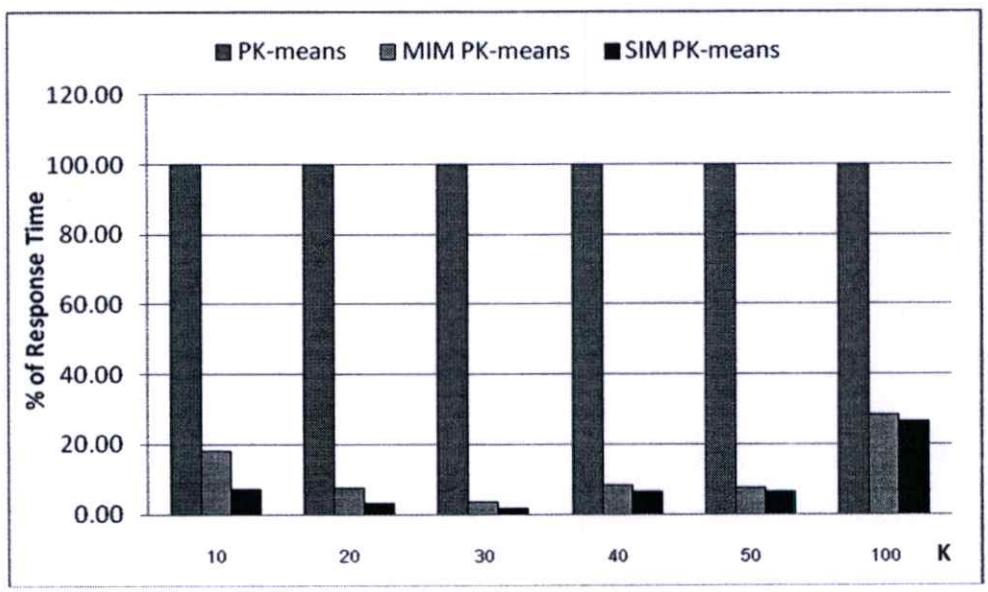
จากรูปที่ 4.13 แสดงผลการเปรียบเทียบความคลาดเคลื่อนของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์กับชุดข้อมูล Intrusion.txt พบว่าขั้นตอนวิธี MIM PK-means และขั้นตอนวิธี SIM PK-means ที่นำเสนอมีค่าความคลาดเคลื่อนของการจัดกลุ่มข้อมูลน้อยกว่าขั้นตอนวิธี PK-means ในทุกกลุ่มที่ต้องการ แสดงว่าขั้นตอนวิธีทั้งสองสามารถจัดกลุ่มข้อมูลได้แม่นยำกว่าขั้นตอนวิธี PK-means เช่นเดียวกับชุดข้อมูล Cloud.txt และ Spam.txt ดังที่กล่าวมาแล้วข้างต้น



รูปที่ 4.14 แนวโน้มของเวลาทั้งหมดที่ใช้ในการประมวลผลของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์กับชุดข้อมูล Intrusion.txt เมื่อใช้หน่วยประมวลผลเท่ากับ 2 หน่วย



รูปที่ 4.15 เปรียบเทียบเวลาทั้งหมดที่ใช้ในการประมวลผลของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์กับชุดข้อมูล Intrusion.txt เมื่อใช้หน่วยประมวลผลเท่ากับ 2 หน่วย



รูปที่ 4.16 เปรียบเทียบเวลาทั้งหมดที่ใช้ในการประมวลผลของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์กับชุดข้อมูล Intrusion.txt เมื่อใช้หน่วยประมวลผลเท่ากับ 4 หน่วย

จากรูปที่ 4.15 และ 4.16 แสดงผลการเปรียบเทียบเวลาทั้งหมดที่ใช้ในการประมวลผลของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานกับชุดข้อมูล Intrusion.txt โดยกำหนดให้เวลาทั้งหมดที่ใช้ในการประมวลผลของขั้นตอนวิธี PK-means คิดเป็น 100 เปอร์เซ็นต์ เมื่อจำนวนหน่วยประมวลผลเท่ากับ 2 และ 4 หน่วย พบว่าขั้นตอนวิธี SIM PK-means ใช้เวลาทั้งหมดในการประมวลผลน้อยที่สุดในทุกกรณี ส่วนขั้นตอนวิธี MIM PK-means ใช้เวลาทั้งหมดในการประมวลผลรองลงมา

บทที่ 5
บทสรุปและแนวทางการพัฒนางานวิจัย

เนื้อหาในบทนี้เป็นบทสรุปของงานวิจัย ซึ่งงานวิจัยนี้ได้มีการออกแบบขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์โดยนำวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่มาประยุกต์ใช้ ทำให้ได้ขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์ที่สามารถจัดกลุ่มข้อมูลได้อย่างรวดเร็ว แม่นยำ และสามารถรองรับข้อมูลขนาดใหญ่ได้ ส่วนสุดท้ายคือ การเสนอแนวทางในการพัฒนางานวิจัย

5.1 บทสรุป

ในงานวิจัยนี้ได้นำเสนอขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานที่มีประสิทธิภาพบนระบบคลัสเตอร์ โดยนำวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่มาประยุกต์ใช้ ซึ่งมีทั้งหมด 2 ขั้นตอนวิธี คือ ขั้นตอนวิธีที่ 1) ขั้นตอนวิธีเคมีนแบบขนานที่กำหนดให้กระบวนการหลักเลือกชุดจุดศูนย์กลางเริ่มต้น (MIM PK-means) และขั้นตอนวิธีที่ 2) ขั้นตอนวิธีเคมีนแบบขนานที่กำหนดให้กระบวนการรองเลือกชุดจุดศูนย์กลางเริ่มต้น (SIM PK-means) เนื่องจากขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนาน (PK-means) ที่มีผู้นำเสนอไว้ก่อนหน้านี้ใช้เวลามากในการจัดกลุ่มข้อมูล และผลลัพธ์ที่ได้มีความคลาดเคลื่อนสูง เพราะขั้นตอนวิธีนี้นำวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบสุ่มมาใช้ ดังนั้นผู้วิจัยจึงออกแบบ และพัฒนาวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบใหม่ที่มีประสิทธิภาพมากกว่า โดยนำแนวคิดของวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบคัดสรรมาศึกษาและพัฒนาต่อทำให้ได้วิธีเลือกแบบใหม่ที่มีความซับซ้อนด้านเวลาเท่ากับ O(KN) และนำมาเปรียบเทียบกับวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบต่างๆ ดังนี้

1) วิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบสุ่ม (Random Method)
2) วิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบคัดสรร (Seeding Method)
3) วิธีเลือกชุดจุดศูนย์กลางเริ่มต้นแบบให้ค่าน้ำหนักดีกำลังสอง (D^2 Weighting Method)

โดยทดลองกับชุดข้อมูล Cloud.txt, Spam.txt และ Intrusion.txt เมื่อจำนวนกลุ่มที่ต้องการเท่ากับ 10, 20, 30, 40, 50 และ 100 และจำนวนหน่วยประมวลผลที่ใช้ในการทำงานเท่ากับ 1 หน่วย ซึ่งการเปรียบเทียบประสิทธิภาพแบ่งออกเป็น 3 ส่วนคือ เวลาที่ใช้ในการประมวลผล (Response Time) จำนวนรอบ และความคลาดเคลื่อน (Error) ของการจัดกลุ่มข้อมูล ซึ่งเวลาที่ใช้ในการประมวลผลจะสอดคล้องกับความซับซ้อนด้านเวลา และจำนวนรอบของการจัดกลุ่มข้อมูล จากผล

การทดลองส่วนใหญ่พบว่า วิธีเลือกแบบใหม่สามารถเลือกชุดจุดศูนย์กลางเริ่มต้นที่ทำให้การจัดกลุ่มข้อมูลสำเร็จได้อย่างรวดเร็วที่สุด เมื่อชุดข้อมูลคือ Cloud.txt และ Intrusion.txt ส่วนผลลัพธ์ที่ได้จากการจัดกลุ่มของชุดข้อมูล Spam.txt มีค่าความคลาดเคลื่อนน้อยที่สุด

สำหรับขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์ทั้งสองวิธีที่นำเสนอคือ ขั้นตอนวิธี MIM PK-means ที่มีความซับซ้อนด้านเวลาเท่ากับ O(KN+RK(N/P)) และขั้นตอนวิธี SIM PK-means ที่มีความซับซ้อนด้านเวลาเท่ากับ O(RK(N/P)) เมื่อเปรียบเทียบกับขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานที่ใช้วิธีเลือกจุดศูนย์กลางเริ่มต้นแบบสุ่ม โดยใช้จำนวนหน่วยประมวลผลเท่ากับ 2 และ 4 หน่วย ผลการทดลองส่วนใหญ่พบว่าขั้นตอนวิธีทั้งสองที่นำเสนอสามารถจัดกลุ่มข้อมูลได้อย่างรวดเร็ว และแม่นยำมากกว่าขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานที่นำมาเปรียบเทียบ และเนื่องจากระบบคลัสเตอร์ประกอบด้วยเครื่องคอมพิวเตอร์ที่มีหน่วยความจำเป็นของตนเองทำให้สามารถรองรับข้อมูลขนาดใหญ่ได้ ซึ่งขั้นตอนวิธีที่นำเสนอมีข้อดี-ข้อเสียดังนี้

ตารางที่ 5.1 ข้อดี-ข้อเสียของขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์ที่นำเสนอ

ขั้นตอนวิธีเคมีนแบบขนาน	ข้อดี	ข้อเสีย
ขั้นตอนวิธี MIM PK-means	1. เหมาะกับชุดข้อมูลขนาดเล็ก 2. ง่ายต่อการพัฒนาเนื่องจากมีขั้นตอนการทำงานที่ง่าย และไม่ซับซ้อน 3. ใช้เวลาในการสื่อสารระหว่างหน่วยประมวลผลตอนเลือกชุดจุดศูนย์กลางเริ่มต้นน้อย	1. ไม่เหมาะกับข้อมูลขนาดใหญ่ เพราะหน่วยประมวลผลหลักต้องเลือกชุดจุดศูนย์กลางเริ่มต้นเพียงลำพังทำให้ใช้เวลามากในส่วนนี้ และส่งผลให้หน่วยประมวลผลอื่นๆ ต้องรอคอยนานก่อนการจัดกลุ่ม
ขั้นตอนวิธี SIM PK-means	1. เหมาะกับชุดข้อมูลขนาดใหญ่ 2. ใช้เวลาในการคำนวณตอนเลือกชุดจุดศูนย์กลางเริ่มต้นน้อย เพราะหน่วยประมวลผลทุกหน่วยช่วยกันทำงาน	1. ไม่เหมาะกับการใช้หน่วยประมวลผล และจำนวนกลุ่มที่ต้องการจำนวนมาก เพราะทำให้เวลาในการสื่อสารระหว่างหน่วยประมวลผลตอนเลือกชุดจุดศูนย์กลางเริ่มต้นมากตามไปด้วย

ดังนั้น สามารถสรุปได้ว่าขั้นตอนวิธีเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์ที่นำเสนอสามารถจัดกลุ่มข้อมูลได้อย่างรวดเร็ว แม่นยำ และยังสามารถรองรับปัญหาที่มีขนาดใหญ่มากกว่าคอมพิวเตอร์เพียงเครื่องเดียวจะสามารถทำงานได้

5.2 แนวทางการพัฒนางานวิจัย

- ลดเวลาที่ใช้ในการสื่อสารระหว่างหน่วยประมวลผล ช่วงที่หน่วยประมวลผลหลักทำการส่งชุดข้อมูลย่อยที่แบ่งได้ตามจำนวนหน่วยประมวลผลที่ใช้ในการทำงานไปยังหน่วยประมวลผลรองจำนวน P-1 ครั้ง ซึ่งขนาดของชุดข้อมูลย่อยเท่ากับ N/P ซึ่ง N คือ ขนาดของชุดข้อมูล โดยออกแบบให้แต่ละหน่วยประมวลผลอ่านชุดข้อมูล และแบ่งข้อมูลด้วยตนเอง ซึ่งชุดข้อมูลย่อยที่ได้รับจะตรงตามลำดับของหน่วยประมวลผลเหล่านั้น จากนั้นทุกหน่วยประมวลผลจะทำการเลือกชุดจุดศูนย์กลางเริ่มต้น และจัดกลุ่มข้อมูลเป็นลำดับถัดไป สุดท้ายคือหน่วยประมวลผลรองทุกหน่วยส่งผลลัพธ์ไปยังหน่วยประมวลผลหลักเพื่อรวบรวม และบันทึกลงไฟล์ข้อมูล

- ลดเวลาที่ใช้ในการสื่อสารระหว่างหน่วยประมวลผล ช่วงที่หน่วยประมวลผลทุกหน่วยคำนวณค่าเฉลี่ยเพื่อกำหนดเป็นชุดจุดศูนย์กลางสำหรับการจัดกลุ่มข้อมูลในรอบถัดไป

- ออกแบบวิธีเลือกชุดจุดศูนย์กลางเริ่มต้นที่มีประสิทธิภาพ และยังสามารถเลือกชุดจุดศูนย์กลางเริ่มต้นที่เหมาะสมกับทุกชุดข้อมูล และทุกกลุ่มที่ต้องการ

เอกสารอ้างอิง

[1] D. Arthur, and S. Vassilvitskii, "**K-means++: The Advantages of Careful Seeding**", SODA, 2007.

[2] D. Arthur and S. Vassilvitskii, "**K-means++ test code**",http://www.stanford.edu/_darthur/kMeanspp Test.zip.

[3] H. BISGIN, "**Parallel Clustering Algorithm With Application to Climatology**", Computational Science and Engineering, December 2007.

[4] M. N. Joshi, "**Parallel *K* - Means Algorithm on Distributed Memory Multiprocessors**", Computer Science Department University of Minnesota, Twin Cities, spring 2003.

[5] C. H. Jun, J. S. Lee, and H. S. Park, "**A K-means-like Algorithm for K-medoids Clustering and Its Performance**", Proceedings of the 36th CIE Conference on Computers & Industrial Engineering, pp.1222-1231, Taipei, Taiwan, Jun. 20-23, 2006.

[6] S. Kantabutra, "**Parallel K-Means Clustering Algorithm on NOW**", September 1999.

[7] S. Kantabutra, P. Kornpitak, and C. Naramittakapong , "**Pipelined K-means Algorithm on COWs**", ISCIT , September 03-05, 2003, Hatyai, Songkhla, Thailand.

[8] W. Liao, "**The Software Package of Parallel K-means**", http://www.ece.northwestern.edu/~wkliao/Kmeans/index.html, 2005.

[9] J. Mao, L. Ou, Z. Xiong, and Y. Zhang, "**The Study of Parallel K-Means Algorithm**", Proceedings of the 6th World Congress on Intelligent Control and Automation, June 21 - 23, 2006, Dalian, China.

[10] R. Ostrovsky, Y. Rabani, L. J. Schulman, and C. Swamy, "**The Effectiveness of Lloyd-Type Methods for the k-Means Problem**", IEEE, 2006.

[11] E. SÜLÜN, "**Improvements in K-means Algorithm to Execute on Large Amounts of Data**", http://library.iyte.edu.tr/tezler/master/bilgisayaryaziliz/T000441.pdf, October 2004.

ภาคผนวก

ผลงานที่ได้รับการตีพิมพ์

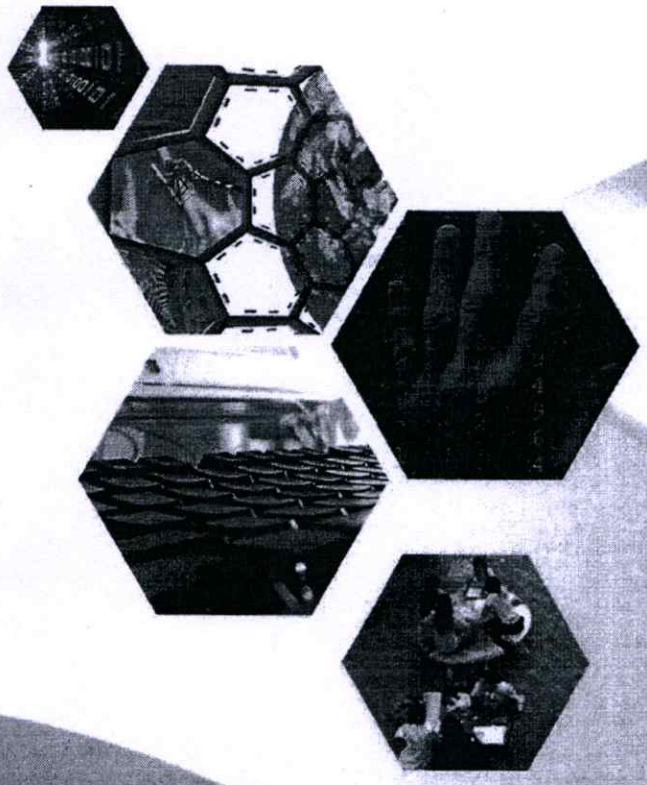
KST Research Center
Knowledge Smart Technologies
Since 2008

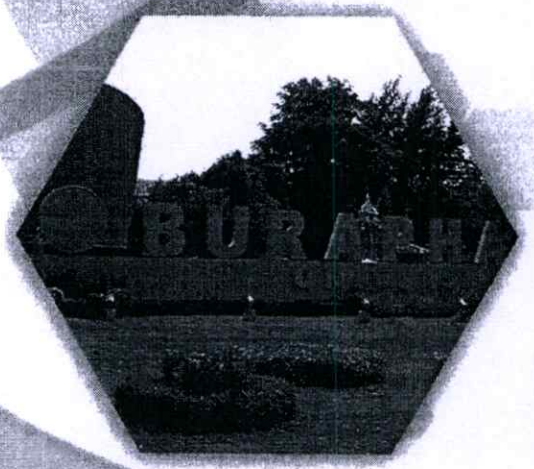
Department of Computer Science
Faculty of Science
Burapha University
Chonburi, THAILAND

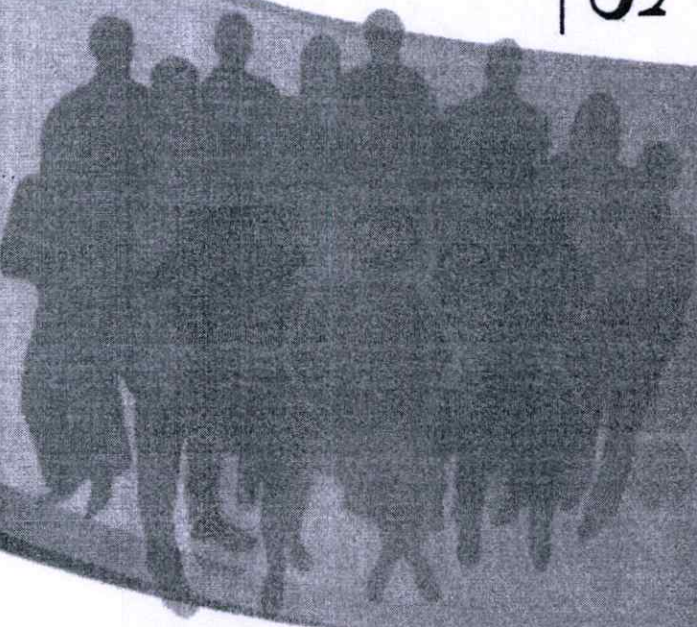
Proceedings of the Conference on Knowledge and Smart Technologies

July 24-25, 2009

PROCEEDINGS







อัลกอริทึมเคมีนคลัสเตอร์ริงแบบขนานที่มีประสิทธิภาพบนระบบคลัสเตอร์
An Efficient Parallel K-means Clustering Algorithm on a Cluster System

นเรศ ผ่องสวัสดิ์กุล[1] และ จีรพร วีระพันธุ์[2]
สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง กรุงเทพฯ 10520
E-Mail: [1]s0067505@kmitl.ac.th และ [2]ksjeerap@kmitl.ac.th

Abstract

K-means clustering algorithm is applied to classify or to group objects into K groups by considering their attributes or features and can be used for developing applications in various fileds. However, the problem of those applications are time consuming and out of memory because the data are too large. In this research we present the efficient K-means clustering algorithm by introducing the new method for selecting initial cluster centers and applying a parallel method to solve the problem. To evaluate the performance of the proposed algorithm, a number of experiments were performed on a cluster system. The investigated results showed that the response time was improved and the system can support more data.

Keyword: parallel clustering algorithm

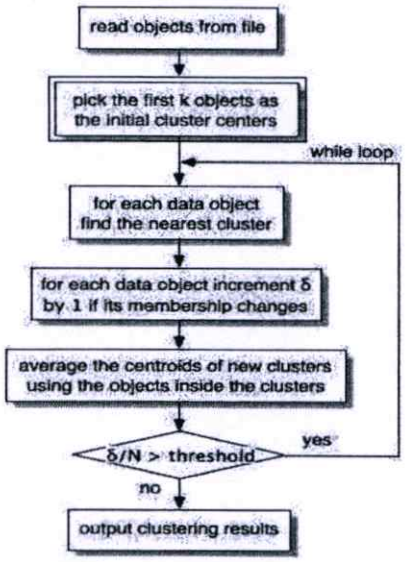
บทคัดย่อ

เคมีนคลัสเตอร์ริงเป็นอัลกอริทึมสำหรับจัดกลุ่มวัตถุต่างๆ ตามจำนวนกลุ่มที่ต้องการ โดยพิจารณาจากคุณลักษณะของวัตถุเหล่านั้นและถูกนำไปพัฒนาแอพพลิเคชันจำนวนมากในหลากหลายสาขา แต่การนำไปประยุกต์ใช้ต่างก็ประสบปัญหาเดียวกันคือ เวลาและหน่วยความจำที่ต้องใช้ในการประมวลผลมากเกินไป เนื่องจากข้อมูลมีขนาดใหญ่ งานวิจัยนี้เรานำเสนอวิธีการเพิ่มประสิทธิภาพของอัลกอริทึมเคมีนคลัสเตอร์ริงโดยนำเสนออัลกอริทึมสำหรับเลือกชุดของจุดศูนย์กลางเริ่มต้น และนำวิธีการประมวลผลแบบขนานบนระบบคลัสเตอร์มาประยุกต์ใช้ ซึ่งผลการทดลองแสดงให้เห็นว่าเวลาที่ต้องใช้สำหรับการประมวลลดลง และสามารถรองรับปัญหาขนาดใหญ่ได้มากกว่าเดิม

คำสำคัญ: อัลกอริทึมเคมีนคลัสเตอร์ริงแบบขนาน

1. บทนำ

อัลกอริทึมเคมีนคลัสเตอร์ริง (K-means Clustering Algorithm) เป็นวิธีการแบ่งกลุ่มข้อมูลของวัตถุทั้งหมดออกตามจำนวนกลุ่มที่ต้องการ K กลุ่ม และค่า K ต้องน้อยกว่าจำนวนของวัตถุทั้งหมด (N) ซึ่งจำนวนกลุ่มต้องเป็นเลขจำนวนเต็มบวก และการจัดกลุ่มวัตถุต้องอาศัยความเหมือนของวัตถุ โดยวัดจากระยะห่างที่น้อยที่สุดระหว่างวัตถุกับจุดศูนย์กลางของกลุ่มทั้งหมด เพื่อจัดวัตถุเข้าสู่กลุ่มต่างๆ ตามจำนวนกลุ่มที่ต้องการ นิยมใช้การวัดระยะแบบยูคลิด (Euclidean distance) และจุดศูนย์กลางเหล่านั้นเป็นตัวแทนของกลุ่มต่างๆ สามารถคำนวณได้จากค่าเฉลี่ยของวัตถุที่อยู่ภายในกลุ่มเดียวกัน ซึ่งวัตถุเหล่านั้นมีคุณลักษณะเหมือนกัน ส่วนวัตถุที่อยู่ต่างกลุ่มก็มีคุณลักษณะที่แตกต่างกัน และอัลกอริทึมเคมีนแบบลำดับ เมื่อใช้หนึ่งหน่วยประมวลผลต้องใช้เวลาสำหรับการประมวลผลเท่ากับ O(RKN) และมีขั้นตอนย่อยในการทำงาน ดังนี้[8]



รูปที่ 1. อัลกอริทึมเคมีน

แอพพลิเคชันจำนวนมากนำเอาอัลกอริทึมเคมีนมาประยุกต์ใช้ในหลากหลายสาขา โดยเฉพาะแอพพลิเคชันในการทำเหมืองข้อมูล (Data Mining) ซึ่งต่างก็ต้องการอัลกอริทึมที่ประมวลผลได้อย่างรวดเร็วกับข้อมูลขนาดใหญ่ ยกตัวอย่างเช่น ระบบสังเกตการณ์โลก (The Earth Observing System: EOS) ขององค์การนาซา (NASA)[6] ระบบนี้สามารถสร้างข้อมูลจำนวนมหาศาล (หน่วยเป็นเทราไบต์ต่อวัน) ข้อมูลเหล่านั้นถูกนำมาใช้ในการระบุความผิดปกติที่เกิดขึ้นบนโลกของเรา ซึ่งการจัดกลุ่มของข้อมูลสามารถทำได้โดยใช้อัลกอริทึมคลัสเตอร์ริง

อย่างไรก็ตาม เมื่อข้อมูลมีขนาดใหญ่จะส่งผลกระทบต่อเวลา และหน่วยความจำที่จำเป็นต้องใช้สำหรับการประมวลผล ดังนั้นแอพพลิเคชันมากมายที่นำอัลกอริทึมเคมีนแบบลำดับมาใช้ต้องประสบกับปัญหาเรื่องเวลาและหน่วยความจำที่ต้องใช้ในการประมวลผลมากเกินไป เนื่องจากข้อมูลมีขนาดใหญ่ และเพื่อลดเวลาในการประมวลผล ได้มีงานวิจัยเกี่ยวกับอัลกอริทึมสำหรับเลือกชุดของจุดศูนย์กลางเริ่มต้น [1,5,10] ดังแสดงในรูปที่ 1 (ขั้นตอนที่ 2) ก่อนการจัดกลุ่มวัตถุด้วยอัลกอริทึมเคมีนคลัสเตอร์ริง ซึ่งช่วยให้การจัดกลุ่มวัตถุรวดเร็วยิ่งขึ้น เนื่องจากจำนวนรอบของการจัดกลุ่มวัตถุลดลง แต่ยังต้องประสบกับปัญหาการมีหน่วยความจำไม่เพียงพอสำหรับการประมวลผล เนื่องจากอาศัยพื้นที่หน่วยความจำบนคอมพิวเตอร์เพียงเครื่องเดียวเท่านั้น และเร็วๆ นี้มีงานวิจัยเกี่ยวกับอัลกอริทึมเคมีนคลัสเตอร์ริงแบบขนานบนเครือข่ายของคอมพิวเตอร์ [3,4,6,8,11] ซึ่งช่วยทำให้เวลาที่จำเป็นต้องใช้สำหรับการประมวลผลลดลง และสามารถประมวลผลข้อมูลที่มีขนาดใหญ่มากกว่าคอมพิวเตอร์เพียงเครื่องเดียวจะสามารถประมวลผลได้ เนื่องจากใช้คอมพิวเตอร์หลายเครื่องในการประมวลผล

งานวิจัยนี้นำเสนอวิธีการเพิ่มประสิทธิภาพของอัลกอริทึมเคมีนด้านความเร็วในการประมวลผล และสามารถประมวลผลข้อมูลที่มีขนาดใหญ่ขึ้นได้ โดยนำเสนออัลกอริทึมสำหรับเลือกชุดของจุดศูนย์กลางเริ่มต้นแบบใหม่ และนำวิธีการประมวลผลแบบขนานมาประยุกต์ใช้ร่วมกัน

หัวข้อต่อไปที่จะกล่าวถึงคือ งานวิจัยที่เกี่ยวข้อง หัวข้อที่ 3 งานวิจัยที่นำเสนอ หัวข้อที่ 4 การวิเคราะห์ประสิทธิภาพของอัลกอริทึม หัวข้อที่ 5 การทดลองและหัวข้อสุดท้ายคือ สรุปผล

2. งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้องที่เราทำการศึกษาแบ่งออกเป็น 2 ส่วนคือ งานวิจัยเกี่ยวกับอัลกอริทึมสำหรับเลือกชุดของจุดศูนย์กลางเริ่มต้น [1,10] และงานวิจัยเกี่ยวกับอัลกอริทึมเคมีนคลัสเตอร์ริงแบบขนาน [6,8,11] ซึ่งในงานวิจัย[10] แสดงวิธีการของลอยด์ (Lloyd-Type Methods) เลือกจุดศูนย์กลาง 2 จุดแรก โดยเลือกคู่ของวัตถุ $x, y \in X$ ด้วยค่าความน่าจะเป็นดังนี้ $\| x - y \|^2$ ส่วนจุดศูนย์กลางจุดถัดไปเลือกจากวัตถุ x ด้วยค่าความน่าจะเป็นดังนี้ $\min_{j \in \{1,\ldots,i\}} \| x - c_j \|^2$ เมื่อ X เป็นชุดของข้อมูลทั้งหมด ส่วน i คือ จำนวนของจุดศูนย์กลางที่ถูกเลือกไว้แล้ว และ c คือจุดศูนย์กลางของกลุ่ม และงานวิจัย[1] เลือกจุดศูนย์กลางจุดแรก โดยสุ่มจากวัตถุทั้งหมด และเลือกจุดศูนย์กลางจุดถัดไปด้วยค่าความน่าจะเป็นดังนี้

$D(x)^2$ เมื่อ $x \in X$ และ $D(x)$ คือ ระยะห่างระหว่างวัตถุกับจุดศูนย์กลางของวัตถุนั้น ซึ่งถูกเรียกว่า "D² Weighting" อัลกอริทึมสำหรับเลือกชุดของจุดศูนย์กลางเริ่มต้นมีความสำคัญมากต่อการจัดกลุ่มวัตถุ เนื่องจากอัลกอริทึมเคมีนต้องจัดกลุ่มวัตถุวนซ้ำไปเรื่อยๆ จนกว่าวัตถุทั้งหมดจะไม่มีการเปลี่ยนกลุ่ม ซึ่งต้องใช้เวลาในการประมวลผลเท่ากับ O(RKN) [6] เมื่อ R คือ จำนวนรอบของการทำงานซ้ำ ส่วน K คือ จำนวนกลุ่มที่ต้องการ และ N คือ จำนวนของวัตถุทั้งหมด จากผลการทดลองพบว่าอัลกอริทึมสำหรับเลือกชุดของจุดศูนย์กลางเริ่มต้นเหล่านี้ช่วยลดจำนวนรอบของการทำงานลง จึงส่งผลให้เวลาที่จำเป็นต้องใช้ในการประมวลผลลดลงตามไปด้วย

ส่วนต่อมาเป็นงานวิจัยเกี่ยวกับอัลกอริทึมเคมีนคลัสเตอร์ริงแบบขนาน ซึ่งในงานวิจัย[6] ออกแบบให้แต่ละโปรเซสเซอร์จัดกลุ่มวัตถุที่อยู่กลุ่มเดียวกันเท่านั้น งานวิจัย [11] ออกแบบให้แต่ละโปรเซสเซอร์จัดกลุ่มวัตถุของทุกกลุ่ม และกระบวนหลักทำหน้าที่เลือกชุดของจุดศูนย์กลางแล้วกระจายไปยังทุกกระบวนการย่อย ซึ่งโปรเซสเซอร์หลักรับผิดชอบงานของทั้ง 2 กระบวนการ และงานวิจัย [8]ได้พัฒนาให้แต่ละกระบวนการทำหน้าที่เลือกชุดของจุดศูนย์กลางเอง ซึ่งวิธีการประมวลผลแบบขนานหรือแบบกระจายมีประโยชน์เป็นอย่างมากในการเพิ่มประสิทธิภาพของอัลกอริทึมเคมีน โดยทำการแบ่งข้อมูลไปยังหลายโปรเซสเซอร์ เพื่อให้โปรเซสเซอร์เหล่านั้นช่วยกันจัดกลุ่มวัตถุพร้อมๆ กันทำให้เวลาที่จำเป็นต้องใช้สำหรับการประมวลผลลดลง และสามารถรองรับปัญหาขนาดใหญ่มากกว่าคอมพิวเตอร์เครื่องเดียวจะสามารถประมวลผลได้

3. งานวิจัยที่นำเสนอ

3.1 อัลกอริทึมสำหรับเลือกชุดของจุดศูนย์กลางเริ่มต้น

แนวคิดของอัลกอริทึมที่นำเสนอสำหรับเลือกชุดของจุดศูนย์กลางเริ่มต้นแบบใหม่คือ เลือกจุดศูนย์กลาง 2 จุดแรกให้เป็นขอบของข้อมูลทั้งหมด ส่วนจุดศูนย์กลางจุดถัดไปทำหน้าที่แบ่งข้อมูลออกทีละครึ่งหนึ่ง โดยเลือกจุดศูนย์กลางของกลุ่มทีละจุดสลับกับการจัดกลุ่มวัตถุ ซึ่งอัลกอริทึมนี้ต้องใช้เวลาสำหรับการประมวลผลเท่ากับ O(KN) และมีขั้นตอนย่อยในการทำงาน ดังนี้

1. เลือกวัตถุที่มีระยะห่างน้อยที่สุดจากพิกัดศูนย์
2. กำหนดจุดศูนย์กลางจากวัตถุที่ถูกเลือกไว้
3. จัดกลุ่มวัตถุทั้งหมดตามระยะห่างที่น้อยที่สุดระหว่างวัตถุกับจุดศูนย์กลางที่ถูกกำหนดไว้
4. เลือกวัตถุที่มีระยะห่างมากที่สุดจากทุกกลุ่ม
5. ทำงานซ้ำขั้นตอนที่ 2 จนกว่าจุดศูนย์กลางจะครบ K จุดศูนย์กลาง
6. คำนวณค่าเฉลี่ยของแต่ละกลุ่มเพื่อกำหนดเป็นชุดของจุดศูนย์กลางเริ่มต้น

3.2 อัลกอริทึมเคมีนแบบขนาน

แนวคิดของอัลกอริทึมเคมีนแบบขนานที่นำเสนอคล้ายกับวิธีการในงานวิจัย[8] คือ ออกแบบให้แต่ละโปรเซสเซอร์จัดกลุ่มวัตถุของทุกกลุ่ม และกระบวนการหลัก (Master Process) ทำหน้าที่เลือกจุดศูนย์กลางเริ่มต้นแบบใหม่ ซึ่งพัฒนาตามแบบจำลองการเขียนโปรแกรมเพียงโปรแกรมเดียว และใช้โปรแกรมนี้ทำงานบนทุกโปรเซสเซอร์กับชุดข้อมูลที่ต่างกัน (SPMD) ดังนี้

Master Process

1. Read objects from file.
2. Randomly form equal subsets follow about number of processors.
3. Select initial centers.
4. Send each subset to each slave process.
5. Broadcast initial centers to all slave processes.
6. Receive result sets from all slave processes.

Slave Process

1. Receive a subset P and set C from master process.
2. Assign internal objects into internal clusters and count of objects moved.
3. Calculate sum and count of objects into internal cluster.
4. Broadcast sum, count of objects and count of objects moved to all slave processes.
5. Calculate new centers by using sum and counts.
6. Repeat step 2 if objects moved.
7. Send result set to master process.

4. การวิเคราะห์ประสิทธิภาพของอัลกอริทึม

การวิเคราะห์เวลาที่เครื่องคอมพิวเตอร์ต้องใช้ในการประมวลผลอัลกอริทึม (Time Complexity) แบ่งเป็น 2 ส่วน คือ เวลาสำหรับการสื่อสารระหว่างโปรเซสเซอร์ (Communication Time: T_{comm}) และเวลาสำหรับการประมวลผล (Computation Time: T_{comp}) ซึ่งอัลกอริทึมเคมีนแบบขนานที่นำเสนอมีเวลาสำหรับการสื่อสาร และเวลาสำหรับการประมวลผลทั้งหมด 10 ขั้นตอน ดังนี้

ขั้นตอน 1: กระบวนการหลักเลือกชุดของจุดศูนย์กลางเริ่มต้น

$$T_{comp1} = KN$$

ขั้นตอน 2: กระบวนการหลักส่งกลุ่มข้อมูลไปยังแต่ละกระบวนการย่อย

$$T_{comm1} = nProc(T_{startup} + N/nProc\ T_{data})$$

ขั้นตอน 3: กระบวนการหลักกระจายชุดของจุดศูนย์กลางเริ่มต้นไปยังทุกกระบวนการย่อย

$$T_{comm2} = T_{startup} + |C|\ T_{data}$$

ขั้นตอน 4: กระบวนการย่อยนำข้อมูลของวัตถุและชุดของจุดศูนย์กลางเริ่มต้นมาจัดกลุ่มและนับจำนวนวัตถุที่มีการเปลี่ยนกลุ่ม

$$T_{comp2} = K|P|$$

ขั้นตอน 5: กระบวนการย่อยคำนวณผลรวมและนับจำนวนสมาชิกของแต่ละกลุ่ม

$$T_{comp3} = |P|$$

ขั้นตอน 6: แต่ละกระบวนการย่อยกระจายผลรวมไปยังทุกกระบวนการย่อย

$$T_{comm3} = T_{startup} + |C|\ T_{data}$$

ขั้นตอน 7: แต่ละกระบวนการย่อยกระจายจำนวนสมาชิกไปยังทุกกระบวนการย่อย

$$T_{comm4} = T_{startup} + K\ T_{data}$$

ขั้นตอน 8: แต่ละกระบวนการย่อยกระจายจำนวนวัตถุที่มีการเปลี่ยนกลุ่มไปยังทุกกระบวนการย่อย

$$T_{comm5} = T_{startup} + T_{data}$$

ขั้นตอน 9: กระบวนการย่อยนำผลรวม และจำนวนสมาชิกของแต่ละกลุ่มมาคำนวณหาค่าเฉลี่ย

$$T_{comp4} = |C|$$

ขั้นตอน 10: แต่ละกระบวนการย่อยส่งผลลัพธ์การจัดกลุ่มกลับไปยังกระบวนการหลัก

$$T_{comm6} = T_{startup} + G\ T_{data}$$

จากอัลกอริทึมเคมีนแบบขนานพบว่ากระบวนการย่อยมีการทำงานซ้ำตั้งแต่ขั้นตอนที่ 2 จนถึงขั้นตอนที่ 5 ซึ่ง TCM เป็นผลรวมของเวลาที่ใช้สำหรับการสื่อสาร และ TCP เป็นผลรวมของเวลาที่ใช้สำหรับการประมวลผล ฉะนั้นเวลาที่เครื่องคอมพิวเตอร์ต้องใช้ในการประมวลผลสามารถคำนวณได้จากสมการ ดังนี้

$$\text{Total Time} = TCM + TCP$$

$$\begin{aligned} TCM &= T_{comm1} + T_{comm2} + T_{comm3} + T_{comm4} + T_{comm5} + T_{comm6} \\ &= nProc\,(T_{startup} + N/nProc\ T_{data}) + T_{startup} + |C|\ T_{data} + R\,(T_{startup} + |C|\ T_{data}) + R\,(T_{startup} + K\ T_{data}) + R\,(T_{startup} + T_{data}) + T_{startup} + G\ T_{data} \\ &= (nProc + 3R + 2)\ T_{startup} + (N + |C| + R\,(|C| + K + 1) + G)\ T_{data} \\ &= O(N + R|C|) \end{aligned}$$

$$\begin{aligned} TCP &= T_{comp1} + T_{comp2} + T_{comp3} + T_{comp4} \\ &= KN + RK|P| + R|P| + R|C| \\ &= KN + R\,(K|P| + |P| + |C|) \\ &= O(KN + RK|P|) \end{aligned}$$

$$\begin{aligned} \text{Total Time} &= O(N + R|C|) + O(KN + RK|P|) \\ &= O(KN + RK|P|) \end{aligned}$$

การวิเคราะห์หน่วยความจำที่เครื่องจำเป็นต้องใช้ในการประมวลผลอัลกอริทึม (Space Complexity) เนื่องจากกระบวนการหลักทำการแบ่งข้อมูลทั้งหมดไปยังทุกกระบวนย่อยแบบสุ่ม ซึ่งแต่ละโปรเซสเซอร์ทำงานอยู่บนเครื่องคอมพิวเตอร์ต่างเครื่องกัน และแต่ละเครื่องมีหน่วยความจำเป็นของตัวเอง ดังนั้นอัลกอริทึมเคมีนแบบขนานจึงสามารถรองรับข้อมูลที่มีขนาดใหญ่ขึ้น และต้องใช้หน่วยความจำทั้งหมดเท่ากับ O(N) [6]

จากอัลกอริทึมเคมีนแบบขนาน $T_{startup}$ เป็นเวลาสำหรับเตรียมการเริ่มต้นในการส่งข้อมูลและ T_{data} เป็นเวลาสำหรับส่งข้อมูลไปยังโปรเซสเซอร์หน่วยอื่น ซึ่งเวลาทั้ง 2 เป็นเวลาคงที่ขึ้นอยู่กับระบบ [6] ส่วน nProc คือจำนวนของโปรเซสเซอร์ทั้งหมดที่ใช้ในการประมวลผล

เซต N คือ ข้อมูลของวัตถุทั้งหมดที่ต้องการจัดกลุ่ม เซต P คือ ข้อมูลของวัตถุที่แต่ละโปรเซสเซอร์รับผิดชอบ ดังนั้น $|P| \approx N/nProc$ เซต G คือ ผลลัพธ์ของการจัดกลุ่มวัตถุในกระบวนการย่อย และเซต C คือ ชุดของจุดศูนย์กลาง ซึ่งประกอบด้วยจุดศูนย์กลางของทุกกลุ่ม

5. การทดลอง

การทดลองในงานวิจัยนี้เป็นการพัฒนาอัลกอริทึมสำหรับเลือกจุดศูนย์กลางเริ่มต้น และอัลกอริทึมเคมีนคลัสเตอร์ริงแบบขนานบนระบบคลัสเตอร์แบบเนื้อเดียว (Homogeneous Cluster System) ซึ่งใช้ระบบการส่งผ่านข้อความ (Message Passing Interface: MPI) เพื่อทำให้กลุ่มของคอมพิวเตอร์สามารถติดต่อสื่อสารและทำงานร่วมกันได้ โดยใช้ภาษาซีในการพัฒนาโปรแกรม ส่วนไฟล์ข้อมูลที่นำมาใช้ในการทดลองมีรายละเอียด ดังนี้

ตาราง 1. ชุดข้อมูลสำหรับการทดลอง

Data set	Quantity	Attributes	Size: KB
Cloud	1,024	10	106
Spam	4,601	58	686
Intrusion	494,019	35	63,539

ส่วนแรกคือ การทดลองวัดประสิทธิภาพของอัลกอริทึมเคมีนแบบลำดับ สำหรับหนึ่งหน่วยประมวลผล โดยพิจารณาจากเวลาที่ต้องใช้สำหรับการประมวลผล และจำนวนรอบของการทำงานซ้ำ ซึ่งนำอัลกอริทึมสำหรับเลือกชุดของจุดศูนย์กลางเริ่มต้นแบบต่างๆ มาประยุกต์ใช้กับอัลกอริทึมเคมีนแบบลำดับ ดังนี้

อัลกอริทึม 1: เลือกชุดของจุดศูนย์กลางเริ่มต้น โดยเลือกวัตถุที่อยู่ลำดับแรกจนถึงลำดับที่ K จากวัตถุทั้งหมด [8]

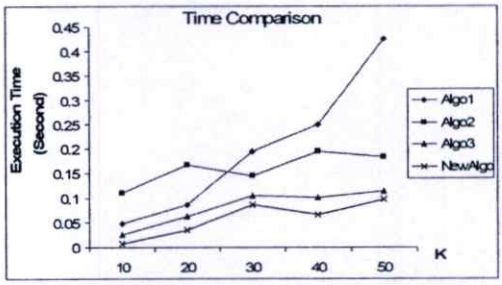
อัลกอริทึม 2: เลือกจุดศูนย์กลาง 2 จุดแรก โดยเลือกคู่ของวัตถุที่มีระยะห่างมากที่สุด และเลือกจุดศูนย์กลางจุดถัดไปจากวัตถุที่มีระยะห่างมากที่สุดระหว่างวัตถุกับจุดศูนย์กลางของวัตถุนั้น [10]

อัลกอริทึม 3: เลือกจุดศูนย์กลางจุดแรก โดยสุ่มจากวัตถุทั้งหมด และเลือกจุดศูนย์กลางจุดถัดไปจากวัตถุที่มีระยะห่างมากที่สุดระหว่างวัตถุกับจุดศูนย์กลางของวัตถุนั้น [1]

อัลกอริทึมที่นำเสนอ: เลือกจุดศูนย์กลางจุดแรกจากวัตถุที่อยู่ใกล้กับพิกัดศูนย์มากที่สุด และเลือกจุดศูนย์กลางจุดถัดไปจากวัตถุที่มีระยะห่างมากที่สุดระหว่างวัตถุกับจุดศูนย์กลางของวัตถุนั้นจนครบถ้วน แล้วคำนวณค่าเฉลี่ยเพื่อกำหนดเป็นชุดของจุดศูนย์กลางเริ่มต้น

ตาราง 2. เวลาที่ต้องใช้สำหรับการประมวลผล

K	Algo1	Algo2	Algo3	NewAlgo
10	0.0469	0.1094	0.0262	0.0078
20	0.0859	0.1680	0.0625	0.0352
30	0.1953	0.1445	0.1055	0.0859
40	0.2500	0.1953	0.1016	0.0664
50	0.4258	0.1836	0.1133	0.0977



รูปที่ 2. เปรียบเทียบเวลาที่ต้องใช้สำหรับการประมวลผล

ตาราง 3. จำนวนรอบของการทำงานซ้ำ

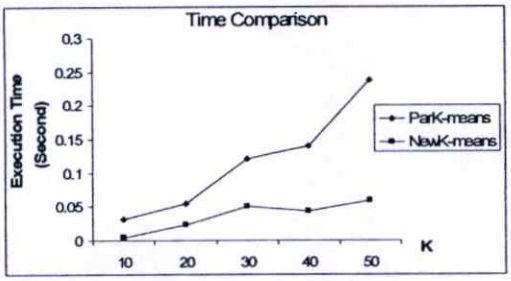
K	Algo1	Algo2	Algo3	NewAlgo
10	32	31	17	2
20	31	36	22	11
30	48	19	24	20
40	46	23	17	11
50	65	16	15	13

จากผลการทดลองของอัลกอริทึมเคมีนแบบลำดับแสดงให้เห็นว่าอัลกอริทึมสำหรับเลือกชุดของจุดศูนย์กลางเริ่มต้นที่นำเสนอ เมื่อถูกนำไปประยุกต์ใช้กับอัลกอริทึมเคมีนแล้วทำให้ประสิทธิภาพของอัลกอริทึมเคมีนดีที่สุด ซึ่งมีเพียงอัลกอริทึมเคมีนแบบที่ 3 เท่านั้นที่เวลาสำหรับการประมวลผล และจำนวนรอบของการทำงานซ้ำถูกบันทึกเป็นค่าเฉลี่ย เนื่องจากจุดศูนย์กลางจุดแรกถูกเลือกโดยการสุ่มทำให้ชุดของจุดศูนย์กลางเริ่มต้นที่ได้รับมาขาดความแน่นอน ส่งผลให้เวลาที่ต้องใช้สำหรับการประมวลผล และจำนวนรอบของการทำงานซ้ำอาจไม่เท่ากันในการประมวลผลแต่ละครั้ง

ส่วนสุดท้ายคือ การทดลองวัดประสิทธิภาพของอัลกอริทึมเคมีนแบบขนาน โดยทำการเปรียบเทียบประสิทธิภาพระหว่างอัลกอริทึมเคมีนแบบขนานแบบเดิมซึ่งใช้อัลกอริทึมสำหรับเลือกชุดของจุดศูนย์กลางเริ่มต้นแบบแรก [8] กับอัลกอริทึมเคมีนแบบขนานแบบใหม่ที่นำเสนอ โดยใช้หน่วยประมวลผลจำนวน 2 หน่วย ในการทดลองครั้งนี้

ตาราง 4. ผลการเปรียบเทียบประสิทธิภาพอัลกอริทึมเคมีนแบบขนาน

K	ParK-means		NewK-means	
	Itera	Time	Itera	Time
10	32	0.0313	2	0.0039
20	31	0.0547	11	0.0234
30	48	0.1211	20	0.0508
40	46	0.1406	11	0.0430
50	65	0.2383	13	0.0586



รูปที่ 3. ผลการเปรียบเทียบประสิทธิภาพอัลกอริทึมเคมีนแบบขนาน

จากผลการทดลองของอัลกอริทึมเคมีนแบบขนานแสดงให้เห็นว่าอัลกอริทึมเคมีนแบบขนานแบบใหม่มีประสิทธิภาพสูงกว่าแบบเดิม เนื่องจากนำอัลกอริทึมสำหรับเลือกชุดของจุดศูนย์กลางเริ่มต้นที่นำเสนอมาประยุกต์ใช้ โดยอัลกอริทึมเคมีนแบบขนานจะมีจำนวนรอบของการทำงานซ้ำเท่ากับอัลกอริทึมเคมีนแบบลำดับเสมอ เมื่อจำนวนกลุ่มที่ต้องการเท่ากัน(K) ถึงแม้ว่าการเลือกชุดของจุดศูนย์กลางเริ่มต้นเป็นความรับผิดชอบของกระบวนการหลัก (Master Process) เท่านั้น แต่เวลาที่ใช้ในการประมวลผลส่วนใหญ่จะอยู่ในกระบวนการย่อย (Slave Process) ซึ่งเป็นผลเนื่องมาจากวิธีการของอัลกอริทึมเคมีนคลัสเตอร์ริง และจำนวนรอบที่ใช้ ดังนั้นวิธีการที่นำเสนอในงานวิจัยนี้ช่วยทำให้อัลกอริทึมเคมีนคลัสเตอร์ริงแบบขนานมีประสิทธิภาพสูงยิ่งขึ้น

6. สรุปผล

งานวิจัยนี้นำเสนอวิธีการเพิ่มประสิทธิภาพของอัลกอริทึมเคมีนคลัสเตอร์ริงให้ดียิ่งขึ้นเพื่อตอบสนองกับความต้องการของแอพพลิเคชันในปัจจุบันที่ต้องทำงานกับข้อมูลขนาดใหญ่ ซึ่งก่อให้เกิดปัญหาที่สำคัญมากคือ เวลาและหน่วยความจำที่ต้องใช้สำหรับประมวลผลมากเกินไป ดังนั้นเราจึงนำเสนออัลกอริทึมสำหรับเลือกชุดของจุดศูนย์กลางเริ่มต้นแบบใหม่ และนำวิธีการประมวลผลแบบขนานมาประยุกต์ใช้ร่วมกัน จากผลการทดลองแสดงให้เห็นว่าเวลาที่ต้องใช้สำหรับการประมวลผลลดลง และสามารถรองรับปัญหาที่มีขนาดใหญ่มากกว่าคอมพิวเตอร์เพียงเครื่องเดียวจะสามารถประมวลผลได้

7. เอกสารอ้างอิง

[1] D. Arthur, and S. Vassilvitskii, "k-means++: The Advantages of Careful Seeding", SODA, 2007.

[2] D. Arthur and S. Vassilvitskii, "K-means++ test code",http://www.stanford.edu/_darthur/kMeanspp Test.zip.

[3] H. BISGIN, "Parallel Clustering Algorithm With Application to Climatology", Computational Science and Engineering, December 2007.

[4] M. N. Joshi, "Parallel *K* - Means Algorithm on Distributed Memory Multiprocessors", Computer Science Department University of Minnesota, Twin Cities, spring 2003.

[5] C. H. Jun, J. S. Lee, and H. S. Park, "A K-means-like Algorithm for K-medoids Clustering and Its Performance", Proceedings of the 36th CIE Conference on Computers & Industrial Engineering, pp.1222-1231, Taipei, Taiwan, Jun. 20-23, 2006.

[6] S. Kantabutra, "Parallel K-Means Clustering Algorithm on NOW", September 1999.

[7] S. Kantabutra, P. Kornpitak, and C. Naramittakapong , "Pipelined K-means Algorithm on COWs", ISCIT , September 03-05, 2003, Hatyai, Songkhla, Thailand.

[8] W. Liao, "The Software Package of Parallel K-means", http://www.ece.northwestern.edu/~wkliao/ Kmeans/index.html, 2005.

[9] J. Mao, L. Ou, Z. Xiong, and Y. Zhang, "The Study of Parallel K-Means Algorithm", Proceedings of the 6th World Congress on Intelligent Control and Automation, June 21 - 23, 2006, Dalian, China.

[10] R. Ostrovsky, Y. Rabani, L. J. Schulman, and C. Swamy, The Effectiveness of Lloyd-Type Methods for the k-Means Problem", IEEE, 2006.

[11] E. SÜLÜN, "Improvements in K-means Algorithm to Execute on Large Amounts of Data", http://library.iyte.edu.tr/tezler/master/bilgisayaryaziliz/ T000441.pdf, October 2004.

ประวัติผู้เขียน

ชื่อ - สกุล	นายนเรศ ผ่องสวัสดิ์กุล
วัน เดือน ปีเกิด	18 ธันวาคม 2527
ที่อยู่	81 ซอย 1 ถ.ท่าแฉลบ ต.ตลาด อ.เมือง จ.จันทบุรี 22000

ประวัติการศึกษา

2550	จบการศึกษาปริญญาวิทยาศาสตรบัณฑิต สาขาวิชาระบบสารสนเทศคอมพิวเตอร์ คณะวิทยาศาสตร์และศิลปศาสตร์ มหาวิทยาลัยบูรพา วิทยาเขตสารสนเทศจันทบุรี
2553	จบการศึกษาหลักสูตรประกาศนียบัตรบัณฑิตวิชาชีพครู คณะครุศาสตร์ มหาวิทยาลัยราชภัฏสวนดุสิต
2550 - ปัจจุบัน	กำลังศึกษาปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง