การศึกษาการจัดกลุ่มข้อมูลโดยใช้ไฮบริดพาร์ทิเคิลสวอมออปติไมเซชั่น

A STUDY OF DATA CLUSTERING USING HYBRID PARTICLE SWARM OPTIMIZATION

โดย

ศิริพร ระเมียดดี

SIRIPORN RAMEADDEE

อาจารย์ที่ปรึกษา

รศ.ดร.อาริต ธรรมโน

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ภาคเรียนที่ 1 ปีการศึกษา 2549

A STUDY OF DATA CLUSTERING USING HYBRID PARTICLE SWARM OPTIMIZATION

SIRIPORN RAMEADDEE

A SYSTEM DEVELOPMENT PROJECT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF SCIENCE PROGRAM IN INFORMATION TECHNOLOGY

FACULTY OF INFORMATION TECNOLOGY

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

1/ 2006

ชื่อหัวข้อ	การศึกษาการจัดกลุ่มข้อมูลโดยใช้ไฮบริดพาร์ทิเคิลสวอมออปติไมเซชั่น
นักศึกษา	นางสาวศิริพร ระเมียดดี
รหัสนักศึกษา	47066421
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
ปีการศึกษา	2549
อาจารย์ที่ปรึกษา	รศ.ดร.อาริต ธรรมโน

บทคัดย่อ

เนื่องจากฐานข้อมูลมีแนวโน้มจะเพิ่มขึ้นต่อไปในอนาคตและเป็นที่ต้องการสำหรับดาต้าไมนิ่ง ซึ่งดาต้าไมนิ่งจะเน้นไปที่เทคนิคและเครื่องมือในการค้นหาสารสนเทศและความรู้ที่เป็นประโยชน์จากข้อมูล ท่ามกลางข้อมูลที่มากมายเทคนิคการจัดกลุ่มจึงเป็นหัวข้อที่น่าสนใจและเติบโตอย่างรวดเร็ว Clustering เป็นกระบวนการจัดกลุ่มข้อมูลที่ขึ้นอยู่กับความคล้ายคลึงกันของข้อมูล มีการพิสูจน์ว่ามีแนวโน้มจะมีเทคนิคใหม่ ๆ เกิดขึ้นเสมอ ในโครงการศึกษาฉบับนี้จะศึกษาถึงหลักการและวิธีการทำงานของอัลกอริทึม k-Means รวมถึงเสนอเทคนิควิธีการใหม่ในการจัดกลุ่มข้อมูลโดยใช้พาร์ทิเคิลสวอมออปติไมเซชั่นโดยจะแสดงให้เห็นว่า PSO สามารถใช้หาตัวแทนของคลัสเตอร์ที่ถูกระบุโดยผู้ใช้ได้อย่างไร และมีการเปรียบเทียบประสิทธิภาพ Hybrid PSO กับการจัดกลุ่มของอัลกอริทึม k-Means ผลลัพธ์ที่ได้จะแสดงให้เห็นว่าเทคนิคการจัดกลุ่มโดยใช้ Hybrid PSO นั้นมีความสามารถเป็นอย่างมากเพื่อที่จะนำไปใช้ในการพัฒนาระบบต่อไป

Title	A Study Of Data Clustering using Hybrid Particle Swarm Optimization
Student	Miss Siriporn Rameaddee
Student ID.	47066421
Degree	Master of Science
Programme	Information Technology Management
Academic	2006
Advisor	Assoc. Prof. Dr. Arit Thammano

ABSTRACT

Due to the database size has the trend to be increased in the future and demand for data mining. Data mining focuses on using efficient techniques and tools to discover useful information and knowledge from data. Among various data mining, Clustering technique is an interesting and fast growing topic. Clustering is the process of grouping data based on similarity. Witness a resurgence of interest in new clustering technique. In this project presents k-Means algorithm and describes method of algorithm. Also proposes new approaches to using PSO to cluster data. It is shown how PSO can be used to find the centroids of a user specified number of clusters. And compared to the performance of k-Means clustering. Result show that Hybrid PSO techniques have much potential to we can bring to implement system.

กิตติกรรมประกาศ

ในโครงการศึกษานี้เป็นการพัฒนาเครื่องมือเพื่อช่วยในการจัดกลุ่มข้อมูล (Clustering) โดยใช้ไฮบริดพาร์ทิเคิลสวอมออปติไมเซชั่น ซึ่งจะไม่สามารถทำให้ประสบความสำเร็จได้ถ้าไม่ได้รับการช่วยเหลือและแรงสนับสนุนมาจากบุคคลที่สำคัญหลาย ๆ ท่าน ดังต่อไปนี้

บิดามารดาผู้อบรมสั่งสอน และบุคคลในครอบครัวที่ให้การสนับสนุนในด้านต่าง ๆ รวมถึงในเรื่องของการศึกษา

รศ.ดร.อาริต ธรรมโน อาจารย์ที่ปรึกษาโครงการ ซึ่งให้ความช่วยเหลือและความกรุณาให้คำแนะนำและเป็นที่ปรึกษา อันเป็นประโยชน์อย่างยิ่งต่อการพัฒนาโครงการศึกษาครั้งนี้

เพื่อน ๆ และรุ่นพี่ทุกคนที่คอยให้ความช่วยเหลือ ให้คำแนะนำในด้านต่าง ๆ ไม่ว่าจะเป็นกำลังใจและคำปรึกษา รวมถึงน้ำใจดี ๆ ที่มีให้กันเสมอมา

ขอขอบคุณบัณฑิตศึกษาและบัณฑิตวิทยาลัย คณะเทคโนโลยีสารสนเทศที่ให้ความช่วยเหลือในเรื่องต่างๆ

และขอขอบพระคุณสถาบันและคณาจารย์ทุกท่านที่คอยช่วยถ่ายทอดวิชาและให้คำปรึกษาที่ดีเสมอมา หวังไว้ว่าโครงการศึกษาฉบับนี้จะให้ประโยชน์แก่ผู้ที่สนใจไม่มากก็น้อย

ศิริพร ระเมียดดี

สารบัญ

หน้า

สารบัญ(ต่อ)

หน้า

สารบัญ(ต่อ)

หน้า

สารบัญ(ต่อ)

หน้า

สารบัญตาราง

สารบัญรูป

สารบัญรูป(ต่อ)

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การศึกษาในเรื่องดาต้าไมนิ่ง (Data Mining) ถือได้ว่าเป็นเทคนิคอย่างหนึ่งที่สามารถวิเคราะห์ข้อมูลต่าง ๆ ได้เป็นอย่างดี เพราะจำนวนข้อมูลในปัจจุบันที่เก็บอยู่ภายในฐานข้อมูลมักมีขนาดใหญ่จนเกินความสามารถที่จะทำการวิเคราะห์ด้วยคน จึงเริ่มใช้การวิเคราะห์ที่เป็นอัตโนมัติในการทำธุรกิจนั้นจำเป็นที่จะต้องศึกษาถึงความต้องการของลูกค้า ดังนั้นดาต้าไมนิ่งจึงเป็นเครื่องมือที่ดีในการพยากรณ์แนวโน้มและพฤติกรรมของข้อมูลรวมถึงเก็บข้อมูลและสารสนเทศต่าง ๆ ไว้เพื่อช่วยในการตัดสินใจรวมถึงทำนายหาลักษณะที่จะเกิดขึ้นภายในอนาคต เพื่อที่สามารถนำมาปรับปรุงกลยุทธ์ต่าง ๆ ขององค์กรได้ เพื่อให้การทำธุรกิจนั้นประสบผลสำเร็จ และเกิดความได้เปรียบทางการค้ากับคู่แข่ง

กระบวนการในการทำดาต้าไมนิ่งมีอยู่หลายเทคนิคด้วยกัน ซึ่งในเรื่องของการจัดกลุ่มข้อมูลเพื่อการวิเคราะห์นั้น ในโครงการฉบับนี้เลือกใช้เทคนิค Clustering ซึ่งเป็นกระบวนการที่ทำการจัดแบ่งกลุ่มข้อมูลเพื่อดูว่าข้อมูลทั้งหมดถูกจัดแบ่งออกเป็นกี่กลุ่ม โดยอาศัยการวิเคราะห์รูปแบบที่คล้ายคลึงกันของข้อมูลหรือรูปแบบที่มีความแตกต่างกันของข้อมูลในการจัดกลุ่ม โดยอาศัยหลักการที่ว่าข้อมูลที่อยู่ในกลุ่มเดียวกันจะมีความคล้ายคลึงกันสูงแต่จะมีความแตกต่างกันอย่างมากเมื่อเปรียบเทียบกับกลุ่มอื่น บ่อยครั้งที่ถือได้ว่าเป็นเทคนิคที่ดีที่สุดที่จะนำมาใช้เป็นวิธีการแรกเมื่อพบว่ามีปริมาณข้อมูลเป็นจำนวนมาก หรือข้อมูลมีความซับซ้อนสูง

ในวิธีการต่าง ๆ ของการจัดกลุ่มข้อมูลนั้น อัลกอริทึม k-Means เป็นที่นิยมมากที่สุดที่ใช้ในการแบ่งส่วนข้อมูลไปเป็นคลัสเตอร์ (Cluster) ซึ่งวิธีนี้ผลลัพธ์สุดท้ายจะได้ค่า Square Error น้อยที่สุดระหว่างข้อมูลกับศูนย์กลางคลัสเตอร์ และอัลกอริทึมยังเหมาะสมกับคลัสเตอร์ที่มีลักษณะเป็นทรงกลม (Sphere) ถึงแม้ว่า k-Means จะมีประโยชน์อย่างกว้างขวางแต่ก็ยังคงมีข้อเสียตรงที่ k-Means จะใช้ค่าเฉลี่ยของกลุ่มเป็นตัวแทน แต่ค่าเฉลี่ยบางครั้งก็อาจไม่ใช่ตัวแทนที่ดีนัก และยังทำงานผิดพลาดกับข้อมูลที่มีลักษณะเป็น Noise และ Outliers ซึ่งผลลัพธ์ที่ได้จากปัญหาเหล่านี้จึงได้ศึกษาการพัฒนาระบบเพื่อการจัดกลุ่มข้อมูลที่ใช้พื้นฐานของวิวัฒนาการโดยใช้ไฮบริดพาร์ทิเคิลสวอมออปติไมเซชั่น (Hybrid Particle Swarm Optimization : Hybrid PSO) ซึ่งมีความสามารถที่สูงกว่าและยังสามารถจัดการกับข้อบกพร่องที่พบในอัลกอริทึม k-Means โดยใช้กระบวนการค้นหาค่าความเหมาะสมเพื่อให้ได้ตำแหน่งที่ดีที่สุด จากวิธีการที่เป็นไปได้ทั้งหมด (Search Space)

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

1. ศึกษาดาต้าไมนิ่งโดยใช้เทคนิคการวิเคราะห์ข้อมูลแบบ Clustering ในการพัฒนาระบบงานเพื่อการจัดกลุ่มข้อมูล
2. ศึกษาอัลกอริทึม k-Means ซึ่งเป็นอัลกอริทึมหนึ่งที่ใช้ในการจัดกลุ่มข้อมูล
3. ศึกษาเทคนิคใหม่ในการจัดกลุ่มข้อมูลโดยใช้วิธีไฮบริดพาร์ทิเคิลสวอมออปติไมเซชั่นเพื่อใช้ในการพัฒนาระบบ
4. ระบบที่พัฒนาขึ้นมานั้นสามารถที่จะนำไปใช้เป็นข้อมูลอย่างหนึ่งในการสนับสนุนการตัดสินใจได้เป็นอย่างดี

1.3 ทฤษฎีหรือแนวคิดที่ใช้ในการศึกษา

ในดาต้าไมนิ่งนั้นมีเทคนิควิธีการมากมายที่สามารถใช้ในการค้นหาข้อมูลและสารสนเทศที่ประโยชน์ การจัดกลุ่มข้อมูลเป็นเทคนิคอย่างหนึ่งที่สามารถทำให้เราได้มาของข้อมูลที่เราสนใจเพื่อสามารถใช้ข้อมูลนั้นเป็นไปวัตถุประสงค์ที่เราตั้งไว้ และยังสามารถใช้ในการทำนาย เพื่อช่วยในการตัดสินใจได้อย่างมีประสิทธิภาพ การผสมผสานการทำงานระหว่างอัลกอริทึม k-Means และพาร์ทิเคิลสวอมออปติไมเซชั่นมีการทำงานหลาย ๆ รอบจนกระทั่งไม่มีการเปลี่ยนแปลง เพื่อให้ได้มาของตัวแทนที่ดีที่สุด เพราะถ้าเราได้ตำแหน่งที่ดีที่สุดมาเป็นตัวแทนของกลุ่ม สิ่งที่ตามมาก็คือการจัดกลุ่มข้อมูลที่มีประสิทธิภาพ เพราะข้อมูลที่เหลือจะถูกกำหนดให้กับตัวแทนที่ดีเหล่านี้

1.4 ขอบเขตของโครงการ

โครงการนี้เป็นการศึกษาและพัฒนาระบบที่ใช้ในการจัดกลุ่มข้อมูลโดยเป็นการนำเอาเทคนิค Clustering ของทางดาต้าไมนิ่งมาประยุกต์ใช้งาน ซึ่งในโครงการศึกษานี้จะเลือกใช้วิธีไฮบริดพาร์ทิเคิลสวอมออปติไมเซชั่น (Hybrid PSO) ซึ่งวิธีนี้เป็นการรวมข้อดีและหลีกเลี่ยงข้อเสียของอัลกอริทึม PSO และ k-Means โดยวิธีการของอัลกอริทึมเป็นการหาค่าความเหมาะสมโดยใช้พื้นฐานของความน่าจะเป็น โดย PSO จะลอกเลียนแบบพฤติกรรมการเรียนรู้ของสิ่งมีชีวิตแบบกลุ่มเพื่อใช้ในการจัดกลุ่มข้อมูลโดยมีขอบเขตของโครงการดังนี้

1. ทำการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม k-Means และไฮบริดพาร์ทิเคิลสวอมออปติไมเซชั่น
2. สามารถพัฒนาโปรแกรมตามทฤษฎี และอัลกอริทึมได้ถูกต้องโดยใช้ Microsoft Visual Basic V.6.0
3. สามารถนำเสนอผลลัพธ์ให้กับผู้ใช้งานและนำไปใช้ประโยชน์ได้ถูกต้อง

1.5 ขั้นตอนการดำเนินงาน

การจะทำโครงการให้สำเร็จลุล่วงตามวัตถุประสงค์ได้ ต้องมีการกำหนดการดำเนินงาน เพื่อให้การทำงานเป็นไปตามลำดับ ซึ่งในโครงการศึกษานี้มีขั้นตอนในการดำเนินงาน ดังนี้

1. กำหนดหัวข้อ เป้าหมาย และวัตถุประสงค์ ตลอดจนขอบเขตของโครงการ โดยระบบงานนี้ได้กำหนดเป้าหมายเพื่อทำการจัดกลุ่มข้อมูลโดยใช้วิธีไฮบริดพาร์ทิเคิลสวอมออปติไมเซชั่น (Hybrid Particle Swarm Optimization : Hybrid PSO)
2. ศึกษาค้นคว้าข้อมูลเกี่ยวกับดาต้าไมนิ่งที่ใช้เทคนิค Clustering โดยเลือกอัลกอริทึมที่เกี่ยวข้อง เช่น อัลกอริทึม k-Means และอัลกอริทึมของพาร์ทิเคิลสวอมออปติไมเซชั่น เพื่อใช้ในการจัดกลุ่มข้อมูลได้อย่างมีประสิทธิภาพ
3. ศึกษาและพัฒนาโปรแกรมที่ใช้ในการจัดกลุ่มโดยใช้อัลกอริทึม k-Means และไฮบริดพาร์ทิเคิลสวอมออปติไมเซชั่น และทำการเปรียบเทียบผลลัพธ์ที่ได้
4. ทดสอบโปรแกรมเพื่อปรับปรุง แก้ไข หากมีข้อผิดพลาดเกิดขึ้น
5. ประเมินผลและวิเคราะห์ผลที่ได้จากการพัฒนาระบบ

1.6 ประโยชน์ที่คาดว่าจะได้รับ

หลังจากที่ได้ทำการศึกษาและพัฒนาระบบงาน คาดว่าจะได้รับประโยชน์ดังต่อไปนี้

1. ทำให้ได้รับความรู้และมีความเข้าใจถึงหลักการ วิธีการต่าง ๆ รวมไปถึงขั้นตอนการทำงานของดาต้าไมนิ่งได้เป็นอย่างดี
2. จากหลักการทำงานของดาต้าไมนิ่งโดยใช้เทคนิค Clustering ด้วยอัลกอริทึมของพาร์ทิเคิลสวอมออปติไมเซชั่น สามารถใช้ในการค้นตำแหน่งที่ดีที่สุดเพื่อเป็นตัวแทนศูนย์กลางกลุ่มในการจัดกลุ่มข้อมูลได้อย่างมีประสิทธิภาพ
3. สามารถนำระบบที่พัฒนาขึ้นไปใช้ในการจัดกลุ่มข้อมูลโดยข้อมูลที่ได้จากการจัดกลุ่มนั้น สามารถนำไปประยุกต์ใช้ก่อให้เกิดประโยชน์ทางด้านต่าง ๆ ได้

บทที่ 2
ดาต้าไมนิ่งและทฤษฎีที่เกี่ยวข้อง

ในฐานข้อมูลที่มีข้อมูลอยู่เป็นจำนวนมหาศาลนั้นยากต่อการที่จะนำมาใช้งานเพื่อให้ได้ประโยชน์อย่างทันท่วงที ดังนั้นเทคโนโลยีที่ช่วยในการสืบค้นหาข้อมูลที่เป็นประโยชน์ และน่าสนใจบนฐานข้อมูลขนาดใหญ่หรือที่เรียกกันว่าดาต้าไมนิ่งถือเป็นศาสตร์ทางคอมพิวเตอร์อย่างหนึ่งที่ได้รับความสนใจอย่างมาก เนื่องจากมีการนำเอาเทคนิคของดาต้าไมนิ่งมาใช้วิเคราะห์สืบค้นหาข้อเท็จจริงที่ซ่อนอยู่ในฐานข้อมูลออกมาประยุกต์ใช้ให้เป็นประโยชน์แก่ธุรกิจซึ่งในปัจจุบันที่มีการแข่งขันกันสูงได้ ซึ่งนอกจากจะช่วยในการตัดสินใจแล้ว ยังอาจได้รับข้อมูลใหม่ ๆ ทำให้การวิเคราะห์ข้อมูลมีประสิทธิภาพสูงขึ้นไปอีก

การศึกษาในเรื่องดาต้าไมนิ่ง (Data Mining) ถือได้ว่าเป็นเทคนิคอย่างหนึ่งที่สามารถวิเคราะห์ข้อมูลต่าง ๆ ได้เป็นอย่างดี เพราะจำนวนข้อมูลในปัจจุบันที่เก็บอยู่ในฐานข้อมูลมักมีขนาดใหญ่จนเกินความสามารถที่จะทำการวิเคราะห์ด้วยคน จึงเริ่มใช้ในการวิเคราะห์โดยอัตโนมัติ นอกจากนั้นดาต้าไมนิ่งยังเป็นเครื่องมือที่ดี ในการพยากรณ์แนวโน้มและพฤติกรรมของข้อมูล รวมถึงเก็บความรู้และข้อมูลต่าง ๆ เพื่อช่วยในการตัดสินใจเนื่องจากดาต้าไมนิ่งสามารถที่จะตอบคำถามทางธุรกิจได้เป็นอย่างดี

2.1 ความหมายของดาต้าไมนิ่ง

ในปัจจุบันนี้ข้อมูลมีความสำคัญต่อการดำเนินธุรกิจต่าง ๆ เป็นอย่างมาก จึงมีความจำเป็นที่จะต้องเก็บข้อมูลที่เกี่ยวข้องกับธุรกิจและนำสารสนเทศมาประยุกต์ เพื่อนำผลที่ได้มาช่วยในการสนับสนุนการตัดสินใจ และการที่มีวิธีการวิเคราะห์ที่ดีและเหมาะสมจะทำให้มีโอกาสเป็นผู้นำ ซึ่งส่งผลให้มีการเพิ่มจำนวนฐานข้อมูลในการเก็บข้อมูลมากขึ้นและทำให้เกิดปัญหาตามมา ทำให้การวิเคราะห์ข้อมูลจำนวนมาก ๆ เป็นเรื่องที่ยาก จึงได้มีการคิดค้นเทคโนโลยีที่เรียกว่า ดาต้าไมนิ่ง (Data Mining) มาช่วยในการวิเคราะห์

ดาต้าไมนิ่ง (Data Mining) คือกระบวนการค้นหาความรู้ที่น่าสนใจ ได้แก่ รูปแบบความสัมพันธ์ การเปลี่ยนแปลง ความผิดปกติ และลักษณะโครงสร้างที่สำคัญซึ่งอยู่ในฐานข้อมูล (Database) แต่ได้ถูกซ่อนอยู่ในฐานข้อมูลที่มีอยู่จำนวนมากนี้ ทำให้ได้สารสนเทศสำคัญ ที่ยังไม่เคยทราบออกมา เพื่อช่วยในการสนับสนุนการตัดสินใจในการบริหารและนำไปประยุกต์ใช้ในธุรกิจต่าง ๆ ได้ง่าย โดยดาต้าไมนิ่งมีข้อดีที่แตกต่างจากเครื่องมือวิเคราะห์อื่น ๆ คือ ดาต้าไมนิ่งสามารถมองความสัมพันธ์ของข้อมูลในหลายมิติและสามารถใช้กับข้อมูลขนาดใหญ่ได้

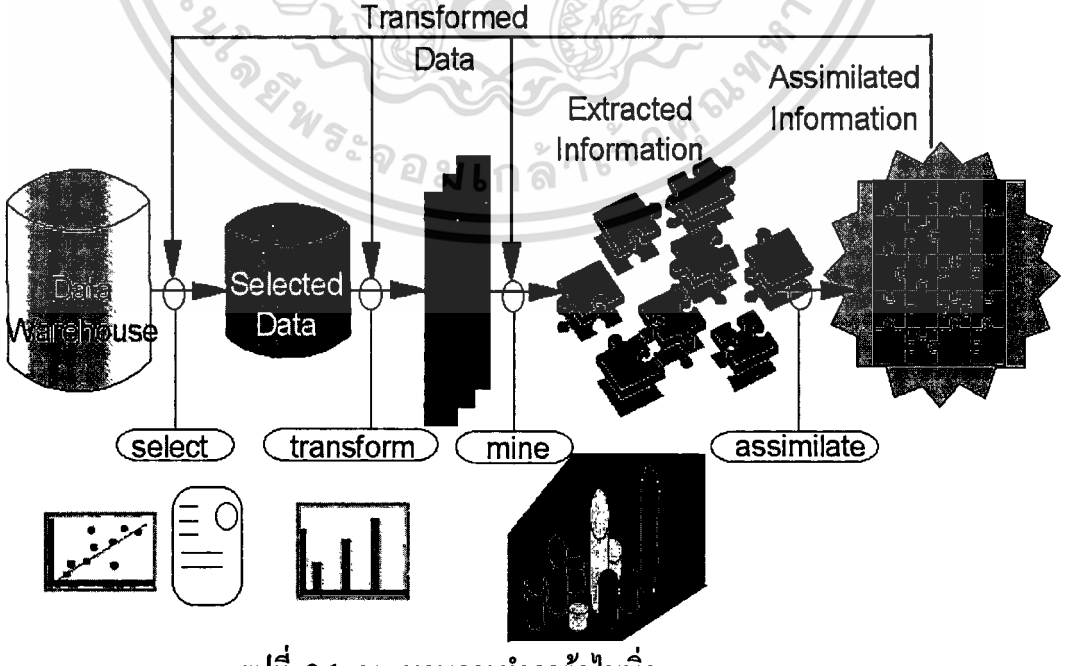
2.2 กระบวนการทำงานของดาต้าไมนิ่ง

เนื่องจากดาต้าไมนิ่งเป็นกระบวนการที่ทำงานกับฐานข้อมูลต่าง ๆ จึงเป็นที่รู้จักอีกชื่อหนึ่งว่า การค้นหาความรู้จากฐานข้อมูล (Knowledge Discovery in Database) ซึ่งฐานข้อมูลแต่ละประเภทมีลักษณะแตกต่างกันออกไป ดังนั้นก่อนทำดาต้าไมนิ่งนั้นต้องมีวิธีการเตรียมข้อมูลและมีกระบวนการในการเตรียมข้อมูลต่าง ๆ ให้มีความพร้อมในการที่จะทำดาต้าไมนิ่ง เพื่อให้ได้ข้อมูลที่เหมาะสมออกมาได้ก่อน ขั้นตอนต่าง ๆ ให้พร้อมจึงจะทำดาต้าไมนิ่งและวิเคราะห์ผลลัพธ์ที่ได้เป็นลำดับสุดท้าย

ในกระบวนการทำดาต้าไมนิ่งประกอบด้วย 4 ขั้นตอนหลัก ๆ คือ การกำหนดจุดประสงค์ทางธุรกิจ การเตรียมข้อมูล การทำดาต้าไมนิ่ง การวิเคราะห์ผลที่ได้จากการทำดาต้าไมนิ่งและการนำมาใช้ โดยหลังจากที่ทำการวิเคราะห์ผลแล้วสามารถกลับไปเริ่มทำขั้นตอนใดใหม่อีกครั้งก็ได้ ดังแสดงได้ดังรูปที่ 2.1

2.2.1 การกำหนดวัตถุประสงค์ทางธุรกิจ (Business Objective Determination)

การกำหนดวัตถุประสงค์ทางธุรกิจเป็นการกำหนดขอบเขตและเป้าหมายในการทำดาต้าไมนิ่ง โดยต้องทำความเข้าใจถึงปัญหาและความต้องการทางธุรกิจ เพราะปัญหาทางธุรกิจบางปัญหาก็ไม่สามารถทำการแก้ไขได้ด้วยการทำดาต้าไมนิ่ง ดังนั้นในขั้นตอนนี้จึงต้องทำการวิเคราะห์ทางธุรกิจรวมทั้งต้องวิเคราะห์ข้อมูลเบื้องต้นว่า เรามีข้อมูลใดอยู่บ้าง เพื่อทำให้การแก้ปัญหาเป็นอย่างถูกต้อง และผลจากการทำดาต้าไมนิ่งนั้นจะดีหรือไม่จำเป็นต้องมีวัตถุประสงค์ที่ชัดเจน ไม่เช่นนั้นแล้วอาจล้มเหลวได้



รูปที่ 2.1 กระบวนการทำดาต้าไมนิ่ง

2.2.2 การเตรียมข้อมูล (Data Preparation)

ขั้นตอนการเตรียมข้อมูลนี้เป็นขั้นตอนที่สำคัญมากและใช้เวลาในการทำงานมากกว่าขั้นตอนอื่น ๆ ในการทำดาต้าไมนิ่ง เนื่องจากเป็นการจัดเตรียมข้อมูลเพื่อส่งต่อไปยังกระบวนการไมนิ่ง ถ้าเรามีการเตรียมข้อมูลที่ไม่ดีหรือเกิดข้อผิดพลาด จะส่งผลให้การทำไมนิ่งนั้นผิดไปจากวัตถุประสงค์ที่ตั้งไว้ ซึ่งในการเตรียมข้อมูลยังแบ่งออกเป็น 3 ขั้นตอนย่อย ดังนี้

2.2.2.1. การเลือกข้อมูล (Data Selection)

เป็นการวิเคราะห์คัดเลือกข้อมูลที่มีประโยชน์และน่าสนใจต่อการใช้งานออกมาจากฐานข้อมูลจากข้อมูลเพื่อทำงานในขั้นตอนต่อไป โดยการเลือกข้อมูลนั้นจะเป็นไปตามวัตถุประสงค์ทางธุรกิจที่กำหนดไว้ในตอนแรก โดยตัวแปรที่ถูกเลือกมานั้นจะมีการกำหนดชนิด ค่ารูปแบบ และลักษณะที่ชัดเจนเอาไว้

2.2.2.2 การเตรียมข้อมูลก่อนการประมวลผล (Data Preprocessing)

ตรวจสอบข้อมูลและแก้ไขเพื่อทำให้ได้ข้อมูลที่มีคุณภาพดี และทำให้ข้อมูลที่ถูกเลือกนั้นถูกต้องครบถ้วนตามที่จะต้องใช้ในการทำดาต้าไมนิ่งในขั้นตอนต่อ ๆ ไป ซึ่งข้อมูลที่ผ่านการเลือกมานั้นจะต้องผ่านการตรวจสอบได้แก่ การตรวจสอบข้อมูล (Data Cleaning) การรวบรวมข้อมูล (Data Integration) การลดจำนวนข้อมูล (Data Reduction)

2.2.2.3. การแปลงข้อมูล (Data Transformation)

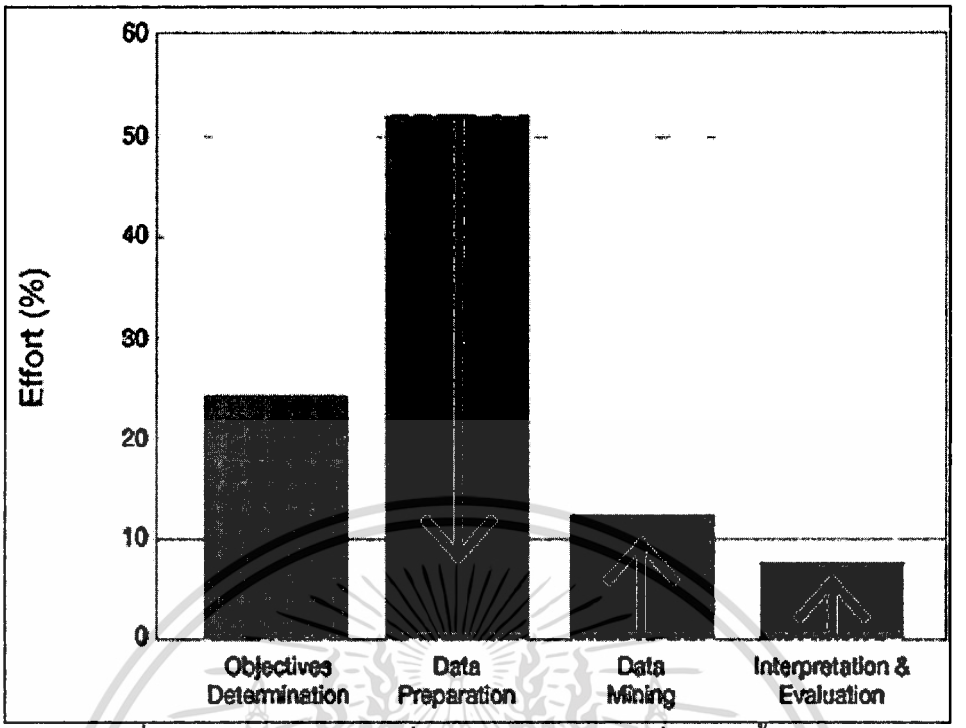
เป็นการแปลงข้อมูลหรือรวบรวมข้อมูลไว้ด้วยกัน ซึ่งมีวัตถุประสงค์คือทำให้การทำดาต้าไมนิ่งมีประสิทธิภาพมากขึ้นและทำให้รูปแบบของข้อมูลสอดคล้องกับเทคนิคที่จะนำมาใช้ เนื่องจากข้อมูลที่จะนำมาใช้ทำนั้น บางครั้งอยู่ในรูปแบบที่ไม่เหมาะสมกับอัลกอริทึมที่เลือกใช้จึงต้องมีการแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมก่อน

2.2.3 การทำดาต้าไมนิ่ง (Data Mining)

เป็นกระบวนการสำคัญในการเลือกเทคนิคดาต้าไมนิ่งที่เหมาะสม ซึ่งในขั้นตอนนี้จะมีความสัมพันธ์กับการวิเคราะห์ข้อมูลและขั้นตอนการแปลงข้อมูลที่ผ่านมา ในส่วนนี้จะเกี่ยวข้องกับอัลกอริทึมหลายแบบ ซึ่งในการเลือกใช้โมเดลนั้นขึ้นอยู่กับวัตถุประสงค์ในการทำไมนิ่ง

2.2.4 การวิเคราะห์ผลลัพธ์ที่ได้จากการทำดาต้าไมนิ่งและการนำความรู้มาใช้ (Analysis Of Result and Assimilation of Knowledge)

การวิเคราะห์ผลที่ได้จากการทำดาต้าไมนิ่งและการนำความรู้ไปใช้เป็นขั้นตอนสุดท้ายในการทำดาต้าไมนิ่ง นักวิเคราะห์จะต้องนำผลที่ได้จากการไมนิ่งมาตีความหมายและสรุปผล เพื่อนำไปเป็นสารสนเทศที่ช่วยในการตัดสินใจ ถ้านักวิเคราะห์เห็นว่าผลที่ได้ไม่เป็นไปตามวัตถุประสงค์ที่วางไว้ ก็สามารถที่จะย้อนกลับไปแก้ไขในขั้นตอนใด ๆ ได้



รูปที่ 2.2 แสดงเปอร์เซ็นต์ที่ใช้ในการทำดาต้าไมนิ่งแต่ละขั้นตอน

2.3 เทคนิคในการวิเคราะห์ข้อมูลของดาต้าไมนิ่ง

ในโครงการพัฒนาระบบนี้ เราจะเน้นถึงเทคนิคในการ Clustering หรือที่เรียกกันว่า Database Segmentation เพื่อใช้ในการจัดกลุ่มข้อมูลเพื่อทำการวิเคราะห์ โดยการทำดาต้าไมนิ่งนั้นมีโมเดลอยู่หลายแบบซึ่งการเลือกใช้โมเดลขึ้นอยู่กับวัตถุประสงค์ในการทำดาต้าไมนิ่ง ซึ่งแบ่งออกเป็น 4 ประเภทหลัก ดังนี้

2.3.1 Predictive Modeling

เป็นโมเดลที่ใช้ในการสร้างแบบจำลองพยากรณ์เพื่อทำนายหาค่าความเป็นไปได้ โดยสังเกตจากข้อมูลที่มีอยู่ซึ่งข้อมูลต้องมีความถูกต้องและสมบูรณ์ จะทำให้แบบจำลองสามารถทำนายผลได้อย่างถูกต้อง ซึ่งแบบจำลองในการพัฒนาจะแบ่งออกเป็นสองช่วงด้วยกันคือ

- Training : เป็นช่วงในการสร้างแบบจำลองขึ้นมาใหม่โดยใช้ข้อมูลในอดีต และใช้ข้อมูลในปริมาณมาก
- Testing : เป็นการตรวจสอบประสิทธิภาพของแบบจำลองใหม่ที่สร้างขึ้นมา ซึ่งใช้ข้อมูลในปริมาณที่ไม่มากนัก

2.3.2 Database Segmentation (Clustering)

มีอีกชื่อหนึ่งว่า Clustering เป็นโมเดลที่ใช้ในการจัดกลุ่มข้อมูล โดยการแบ่งกลุ่มข้อมูลจะแบ่งตามลักษณะที่เหมือนกันของข้อมูล เช่น การแบ่งตามอายุ รายได้ ซึ่งในการแบ่งกลุ่มข้อมูลนั้น

ไม่สามารถที่จะกำหนดได้ว่าข้อมูลควรที่จะอยู่ในกลุ่มใด แต่จะเป็นการกำหนดกลุ่มจากธรรมชาติของข้อมูลเองมากกว่า โดยไม่มีการใช้อคติหรือประสบการณ์มาช่วยในการตัดสินใจ

การจัดกลุ่มข้อมูล คือการรวบรวมข้อมูลที่มีความคล้ายคลึงกันไว้ในกลุ่มเดียวกัน และข้อมูลที่ไม่คล้ายคลึงกันจะอยู่ในกลุ่มอื่น ซึ่งข้อมูลแต่ละตัวจะต้องอยู่ในกลุ่มใดกลุ่มหนึ่งเพียงกลุ่มเดียว ซึ่งข้อมูลแต่กลุ่มที่ได้นั้นจะเรียกว่า "Cluster" นั่นคือลักษณะการแบ่งแยกข้อมูลบนพื้นฐานของความคล้ายคลึงกันในตัวเอง วัตถุประสงค์ของการจัดกลุ่มคือเพื่อจัดกลุ่มข้อมูลซึ่งเป็นประโยชน์ในงานด้านต่าง ๆ เช่น การตลาด การแพทย์ การปกครอง เป็นต้น

2.3.3 Link Analysis

เป็นโมเดลที่ใช้วิคราะห์หาความสัมพันธ์ (Association) ระหว่างข้อมูลเพื่อที่จะดูว่าแต่ละกลุ่มของข้อมูลมีความสัมพันธ์กันในลักษณะใด ซึ่งเป็นแบบจำลองที่ได้รับความนิยมในการหาความสัมพันธ์ระหว่างลูกค้ากับสินค้า

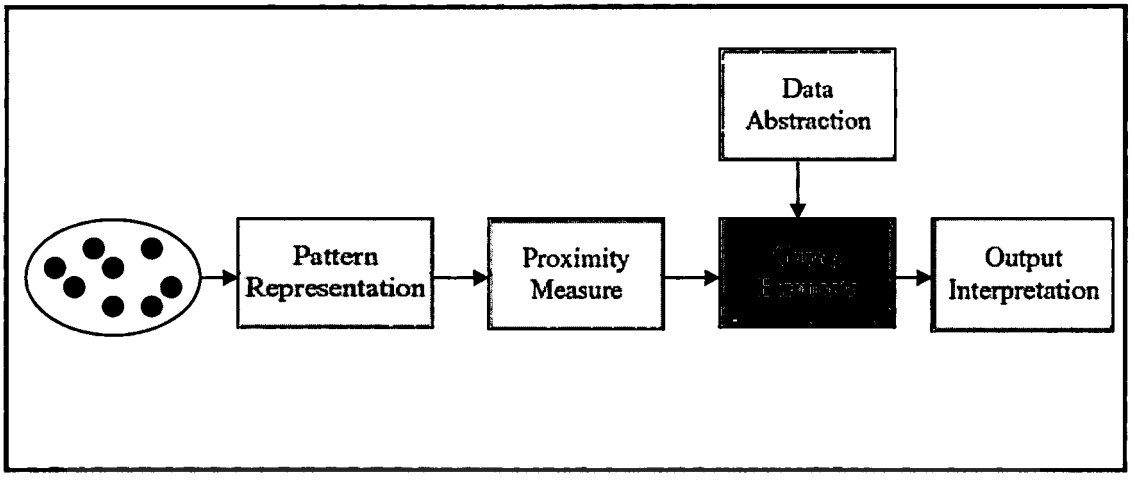
2.3.4 Deviation Detection

เป็นวิธีการหาค่าที่แตกต่างไปจากค่ามาตรฐาน โดยทั่วไปมักใช้วิธีการทางสถิติหรืออาศัยการวาดกราฟหรือแสดงผลในรูปกราฟฟิกทำให้เห็นภาพมากขึ้น แล้วดูการกระจายข้อมูลว่ามีการกระจายออกไปจากกลุ่มหรือไม่ มักใช้ในการตรวจจับสิ่งผิดปกติต่าง ๆ ซึ่งเทคนิคนี้ถูกใช้ในงานด้านการตรวจสอบการปลอมบัตรเครดิตหรือการจับการโกง เป็นต้น จากเทคนิคที่กล่าวมายังไม่มีเทคนิคใดเลยที่จะสามารถแก้ปัญหาของดาต้าไมนิ่งได้ทุกปัญหา ดังนั้นการเลือกเทคนิคของดาต้าไมนิ่งให้เหมาะสมกับงานแต่ละอย่างจะนำไปสู่วิธีการแก้ปัญหาของดาต้าไมนิ่งได้ดีที่สุด

2.4 โมเดลการจัดกลุ่มข้อมูล (Database Segmentation : Clustering)

กระบวนการในการจัดกลุ่มทางกายภาพหรือนามธรรม โดยจัดให้สิ่งที่มีลักษณะคล้ายคลึงกันมาอยู่ในประเภทเดียวกันเราเรียกว่า Clustering

มนุษย์มีความสามารถในการจัดกลุ่ม พวกเราสามารถแยกสีส้มออกจากสีเขียว แยกวงกลมออกจากสามเหลี่ยม แยกดอกไม้ออกจากผลไม้ พวกเราจัดกลุ่มสิ่งต่าง ๆ ตามธรรมชาติโดยการเปรียบเทียบสิ่งต่าง ๆ ตามลักษณะเฉพาะที่มีความเด่นมากที่สุด ตัวอย่างเช่น สี ขนาด รูปร่าง น้ำหนัก เป็นต้น การจัดกลุ่มเป็นเทคนิคทั่วไปทางด้านสถิติที่เกี่ยวข้องกับการวิเคราะห์ข้อมูล ซึ่งอัลกอริทึมของการจัดกลุ่มเป็นเทคนิคสำคัญที่เกี่ยวข้องกับการวิเคราะห์ข้อมูล ซึ่งถูกใช้ในหลาย ๆ สาขา ประกอบไปด้วย การจดจำรูปแบบ (Pattern Recognition), การประมวลผลภาพ (Image Processing), การวิเคราะห์ข้อมูล (Data Analysis), การวิจัยตลาด (Market Research) และทางด้านวิศวกรรมอื่น ๆ



รูปที่ 2.3 ลำดับการทำงานในการจัดกลุ่มข้อมูล

ในการจัดกลุ่มข้อมูลนี้เป็นเครื่องมืออันหนึ่งที่จะใช้ช่วยให้เราเข้าใจถึงการกระจายของข้อมูลหรือให้เรามองเห็นลักษณะเฉพาะของแต่ละกลุ่ม และมุ่งเจาะลึกไปที่กลุ่มที่เราสนใจจะศึกษา ในอีกทางเลือกหนึ่งคือการจัดกลุ่มข้อมูลนี้อาจจะถูกใช้ในขั้นตอนการทำงานของอัลกอริทึมบางอย่าง เช่น การแสดงลักษณะเฉพาะของข้อมูล และการแยกประเภทของข้อมูลออกเป็นกลุ่ม ๆ เพื่อจะใช้ในการค้นหากลุ่มของข้อมูล ซึ่งการจัดกลุ่มนั้นกลายมาเป็นหัวข้อสำคัญในการสนับสนุนการวิจัยซึ่งประกอบงานทางด้านดาต้าไมนิ่ง (Data Mining) สถิติ (Statistics) สื่อเพื่อการเรียนรู้ (Machine Learning) เทคโนโลยีฐานข้อมูล (Spatial Database Technology) ชีววิทยา (Biology) และการตลาด (Marketing) ซึ่งจะมีการสะสมข้อมูลขนาดใหญ่อยู่ในฐานข้อมูล

อัลกอริทึมในการจัดกลุ่มสามารถแบ่งเป็น 2 ประเภทหลักคือการจัดกลุ่มนั้นเป็น Supervised Learning หรือเป็น Unsupervised Learning

- Supervised Learning : การเรียนรู้แบบมีการชี้นำ เป็นเทคนิคในการสร้างกลุ่มข้อมูลที่ตามจุดมุ่งหมายที่วางเอาไว้ โดยมีการกำหนดรูปแบบข้อมูลเข้า กลุ่มข้อมูล และจำนวนกลุ่มไว้ล่วงหน้าแล้ว ตัวอย่างเช่นจัดกลุ่มนักเรียนว่า ดีมาก ดี ปานกลาง หรือไม่ดี (4 กลุ่มข้อมูล) โดยพิจารณาจากประวัติและผลการเรียน เป็นต้น
- Unsupervised Learning : การเรียนรู้แบบไม่มีการชี้นำ เป็นเทคนิคการจัดกลุ่มที่ดึงเอาข้อมูลที่เราสนใจที่มีลักษณะเหมือนกันออกมาเพื่อใช้ช่วยในการจัดกลุ่มข้อมูล ตัวอย่างเช่น ทำการจัดกลุ่มข้อมูลโดยขึ้นอยู่กับระยะทาง เป็นต้น

หลาย ๆ อัลกอริทึมในการจัดกลุ่มเป็นการเรียนรู้แบบไม่มีการชี้นำจะถูกพัฒนาขึ้นซึ่งส่วนใหญ่จะเป็นอัลกอริทึมที่จัดกลุ่มไปเป็นคลัสเตอร์โดยขึ้นอยู่กับลักษณะของข้อมูลเข้า ซึ่งความต้องการทั่วไปของอัลกอริทึมที่มีการจัดกลุ่มที่ดีในดาต้าไมน์นิ่ง ได้แก่

- ความสามารถในการปรับตัว (Versatility) : มีหลาย ๆ อัลกอริทึมที่ถูกออกแบบให้กับข้อมูลที่มีลักษณะ Interval-based อย่างไรก็ตามการใช้งานอาจจะต้องการข้อมูลในชนิดอื่นด้วย อาทิเช่นข้อมูลที่เป็นแบบ Binary, Nominal, Ordinal หรือข้อมูลแบบทุกชนิดรวมกัน
- การค้นพบคลัสเตอร์ในรูปร่างที่หลากหลาย : หลาย ๆ อัลกอริทึมจะพิจารณาตามหลักของ Euclidean หรือ Manhattan เช่นมีแนวโน้มว่าจะพบคลัสเตอร์เป็นรูปทรงกลมด้วยขนาดและความหนาแน่นที่คล้ายกัน อย่างไรก็ตามคลัสเตอร์ที่ค้นพบอาจมีรูปร่างใด ๆ ที่ไม่เฉพาะเจาะจงซึ่งเป็นสิ่งสำคัญในการพัฒนาอัลกอริทึม
- ความต้องการข้อมูลเข้าเป็นค่าที่แน่นอน : หลาย ๆ อัลกอริทึมในการจัดกลุ่มต้องการค่าพารามิเตอร์ที่ถูกระบุโดยผู้ใช้เช่น จำนวนของคลัสเตอร์ควรกำหนดให้เหมาะกับการใช้วิเคราะห์ข้อมูล อย่างไรก็ตามในเซตข้อมูลขนาดใหญ่ซึ่งเป็นที่น่าพอใจว่า วิธีการต้องการแนวทางจำกัดจากผู้ใช้งานเพื่อที่จะหลีกเลี่ยงข้อมูลที่เอนเอียง
- การจัดการกับข้อมูลที่ไม่มีคุณภาพ : ฐานข้อมูลส่วนใหญ่จะมีข้อมูลที่นอกประเด็นอาจขาดหาย ไม่ทราบข้อมูล ข้อมูลที่ไม่ถูกต้อง บางอัลกอริทึมเมื่อได้รับข้อมูลเหล่านี้อาจนำไปสู่คลัสเตอร์ที่ไม่มีคุณภาพดังนั้นอัลกอริทึมในการจัดกลุ่มควรมีความสามารถในการจัดการกับค่าเบี่ยงเบนเหล่านี้เพื่อปรับปรุงให้ได้คลัสเตอร์ที่มีคุณภาพ
- ไม่มีผลต่อลำดับของข้อมูลเข้า : บางอัลกอริทึมจะมีผลกับลำดับของข้อมูลเข้า เช่น ในเซตของข้อมูลชุดเดียวกัน เมื่อเสนอในอัลกอริทึมด้วยลำดับที่แตกต่างกันทำให้เกิดคลัสเตอร์ที่แตกต่างกัน ซึ่งเป็นสิ่งสำคัญที่จะทำการพัฒนาอัลกอริทึมในเรื่องวิธีในการจัดกลุ่มที่จะให้ผลลัพธ์ที่แน่นอนโดยไม่คำนึงถึงลำดับของข้อมูลที่ถูกเสนอ
- ข้อมูลที่มีหลายมิติ : ในเซตข้อมูลที่มีขนาดใหญ่จะมีจำนวนของแอตทริบิวต์มากหรือมีมิติที่หลากหลาย ซึ่งในหลาย ๆ อัลกอริทึมของการจัดกลุ่มยังไม่สามารถจัดการกับจำนวนของมิติที่หลากหลาย (มากกว่า 8-10 มิติ) ได้จึงเป็นสิ่งที่ท้าทายสำหรับเซตข้อมูลที่มีมิติสูง ๆ โดยเฉพาะอย่างยิ่งเมื่อต้องพิจารณาข้อมูลที่มีน้อยมาก ๆ และมีความบิดเบือนสูง
- การอธิบายและความสะดวกกับการใช้งาน : ผู้ใช้คาดหวังว่าผลลัพธ์จากการจัดกลุ่มนั้นสามารถที่จะอธิบายทำให้เข้าใจได้และสามารถใช้ในการทำงานได้ ซึ่งเป็นสิ่งสำคัญที่ต้องศึกษาว่าเป้าหมายของการใช้งานจะมีผลอย่างไรในการเลือกใช้วิธีการจัดกลุ่มด้วยเทคนิคต่าง ๆ

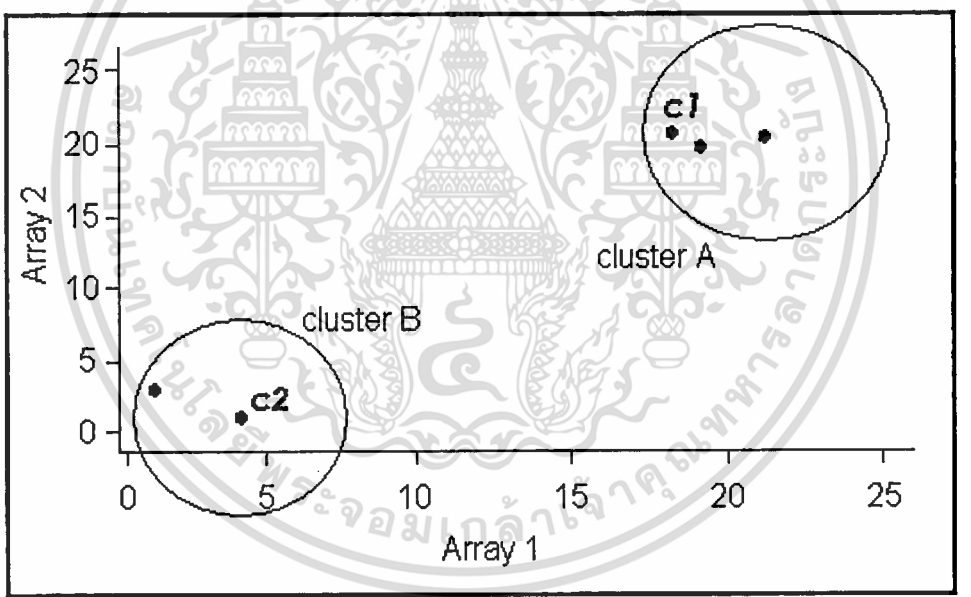
ที่ผ่านมายังไม่มีอัลกอริทึมใดอัลกอริทึมเดียวที่สามารถมีครบทุกความต้องการที่กล่าวมา ดังนั้นจึงเป็นเรื่องสำคัญในการทำความเข้าใจในลักษณะของแต่ละอัลกอริทึมเป็นการช่วยให้เลือกใช้ได้ตรงกับปัญหาที่เกิดขึ้น

2.4.1 Partitioning Method

หลักการทำงานในการจัดกลุ่มของวิธีนี้คือ เริ่มต้นจากการกำหนดกลุ่มที่ต้องการจะแบ่งโดยสมมติว่าในฐานข้อมูลมีข้อมูลทั้งหมดมี n ออบเจกต์ แล้วทำการแบ่งกลุ่มออกเป็น k พาร์ทิชัน ซึ่งแต่ละพาร์ทิชันก็คือตัวแทนของคลัสเตอร์นั่นเอง โดยมีเงื่อนไขดังนี้

- โดยที่ k ต้องน้อยกว่าหรือเท่ากับ n
- แต่ละกลุ่มต้องมีสมาชิกอย่างน้อยหนึ่งออบเจกต์
- แต่ละออบเจกต์ต้องอยู่ในกลุ่มเพียงกลุ่มเดียว

จากนั้นใช้เทคนิคที่เรียกกันว่า Iterative Relocation ซึ่งเป็นความพยายามที่จะทำการปรับปรุงการจัดกลุ่มข้อมูลโดยการนำข้อมูลแต่ละตัวมาลองทดสอบย้ายกลุ่มไปแต่ละกลุ่ม และพิจารณาค่าแตกต่างจากจุดศูนย์กลาง แล้วให้ข้อมูลตัวนั้นอยู่ที่กลุ่มที่ให้ค่าแตกต่างจากจุดศูนย์กลางน้อยที่สุด ซึ่งผลลัพธ์ที่ได้นั้นคือข้อมูลที่อยู่ภายในกลุ่มเดียวกันจะมีลักษณะที่คล้ายคลึงกัน ส่วนต่างกลุ่มกันก็จะมีลักษณะที่แตกต่างกันออกไป ดังแสดงไว้ดังรูปที่ 2.4



รูปที่ 2.4 แสดงผลลัพธ์ที่ได้จาก Partition Method

วิธีการจัดกลุ่มแบบนี้เป็นที่นิยมมากเพราะมีอัลกอริทึมที่เหมาะกับงานประเภทนี้เป็นจำนวนมาก เช่น อัลกอริทึม k-Means โดยมีหลักการทำงานคือ ในแต่ละกลุ่มจะแสดงค่า mean ของข้อมูลในแต่ละคลัสเตอร์ขึ้นมา ซึ่งอัลกอริทึมนี้เหมาะกับการจัดการฐานข้อมูลที่มีขนาดเล็กจนถึงขนาดกลาง โดยจะทำงานได้ดีในคลัสเตอร์ที่มีรูปร่างทรงกลม แต่ไม่เหมาะที่จะจัดการกับฐานข้อมูลขนาดใหญ่มาก ๆ หรือมีรูปทรงที่ซับซ้อน จึงมีอัลกอริทึมอื่น ๆ ขึ้นมาเพื่อจัดการกับข้อจำกัดเหล่านี้ ได้แก่ k-Medoids, CLARA, CLARANS เป็นต้น

2.4.2 Hierachycal Method

เป็นการจัดกลุ่มข้อมูลโดยสร้างเป็นชั้น ๆ แบบลำดับขั้นจากข้อมูลทั้งหมดที่มีอยู่ ซึ่งวิธีนี้ไม่เป็นที่นิยมมากนัก เพราะมีค่าใช้จ่ายในการคำนวณสูงและเสียเวลา ส่วนใหญ่จะใช้กับข้อมูลที่ไม่ใช่ตัวเลข เช่น สัญลักษณ์

2.4.3 Density-based Method

วิธีการนี้มีการพัฒนาความเข้าใจทางด้านความหนาแน่น (Density) ซึ่งแนวคิดของวิธีนี้คือจะปล่อยให้ คลัสเตอร์ขยายไปเรื่อย ๆ จนกว่าค่าความหนาแน่น (จำนวนออบเจกต์) ของคลัสเตอร์ที่อยู่ใกล้กันมีค่ามากกว่าค่าหนึ่งที่กำหนดไว้ตอนแรก จะเห็นได้ว่าวิธีนี้สามารถกำจัดสิ่งรบกวนต่าง ๆ ออกไปได้ และทำงานกับคลัสเตอร์ที่มีรูปทรงใด ๆ ก็ได้

2.4.4 Grid-based Method

วิธีการนี้จะทำการ Quantize ที่ว่างของออบเจกต์ลงในเซลล์จำนวนหนึ่ง ซึ่งอยู่ในรูปโครงสร้างแบบตาราง (Grid) ซึ่งในทุก ๆ ขั้นตอนในการจัดกลุ่มจะถูกจัดการลงบนโครงสร้างแบบตารางซึ่งข้อดีของวิธีนี้คือการทำงานใช้เวลาไม่มากนักและไม่ขึ้นกับจำนวนของข้อมูล จะขึ้นกับเพียงจำนวนเซลล์ในแต่ละส่วนในที่ว่างของ Quantize

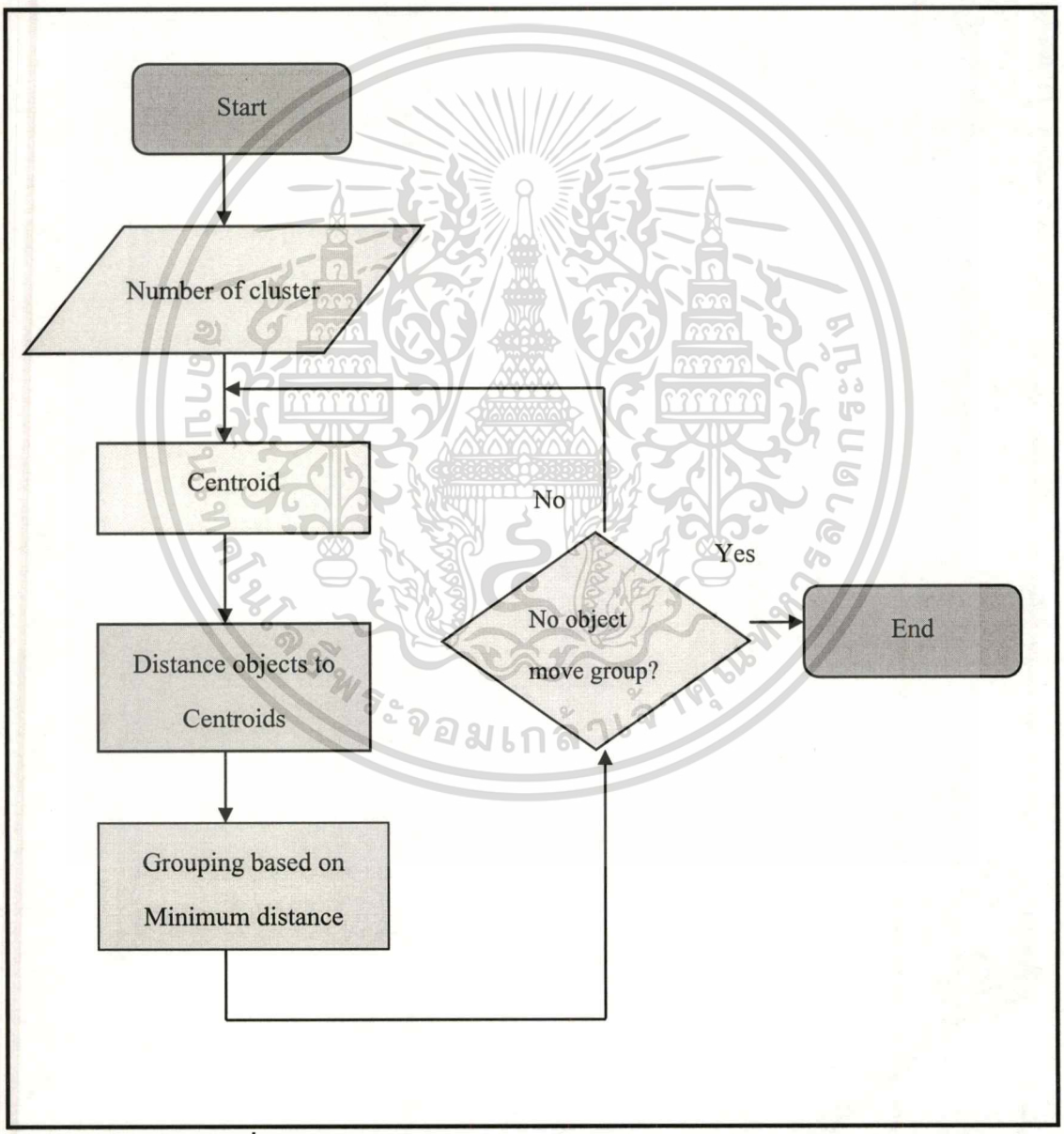
2.4.5 Model-based Method

สมมติฐานของวิธีนี้คือ สร้างโมเดลสำหรับแต่ละคลัสเตอร์ขึ้นมา และทำการหาข้อมูลที่เหมาะสมที่สุดให้กับโมเดล อัลกอริทึมในวิธีนี้อาจจะค้นหาคลัสเตอร์โดยการสร้างฟังก์ชันความหนาแน่นที่ส่งผลกับระยะการกระจายของตำแหน่งข้อมูล ซึ่งจะนำไปสู่การทำงานที่มีความเป็นอัตโนมัติมากขึ้น

2.5 เนื้อหาและหลักการของอัลกอริทึม k-Means

อัลกอริทึม k-Means เป็นอัลกอริทึมหนึ่งในวิธีการแบบ Partitioning Method ที่นิยมใช้ในการจัดกลุ่มข้อมูลเพราะมีความง่ายและตรงไปตรงมาโดยใช้พื้นฐานจากการวิเคราะห์ความแตกต่างโดยคลัสเตอร์จะเป็นกลุ่มของข้อมูลและมีการกำหนดจำนวนกลุ่มของคลัสเตอร์ไว้ก่อน โดยหนึ่งในการทำงานที่สำคัญของอัลกอริทึมคือการวัดความคล้ายคลึง (Similarity Measure) ซึ่งจะถูกใช้ในการพิจารณาว่าข้อมูลมีความเหมือนหรือแตกต่างกัน โดยคำนวณระยะห่างระหว่างศูนย์กลางกลุ่มกับข้อมูลสมาชิกได้จากสูตร Euclidean Distance Method ซึ่งใช้วัดความคล้ายคลึงของข้อมูล โดยข้อมูลภายในคลัสเตอร์เดียวกันจะมีความเหมือนกันและแตกต่างกันมากในระหว่างคลัสเตอร์ ซึ่งในแต่ละคลัสเตอร์จะมีศูนย์กลาง (Centroid) อันเดียวร่วมกันซึ่งเป็นจุดศูนย์กลาง (Midpoint) ของคลัสเตอร์โดยศูนย์กลางดังกล่าวจะเป็นค่าเฉลี่ยของข้อมูลทั้งหมดภายในคลัสเตอร์เดียวกัน ซึ่งวิธีนี้จะทำให้ได้ค่า Square Error น้อยที่สุดระหว่างข้อมูลกับศูนย์กลางของคลัสเตอร์

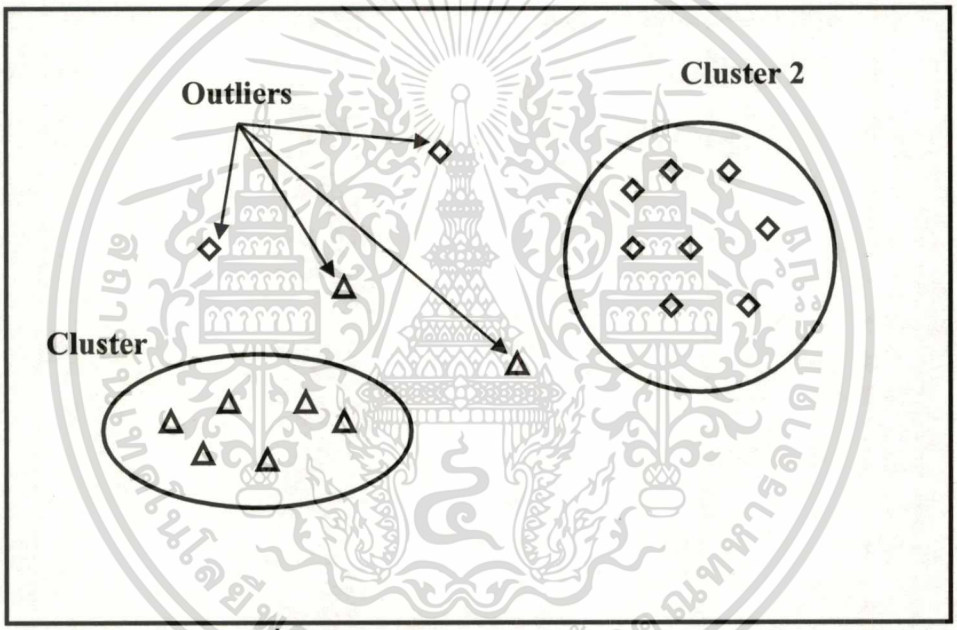
อัลกอริทึม k-Means จะเริ่มต้นด้วยการสุ่มศูนย์กลางเริ่มต้นของคลัสเตอร์และมีการกำหนดข้อมูลให้กับคลัสเตอร์โดยพิจารณาจากความคล้ายคลึงกันระหว่างข้อมูลกับศูนย์กลางของคลัสเตอร์ โดยข้อมูลที่อยู่ใกล้กับศูนย์กลางของคลัสเตอร์ใดมากที่สุดจะถูกนำไปรวมกับคลัสเตอร์นั้นซึ่งเป็นการพยายามปรับปรุงการแบ่งกลุ่มข้อมูลแบบวนรอบไปเรื่อยๆ (Iterative) ด้วยการเคลื่อนย้ายข้อมูลจากกลุ่มหนึ่งไปยังอีกกลุ่มหนึ่งที่ซึ่งข้อมูลนั้นจะใกล้กับศูนย์กลางกลุ่มจนกระทั่งได้ข้อมูลที่มีความคล้ายคลึงกันในกลุ่มเดียวกันแต่จะแตกต่างกันระหว่างกลุ่ม โดยขั้นตอนทำงานของอัลกอริทึม k-Means นั้นสามารถอธิบายได้ดังรูปที่ 2.5



รูปที่ 2.5 อธิบายภาพรวมการทำงานของอัลกอริทึม k-Means

2.5.1 อัลกอริทึม k-Means สำหรับการจัดกลุ่ม

อัลกอริทึม k-Means สำหรับการจัดกลุ่ม อัลกอริทึมจะทำการสุ่มตัวแปร k และแบ่ง n ออบเจกต์ให้อยู่ใน k กลุ่ม ดังนั้นผลลัพธ์ที่ได้คือภายในคลัสเตอร์จะมีความคล้ายคลึงกันสูง แต่แตกต่างกันระหว่างคลัสเตอร์ ความคล้ายคลึงกันของคลัสเตอร์จะถูกวัดจากค่าเฉลี่ยของออบเจกต์ในคลัสเตอร์นั้น หลักการทำงานของอัลกอริทึมนี้ขั้นแรกจะทำการสุ่มเลือกค่า k ของออบเจกต์ แต่ละการเริ่มต้นจะแทนด้วยค่าเฉลี่ยหรือค่ากลางของคลัสเตอร์ สำหรับแต่ละออบเจกต์ที่ยังเหลืออยู่จะถูกกำหนดไปเป็นคลัสเตอร์ ซึ่งจะมีความคล้ายคลึงกันมากที่สุดโดยพิจารณาจากระยะทางระหว่างออบเจกต์กับค่าเฉลี่ยของคลัสเตอร์นั้น จากนั้นจะคำนวณค่าเฉลี่ยในแต่ละคลัสเตอร์ใหม่ซึ่งการทำงานนี้จะทำซ้ำจนกระทั่งไม่มีการเปลี่ยนแปลงหรือเป็นไปตามหลักการที่เหมาะสม



รูปที่ 2.6 แสดงถึงการนำข้อมูลไปจัดกลุ่ม

อัลกอริทึม k-Means ได้กำหนดสัญลักษณ์ต่าง ๆ ไว้ดังต่อไปนี้

โดย n คือ ขนาดของอินพุท นั่นคือจำนวนข้อมูลของคลัสเตอร์

x_i คือ ค่าข้อมูล เมื่อ i = 1,…,n

m คือ ค่าเฉลี่ยของกลุ่ม นั่นคือศูนย์กลางของคลัสเตอร์

$d(x_i, m)$ คือ ผลต่างรวมของระยะทางของข้อมูลจากทุก ๆ กลุ่มกับศูนย์กลางกลุ่ม

สรุปการทำงานของอัลกอริทึม k-Means ได้ดังนี้

1. ทำการสุ่มค่าศูนย์กลางเริ่มต้นของคลัสเตอร์
2. ทำซ้ำ

(a) สำหรับแต่ละเวกเตอร์ข้อมูลจะกำหนดให้กับเวกเตอร์ศูนย์กลางที่อยู่ใกล้ที่สุด โดยระยะทางไปยังเวกเตอร์ศูนย์กลางคำนวณได้จาก

$$d(x_i, m) = \sqrt{\sum_{i=1}^{n} (x_i - m)^2} \tag{2.1}$$

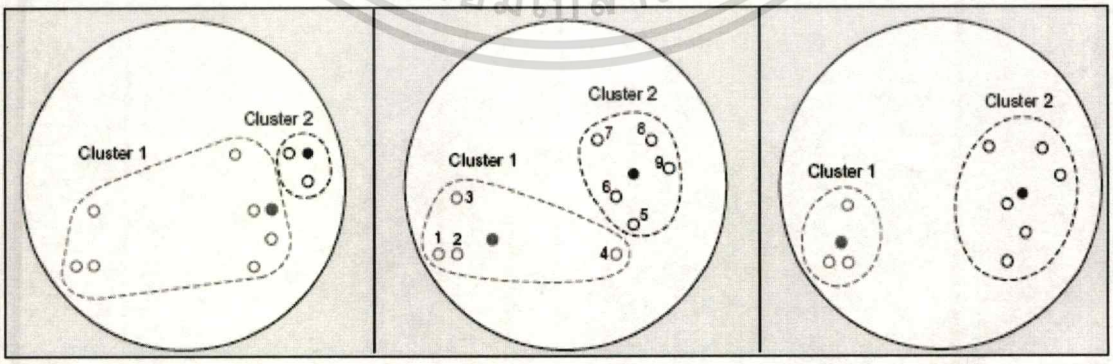
ที่ซึ่ง i เป็นเครื่องหมายที่แสดงถึงขนาด

(b) คำนวณค่าเฉลี่ยใหม่สำหรับเวกเตอร์ศูนย์กลางของคลัสเตอร์ซึ่งสามารถ โดยคำนวณได้จาก

$$m = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.2}$$

ทำซ้ำจนกระทั่งเป็นไปตามหลักการที่เหมาะสม

อัลกอริทึมจะพยายามลดค่า Square Error จนผลลัพธ์ของการจัดกลุ่มจะได้ค่าที่น้อยที่สุด การทำงานของอัลกอริทึมนี้จะหยุดเมื่อได้ค่าใดค่าหนึ่งที่เป็นไปตามหลักการที่เหมาะสมอันได้แก่ ค่า Squere Error ไม่เปลี่ยนแปลงหรือเบนเข้าหาค่าค่าหนึ่งที่กำหนดไว้ หรือเมื่อค่าสูงสุดในการทำซ้ำมีค่ามากเกินหรือเมื่อทำซ้ำจนมีการเปลี่ยนแปลงเพียงเล็กน้อยของเวกเตอร์ศูนย์กลางมากกว่าจำนวนของการทำซ้ำ รวมทั้งเมื่อไม่มีการเปลี่ยนแปลงของสมาชิกในคลัสเตอร์ ซึ่งการจัดกลุ่มโดยใช้อัลกอริทึม k-Means แสดงได้ดังรูปที่ 2.7



รูปที่ 2.7 แสดงการจัดกลุ่มของอัลกอริทึม k-Means

2.5.2 ประสิทธิภาพของอัลกอริทึม k-Means

อัลกอริทึม k-Means จะทำงานได้อย่างมีประสิทธิภาพกับข้อมูลที่มีลักษณะเป็นทรงกลม ข้อดีของอัลกอริทึมนี้คือ ง่าย ทำงานได้เร็วแม้จะมีข้อมูลขนาดใหญ่ แต่จะเหมาะสำหรับการทำงานกับข้อมูลประเภท Numeric หรือข้อมูลเชิงตัวเลขเท่านั้น หากไม่ใช้ข้อมูลประเภทนี้ก็จะต้องแปลง (Map) ให้เป็นตัวเลขก่อน เช่น เพศชายแปลงเป็น 0 หรือเพศหญิงแปลงเป็น 1 เป็นต้น

ถึงแม้ว่าอัลกอริทึม k-Means จะมีประโยชน์อย่างกว้างขวางในการจัดกลุ่มแต่ก็ยังคงมีข้อเสียอีกมากมาย เนื่องจากค่า Means ที่คำนวณอาจคลาดเคลื่อนได้ง่ายและไม่สมเหตุสมผลหากมีข้อมูลที่มีลักษณะเป็น Noise และ Outliers อยู่แม้จะมีจำนวนน้อยก็ตาม และการทำให้ได้มาของค่า Square Error ที่น้อยที่สุดนั้นเป็นไปได้ที่จะใช้เวลาในการทำงานนาน และประสิทธิภาพของอัลกอริทึมยังมีข้อจำกัดในเรื่องการกำหนดจำนวนกลุ่มในตอนแรก ซึ่งการสุ่มศูนย์กลางเริ่มต้นโดยถ้าศูนย์กลางเริ่มต้นของการสุ่มครั้งหนึ่งไม่เหมือนกับการสุ่มอีกครั้งหนึ่งจะส่งให้ผลการจัดกลุ่มนั้นแตกต่างกัน และถ้าการสุ่มศูนย์กลางครั้งแรกไม่ใช่ศูนย์กลางที่ดี จะทำให้การจัดการกลุ่มข้อมูลได้ผลไม่ดี อย่างไรก็ตามอัลกอริทึม k-Means ก็สามารถทำงานได้เป็นอย่างดีในการหาศูนย์กลางของการจัดกลุ่มที่ดีที่สุด

ประสิทธิภาพของอัลกอริทึมนี้คือ $O(nkt)$

โดย n คือ จำนวนของออบเจกต์

k คือ จำนวนของกลุ่ม

t คือ จำนวนการทำซ้ำของอัลกอริทึมในการปรับปรุงคลัสเตอร์

สำหรับการวิเคราะห์การทำงานของอัลกอริทึม k-Means นี้จึงขึ้นอยู่กับจำนวนข้อมูลทั้งหมด จำนวนกลุ่มทั้งหมดที่ต้องการ และจำนวนรอบการทำงานที่ใช้ในการปรับปรุงคลัสเตอร์

2.5.3 Distance Function

Distance Function เป็นฟังก์ชันที่ใช้กำหนดความคล้ายคลึงกันในกลุ่มของข้อมูล โดยพิจารณาจากระยะทางระหว่างข้อมูล 2 ตัวซึ่งจะอยู่ในรูปของ Metric ซึ่งตัวอย่างของ Distance Function ที่ใช้ในโครงการศึกษานี้คือ Euclidean Distance Function ซึ่งมีสูตรดังนี้

$$d = \sqrt{\sum_{i=1}^{n} (x_i - m)^2} \tag{2.3}$$

โดย d คือ ระยะทาง

x_i คือ ตำแหน่งของข้อมูล

n คือ จำนวนข้อมูลสมาชิก

m คือ ตำแหน่งค่าเฉลี่ยของกลุ่มข้อมูล

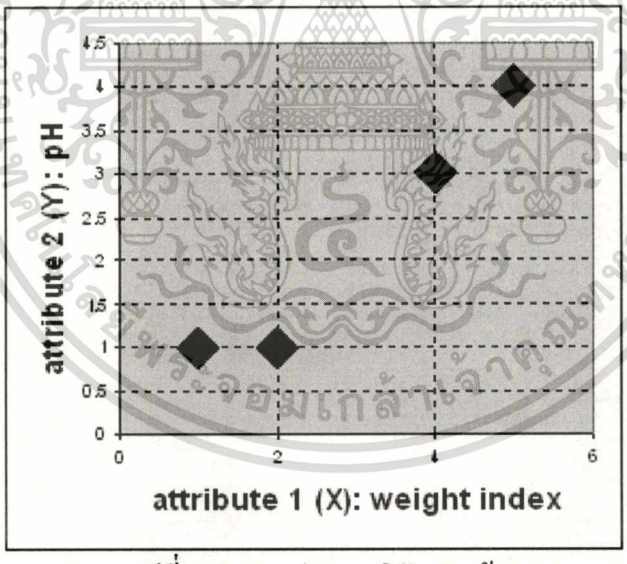
2.5.4 ตัวอย่างการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม k-Means

สมมติเรามีออบเจกต์อยู่ 4 ออบเจกต์ ซึ่งแต่ละออบเจกต์มี 2 แอตทริบิวต์ (Attribute) ดังแสดงไว้ในตาราง 2.1 ซึ่งจุดประสงค์ของตัวอย่างนี้คือต้องการจัดกลุ่มยา (Medicine) เป็น 2 กลุ่ม (k=2) โดยใช้ค่า pH และ Weight index ในการจัดกลุ่มของยา

ตารางที่ 2.1 ตัวอย่างของข้อมูลยาที่จะใช้ในการจัดกลุ่ม

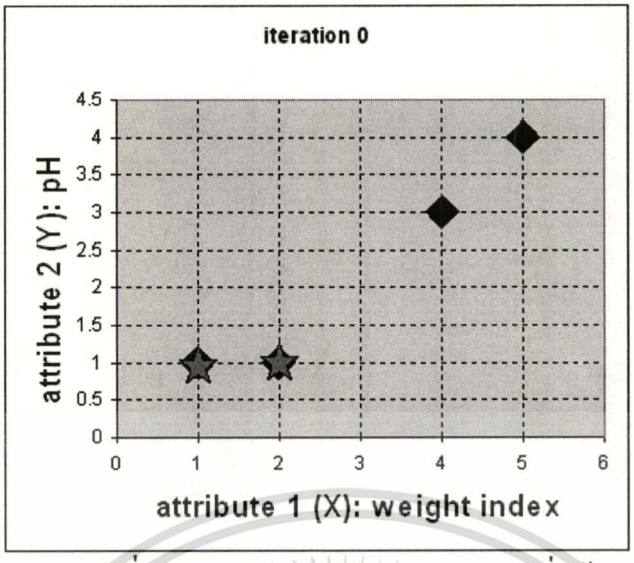
Object	Attribute 1 (X) : Weight index	Attribute 2 (Y) : pH
Medicine A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4

ตัวยาแต่ละตัวจะแทน 1 จุดที่ประกอบด้วย 2 แอตทริบิวต์ (X,Y) ซึ่งสามารถเขียนกราฟแสดงพิกัดได้ดังรูปที่ 2.8



รูปที่ 2.8 กราฟแสดงพิกัดของข้อมูลยา

<u>ขั้นที่ 1</u> กำหนดค่าของ centroid เริ่มต้น สมมติเราใช้ Medicine A และ Medicine B เป็น centroid เริ่มต้น เราจะให้ c_1 และ c_2 แทนด้วยพิกัดของ centroid ดังนั้น $c_1 = (1,1)$ และ $c_2 = (2,1)$



รูปที่ 2.9 กราฟแสดงพิกัดของ centroid เริ่มต้น

<u>ขั้นที่ 2</u> เป็นการหาค่า distance หรือระยะระหว่างออบเจกต์แต่ละตัวเปรียบเทียบกับค่า centroid ในแต่ละกลุ่ม ซึ่งเราจะใช้ Euclidean Distance Method ในการหาดังนั้นเราจะได้ Distance Matrix ที่ Iteration 0 : (D^0) ดังนี้

$$D^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 3.83 & 4.24 \end{bmatrix} \quad \begin{matrix} c_1 = (1,1) & group-1 \\ c_2 = (2,1) & group-2 \end{matrix}$$

A B C D

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \begin{matrix} X \\ Y \end{matrix}$$

แต่ละ Column ใน Distance Matrix แสดงถึงออบเจกต์ แถวแรกของ Distance matrix เป็นระยะระหว่างออบเจกต์ แต่ละออบเจกต์เปรียบเทียบกับ centroid ตัวแรก และแถวที่สองก็คือระยะระหว่างแต่ละออบเจกต์ เปรียบเทียบกับ centroid ตัวที่สอง

ตัวอย่างเช่น ระยะทางระหว่าง Medicine C = (4,3) กับ centroid ตัวแรกคือ $c_1 = (1,1)$ มีค่า Distance = $\sqrt{(4-1)^2 + (3-1)^2} = 3.61$ และระยะระหว่าง Medicine C = (4,3) กับ centroid ตัวที่สองคือ $c_2 = (2,1)$ มีค่า Distance = $\sqrt{(4-2)^2 + (3-1)^2} = 2.83$ เป็นต้น

<u>ขั้นที่ 3</u> เป็นการจัดกลุ่มให้กับออบเจกต์แต่ละตัว นั่นคือเรากำหนดแต่ละออบเจกต์ไปไว้ในกลุ่มที่มีค่า Distance น้อยที่สุด ดังนั้น Medicine A จะถูกกำหนดให้อยู่ในกลุ่มที่ 1 Medicine B อยู่ในกลุ่มที่ 2 Medicine C อยู่ในกลุ่มที่ 2 และ Medicine D อยู่ในกลุ่มที่ 2 เช่นเดียวกัน ซึ่งออบเจกต์ที่อยู่ในกลุ่มดนั้นแทนด้วย 1 แสดงใน Group Matrix ดังนี้

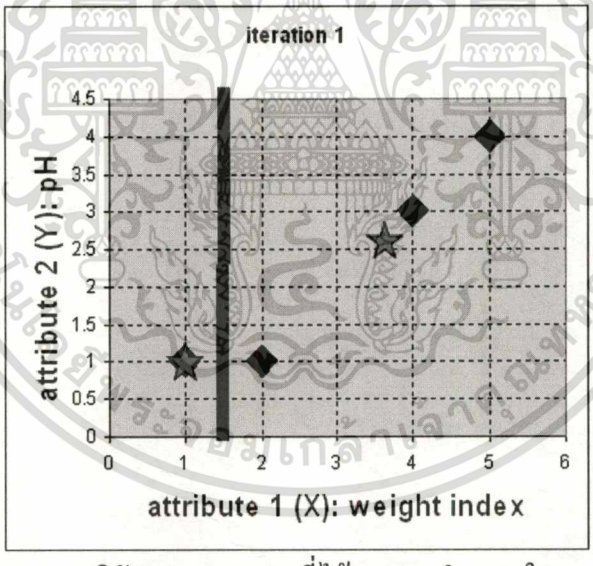
$$G^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \quad \begin{matrix} group-1 \\ group-2 \end{matrix}$$

A B C D

<u>ขั้นที่ 4</u> ใน Iteration-1 นี้จะเป็นการหาค่า centroid ใหม่ ซึ่งเมื่อเราทราบสมาชิกในแต่ละกลุ่มแล้ว ในขั้นตอนนี้เราจะทำการคำนวณหาค่า centroid ของแต่ละกลุ่มใหม่ ซึ่งในกลุ่มที่ 1 นั้นมีสมาชิกเพียงหนึ่งตัว ดังนั้นค่า centroid ก็ยังคงเป็นตัวเดิมคือ $c_1 = (1,1)$ ในกลุ่มที่ 2 ซึ่งมีสมาชิกอยู่ 3 ตัว ดังนั้นค่า centroid คือการหาค่าเฉลี่ยของพิกัดระหว่างสมาชิก 3 ตัว เพราะฉะนั้นจะได้ centroid ใหม่ของกลุ่มที่ 2 ดังนี้

$$c_2 = \left(\frac{2+4+5}{3}, \frac{1+3+4}{3} \right) = (3.7,2.7)$$

ซึ่ง centroid ใหม่ที่ได้มาจากการคำนวณของทั้งสองกลุ่มสามารถแสดงได้ดังรูปที่ 2.10



รูปที่ 2.10 แสดงพิกัดของ centroid ที่ได้จากการคำนวณในการทำรอบที่ 1

<u>ขั้นที่ 5</u> ใน Iteration-1 จะเป็นการคำนวณหา Distance ของทุก ๆ ออบเจกต์เปรียบเทียบกับ centroid ใหม่คล้ายในขั้นที่ 2 ซึ่ง Distance matrix ในการทำรอบที่ 1 คือ

$$D^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \quad \begin{matrix} c_1 = (1,1) & group-1 \\ c_2 = (1.4,2.7) & group-2 \end{matrix}$$

A B C D

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \begin{matrix} X \\ Y \end{matrix}$$

ขั้นที่ 6 Iteration-1 จะเป็นการจัดกลุ่มของออบเจกต์ เหมือนในขั้นที่ 3 ซึ่งเราจะทำการกำหนดแต่ละออบเจกต์เข้าไปในกลุ่มที่มี Distance น้อยที่สุด จาก Distance matrix ที่ได้จากการคำนวณใหม่นี้เราจะย้าย Medicine B ไปไว้กลุ่มที่ 1 ในขณะที่ออบเจกต์ตัวอื่น ๆ ก็ยังอยู่ในกลุ่มเหมือนเดิม ซึ่งจะได้ Group matrix ดังนี้

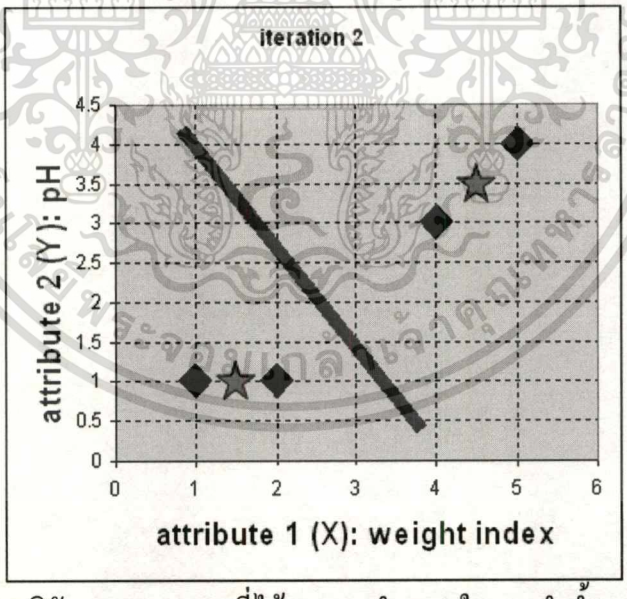
$$G^1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad \begin{matrix} group-1 \\ group-2 \end{matrix}$$

$$\mathbf{A \quad B \quad C \quad D}$$

ขั้นที่ 7 Iteration-2 เป็นการหาค่า centroid ใหม่ ซึ่งเราจะทำซ้ำขั้นที่ 4 เพื่อคำนวณหาพิกัดของ centroid ใหม่ โดยจะพิจารณาจากการจัดกลุ่มในรอบที่แล้ว ซึ่งทั้งในกลุ่มที่ 1 และ 2 ต่างก็มีสมาชิก 2 ตัว ดังนั้น centroid ใหม่คำนวณได้ดังนี้

$$c_1 = \left(\frac{1+2}{2}, \frac{1+1}{2}\right) = (1.5,1) \quad \text{และ} \quad c_2 = \left(\frac{4+5}{2}, \frac{3+4}{2}\right) = (4.5,3.5)$$

ซึ่ง centroid ใหม่ที่ได้มาจากการคำนวณของทั้งสองกลุ่มสามารถแสดงได้ดังรูปที่ 2.11



รูปที่ 2.11 พิกัดของ centroid ที่ได้จากการคำนวณในการทำซ้ำรอบที่ 2

ขั้นที่ 8 Iteration-2 ทำซ้ำในขั้นที่ 2 อีกครั้ง เราจะได้ Distance Matrix ใหม่ในการทำซ้ำรอบที่ 2 เป็นดังนี้

$$D^0 = \begin{bmatrix} 0.5 & 0.5 & 3.2 & 4.61 \\ 4.3 & 3.54 & 0.71 & 0.71 \end{bmatrix} \quad \begin{matrix} c_1 = (1.5,1) & group-1 \\ c_2 = (4.5,3.5) & group-2 \end{matrix}$$

$$\begin{matrix} \mathbf{A} & \mathbf{B} & \mathbf{C} & \mathbf{D} \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \qquad \begin{matrix} X \\ Y \end{matrix}$$

<u>ขั้นที่ 9</u> Iteration-2 เป็นการกำหนดแต่ละออบเจกต์เข้าไปอยู่ในกลุ่มที่มี Distance น้อยที่สุด

$$G^2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \qquad \begin{matrix} group-1 \\ group-2 \end{matrix}$$

$$\mathbf{A} \quad \mathbf{B} \quad \mathbf{C} \quad \mathbf{D}$$

ผลลัพธ์ที่ได้คือ $G^2 = G^1$ นั่นคือ Group matrix ของการทำซ้ำในรอบที่ 2 และรอบที่ 1 ให้ผลลัพธ์เหมือนเดิม ซึ่งเป็นการเปรียบเทียบการจัดกลุ่มของรอบสุดท้ายและในรอบนี้แสดงให้เห็นว่าออบเจกต์ไม่ได้มีการเปลี่ยนกลุ่มแล้ว ด้วยเหตุนี้การทำงานของอัลกอริทึม k-Means ได้มาถึงจุดที่ไม่มีการเปลี่ยนแปลงแล้ว ดังนั้นเราจะเอากลุ่มที่ได้จากการจัดกลุ่มครั้งสุดท้ายมาเป็นผลลัพธ์ ดังแสดงในตารางที่ 2.2

ตาราง 2.2 ผลลัพธ์ที่ได้จากการจัดกลุ่มด้วยอัลกอริทึม k-Means

Object	Attribute 1(X):Weight index	Attribute 2 (Y):pH	Cluster
Medicine A	1	1	1
Medicine B	2	1	1
Medicine C	4	3	2
Medicine D	5	4	2

ถึงอัลกอริทึม k-Means จะมีความสามารถในการจัดกลุ่มกับเซตข้อมูลขนาดใหญ่ ซึ่งสามารถใช้เวลาทำงานได้อย่างรวดเร็วในการหาตัวแทนศูนย์กลางของคลัสเตอร์นั่นคือจะใช้ค่าเฉลี่ยอย่างที่ได้กล่าวมาแล้ว แต่การจัดกลุ่มที่ได้อาจมีการผิดพลาดซึ่งอาจเกิดจากการพบว่ามีค่า Noise หรือ Outliers ซึ่งค่านี้จะทำให้การจัดกลุ่มข้อมูลผิดพลาดไป ดังนั้นจะมีวิธีการใดที่สามารถจัดการกับข้อบกพร่องเหล่านี้ของ k-Means ได้ เพื่อให้การจัดกลุ่มข้อมูลเป็นไปอย่างมีประสิทธิภาพและพบข้อบกพร่องน้อยสุด

บทที่ 3
ไฮบริดพาร์ทิเคิลสวอมออปติไมเซชั่น

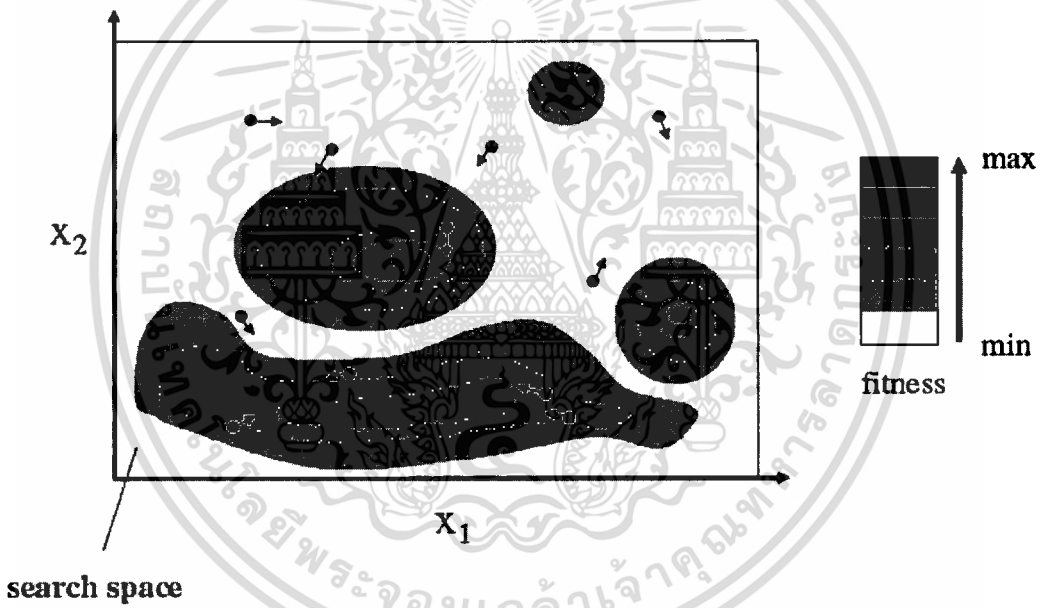
จากที่กล่าวมาแล้วว่าวิธีการวิเคราะห์การจัดกลุ่มโดยใช้อัลกอริทึม k-Means นั้นใช้หลักการทำงานขั้นพื้นฐานเท่านั้น โดย k-Means จะใช้ค่าเฉลี่ยของกลุ่มเป็นตัวแทน แต่ค่าเฉลี่ยบางครั้งก็อาจไม่ใช่ตัวแทนที่ดีนัก อาจทำให้ผลลัพธ์ที่ได้ยังไม่ดีที่สุด และมีการคำนวณที่สิ้นเปลืองในการที่จะทำให้ได้มาของค่า Square Error ที่น้อยที่สุดและยังเกิดข้อผิดพลาดกับข้อมูลที่มีลักษณะเป็น Noise และ Outliers รวมทั้งยังไม่มีประสิทธภาพในเซตข้อมูลที่มีคลัสเตอร์ที่ซับซ้อน โดยผลลัพธ์ที่ได้จากปัญหาเหล่านี้จึงได้มีการจัดหาวิธีการจัดกลุ่มที่ใช้เทคนิคที่มีความน่าสนใจและมีการพัฒนากันอย่างต่อเนื่องคือ พาร์ทิเคิลสวอมออปติไมเซชั่น (Particle Swarm Optimization : PSO) โดยใช้พื้นฐานของวิวัฒนาการที่มีความสามารถที่สูงกว่าในการหาค่าที่เหมาะสมเพื่อให้ได้ผลลัพธ์ที่ดีที่สุดจากวิธีการที่เป็นไปได้ทั้งหมด (Search Space) เพื่อใช้เป็นตัวแทนศูนย์กลางของกลุ่ม

พาร์ทิเคิลสวอมออปติไมเซชั่น (PSO) เป็นอัลกอริทึมที่ตั้งอยู่บนรากฐานของจำนวนประชากร โดยการทำงานของอัลกอริทึมจะลอกเลียนแบบการรวมกลุ่มการบินของนกหรือการว่ายเป็นฝูงของปลาจนได้มาของระบบวิวัฒนาการด้วยตนเองซึ่งสามารถใช้ในการค้นหาวิธีการที่ดีที่สุดจากวิธีการทั้งหมดได้โดยอัตโนมัติ จากปัญหาที่แตกต่างกันนั้นจึงมีการตัดสินให้ใช้วิธีการค้นหาโดยการใช้ฟังก์ชั่นความเหมาะสม (Fitness Function) ซึ่งไม่เหมือนกับอัลกอริทึมในการเรียนรู้ที่เป็นวิวัฒนาการอื่น ๆ ปัจจุบันมีผู้แต่งหรือผู้ที่สนใจมากมายเกี่ยวกับ PSO ได้ศึกษาและทำการพัฒนาการทำงานของอัลกอริทึมอย่างต่อเนื่อง ซึ่งประกอบไปด้วยแนวคิดที่ง่ายซึ่งต้องการตัวแปรในจำนวนที่น้อยในการตัดสินใจหรือใช้โค้ด (Code) ของคอมพิวเตอร์เพียงไม่กี่แถวในการทำงานและใช้หน่วยความจำและความเร็วจากคอมพิวเตอร์ที่ไม่จำเป็นต้องสูงมากนัก รวมถึงต้องการการทำงานทางด้านคณิตศาสตร์ที่ไม่ซับซ้อน ดังนั้น PSO จึงง่ายในการดำเนินการและมีลักษณะของรูปแบบที่มั่นคงและมีผลการคำนวณที่ถูกต้องและมีประสิทธิภาพ

3.1 เนื้อหาและหลักการของพาร์ทิเคิลสวอมออปติไมเซชั่น (PSO)

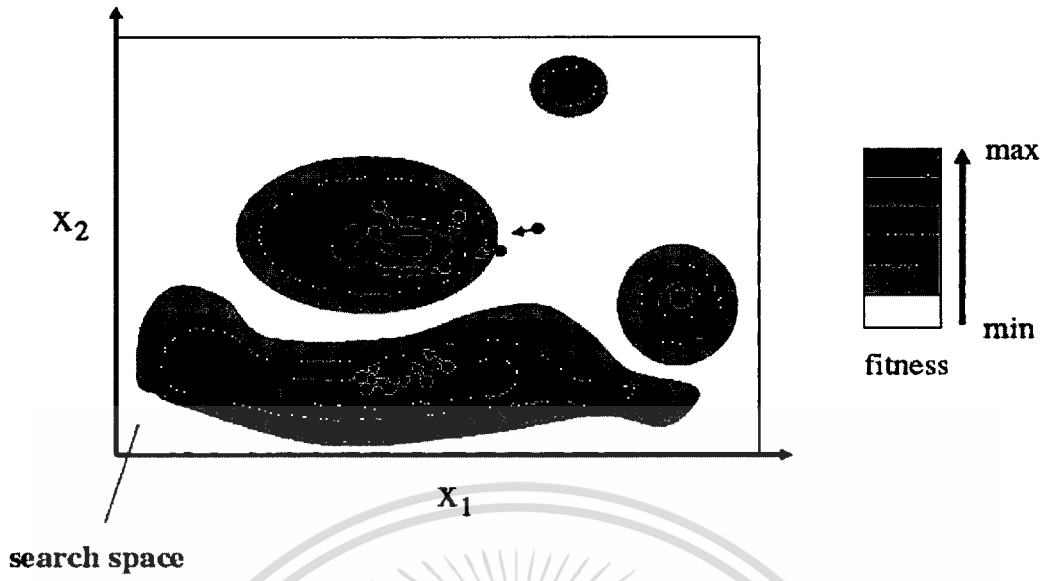
พาร์ทิเคิลสวอมออปติไมเซชั่น (Particle Swarm Optimization : PSO) เป็นเทคนิควิธีการขุองอัลกอริทึมเพื่อใช้สำหรับค้นหาค่าที่เหมาะสมของปัญหาใด ๆ โดยใช้วิธีของกลุ่มประชากรในการค้นหา และใช้พื้นฐานของความน่าจะเป็นซึ่งถูกคิดค้นพัฒนาในปี 1995 โดย James Kennedy และ Russell Eberhart เป็นวิธีที่ได้มาจากพฤติกรรมการเรียนรู้การอยู่ร่วมกันของสิ่งมีชีวิตแบบกลุ่ม ซึ่งสามารถทำให้เข้าใจง่าย เช่น การเรียนรู้การอาหารของนก หรือการว่ายเป็นฝูงของปลา เป็นต้น

เมื่อเราพิจารณา PSO แล้วจะพบว่าเป็นอัลกอริทึมที่มีแนววิธีการที่คล้ายกันกับจีเนอติกอัลกอริทึม (Genetic Algorithm : GA) ตรงที่ PSO จะใช้วิธีการสุ่มสร้างประชากรจำนวนหนึ่งขึ้นมาที่เรียกว่า พาร์ทิเคิล (Particle) แล้วทำการคำนวณฟังก์ชั่นความเหมาะสม (Fitness Function) หรือฟังก์ชั่นเป้าหมาย (Objective Function) ทำให้ได้มาของตำแหน่งพาร์ทิเคิลและค่าความเหมาะสม (Fitness Value) ที่ดีที่สุด โดยได้มาจากการใช้ประชากรเดิมทำการเคลื่อนที่ผ่านพื้นที่ทั้งหมด และปรับทิศทางไปยังตำแหน่งของพาร์ทิเคิลตัวที่มีค่าความเหมาะสมมากกว่าตัวพาร์ทิเคิลเอง เป็นจำนวนหลาย ๆ รอบ จนกระทั่งได้ตำแหน่งที่ดีที่สุดเท่าที่พาร์ทิเคิลชุดนี้จะหาออกมาได้ สำหรับส่วนที่ไม่เหมือนกับ GA ตรงที่ PSO ไม่มีตัวดำเนินการ (Operation) เช่น Crossover และ Mutation เป็นต้น สำหรับข้อดีของ PSO เมื่อเปรียบเทียบกับ GA ได้แก่ ง่ายในการสร้าง และใช้ตัวแปรน้อยกว่า จึงทำให้การปรับเงื่อนไขทำได้ดีกว่าในการลู่เข้าสู่เป้าหมาย



รูปที่ 3.1 แสดงการเริ่มต้นในการค้นหาตำแหน่งของแต่ละพาร์ทิเคิล

วิธีการทำงานของ PSO นั้นถือได้ว่าเป็นการประยุกต์จากการลอกเลียนแบบธรรมชาติมา ซึ่งสามารถอธิบายได้โดยใช้สถานการณ์ต่อไปนี้ ถ้าฝูงนกกำลังค้นหาอาหารอย่างสุ่มอยู่ในพื้นที่หนึ่ง ซึ่งฝูงนกก็คือพาร์ทิเคิลและแหล่งอาหารจะเป็นตำแหน่งที่มีค่าความเหมาะสมสูงที่สุดดังแสดงในรูปที่ 3.1 โดยนกแต่ละตัวไม่รู้ว่าอาหารอยู่ที่ใดในบริเวณนั้น แต่จะรู้โดยการบินลงไปทีละรอบ ซึ่งในที่นี้ถ้าหากมีนกอยู่หลายตัว ก็จะลงไปสำรวจได้หลายบริเวณในแต่ละรอบ และนกแต่ละตัวต้องจดจำตำแหน่งที่ดีที่สุดที่เคยได้ไปมารวมถึงตำแหน่งที่ดีที่สุดจากนกตัวอื่น หลังจากนั้นจึงรู้ว่านกตัวใดอยู่ในตำแหน่งใกล้แหล่งอาหารมากที่สุด (ตำแหน่งที่ดีที่สุดที่ค้นพบ) ซึ่งนกตัวอื่นก็จะตามนกตัวนั้นไปจนกระทั่งเจอแหล่งอาหารดังรูปที่ 3.2



รูปที่ 3.2 แสดงการเคลื่อนที่ของพาร์ทิเคิลเมื่อค้นพบตำแหน่งที่ดีที่สุด

3.2 การทำงานของอัลกอริทึม PSO

ในระบบของ PSO นั้นกลุ่มของสิ่งมีชีวิตชนิดเดียวกันจะเรียกว่า พาร์ทิเคิล (Particles) ที่เคลื่อนที่ผ่านพื้นที่ทั้งหมด แต่ละพาร์ทิเคิลจะแทนด้วยวิธีการที่เป็นไปได้ และตำแหน่งของพาร์ทิเคิลจะได้รับอิทธิพลจากตำแหน่งที่ดีที่สุดซึ่งได้มาจากการไปด้วยตัวเอง (พบด้วยตัวเอง) หรือได้มาจากตำแหน่งของพาร์ทิเคิลที่ดีที่สุดที่อยู่ในบริเวณอื่น (พาร์ทิเคิลใกล้เคียงอื่น ๆ เป็นผู้พบ) จุดมุ่งหมายของ PSO คือค้นหาตำแหน่งพาร์ทิเคิลที่ดีที่สุดที่ได้จากฟังก์ชั่นเป้าหมาย ดังนั้นประสิทธิภาพของแต่ละพาร์ทิเคิลจะขึ้นอยู่กับการทำให้พาร์ทิเคิลได้ค่าที่ดีที่สุดจากพาร์ทิเคิลทั้งหมด โดยได้จากการใช้ฟังก์ชันเป้าหมาย (Fitness Function) และกำหนดให้เป็นค่าความเหมาะสม (Fitness Value)

พาร์ทิเคิลแต่ละตัวนั้นต้องทราบตำแหน่ง (Position) ที่มีค่าความเหมาะสมสูงที่สุดจากทั้งหมดของกลุ่มซึ่งตำแหน่งนี้จะถูกเรียกว่า global best *(gbest)* และแต่ละพาร์ทิเคิลจะมีการเก็บค่าความเหมาะสมสูงที่สุดของตัวเองไว้โดยตำแหน่งของค่าเหล่านี้จะเป็น personel best *(pbest)* ของตัวเอง เพื่อใช้ในการปรับตำแหน่งถัดไปของแต่ละพาร์ทิเคิลเอง อัลกอริทึมพื้นฐานจะมีส่วนเกี่ยวข้องกับการจัดการกับจำนวนประชากรของพาร์ทิเคิลตลอดการค้นหา จนตำแหน่งที่ดีที่สุดและเหมาะสมที่สุดจะถูกค้นพบ และแต่ละครั้งในการทำซ้ำ ทุก ๆ พาร์ทิเคิลจะมีการปรับขนาดของเวกเตอร์ความเร็ว (Velocity Vector) ตามการชักจูงจากทั้งวิธีที่ดีที่สุดของพาร์ทิเคิลเอง *(pbest)* และวิธีที่ดีที่สุดของบริเวณใกล้เคียง *(gbest)* จากนั้นตำแหน่งใหม่จะถูกคำนวณเพื่อตรวจสอบ ทำให้ได้มาของผลลัพธ์ที่ดีที่สุด แต่ละพาร์ทิเคิลในกลุ่มจะทำการเปลี่ยนแปลงตำแหน่งโดยใช้ข้อมูลดังนี้

- ตำแหน่งปัจจุบันของพาร์ทิเคิล
- อัตราความเร็วปัจจุบันของพาร์ทิเคิล
- ตำแหน่ง *pbest* ของพาร์ทิเคิล
- ตำแหน่ง *gbest* ของพาร์ทิเคิลทั้งหมด
- ระยะทางระหว่างตำแหน่งปัจจุบันและ *pbest*
- ระยะทางระหว่างตำแหน่งปัจจุบันและ *gbest*

การทำงานของอัลกอริทึม PSO จะเริ่มต้นโดยกำหนดค่าเริ่มต้นของ PSO อย่างสุ่ม ต่อจากนั้นจะคำนวณหาแต่ละค่าที่เหมาะสมของแต่ละพาร์ทิเคลิจากฟังก์ชั่นความเหมาะสม หลังจากนั้นจะสำรวจแล้วเก็บค่าที่ดีที่สุดที่กล่าวไว้ข้างต้นซึ่งมีสองค่าด้วยกัน โดยค่าแรกได้แก่ ค่าความเหมาะสมเท่าที่มีในขณะนั้นของแต่ละพาร์ทิเคิลเองซึ่งจะเรียกว่า *pbest* และค่าที่สองได้แก่ค่าความเหมาะสมที่ดีที่สุดที่ครอบคลุมถึงพาร์ทิเคิลทั้งหมดรวมถึงรอบทุกรอบที่ผ่านมาด้วย ซึ่งจะเรียกค่านี้ว่า *gbest* หลังจากหาค่าที่ดีที่สุดได้ทั้ง 2 ค่าแล้ว แต่ละพาร์ทิเคิลจะทำการปรับค่าความเร็วและตำแหน่งโดยใช้สมการ (3.1) และ (3.2) ตามลำดับ ซึ่งกระบวนการนี้จะทำซ้ำจนกระทั่งจำนวนของการทำซ้ำมากเกินไปหรือมากกว่าจำนวนที่กำหนดไว้หรือเมื่อเปลี่ยนอัตราเร็วจนมีค่าเข้าใกล้ 0

$$v(t+1) = v(t) + C_1 rand_1(pbest(t) - present(t)) + C_2 rand_2(gbest(t) - present(t)) \quad (3.1)$$

$$present(t+1) = present(t) + v(t+1) \quad (3.2)$$

โดย $present(t)$ คือ ตำแหน่งปัจจุบันของพาร์ทิเคิล

$v(t)$ คือ อัตราความเร็วของพาร์ทิเคิล ณ เวลา t

t คือ จำนวนรอบในการทำงาน

C_1, C_2 คือ ค่าแฟกเตอร์ของการเรียนรู้ นิยมให้ $C_1 = C_2 = 2$

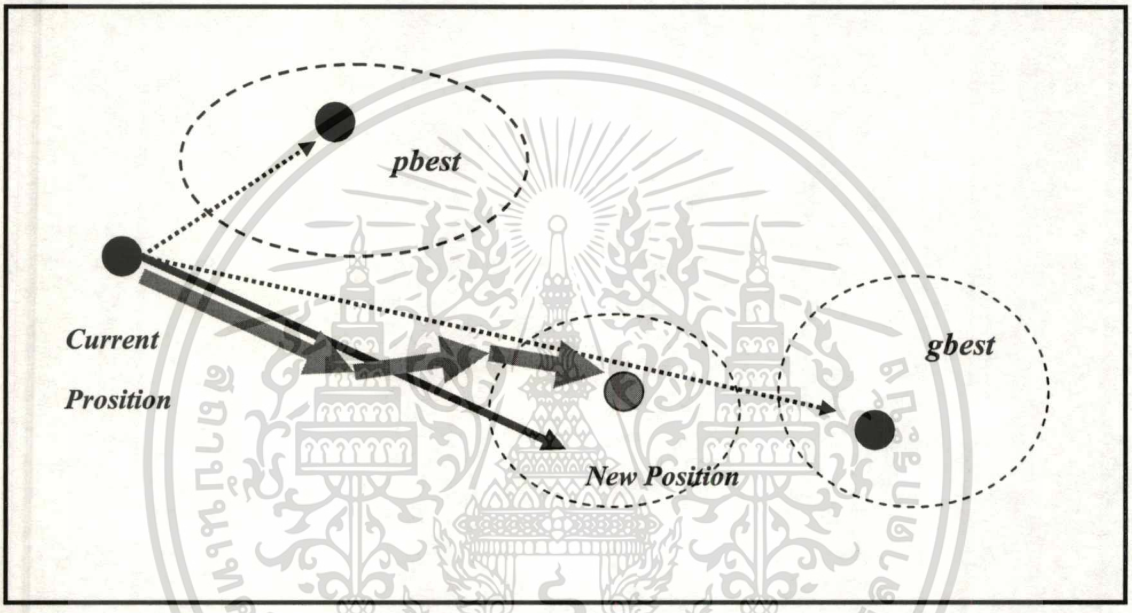
$rand_1, rand_2$ คือ ค่าที่ได้จากการสุ่มมีค่าระหว่าง 0 ถึง 1

pbest คือ ตำแหน่งที่มีค่าความเหมาะสมดีที่สุดของพาร์ทิเคิลเอง

gbest คือ ตำแหน่งที่มีค่าความเหมาะสมที่ดีที่สุดจากพาร์ทิเคิลทั้งหมด

โดยที่ค่า C_1 มีผลต่ออัตราความเร็วในการเข้าหาคำตอบที่ดีเท่าที่รู้ (*pbest*) ของตำแหน่งปัจจุบัน ส่วนค่า C_2 มีผลต่อการลู่เข้าสู่คำตอบที่ดีที่สุด (*gbest*) จากตำแหน่งปัจจุบัน จะเห็นได้ว่าตัวแปรทั้ง 2 มีผลอย่างมากต่อการลู่เข้าสู่คำตอบ และอัตราความเร็วในการลู่เข้าสู่เป้าหมาย ดังนั้นการเลือกค่าที่เหมาะสมสำหรับแต่ละตัวแปรจึงมีความสำคัญอย่างยิ่ง

ถ้าหากค่า C_1 และ C_2 มีค่าสูงทั้งคู่ การลู่เข้าสู่คำตอบจะเร็ว แต่ถ้า C_1 และ C_2 มีค่าต่ำทั้งคู่ แต่ละพาร์ทิเคิลจะลู่เข้าสู่คำตอบอย่างประสานงานกัน แม้ว่าความเร็วจะช้า แต่ก็จะได้รับคำตอบที่เหมาะสมที่สุด ถ้าค่า C_1 สูงกว่าค่า C_2 ทำให้การค้นหากระจัดกระจาย ไม่เป็นรูปแบบที่สอดคล้องกัน ทำให้ไม่สามารถลู่เข้าสู่คำตอบได้ แต่ถ้าค่า C_2 สูงกว่าค่า C_1 จะเป็นการลู่เข้าสู่คำตอบอย่างรวดเร็วแต่ส่วนใหญ่คำตอบที่ได้จะเป็นคำตอบที่เหมาะสมเฉพาะที่ (Local) ไม่ใช่คำตอบที่เหมาะสมที่สุดของทั้งหมด (Global) ดังนั้นการสุ่มค่า C_1 และ C_2 ที่เหมาะสม จะขึ้นอยู่กับปัญหา โดยปกติมีค่าอยู่ระหว่าง 0 ถึง 4 แต่ที่นิยมใช้กันคือ $C_1 = C_2 = 2$



รูปที่ 3.3 แสดงลักษณะการปรับค่าความเร็วและตำแหน่งของพาร์ทิเคิล

พิจารณาการหาตำแหน่งใหม่ของแต่ละพาร์ทิเคิลจะประกอบไปด้วย 2 ขั้นตอนหลักคือ ขั้นตอนการหาความเร็วใหม่ และขั้นตอนการปรับตำแหน่งดังแสดงในรูปที่ 3.3 ซึ่งเอาความเร็วใหม่ไปบวกกับตำแหน่งเดิม โดยเริ่มจากสมการ (3.1) ในการคำนวณหาความเร็วที่จะต้องเปลี่ยนให้กับพาร์ทิเคิลก่อน แล้วจึงนำค่าความเร็วที่หาได้มารวมกับตำแหน่งเดิมของพาร์ทิเคิลนั้นซึ่งเป็นไปตามสมการ (3.2) ตามลำดับ สุดท้ายจะได้ค่าตำแหน่งใหม่ของทุกพาร์ทิเคิล จากนั้นเริ่มคำนวณหาค่าความเหมาะสมในรอบต่อไปจนกระทั่งได้ค่าเหมาะสมที่ดีที่สุดซึ่งก็คือตำแหน่งที่ดีที่สุดนั่นเอง

นอกจากนี้ในการปรับความเร็ว และตำแหน่งของแต่ละพาร์ทิเคิลนั้นเราต้องพิจารณาถึงค่าความเร็วสูงสุดที่เราจะอนุญาต โดยค่านี้เป็นค่าที่มีความสำคัญมาก เนื่องจากหากเรากำหนดค่านี้ให้มีค่ามากเกินไป ก็จะทำให้มีโอกาสมากที่จะทำให้พาร์ทิเคิลนั้นวิ่งผ่านจุดที่มีค่าความเหมาะสมมากที่สุดไป

```
For each particle /*กำหนดค่าเริ่มต้น*/
    Initialize particle
End
Do
    For each particle /*ประเมินค่าความเหมาะสมของแต่ละ particle*/
        Calculate fitness value
        If the fitness value is better than the best fitness value (pBest)
        in history then set current value as the new pBest
    End
    Choose the particle with the best fitness value of all the particles as the gBest
    For each particle /*ปรับข้อมูลและความเร็วในการมุ่งสู่เป้าหมาย*/
        Calculate particle velocity according equation (3.1)
        Update particle position according equation (3.2)
While maximum iterations or minimum error criteria is not attained
```

รูปที่ 3.4 แสดง Pseudo-code ของอัลกอริทึม PSO

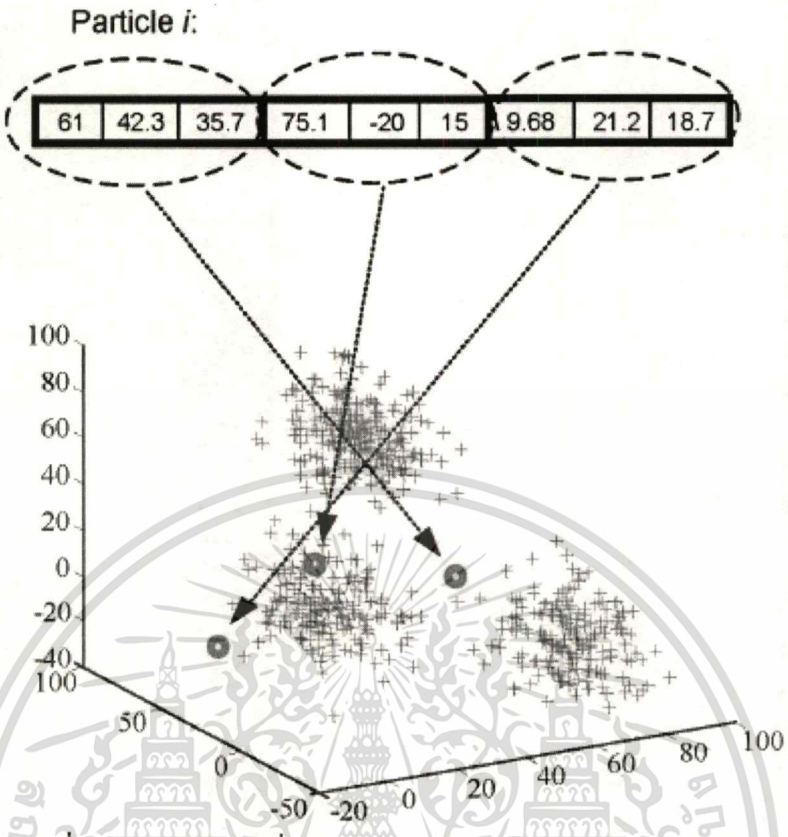
3.2.1 สรุปการทำงานอัลกอริทึม PSO

จากที่ได้กล่าวมาในส่วนการทำงานของพาร์ทิเคิลสวอมออปติไมเซชั่น เราสามารถสรุปการทำงานของอัลกอริทึม PSO ได้ดังนี้

3.2.1.1 กำหนดค่าเริ่มต้นต่าง ๆ ของแต่ละพาร์ทิเคิล

โดยทำการสร้างประชากรของพาร์ทิเคิลขึ้นมา พร้อมทั้งทำการสุ่มตำแหน่งและความเร็วของแต่ละพาร์ทิเคิล จากนั้นกำหนดค่า *pbest* ของตำแหน่งปัจจุบันสำหรับแต่ละพาร์ทิเคิล และหาค่า *pbest* ที่ดีที่สุดเพื่อกำหนดเป็น *gbest* โดยที่ตำแหน่งและค่าที่ดีที่สุดจะถูกเก็บไว้

จากรูปที่ 3.5 แสดงตัวอย่างการเริ่มต้นการสร้างประชากรในแต่ละพาร์ทิเคิล จากรูปนั้น กำหนดให้ n = 3 และ k = 3 นั่นคือพื้นที่ทั้งหมดเป็นแบบ 3 มิติ และต้องการจัดกลุ่มเป็น 3 กลุ่ม โดยพาร์ทิเคิลนี้ถูกแทนด้วยศูนย์กลางของคลัสเตอร์ดังนี้ [(61,42.3,35.7),(75.1,-20,15) และ (9.68,21.2,18.7)]



รูปที่ 3.5 ตัวอย่างการเริ่มต้นการสร้างประชากรใน 1 พาร์ทิเคิล

3.2.1.2 คำนวณหาตำแหน่งของแต่ละพาร์ทิเคิล

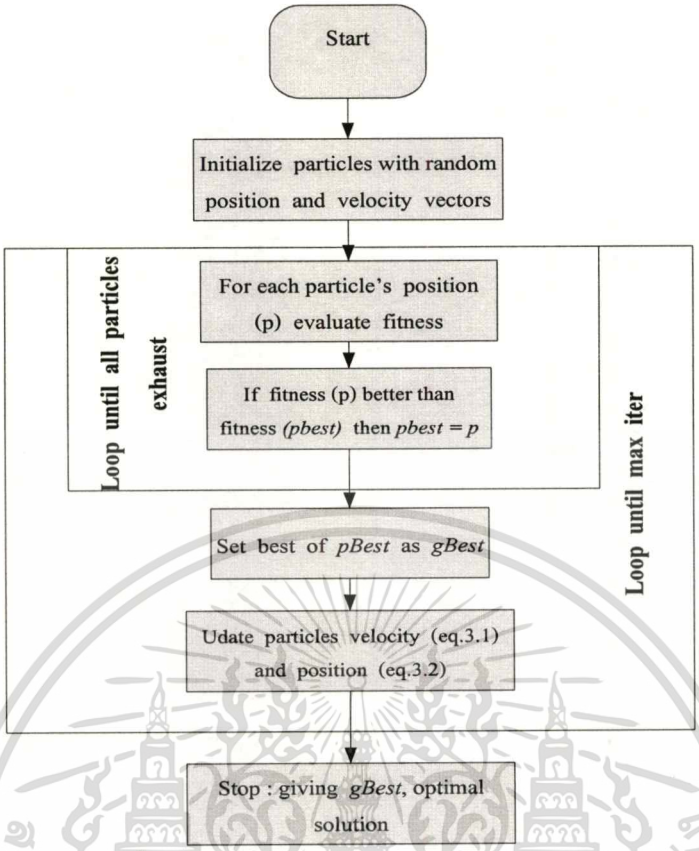
ค่าความเหมาะสมจะถูกคำนวณในแต่ละพาร์ทิเคิล ถ้าค่าที่ได้ดีกว่า *pbest* ปัจจุบันของพาร์ทิเคิล ค่า *pbest* นั้นจะถูกแทนด้วยค่าปัจจุบันนั้น และถ้าค่าที่ดีที่สุดของ *pbest* ดีกว่าค่า *gbest* ปัจจุบัน *gbest* จะถูกแทนด้วยค่าที่ดีที่สุดนั้น โดยตำแหน่งและค่าที่ดีที่สุดจะถูกเก็บไว้

3.2.1.3 ปรับปรุงค่าความเร็วและตำแหน่ง

ทำการเปลี่ยนแปลงค่าความเร็วและตำแหน่งปัจจุบันของแต่ละพาร์ทิเคิล โดยใช้สมการ (3.1) และ (3.2) ตามลำดับ

3.2.1.4 ตรวจสอบเงื่อนไขในการจบทำงาน

ตรวจสอบว่า จำนวนรอบในการทำงานปัจจุบันไปถึงจำนวนรอบในการทำงานสูงสุดที่กำหนดไว้ล่วงหน้าหรือยัง ถ้าใช่ก็จบการทำงานแต่ถ้าไม่ใช่ให้กลับไปทำที่ขั้นตอนที่ 2 ซึ่งขั้นตอนทั้งหมดสามารถแสดงเป็น Flow chart ได้ดังรูปที่ 3.6



รูปที่ 3.6 แสดง Flow Chart ของอัลกอริทึม PSO

PSO เป็นประโยชน์อย่างมากที่ใช้ในการค้นหาเช่นเดียวกับจีเนอติกและเป็นวิธีการค้นหาที่เข้าใกล้ตำแหน่งที่ดีที่สุดโดยใช้ *pbest* และ *gbest*

3.3 การจัดกลุ่มข้อมูลโดยใช้พาร์ทิเคิลสวอมออปติไมเซชั่น

เช่นเดียวกันกับอัลกอริทึมการจัดกลุ่มที่ใช้การแบ่งส่วนอื่น ๆ เป้าหมายการจัดกลุ่มของ PSO คือค้นหาศูนย์กลางที่เหมาะสมของคลัสเตอร์เพื่อทำให้ภายในคลัสเตอร์มีความคล้ายคลึงกันมากที่สุดและแตกต่างกันระหว่างคลัสเตอร์ ในการจัดกลุ่มนั้นพาร์ทิเคิลจะหมายถึงจำนวนของการจัดกลุ่มที่เป็นไปได้สำหรับพาร์ทิเคิลปัจจุบัน ดังนั้นหนึ่งพาร์ทิเคิลก็คือหนึ่งวิธีการที่เป็นไปได้สำหรับการจัดกลุ่มโดยหนึ่งพาร์ทิเคิลนั้นแทนด้วย N_c ซึ่งเป็นเวกเตอร์ศูนย์กลางของคลัสเตอร์ โดยแต่ละพาร์ทิเคิล x_i ถูกสร้างดังสมการนี้

$$x_i = \left(m_{i1},...,m_{ij},...m_{iN_c}\right) \tag{3.3}$$

โดย m_{ij} คือ เวกเตอร์ศูนย์กลางของคลัสเตอร์ j^{th} ของพาร์ทิเคิล i^{th} ในคลัสเตอร์ C_{ij}

N_c คือ จำนวนของคลัสเตอร์

3.3.1 การคำนวณค่าความเหมาะสม (Fitness Value)

จากที่กล่าวมาแล้วว่าประสบการณ์ของพาร์ทิเคิลเองหรือของพาร์ทิเคิลบริเวณใกล้เคียงทำให้พาร์ทิเคิลต้องเปลี่ยนแปลงเพื่อปรับปรุงตำแหน่งของศูนย์กลางคลัสเตอร์ในแต่ละการทำซ้ำ จาก Euclidean Distance จะมี Clustering Metric ที่ได้ทำการรวมผลต่างของระยะทางที่คำนวณได้จากแต่ละกลุ่ม สำหรับ k กลุ่มข้อมูล $C_1, C_2, ..., C_k$ เป็นดังนี้

$$U(C_1, C_2, ..., C_k) = \sum_{1}^{k} \sum_{j \in c_i} \left\| x_j - m_i \right\| \tag{3.4}$$

โดย U คือ Clustering Metric

C_i คือ กลุ่มข้อมูล i โดย i = 1,2,...,k

x_j คือ ข้อมูลสมาชิกของแต่ละกลุ่ม i

m_i คือ ศูนย์กลางของกลุ่มข้อมูลที่ i

ดังนั้นฟังก์ชั่นค่าความเหมาะสมของพาร์ทิเคิลสามารถหาได้จาก

$$f = \frac{1}{U} \tag{3.5}$$

ค่าความเหมาะสมหาได้จากสมการ (3.5) เพราะจะทำให้ได้ค่าความเหมาะสมมีค่ามากที่สุด เนื่องมาจากระยะห่างระหว่างศูนย์กลางของกลุ่มกับข้อมูลสมาชิกควรมีค่าห่างกันน้อยที่สุด

3.3.2 การจัดกลุ่มโดยใช้อัลกอริทึม PSO

ในส่วนนี้จะเสนออัลกอริทึมมาตรฐานของ PSO สำหรับการจัดกลุ่มข้อมูล โดยการใช้พื้นฐานการทำงานของ PSO นั้น ข้อมูลสามารถนำมาถูกจัดกลุ่มได้ ซึ่งอัลกอริทึมในการจัดกลุ่มโดยใช้พาร์ทิเคิลสวอมออปติไมเซชั่นมีการทำงานดังนี้

1. เริ่มต้นจากแต่ละพาร์ทิเคิลจะทำการสุ่มเลือก N_c ซึ่งเป็นศูนย์กลางกลุ่ม
2. สำหรับ $t = 1$ ถึง $t_{\max}$ ให้ทำดังนี้
   (a) ในแต่ละพาร์ทิเคิล i และแต่ละเวกเตอร์ข้อมูล x_i ให้ทำดังนี้
      i. คำนวณ Euclidean distance จาก $d(x_i, m)$ ไปยังศูนย์กลางทั้งหมดของกลุ่ม
      ii. กำหนด x_i ให้กับคลัสเตอร์โดยที่ $d(x_i, m) = \min_{\forall c=1,...,N_c} \{d(x_i, m)\}$
      iii. คำนวณค่าความเหมาะสมจากสมการที่ (3.5)
   (b) หาตำแหน่ง *pbest* ของแต่ละพาร์ทิเคิลและ *gbest*
   (c) เปลี่ยนค่าใหม่ของศูนย์กลางเวกเตอร์โดยการใช้สมการ (3.1) และ (3.2)

โดยที่ $t_{\max}$ คือ จำนวนมากที่สุดของการทำซ้ำ

การศึกษาถึงอัลกอริทึม PSO ที่ใช้ในการจัดกลุ่มโดยมีพื้นฐานแบบวิวัฒนาการโดยการจำลองการทำงานเหมือนกับการอยู่เป็นกลุ่มของสิ่งมีชีวิต เช่นเดียวกันอัลกอริทึมในการจัดกลุ่ม PSO จะมีการคำนวณหาค่าความเหมาะสมโดยการปรับปรุงค่าความเร็วและตำแหน่งเพื่อให้ได้ตำแหน่งของพาร์ทิเคิลที่ดีที่สุดออกมา ซึ่งตำแหน่งที่ดีที่สุดนี้จะถูกกำหนดให้เป็นตัวแทนศูนย์กลางของคลัสเตอร์ จากนั้นกำหนดข้อมูลที่เหลือให้กับศูนย์กลางเหล่านี้ โดยจะกำหนดข้อมูลให้กับศูนย์กลางที่อยู่ใกล้ที่สุด สำหรับเซตข้อมูลขนาดใหญ่ก็นิยมใช้อัลกอริทึม PSO เพราะสามารถนำไปสู่การค้นหาการจัดกลุ่มที่ดีที่สุด

ข้อได้เปรียบในการใช้อัลกอริทึมพาร์ทิเคิลสวอมออปติไมเซชั่นก็เพื่อเป็นการค้นหาเพื่อให้ได้การจัดกลุ่มที่ดีที่สุดถูกดำเนินการ โดยวิธีการค้นหาที่ใช้พื้นฐานของจำนวนประชากรนี้ช่วยในการลดผลกระทบจากเงื่อนไขเริ่มต้นของอัลกอริทึม k-Means โดยเฉพาะในเรื่องกลุ่มข้อมูลที่มีขนาดใหญ่ และสามารถจัดการกับการผิดพลาดที่เกิดจากข้อมูลที่เป็น Noise และ Outliers ซึ่งอัลกอริทึมของ k-Means ไม่สามารถที่จะจัดการได้ เพราะอัลกอริทึมของ PSO ทุกพาร์ทิเคิลจะเคลื่อนที่ผ่านทุกตำแหน่งที่เป็นไปได้ ดังนั้นข้อมูลที่เป็น Noise และ Outliers นั้นก็อาจเป็นตำแหน่งหนึ่งที่ถูกพิจารณาด้วยพาร์ทิเคิลใด ๆ ที่ไปค้นพบข้อมูลในตำแหน่งนั้น สำหรับเวลาในการหาตัวแทนของกลุ่มนั้นอาจเสียเวลาในการหาตำแหน่ง *pbest* และ *gbest* อาจใช้เวลานานกว่าอัลกอริทึม k-Means เพราะอัลกอริทึม k-Means จะใช้ค่าเฉลี่ยของกลุ่มเป็นตัวแทน แต่ค่าเฉลี่ยบางครั้งก็อาจไม่ใช่ตัวแทนที่ดีนัก เมื่อเทียบกับตัวแทนที่ได้มาจากอัลกอริทึม PSO

การจัดกลุ่มข้อมูลโดยใช้พาร์ทิเคิลสวอมออปติไมเซชั่นเป็นอัลกอริทึมที่ใช้พื้นฐานเดียวกันกับอัลกอริทึม PSO ที่มีการทำงานหลาย ๆ รอบจนกระทั่งเป็นไปตามหลักการที่เหมาะสม เพื่อทำให้ได้มาของตัวแทนที่ดีที่สุดของกลุ่มซึ่งอยู่ในตำแหน่งที่มีค่าความเหมาะสมมากที่สุด เมื่อเราได้ตำแหน่งที่ดีที่สุดมาเป็นตัวแทนของกลุ่มแล้ว สิ่งที่ตามมาก็คือการจัดกลุ่มข้อมูลที่มีประสิทธิภาพ เพราะข้อมูลที่เหลือจะถูกกำหนดให้กับตัวแทนที่ดีเหล่านี้

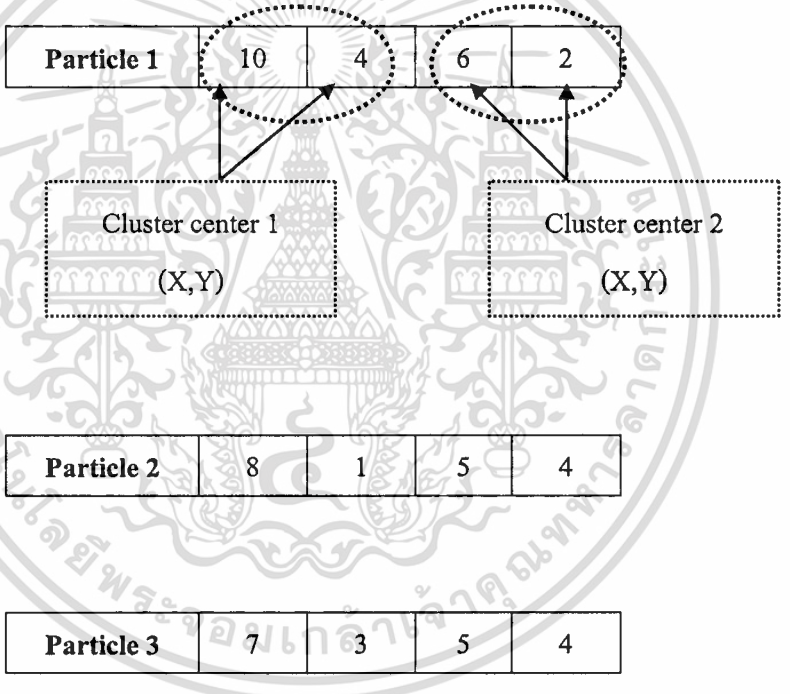
3.4 ตัวอย่างการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม PSO

สมมติว่ามีออบเจกต์ 6 ออบเจกต์ ซึ่งแต่ละออบเจกต์มี 2 แอตทริบิวต์ (X,Y) ซึ่งเป็นข้อมูล 2 มิติ ดังแสดงในตารางที่ 3.1 ให้ทำการจัดกลุ่มโดยใช้อัลกอริทึม PSO

ตารางที่ 3.1 ตัวอย่างข้อมูลที่จะใช้ในการจัดกลุ่มโดยใช้ PSO โดยให้ k = 2 และ Particle = 3

Object	X	Y
1	10	4
2	5	4
3	8	1
4	4	1
5	6	2
6	7	3

ขั้นที่ 1 แต่ละพาร์ทิเคิลทำการสร้างประชากรและสุ่มเลือกศูนย์กลางของคลัสเตอร์ได้ดังนี้



ขั้นที่ 2 รอบที่ t = 1 ในแต่ละพาร์ทิเคิลให้คำนวณหา Euclidean Distance ดังนี้

Particle 1 คำนวณหาระยะทางระหว่างข้อมูลแต่ละออบเจกต์กับศูนย์กลางกลุ่มตัวที่ 1 และกลุ่มที่ 2 ได้ดังนี้ จากศูนย์กลางกลุ่มที่ 1 เท่ากับ

10	4	6	2

ในกลุ่มที่ 1 มีศูนย์กลางเท่ากับ (10,4) คำนวณระยะทางระหว่างศูนย์กลางกับข้อมูลสมาชิกเป็นดังนี้

ออบเจกต์ตัวที่ 1. $\begin{bmatrix}10\\4\end{bmatrix}\begin{bmatrix}10\\4\end{bmatrix} = d_1 = \sqrt{(10-10)^2+(4-4)^2} = 0$

ออบเจกต์ตัวที่ 2. $\begin{bmatrix}10\\4\end{bmatrix}\begin{bmatrix}5\\4\end{bmatrix} = d_2 = \sqrt{(10-5)^2+(4-4)^2} = 5$

ออบเจกต์ตัวที่ 3. $\begin{bmatrix}10\\4\end{bmatrix}\begin{bmatrix}8\\1\end{bmatrix} = d_3 = \sqrt{(10-8)^2+(4-1)^2} = 3.61$

ออบเจกต์ตัวที่ 4. $\begin{bmatrix}10\\4\end{bmatrix}\begin{bmatrix}4\\1\end{bmatrix} = d_4 = \sqrt{(10-4)^2+(4-1)^2} = 6.71$

ออบเจกต์ตัวที่ 5. $\begin{bmatrix}10\\4\end{bmatrix}\begin{bmatrix}6\\2\end{bmatrix} = d_5 = \sqrt{(10-6)^2+(4-2)^2} = 4.47$

ออบเจกต์ตัวที่ 6. $\begin{bmatrix}10\\4\end{bmatrix}\begin{bmatrix}7\\3\end{bmatrix} = d_6 = \sqrt{(10-7)^2+(4-3)^2} = 3.16$

ในกลุ่มที่ 2 มีศูนย์กลางเท่ากับ (6,2) คำนวณระยะทางระหว่างศูนย์กลางกับข้อมูลสมาชิกเป็นดังนี้

ออบเจกต์ตัวที่ 1. $\begin{bmatrix}6\\2\end{bmatrix}\begin{bmatrix}10\\4\end{bmatrix} = d_1 = \sqrt{(6-10)^2+(2-4)^2} = 4.47$

ออบเจกต์ตัวที่ 2. $\begin{bmatrix}6\\2\end{bmatrix}\begin{bmatrix}5\\4\end{bmatrix} = d_2 = \sqrt{(6-10)^2+(2-4)^2} = 2.24$

ออบเจกต์ตัวที่ 3. $\begin{bmatrix}6\\2\end{bmatrix}\begin{bmatrix}8\\1\end{bmatrix} = d_3 = \sqrt{(6-8)^2+(2-1)^2} = 2.24$

ออบเจกต์ตัวที่ 4. $\begin{bmatrix}6\\2\end{bmatrix}\begin{bmatrix}4\\1\end{bmatrix} = d_5 = \sqrt{(6-4)^2+(2-1)^2} = 2.24$

ออบเจกต์ตัวที่ 5. $\begin{bmatrix}6\\2\end{bmatrix}\begin{bmatrix}6\\2\end{bmatrix} = d_4 = \sqrt{(6-6)^2+(2-2)^2} = 0$

ออบเจกต์ตัวที่ 6. $\begin{bmatrix}6\\2\end{bmatrix}\begin{bmatrix}7\\3\end{bmatrix} = d_6 = \sqrt{(6-7)^2+(5-3)^2} = 1.41$

ทำการจัดกลุ่มข้อมูล หากสมาชิกใกล้กับศูนย์กลางกลุ่มใดก็จะนำไปรวมกับกลุ่มนั้น ผลลัพธ์จากการคำนวณได้การจัดกลุ่มข้อมูลดังนี้

Particle 1

กลุ่มที่ 1 : (10,4)

กลุ่มที่ 2 : (5,4),(8,1),(4,1)(6,2),(7,3)

ผลต่างระยะทางรวมจากแต่ละกลุ่มข้อมูล (Euclidean Distance) เท่ากับ 0 และ 8.13 ตามลำดับ

ดังนั้นคำนวณค่า Clustering Metric จะได้ U = 0+8.13 = 8.13

Particle 2 มีศูนย์กลางกลุ่มดังนี้

8	1	5	4

ทำการคำนวณเช่นเดียวกับ particle 1 จะได้ผลลัพธ์ดังนี้

กลุ่มที่ 1 : (10,4),(8,1),(6,2),(7,3)

กลุ่มที่ 2 : (5,4),(4,1)

ผลต่างระยะทางรวมจากแต่ละกลุ่มข้อมูล (Euclidean Distance) เท่ากับ 8.09 และ 3.16 ตามลำดับ

ดังนั้นคำนวณค่า Clustering Metric จะได้ $U = 8.09 + 3.16 = 11.25$

Particle 3 มีศูนย์กลางกลุ่มดังนี้

7	3	5	4

ทำการคำนวณเช่นเดียวกับ particle 1 จะได้ผลลัพธ์ดังนี้

กลุ่มที่ 1 : (10,4),(8,1),(6,2),(7,3)

กลุ่มที่ 2 : (5,4),(4,1)

ผลต่างระยะทางรวมจากแต่ละกลุ่มข้อมูล (Euclidean Distance) เท่ากับ 6.81 และ3.16 ตามลำดั

ดังนั้นคำนวณค่า Clustering Metric จะได้ $U = 6.81 + 3.16 = 9.97$

ขั้นที่ 3 รอบที่ t = 1 คำนวณหาค่าความเหมาะสม (Fitness Value) ในแต่ละพาร์ทิเคิล เพื่อหา *pbest* และ *gbest* โดยใช้สมการ (3.4)

Particle 1 $f = 1/U = 1/8.13 = 0.123$ #

Particle 2 $f = 1/U = 1/11.25 = 0.089$

Particle 3 $f = 1/U = 1/9.97 = 0.100$

ดังนั้นในรอบแรกค่า *pbest* จะเป็นตำแหน่งศูนย์กลางเดิมของแต่ละพาร์ทิเคิล สำหรับ *gbest* จะได้ *gbest* คือตำแหน่งของ **Particle 1** เพราะว่ามีค่าความเหมาะสมสูงที่สุดจากพาร์ทิเคิลทั้งหมด

ขั้นที่ 4 รอบที่ t = 1 ทำการปรับปรุงตำแหน่งของศูนย์กลางของคลัสเตอร์โดยใช้สมการการเปลี่ยนความเร็วและตำแหน่ง ดังสมการที่ (3.1) และ (3.2) ตามลำดับ

โดยกำหนดให้

$v_{i,j}(1) = 0, c_1, c_2 = 2, rand_1 = 0.25, rand_2 = 0.05$

Particle 1 รอบที่ t =1

หาค่าความเร็วได้จาก

$$v(t+1) = v(t) + C_1 rand_1(pbest(t) - present(t)) + C_2 rand_2(gbest(t) - present(t))$$

$$v(2) = (0) + (2)(0.25)\left(\begin{bmatrix}10\\4\\6\\2\end{bmatrix} - \begin{bmatrix}10\\4\\6\\2\end{bmatrix}\right) + (2)(0.05)\left(\begin{bmatrix}10\\4\\6\\2\end{bmatrix} - \begin{bmatrix}10\\4\\6\\2\end{bmatrix}\right) = 0$$

หาตำแหน่งใหม่ได้จาก

$$present(t+1) = present(t) + v(t+1)$$

$$x_1(2) = \left(\begin{bmatrix}10\\4\\6\\2\end{bmatrix} + [0]\right) = \begin{bmatrix}10\\4\\6\\2\end{bmatrix}$$

ดังนั้นตำแหน่งใหม่ของ Particle 1 คือ

10	4	6	2

Particle 2 รอบที่ t =1 สามารถหาความเร็วและตำแหน่งใหม่ได้เช่นเดียวกับ Particle 1

หาค่าความเร็วได้จาก

$$v(t+1) = v(t) + C_1 rand_1(pbest(t) - present(t)) + C_2 rand_2(gbest(t) - present(t))$$

$$v(2) = (0) + (2)(0.25)\left(\begin{bmatrix}8\\1\\5\\4\end{bmatrix} - \begin{bmatrix}8\\1\\5\\4\end{bmatrix}\right) + (2)(0.05)\left(\begin{bmatrix}10\\4\\6\\2\end{bmatrix} - \begin{bmatrix}8\\1\\5\\4\end{bmatrix}\right) = 0.1\begin{bmatrix}2\\3\\1\\-2\end{bmatrix} = \begin{bmatrix}0.2\\0.3\\0.1\\-0.2\end{bmatrix}$$

หาตำแหน่งใหม่ได้จาก

$$present(t+1) = present(t) + v(t+1)$$

$$x_2(2) = \left(\begin{bmatrix}8\\1\\5\\4\end{bmatrix} + \begin{bmatrix}0.2\\0.3\\0.1\\-0.2\end{bmatrix}\right) = \begin{bmatrix}8.2\\1.3\\5.1\\3.8\end{bmatrix}$$

ดังนั้นตำแหน่งใหม่ของ Particle 2 คือ

8.2	1.3	5.1	3.8

Particle 3 รอบที่ t =1 สามารถหาความเร็วและตำแหน่งใหม่ได้เช่นเดียวกับ Particle 1

หาค่าความเร็วได้จาก

$$v(t+1) = v(t) + C_1 rand_1(pbest(t) - present(t)) + C_2 rand_2(gbest(t) - present(t))$$

$$v(2) = (0) + (2)(0.25)\left(\begin{bmatrix}7\\3\\5\\4\end{bmatrix} - \begin{bmatrix}7\\3\\5\\4\end{bmatrix}\right) + (2)(0.05)\left(\begin{bmatrix}10\\4\\6\\2\end{bmatrix} - \begin{bmatrix}7\\3\\5\\4\end{bmatrix}\right) = 0.1\begin{bmatrix}3\\1\\1\\-2\end{bmatrix} = \begin{bmatrix}0.3\\0.1\\0.1\\-0.2\end{bmatrix}$$

หาตำแหน่งใหม่ได้จาก

$$present(t+1) = present(t) + v(t+1)$$

ดังนั้นตำแหน่งใหม่ของ Particle 3 คือ

$$x_3(2) = \left(\begin{bmatrix}7\\3\\5\\4\end{bmatrix} + \begin{bmatrix}0.3\\0.1\\0.1\\-0.2\end{bmatrix}\right) = \begin{bmatrix}7.3\\3.1\\5.1\\3.8\end{bmatrix}$$

7.3	3.1	5.1	3.8

เมื่อได้ตำแหน่งศูนย์กลางกลุ่มใหม่ของแต่ละพาร์ทิเคิลแล้ว ให้กลับไปทำซ้ำในขั้นตอนที่ 2 เพื่อหา Distance จากนั้นทำการจัดกลุ่มข้อมูลให้กับศูนย์กลางใหม่นี้ หากสมาชิกใกล้กับศูนย์กลางกลุ่มใดก็จะนำไปรวมกับกลุ่มนั้น ผลลัพธ์จากการคำนวณได้การจัดกลุ่มข้อมูลดังนี้

ขั้นที่ 2 รอบที่ t = 2

Particle 1 มีศูนย์กลางเป็นตัวเดิม

10	4	6	2

กลุ่มที่ 1 : (10,4)

กลุ่มที่ 2 : (5,4),(8,1),(4,1)(6,2),(7,3)

ผลต่างระยะทางรวมจากแต่ละกลุ่มข้อมูล (Euclidean Distance) เท่ากับ 0 และ 8.13 ตามลำดับ ดังนั้นคำนวณค่า Clustering Metric จะได้ $U = 0+8.13 = 8.13$

Particle 2 มีศูนย์กลางกลุ่มใหม่ดังนี้

8.2	1.3	5.1	3.8

กลุ่มที่ 1 : (10,4),(8,1),

กลุ่มที่ 2 : (5,4),(4,1),(6,2),(7,3)

ผลต่างระยะทางรวมจากแต่ละกลุ่มข้อมูล (Euclidean Distance) เท่ากับ 3.6 และ 7.29 ตามลำดับ ดังนั้นคำนวณค่า Clustering Metric จะได้ $U = 3.6 + 7.29 = 10.89$

Particle 3 มีศูนย์กลางกลุ่มใหม่ดังนี้

7.3	3.1	5.1	3.8

กลุ่มที่ 1 : (10,4),(8,1),(6,2),(7,3)

กลุ่มที่ 2 : (5,4),(4,1)

ผลต่างระยะทางรวมจากแต่ละกลุ่มข้อมูล (Euclidean Distance) เท่ากับ 7.08 และ3.22 ตามลำดับ ดังนั้นคำนวณค่า Clustering Metric จะได้ $U = 7.08 + 3.22 = 10.3$

ขั้นที่ 3 รอบที่ t = 2 คำนวณหาค่าความเหมาะสม (Fitness Value) ในแต่ละพาร์ทิเคิล เพื่อหา *pbest* และ *gbest* ตัวใหม่

Particle 1 $f = 1/U = 1/8.13 = 0.123$ #

Particle 2 $f = 1/U = 1/10.89 = 0.092$

Particle 3 $f = 1/U = 1/10.3 = 0.097$

ดังนั้นในรอบที่ t = 2 นี้ค่า *pbest* ของ Particle 1 และ Particle 3 จะเป็นตำแหน่งศูนย์กลางเดิมของแต่ละพาร์ทิเคิล (ตำแหน่งเดิมในรอบที่ t=1) เพราะค่าความเหมาะสมใหม่มีค่าน้อยลง ดังนั้นจะไม่เปลี่ยนตำแหน่งไป ส่วน *pbest* ของ Particle 2 มีการเปลี่ยนตำแหน่งเพราะค่าความเหมาะสมที่คำนวณได้มีค่ามากกว่าเดิม และสำหรับ *gbest* จะได้ *gbest* คือตำแหน่งของ **Particle 1** เพราะว่ามีค่าความเหมาะสมสูงที่สุดจากพาร์ทิเคิลทั้งหมด

สรุปตำแหน่ง *pbest* และ *gbest* สำหรับแต่ละพาร์ทิเคิลใน **รอบที่ t = 1** ได้ดังนี้

pbest ของ Particle 1 คือ

10	4	6	2

pbest ของ Particle 2 คือ

8.2	1.3	5.1	3.8

pbest ของ Particle 2 คือ

7	3	5	4

gbest ของ Particle ทั้งหมด คือ

10	4	6	2

ขั้นที่ 4 รอบที่ t = 2 ทำการปรับปรุงค่าความเร็วและตำแหน่งของศูนย์กลางของคลัสเตอร์โดยใช้สมการการเปลี่ยนความเร็วและตำแหน่ง

กำหนดให้

$$v_{i,j}(2) = 0, c_1, c_2 = 2, rand_1 = 0.3, rand_2 = 0.5$$

Particle 1 รอบที่ t = 2

หาค่าความเร็วได้จาก

$$v(t+1) = v(t) + C_1 rand_1(pbest(t) - present(t)) + C_2 rand_2(gbest(t) - present(t))$$

$$v(3) = (0) + (2)(0.3)\left(\begin{bmatrix}10\\4\\6\\2\end{bmatrix} - \begin{bmatrix}10\\4\\6\\2\end{bmatrix}\right) + (2)(0.5)\left(\begin{bmatrix}10\\4\\6\\2\end{bmatrix} - \begin{bmatrix}10\\4\\6\\2\end{bmatrix}\right) = 0$$

หาตำแหน่งใหม่ได้จาก

$$present(t+1) = present(t) + v(t+1)$$

$$x_1(3) = \left(\begin{bmatrix}10\\4\\6\\2\end{bmatrix} + [0]\right) = \begin{bmatrix}10\\4\\6\\2\end{bmatrix}$$

ดังนั้นตำแหน่งใหม่ของ Particle 1 คือ

10	4	6	2

Particle 2 รอบที่ t =2 สามารถหาความเร็วและตำแหน่งใหม่ จากรอบที่แล้วจะค่า $v(2) = \begin{bmatrix} 0.2 \\ 0.3 \\ 0.1 \\ -0.2 \end{bmatrix}$

หาค่าความเร็วได้จาก

$$v(t+1) = v(t) + C_1 rand_1(pbest(t) - present(t)) + C_2 rand_2(gbest(t) - present(t))$$

$$v(3) = \begin{bmatrix} 0.2 \\ 0.3 \\ 0.1 \\ -0.2 \end{bmatrix} + (2)(0.3)\left(\begin{bmatrix} 8.2 \\ 1.3 \\ 5.1 \\ 3.8 \end{bmatrix} - \begin{bmatrix} 8.2 \\ 1.3 \\ 5.1 \\ 3.8 \end{bmatrix}\right) + (2)(0.5)\left(\begin{bmatrix} 10 \\ 4 \\ 6 \\ 2 \end{bmatrix} - \begin{bmatrix} 8.2 \\ 1.3 \\ 5.1 \\ 3.8 \end{bmatrix}\right) = \begin{bmatrix} 0.2 \\ 0.3 \\ 0.1 \\ -0.2 \end{bmatrix} + \begin{bmatrix} 1.8 \\ 2.7 \\ 0.9 \\ -1.8 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 1 \\ -2 \end{bmatrix}$$

หาตำแหน่งใหม่ได้จาก

$$present(t+1) = present(t) + v(t+1)$$

$$x_2(3) = \left(\begin{bmatrix} 8.2 \\ 1.3 \\ 5.1 \\ 3.8 \end{bmatrix} + \begin{bmatrix} 2 \\ 3 \\ 1 \\ -2 \end{bmatrix}\right) = \begin{bmatrix} 8.4 \\ 4.3 \\ 6.1 \\ 1.8 \end{bmatrix}$$

ดังนั้นตำแหน่งใหม่ของ Particle 2 คือ

8.4	4.3	6.1	1.8

Particle 3 รอบที่ t =2 สามารถหาความเร็วและตำแหน่งใหม่ จากรอบที่แล้วจะค่า

หาค่าความเร็วได้จาก

$$v(t+1) = v(t) + C_1 rand_1(pbest(t) - present(t)) + C_2 rand_2(gbest(t) - present(t))$$

$$v(3) = \begin{bmatrix} 0.3 \\ 0.1 \\ 0.1 \\ -0.2 \end{bmatrix} + (2)(0.3)\left(\begin{bmatrix} 7 \\ 3 \\ 5 \\ 4 \end{bmatrix} - \begin{bmatrix} 7.3 \\ 3.1 \\ 5.1 \\ 3.8 \end{bmatrix}\right) + (2)(0.5)\left(\begin{bmatrix} 10 \\ 4 \\ 6 \\ 2 \end{bmatrix} - \begin{bmatrix} 7.3 \\ 3.1 \\ 5.1 \\ 3.8 \end{bmatrix}\right) = \begin{bmatrix} 2.52 \\ 0.84 \\ 0.84 \\ -1.68 \end{bmatrix}$$

หาตำแหน่งใหม่ได้จาก

$$present(t+1) = present(t) + v(t+1)$$

$$x_3(3) = \left(\begin{bmatrix} 7.3 \\ 3.1 \\ 5.1 \\ 3.8 \end{bmatrix} + \begin{bmatrix} 2.52 \\ 0.84 \\ 0.84 \\ -1.68 \end{bmatrix}\right) = \begin{bmatrix} 9.82 \\ 3.94 \\ 5.94 \\ 2.12 \end{bmatrix}$$

ดังนั้นตำแหน่งใหม่ของ Particle 3 คือ

9.82	3.94	5.94	2.12

เมื่อได้ตำแหน่งศูนย์กลางกลุ่มใหม่ของแต่ละพาร์ทิเคิลแล้ว ให้กลับไปทำซ้ำในขั้นตอนที่ 2 เพื่อหา Distance จากนั้นทำการจัดกลุ่มข้อมูลให้กับศูนย์กลางใหม่นี้ หากสมาชิกใกล้กับศูนย์กลางกลุ่มใดก็จะนำไปรวมกับกลุ่มนั้น ทำซ้ำจนกระทั่งถึงจำนวนรอบมากที่สุดที่กำหนดไว้ สุดท้ายจะได้การจัดกลุ่มที่ภายในคลัสเตอร์มีความคล้ายคลึงกันมากที่สุดและแตกต่างกันระหว่างคลัสเตอร์ ซึ่งสรุปได้ดังต่อไปนี้

pbest ของ Particle 1 คือ

10	4	6	2

pbest ของ Particle 2 คือ

8.4	4.3	6.1	1.8

pbest ของ Particle 3 คือ

9.82	3.94	5.94	2.12

gbest ของ Particle ทั้งหมด คือ

10	4	6	2

Particle 1 มีศูนย์กลางเป็นตัวเดิม

10	4	6	2

กลุ่มที่ 1 : (10,4)

กลุ่มที่ 2 : (5,4),(8,1),(4,1)(6,2),(7,3)

ผลต่างระยะทางรวมจากแต่ละกลุ่มข้อมูล (Euclidean Distance) เท่ากับ 0 และ 8.13 ตามลำดับ

ดังนั้นคำนวณค่า Clustering Metric จะได้ fitness = 1/8.13 = 0.123

Particle 2 มีศูนย์กลางกลุ่มใหม่ดังนี้

8.4	4.3	6.1	1.8

กลุ่มที่ 1 : (10,4)

กลุ่มที่ 2 : (5,4),(8,1),(4,1)(6,2),(7,3)

ผลต่างระยะทางรวมจากแต่ละกลุ่มข้อมูล (Euclidean Distance) เท่ากับ 0 และ 8.13 ตามลำดับ

ดังนั้นคำนวณค่า Clustering Metric จะได้ fitness = 1/10.11994 = 0.0988

Particle 3 มีศูนย์กลางกลุ่มใหม่ดังนี้

9.82	3.94	5.94	2.12

กลุ่มที่ 1 : (10,4),(8,1),(6,2),(7,3)

กลุ่มที่ 2 : (5,4),(4,1)

ผลต่างระยะทางรวมจากแต่ละกลุ่มข้อมูล (Euclidean Distance) เท่ากับ 7.08 และ3.22 ตามลำดับ

ดังนั้นคำนวณค่า Clustering Metric จะได้ fitness = 1/10.3 = 0.09787

จากผลลัพธ์ที่ได้เราจะเลือก particle ที่มีค่า fitness มากที่สุดซึ่งตัวนั้นจะเป็น *gbest* เพื่อใช้ในการจัดกลุ่มข้อมูลสุดท้ายที่เราได้ การที่เราเปลี่ยนตำแหน่งเมื่อได้ค่า *pbest* หรือ *gbest* ที่มากกว่าตำแหน่งเดิม เป็นเพราะว่าเราจะได้ย้ายตำแหน่งศูนย์กลางของกลุ่มไปยังตำแหน่งที่ดีกว่า จนสุดท้ายจะได้ตำแหน่งที่มีค่าความเหมาะสมสูงที่สุด เมื่อทำการจัดกลุ่มข้อมูลจะให้ข้อมูลที่มีประสิทธิภาพเนื่องจากมีศูนย์กลางของกลุ่มที่ดี ทั้งนี้ตำแหน่งใหม่ที่ได้มา อาจมาจากการที่พาร์ทิเคิลไปพบด้วยตนเองหรือพาร์ทิเคิลอื่นเป็นผู้พบ เมื่อพบตำแหน่งที่ดีพาร์ทิเคิลอื่น ๆ ที่เคลื่อนที่ไปทางที่ผิด ก็จะได้รับการเปลี่ยนตำแหน่งไปยังตำแหน่งที่มีค่าความเหมาะสมที่ดีกว่าเดิม

3.5 การผสมของอัลกอริทึมในการจัดกลุ่มระหว่าง PSO และ k-Means (The Hybrid PSO Clustering)

อัลกอริทึมของ PSO จะมีประสิทธิภาพถ้าให้เวลาในการทำงานที่เพียงพอและสามารถได้ผลลัพธ์จากการจัดกลุ่มจากเซตข้อมูลที่มีมิติต่ำได้ดีกว่าอัลกอริทึมของ k-Means อย่างไรก็ตามในเซตข้อมูลที่มีขนาดใหญ่ทำให้ในขั้นตอนในการค้นหา pbest และ gbest เป็นสาเหตุให้ PSO ต้องการการทำซ้ำที่มากกว่าเพื่อค้นหาวิธีที่ดีสุด ถึงแม้ PSO จะมีการทำงานในการจัดกลุ่มที่ดีกว่า k-Means แต่ในเรื่องของเวลาการทำงาน k-Means จะมีประสิทธิภาพกว่าในเซตข้อมูลขนาดใหญ่ ด้วยเหตุผลนี้อัลกอริทึมในการจัดกลุ่มของ PSO สามารถปรับปรุงให้มีประสิทธิภาพขึ้นได้โดยการผสมกันของอัลกอริทึม PSO และอัลกอริทึม k-Means (Hybrid PSO) โดยอัลกอริทึม Hybrid PSO จะแบ่งการทำงานเป็นสองส่วนคือส่วนของอัลกอริทึม PSO และ k-Means โดยเริ่มต้นการทำงานที่ส่วนของอัลกอริทึม PSO ซึ่งจะทำงานเป็นช่วงเวลาให้ไม่นานนักในการค้นหาตำแหน่งที่ดีที่สุดเพื่อหลีกเลี่ยงการกินเวลาในการคำนวณที่ซับซ้อน จากนั้นผลลัพธ์ที่จากส่วนของ PSO จะถูกใช้เป็นศูนย์กลางเริ่มต้นของอัลกอริทึม k-Means โดยที่อัลกอริทึม k-Means จะค้นหาผลลัพธ์สุดท้ายซึ่งเป็นผลลัพธ์ที่ดีที่สุดออกมา

โดยอัลกอริทึม Hybrid PSO สามารถสรุปการทำงานได้ดังนี้

1. เริ่มต้นที่การจัดกลุ่มโดยใช้อัลกอริทึม PSO จนกระทั่งสิ้นสุดการทำงานซึ่งเป็นตามหลักการที่เหมาะสม
2. จากนั้นผลลัพธ์ของการจัดกลุ่มที่ได้จากอัลกอริทึม PSO จะถูกใช้เป็นค่าศูนย์กลางเริ่มต้นในส่วนของอัลกอริทึม k-Means
3. เริ่มต้นการทำงานของอัลกอริทึม k-Means จนกระทั่งค่าศูนย์กลางที่ได้ไม่มีการเปลี่ยนแปลงหรือเป็นไปตามหลักการที่เหมาะสม

จากการศึกษาถึงอัลกอริทึม PSO ที่ใช้ในการจัดกลุ่มโดยมีพื้นฐานแบบวิวัฒนาการโดยการจำลองการทำงานเหมือนการรวมกลุ่มในการบินหาอาหารของนกหรือการว่ายน้ำเป็นฝูงของปลา โดยอัลกอริทึมจะมีการคำนวณค่าความเหมาะสมเพื่อให้ได้ตำแหน่งของพาร์ทิเคิลที่ดีที่สุดออกมา สำหรับเซตข้อมูลขนาดใหญ่ก็นิยมใช้อัลกอริทึม PSO เพราะสามารถนำไปสู่การค้นหาการจัดกลุ่มที่ดีที่สุดแต่ยังคงต้องการจำนวนการทำซ้ำที่มากและมีการคำนวณที่มากกว่าอัลกอริทึมของ k-Means สำหรับ k-Means จะเป็นการทำงานที่ง่ายและเร็วกว่า PSO โดยการทำงานของ k-Means นั้นเป็นการปรับปรุงการแบ่งกลุ่มข้อมูลโดยการเคลื่อนย้ายไปยังกลุ่มที่ซึ่งข้อมูลนั้นจะมีค่าใกล้กับศูนย์กลางกลุ่มนั้น จากนั้นมีการพัฒนาเทคนิคที่เป็นการผสมผสานของอัลกอริทึม k-Means กับ PSO ซึ่งเป็นการรวมข้อดีและหลีกเลี่ยงข้อเสียของทั้งสองอัลกอริทึมก็คือวิธีการค้นหาด้วยตนเองที่มีประสิทธิภาพและได้วิธีการที่ดีที่สุดหรือวิธีการที่ใกล้เคียงที่สุดในการค้นหาคำตอบที่ซับซ้อน รวมถึงการทำงานที่รวดเร็ว โดยผลลัพธ์ของอัลกอริทึม PSO จะเป็นใช้เป็นค่าเริ่มต้นในการทำงานของอัลกอริทึม k-Means ซึ่งวิธีการเหล่านี้ให้ผลลัพธ์ของการจัดกลุ่มข้อมูลที่มีประสิทธิภาพ

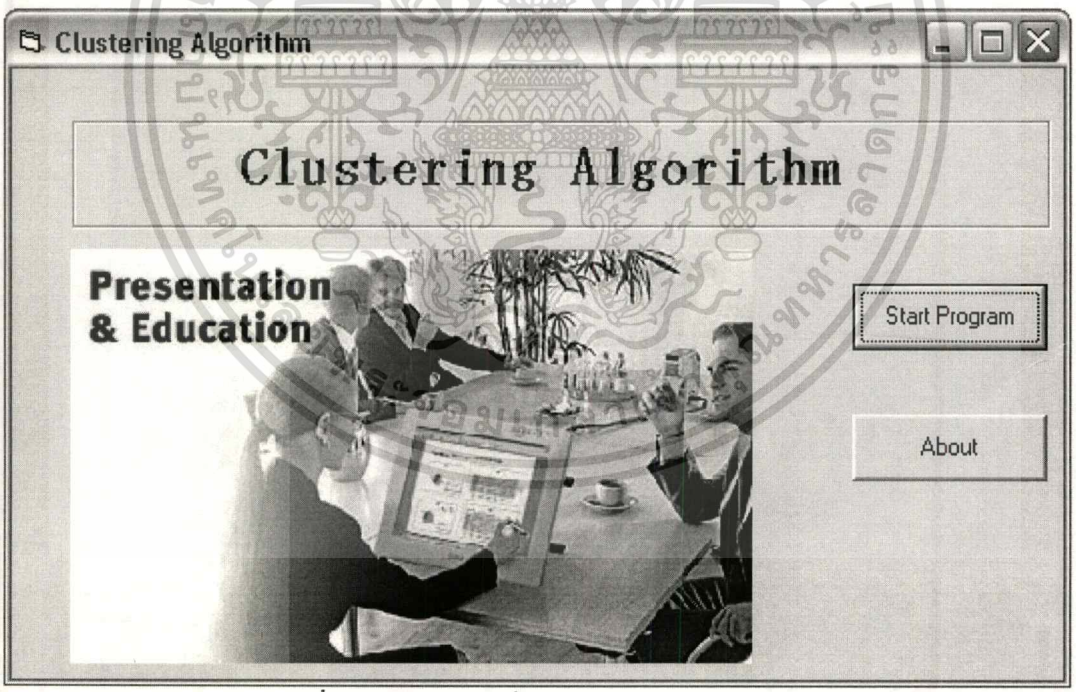
บทที่ 4
การสร้างและทดสอบระบบ

ในบทนี้จะกล่าวถึงรายละเอียดของการพัฒนาระบบการจัดกลุ่มข้อมูล ทำการเปรียบเทียบประสิทธิภาพของอัลกอริทึมที่ใช้ในการจัดกลุ่มข้อมูลระหว่างอัลกอริทึม k-Means และอัลกอริทึม Hybrid PSO ที่ใช้ในการจัดกลุ่มข้อมูล โดยจะอธิบายขั้นตอนต่าง ๆ ในการพัฒนาระบบดังต่อไปนี้

4.1 การทำงานของโปรแกรม

ลักษณะการทำงานของโปรแกรมนี้จะเป็นไปในลักษณะลำดับตามเมนูเพื่อให้ผู้ใช้งานสามารถเข้าใจ และสามารถใช้งานได้อย่างถูกต้อง และง่ายตามขั้นตอน โดยจะมีขั้นตอนการทำงานดังต่อไปนี้

4.1.1 หน้าที่ 1 : Clustering Algorithm



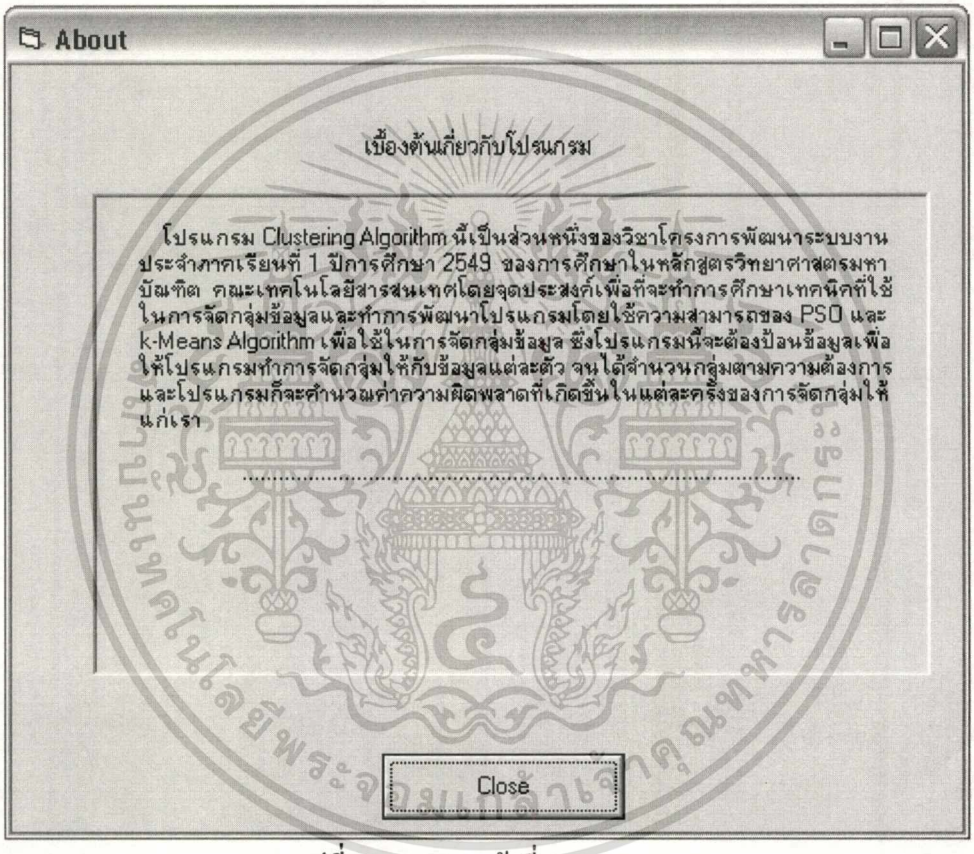
รูปที่ 4.1 แสดงหน้าที่ 1 : Clustering Algorithm

เมื่อเริ่มต้นการทำงานในส่วนนี้จะปรากฏหน้าจอดังรูปที่ 4.1 จะทำหน้าที่บอกให้ผู้ใช้งานทราบว่าขณะนี้โปรแกรมนี้ได้เปิดพร้อมที่จะใช้งานเรียบร้อยแล้ว นอกจากนี้ยังมี About เอาไว้ให้ผู้

ที่ยังไม่มีความเข้าใจในตัวโปรแกรมได้อ่านเพื่อทำความเข้าใจก่อนใช้งาน โดยมีสิ่งที่ผู้ใช้งานต้องพิจารณาคือ

- Start : ใช้สำหรับกดเมื่อผู้ใช้งานต้องการเริ่มใช้งานโปรแกรม
- About : ใช้สำหรับกดเมื่อผู้ใช้ต้องการทราบอธิบายภาพรวมการทำงานและลักษณะที่สำคัญของโปรแกรม

4.1.2 หน้าที่ 1.1 : หน้าจอ About



รูปที่ 4.2 แสดงหน้าที่ 1.1 : About

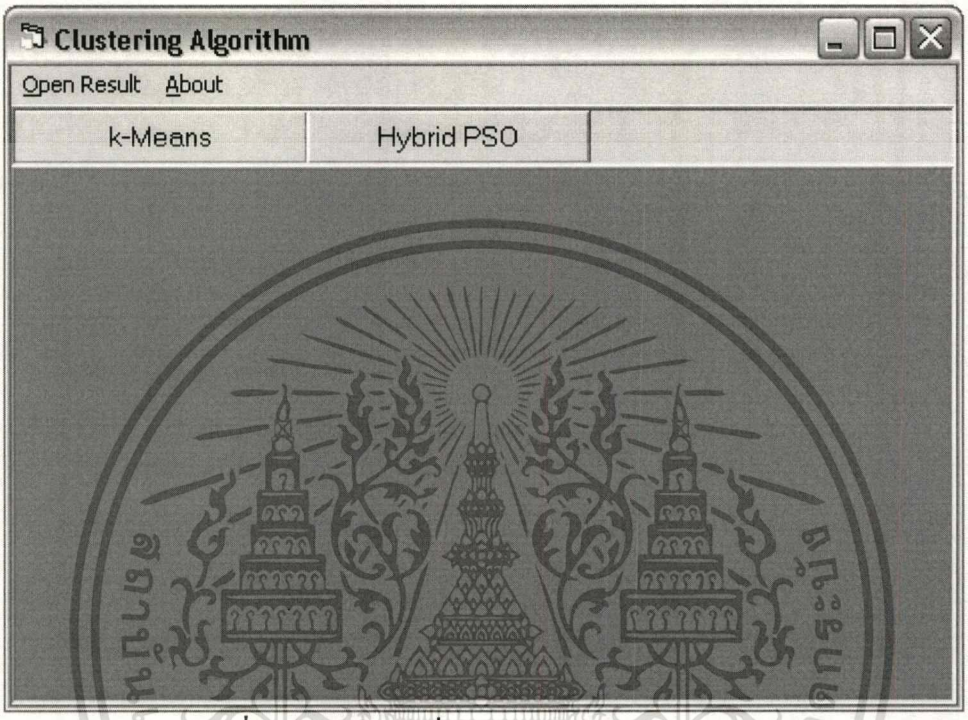
เป็นการแสดงข้อมูลเบื้องต้นเกี่ยวกับที่มา ความสามารถและภาพรวมในการทำงานของโปรแกรมที่ใช้ในการจัดกลุ่มข้อมูล เพื่อใช้สร้าง และเพิ่มความเข้าใจให้กับผู้ใช้งานที่ไม่เคยเห็นโปรแกรมมาก่อน หรือไม่มีพื้นฐานความรู้ที่เกี่ยวข้อง

- Close : ใช้สำหรับกดเมื่อผู้ใช้งานต้องการปิดหน้าต่าง About เมื่อทำการอ่านเรียบร้อยแล้ว

4.1.3 หน้าที่ 2 : หน้าจอหลักของโปรแกรม

การพัฒนาระบบงานนี้ แบ่งอัลกอริทึมที่ใช้ในการจัดกลุ่มออกเป็น 2 ส่วนหลักด้วยกันคือ

1. k-Means Algorithm
2. Hybrid PSO Algorithm



รูปที่ 4.3 แสดงหน้าที่ 2 : หน้าจอหลักของโปแกรม

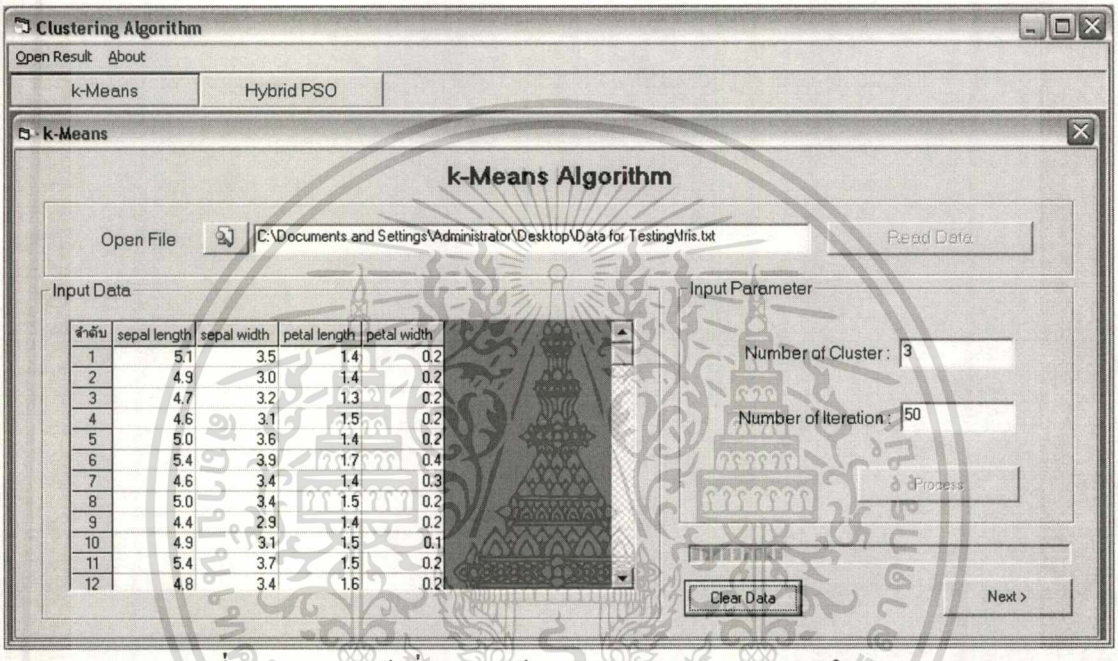
เมื่อเริ่มกดปุ่ม Start ที่หน้าจอแรกผู้ใช้งานก็จะพบกับหน้าจอหลักของโปรแกรม ผู้ใช้งานจะต้องทำหน้าที่วางแผนที่จะเลือกลักษณะการใช้งานโปรแกรม โดยทั่วไปแล้วโปรแกรมนี้จะแบ่งอัลกอริทึมออกเป็น 2 ส่วนหลักดังที่กล่าวไว้ โดยมีสิ่งที่ผู้ใช้งานต้องพิจารณาคือ

- k-Means : ใช้สำหรับกดเมื่อต้องการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม k-Means
- Hybrid PSO : ใช้สำหรับเมื่อต้องการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม Hybrid PSO
- Open Result : ไว้เปิดไฟล์ผลลัพธ์ excel ที่เก็บไว้ จะประกอบไปด้วยส่วนของการ Open File และ Exit
- About : ใช้สำหรับกดเมื่อผู้ใช้ต้องการทราบถึงภาพรวมการทำงานและลักษณะที่สำคัญของโปรแกรม

4.1.4 หน้าที่ 2.1 : หน้าจอการทำงานของอัลกอริทึม k-Means

การทำงานของอัลกอริทึม k-Means แบ่งการทำงานออกเป็น 3 ส่วนหลัก คือ

1. Open File : ส่วนที่ให้ผู้ใช้เปิดไฟล์ข้อมูลที่ต้องการนำมาจัดกลุ่ม
2. Input Data : ส่วนที่ใช้แสดงค่าของข้อมูลทั้งหมดที่จะนำมาจัดกลุ่ม
3. Input Parameter : เป็นส่วนที่ผู้ใช้จะต้องระบุค่าพารามิเตอร์ทั้งหมด ก่อนที่จะทำการประมวลผล



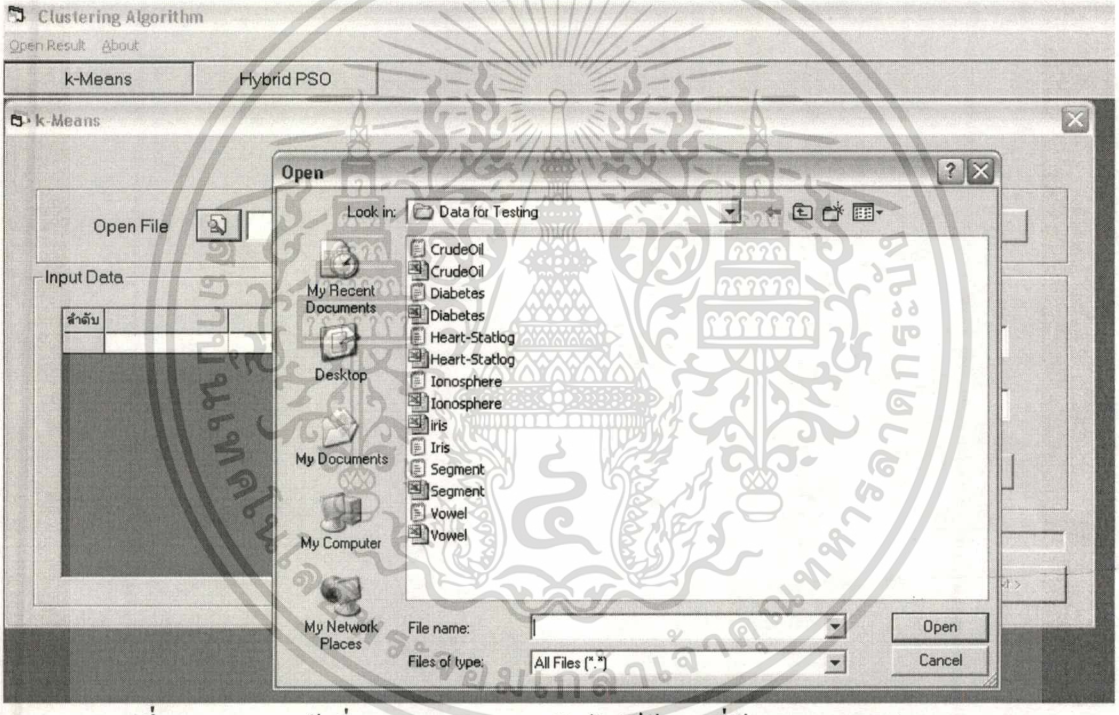
รูปที่ 4.4 แสดงหน้าที่ 2.1 : หน้าจอการทำงานของอัลกอริทึม k-Means

จากหน้าจอหลักของโปรแกรม เมื่อผู้ใช้เลือกอัลกอริทึม k-Means เพื่อใช้ประมวลผลในการจัดกลุ่มข้อมูล โดยมีสิ่งที่ผู้ใช้งานต้องพิจารณาคือ

- Open File : ใช้สำหรับกดเมื่อผู้ใช้งานต้องการเลือกไฟล์ข้อมูลที่จะใช้ในการจัดกลุ่มข้อมูล
- Read Data : ใช้สำหรับกดเมื่อผู้ใช้งานต้องการที่จะให้โปรแกรมโหลดนำค่าข้อมูลเข้า ที่ผู้ใช้ได้เลือกมาในการประมวลผล
- Input Data : ส่วนที่แสดงค่าของข้อมูลทั้งหมด ซึ่งข้อมูลเหล่านี้จะถูกใช้ประมวลผลในการจัดกลุ่มข้อมูล
- Number of Cluster : สำหรับผู้ใช้ระบุจำนวนกลุ่มข้อมูลที่ต้องการจัดกลุ่ม
- Number of Iteration : สำหรับผู้ใช้ระบุจำนวนรอบที่ต้องการให้จัดกลุ่ม

- Process : ใช้สำหรับกดเมื่อผู้ใช้ต้องการให้โปรแกรมทำการประมวลผลตามค่าพารามิเตอร์ที่รับเข้ามา แต่ถ้าใส่จำนวนพารามิเตอร์ไม่ครบ โปรแกรมจะทำการเตือนให้มีการใส่พารามิเตอร์ให้ครบ
- Clear Data : ใช้สำหรับกดเมื่อผู้ใช้ต้องการให้โปรแกรมเคลียร์ข้อมูลและค่าพารามิเตอร์ทั้งหมดที่จะใช้ในการจัดกลุ่มข้อมูล
- Next : ใช้สำหรับกดผู้ใช้ต้องการไปยังหน้าจอที่แสดงผลลัพธ์ทั้งหมด ที่ได้จากการจัดกลุ่ม

4.1.5 หน้าที่ 2.1.1 : หน้าจอแสดงการเปิดไฟล์ข้อมูล

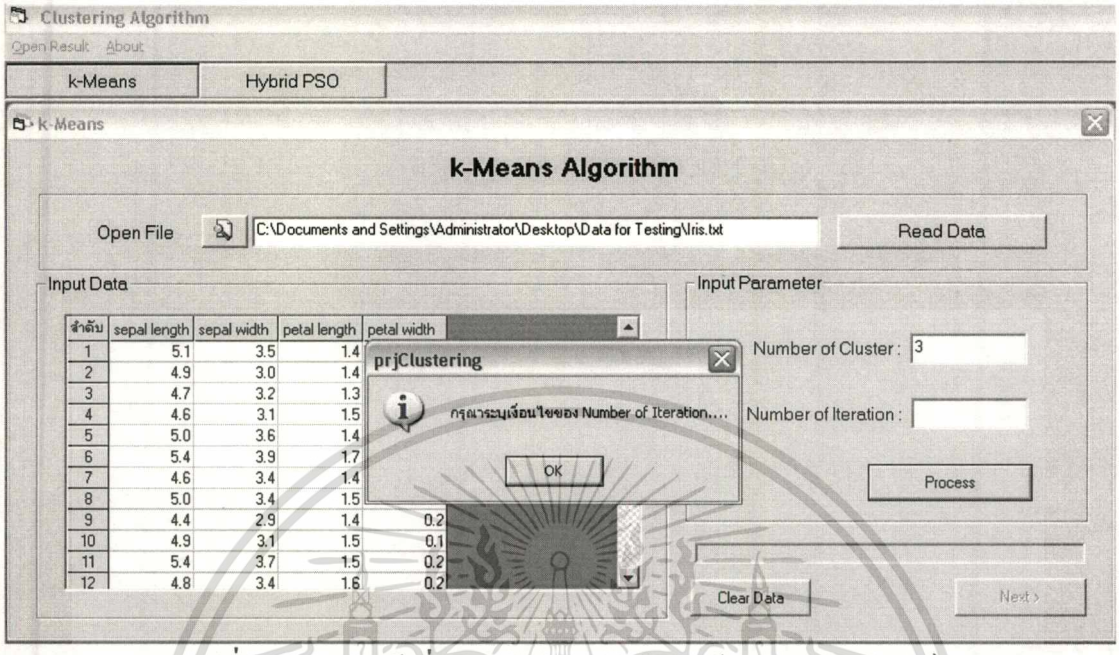


รูปที่ 4.5 แสดงหน้าที่ 2.1.1 : แสดงการเปิดไฟล์ข้อมูลที่ต้องการนำมาประมวลผล

หน้าจอนี้จะเกิดขึ้นเมื่อผู้ใช้กดปุ่ม Open File จากหน้าจอการทำงานของอัลกอริทึม k-Means โดยสิ่งที่ผู้ใช้งานต้องพิจารณาคือ

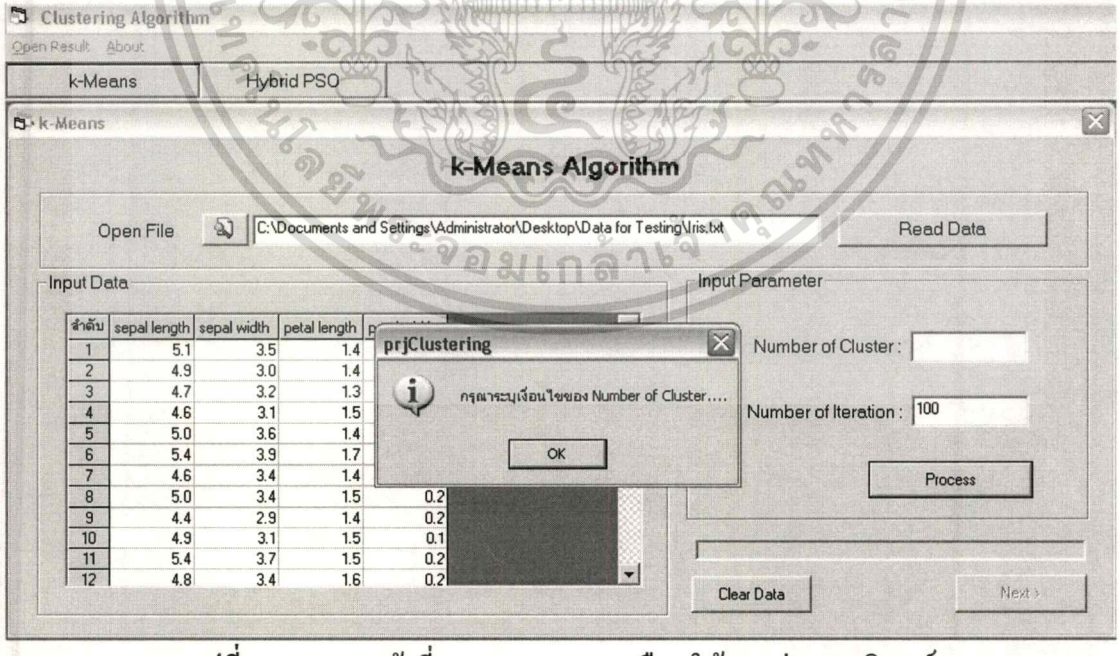
- Open : ใช้สำหรับกดเมื่อต้องการเลือกไฟล์ข้อมูลที่จะใช้ในการจัดกลุ่ม
- Cancel : ใช้สำหรับกดเมื่อผู้ใช้งานไม่ต้องการเลือกไฟล์

4.1.6 หน้าที่ 2.1.2 : หน้าจอแสดงการเตือนให้ระบุค่าจำนวนที่ต้องการทำซ้ำ



รูปที่ 4.6 แสดงหน้าที่ 2.1.2 : แสดงการเตือนให้ระบุค่าพารามิเตอร์

4.1.7 หน้าที่ 2.1.3 : หน้าจอแสดงการเตือนให้ระบุค่าจำนวนที่ต้องการจัดกลุ่ม

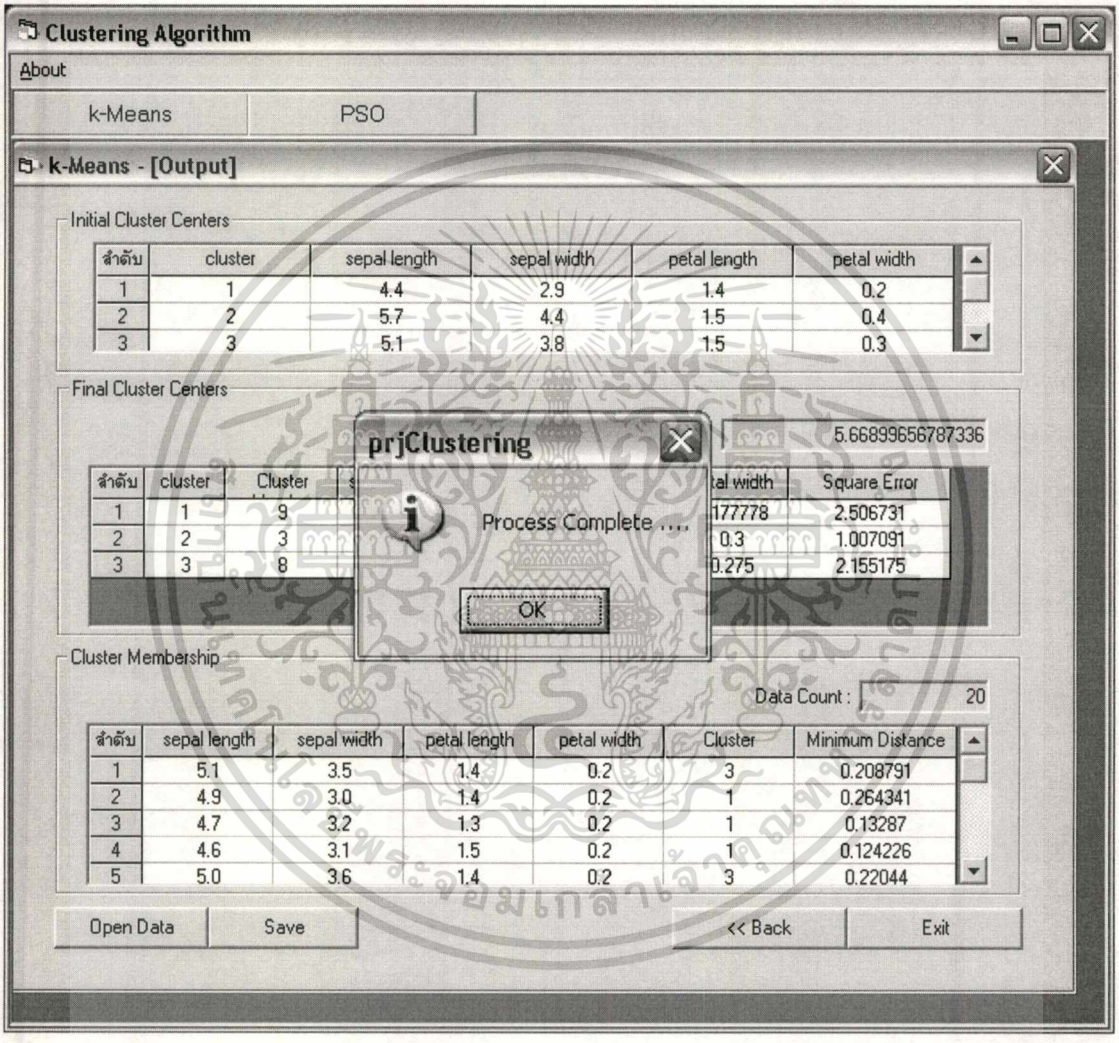


รูปที่ 4.7 แสดงหน้าที่ 2.1.3 : แสดงการเตือนให้ระบุค่าพารามิเตอร์

หน้าจอที่แสดงดังรูปที่ 4.6 และ 4.7 นี้จะเกิดขึ้นเมื่อผู้ใช้ระบุค่าพารามิเตอร์ไม่ครบ โดยจะขึ้นเตือนค่าที่ผู้ใช้ไม่ได้ทำการระบุ โดยสิ่งที่ผู้ใช้งานต้องพิจารณาคือ

- OK : ใช้สำหรับกดเมื่อผู้ใช้รับทราบว่าต้องทำการระบุค่าพารามิเตอร์ใดที่ขาดหาย

4.1.8 หน้าที่ 2.1.4 : หน้าจอแสดงเมื่ออัลกอริทึม k-Means ประมวลผลสำเร็จ



รูปที่ 4.8 แสดงหน้าที่ 2.1.4 : แสดงหน้าจอเมื่ออัลกอริทึม k-Means ทำงานสำเร็จ

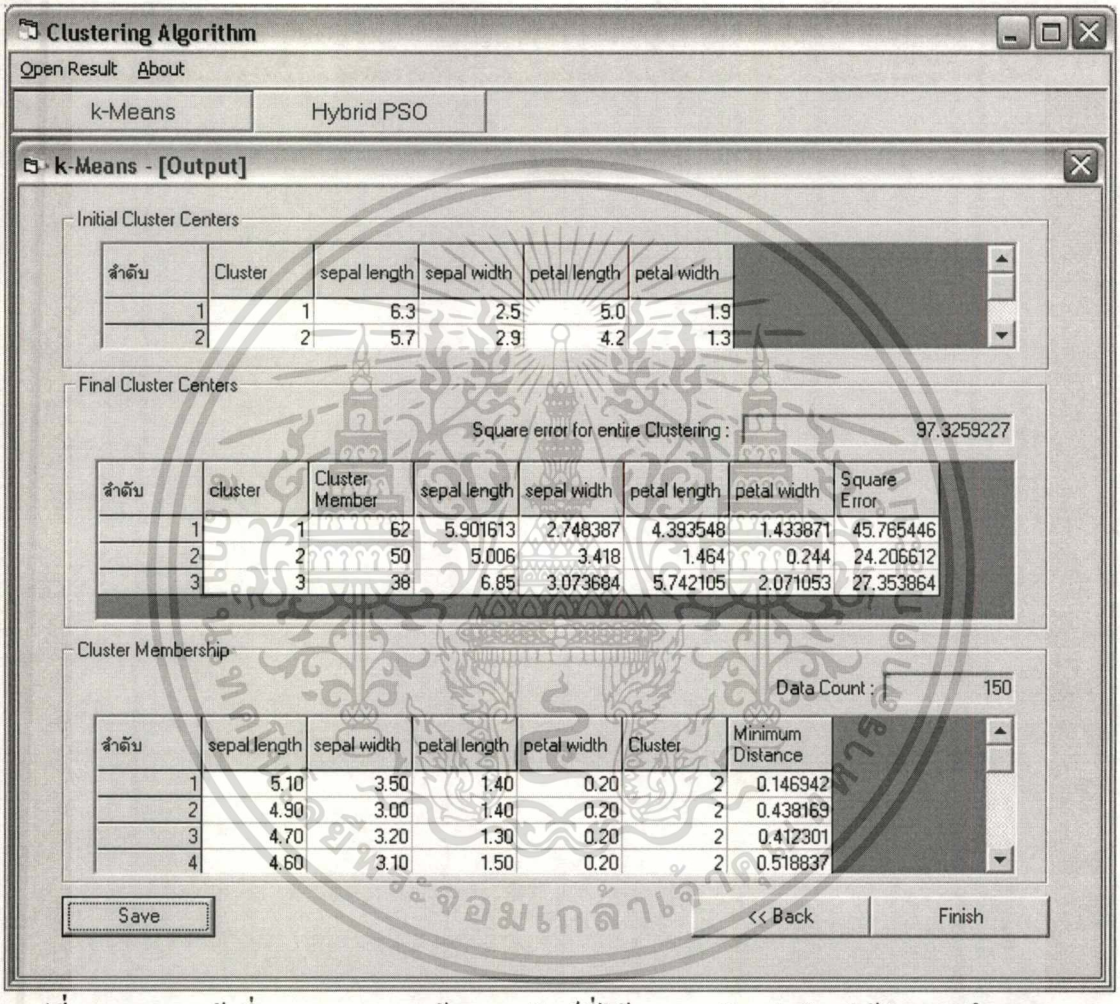
หน้าจอนี้จะเกิดขึ้นเมื่อโปรแกรมทำการประมวลผลจนสำเร็จจนได้ผลลัพธ์ในการจัดกลุ่มออกมา โดยจะมีข้อความแสดงให้ผู้ใช้ทราบ โดยสิ่งที่ผู้ใช้งานต้องพิจารณาคือ

- OK : แสดงให้ทราบถึงจำนวนรอบที่ใช้ในการประมวลผลและใช้สำหรับกดเมื่อผู้ใช้รับทราบว่าโปรแกรมทำงานสำเร็จ และต้องการที่จะดูผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูล

4.1.9 หน้าที่ 2.1.5 : หน้าจอแสดงผลลัพธ์ที่ได้จากการจัดกลุ่มโดยใช้อัลกอริทึม k-Means

โดยหน้าจอที่แสดงผลลัพธ์ของการจัดกลุ่มนี้สามารถแบ่งออกได้เป็น 3 ส่วนหลัก คือ

1. Initial Cluster Centers : ส่วนที่ใช้แสดงค่าศูนย์กลางเริ่มต้นของแต่ละคลัสเตอร์
2. Final Cluster Centers : ส่วนที่ใช้แสดงค่าศูนย์กลางสุดท้ายของแต่ละคลัสเตอร์
3. Cluster Membership : ส่วนที่ใช้แสดงผลลัพธ์ในการจัดกลุ่มของข้อมูลทั้งหมด



รูปที่ 4.9 แสดงหน้าที่ 2.1.5 : แสดงหน้าจอผลลัพธ์ที่ได้จากการจัดกลุ่มโดยใช้อัลกอริทึม k-Means

หน้าจอนี้จะเกิดขึ้นเมื่อโปรแกรมทำการประมวลผลข้อมูลผ่านอัลกอริทึม k-Means สำเร็จแล้ว โดยจะแสดงผลลัพธ์ที่ได้จากการจัดกลุ่มทั้งหมด โดยสิ่งที่ผู้ใช้งานต้องพิจารณาคือ

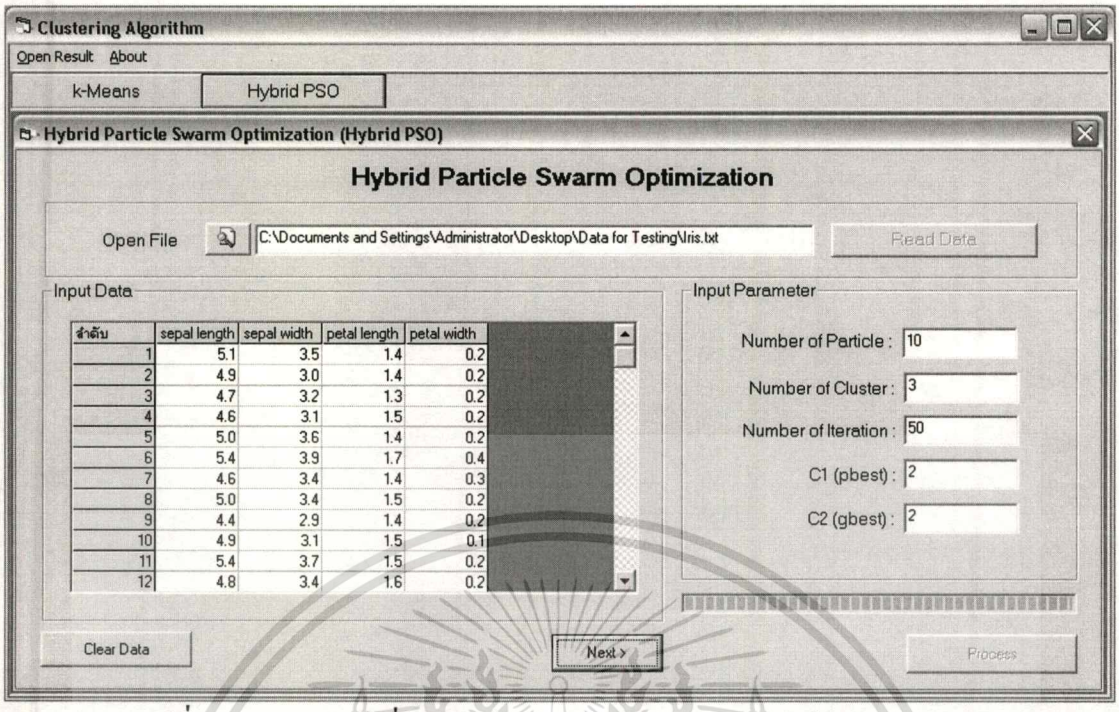
- Initial Cluster Centers : สำหรับใช้ในการแสดงค่าศูนย์กลางเริ่มต้นของแต่ละคลัสเตอร์ที่ใช้ในการจัดกลุ่ม ซึ่งผลลัพธ์ในส่วนนี้โปรแกรมจะทำการสุ่มข้อมูลตามค่าพารามิเตอร์ที่ได้จากการระบุโดยผู้ใช้

- Final Cluster Centers : สำหรับใช้ในการแสดงค่าศูนย์กลางสุดท้ายของแต่ละคลัสเตอร์ที่ได้จากการจัดกลุ่มข้อมูลและจำนวนข้อมูลในแต่ละกลุ่ม รวมทั้งแสดงค่า Square Error ของแต่ละกลุ่มข้อมูล
- Square Error for Entire Clustering : สำหรับใช้แสดงผลรวมของค่า Square Error ของทุกกลุ่มที่ได้จากการจัดกลุ่มข้อมูลในรอบสุดท้าย
- Cluster Membership : สำหรับใช้แสดงผลลัพธ์ในการจัดกลุ่มของข้อมูลทั้งหมด โดยแสดงว่าข้อมูลแต่ละแถวนั้นถูกจัดกลุ่มให้อยู่ในคลัสเตอร์ใด รวมทั้งยังบอกค่า Minimum distance ซึ่งเป็นค่าที่บอกว่าข้อมูลแต่ละแถวอยู่ห่างจากศูนย์กลางคลัสเตอร์เท่าไหร่
- Data Count : สำหรับใช้แสดงจำนวนของแถวหรือเรคอร์ดทั้งหมดในการจัดกลุ่ม
- Save : สำหรับใช้บันทึกผลลัพธ์ที่ได้จากการจัดกลุ่มซึ่งจะเป็นไฟล์ประเภท Excel
- Back : สำหรับใช้เมื่อต้องการกลับไปยังหน้าจอหลักของอัลกอริทึม k-Means
- Exit : สำหรับใช้เมื่อต้องการออกจากโปรแกรม

4.1.10 หน้าที่ 2.2 : หน้าจอการทำงานของอัลกอริทึม Hybrid PSO

การทำงานของอัลกอริทึม Hybrid PSO แบ่งการทำงานออกเป็น 3 ส่วนหลัก คือ

1. Open File : ส่วนที่ให้ผู้ใช้เปิดไฟล์ข้อมูลที่ต้องการนำมาจัดกลุ่ม
2. Input Data : ส่วนที่ใช้แสดงค่าของข้อมูลทั้งหมดที่จะนำมาจัดกลุ่ม
3. Input Parameter : เป็นส่วนที่ผู้ใช้จะต้องระบุค่าพารามิเตอร์ทั้งหมด ก่อนที่จะทำการประมวลผล



รูปที่ 4.10 แสดงหน้าที่ 2.2 : หน้าจอการทำงานของอัลกอริทึม Hybrid PSO

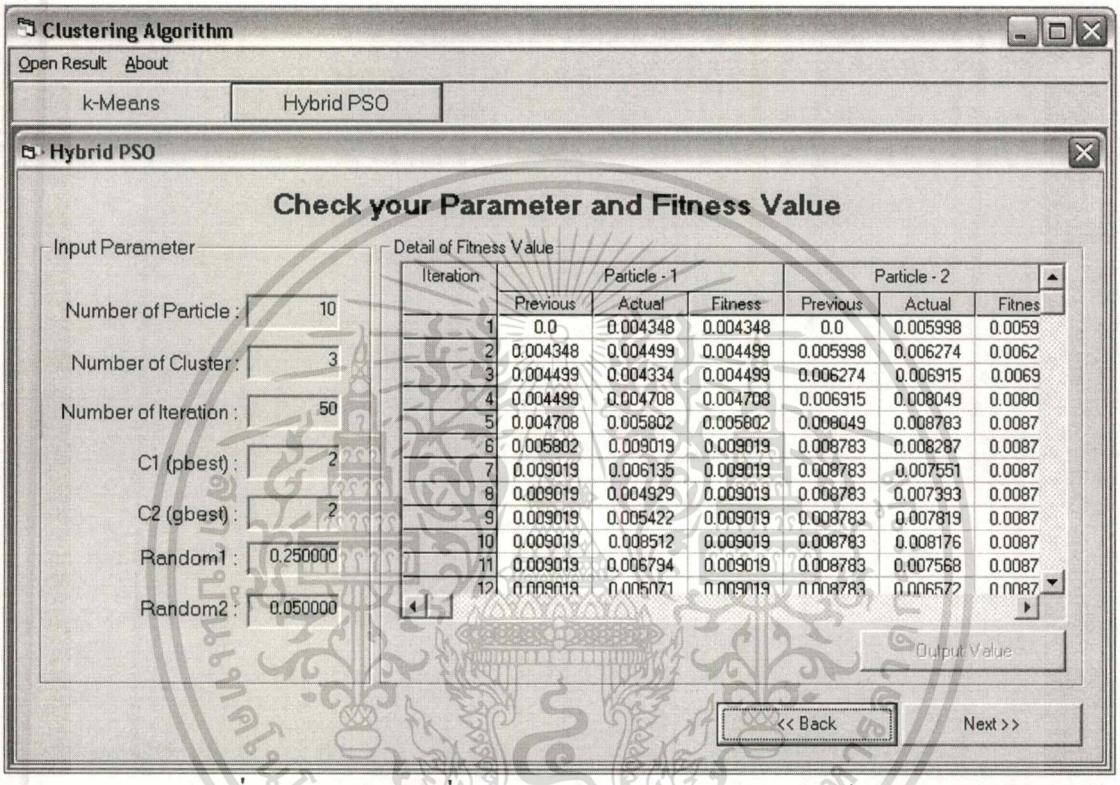
จากหน้าจอหลักของโปรแกรม เมื่อผู้ใช้เลือกอัลกอริทึม Hybrid PSO เพื่อใช้ประมวลผลในการจัดกลุ่มข้อมูล โดยมีสิ่งที่ผู้ใช้งานต้องพิจารณานอกเหนือจาก k-Means คือ

- Number of Particle : สำหรับผู้ใช้ระบุค่าจำนวนพาร์ทิเคิลที่ต้องการ
- Number of Cluster : สำหรับผู้ใช้ระบุจำนวนกลุ่มข้อมูลที่ต้องการจัดกลุ่ม
- Number of Iteration : สำหรับผู้ใช้ระบุจำนวนครั้งที่ต้องการให้พาร์ทิเคิลทำการปรับตำแหน่งของมัน
- Process : เมื่อผู้ใช้ต้องการให้โปรแกรมทำการประมวลผลตามค่าพารามิเตอร์ที่รับเข้ามา แต่ถ้าใส่จำนวนพารามิเตอร์ไม่ครบ โปรแกรมจะทำการเตือนให้มีการใส่พารามิเตอร์ให้ครบ
- Clear Data : เมื่อผู้ใช้ต้องการให้โปรแกรมเคลียร์ข้อมูลและค่าพารามิเตอร์ทั้งหมดที่จะใช้ในการจัดกลุ่มข้อมูล
- Next : เมื่อผู้ใช้ต้องการไปยังหน้าจอที่แสดงค่าพารามิเตอร์และค่าฟิตเนส

4.1.11 หน้าที่ 2.2.1 : หน้าจอที่ใช้ในการตรวจสอบค่าพารามิเตอร์และค่าฟิตเนส

โดยหน้าจอที่แสดงนี้สามารถแบ่งออกได้เป็น 2 ส่วนหลัก คือ

1. Input of Parameter: ส่วนที่ใช้แสดงค่าพารามิเตอร์ที่ได้จากการระบุโดยผู้ใช้
2. Detail of Fitness Value : ส่วนที่ใช้แสดงค่าฟิตเนสที่ได้จากการประมวลผลของแต่ละพาร์ทิเคิลในแต่ละรอบการทำงาน



รูปที่ 4.11 แสดงหน้าที่ 2.2.1 : หน้าจอแสดงค่าพารามิเตอร์และค่าฟิตเนส

หน้าจอนี้จะเกิดขึ้นเมื่อโปรแกรมทำการประมวลผลข้อมูลผ่านอัลกอริทึม Hybrid PSO สำเร็จแล้ว โดยสิ่งที่ผู้ใช้งานต้องพิจารณานอกเหนือจากที่กล่าวมาคือ

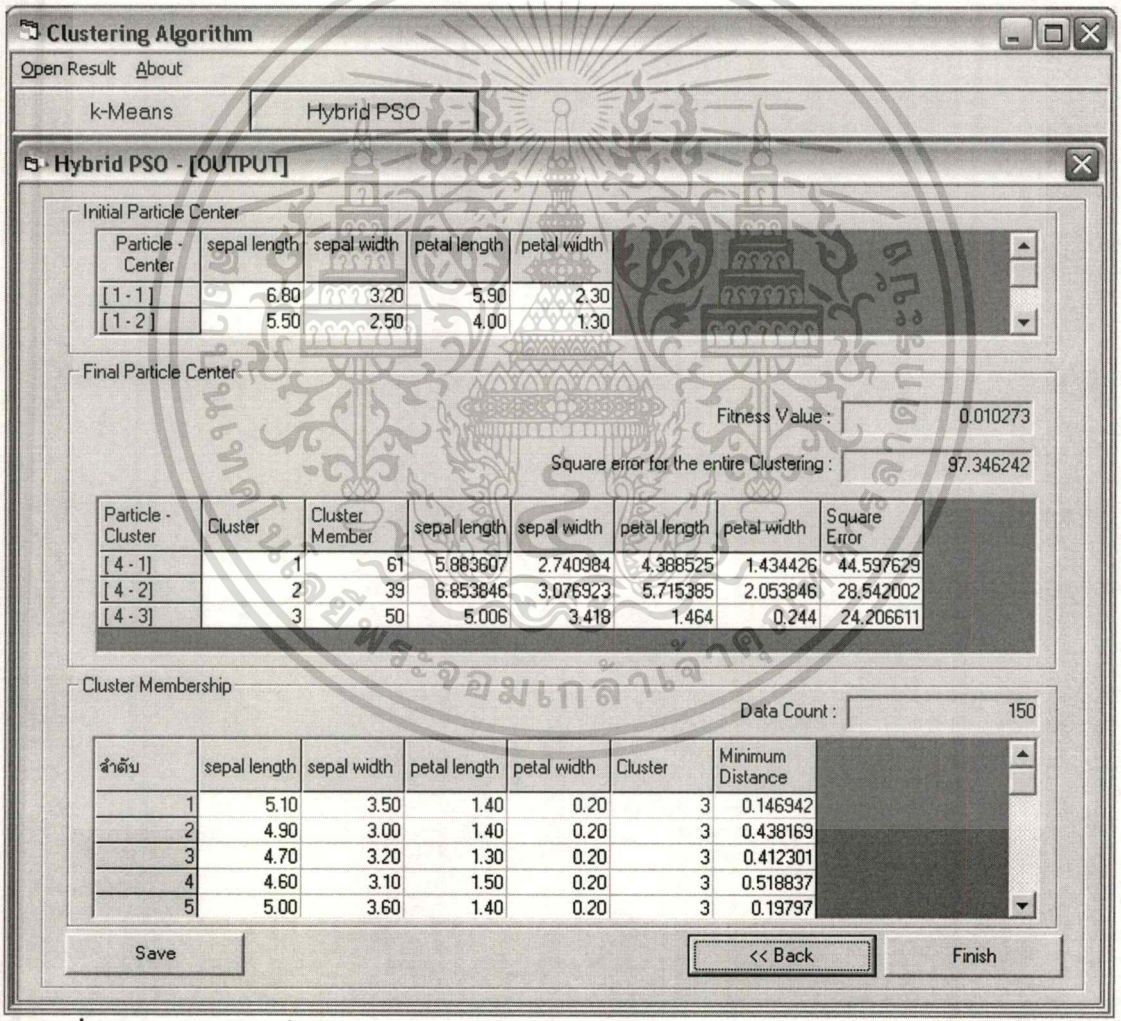
- Random1,Random2 : แสดงค่าที่จากการสุ่มของโปรแกรม โดยมีค่าระหว่าง 0 ถึง 1 ซึ่งจะใช้ในการปรับตำแหน่งในรอบสุดท้ายของแต่ละพาร์ทิเคิล
- Previous : สำหรับแสดงค่าฟิตเนสที่คำนวณได้ในรอบที่ผ่านมาของพาร์ทิเคิล โดยในรอบแรกของแต่ละพาร์ทิเคิลจะกำหนดให้เท่ากับ 0
- Actual : สำหรับแสดงค่าฟิตเนสที่คำนวณได้ในรอบปัจจุบันของพาร์ทิเคิล
- Fitness : สำหรับแสดงค่าฟิตเนสที่ดีที่สุด ซึ่งได้จากการเปรียบเทียบค่าฟิตเนสกันของพาร์ทิเคิลในรอบที่ผ่านมากับรอบปัจจุบัน
- Output Value : สำหรับใช้เมื่อต้องการแสดงค่าผลลัพธ์ที่ได้

- Back : สำหรับใช้เมื่อต้องการกลับไปยังหน้าจอหลักของอัลกอริทึม Hybrid PSO
- Next : สำหรับใช้เมื่อต้องการไปยังหน้าจอแสดงผลลัพธ์ที่ได้จัดกลุ่มจากการจัดกลุ่มโดยใช้อัลกอริทึม Hybrid PSO

4.1.12 หน้าที่ 2.1.2 : หน้าจอแสดงผลลัพธ์การจัดกลุ่มโดยใช้อัลกอริทึม Hybrid PSO

โดยหน้าจอที่แสดงผลลัพธ์ของการจัดกลุ่มนี้สามารถแบ่งออกได้เป็น 3 ส่วนหลัก คือ

1. Initial Particle Centers : ส่วนที่ใช้แสดงค่าศูนย์กลางเริ่มต้นของแต่ละพาร์ทิเคิล
2. Final Particle Centers : ส่วนที่ใช้แสดงค่าศูนย์กลางสุดท้ายที่ดีที่สุดของพาร์ทิเคิล
3. Cluster Membership : ส่วนที่ใช้แสดงผลลัพธ์ในการจัดกลุ่มของข้อมูลทั้งหมด



รูปที่ 4.12 แสดงหน้าที่ 2.2.2 : แสดงหน้าจอผลลัพธ์การจัดกลุ่มโดยใช้อัลกอริทึม Hybrid PSO

หน้าจอนี้จะเกิดขึ้นเมื่อโปรแกรมทำการประมวลผลข้อมูลผ่านอัลกอริทึม Hybrid PSO สำเร็จแล้ว รวมถึงผู้ใช้ได้ทำการตรวจสอบค่าพารามิเตอร์และค่าฟิตเนส หน้าจอจะแสดงผลลัพธ์ที่ได้จากการจัดกลุ่มทั้งหมด โดยสิ่งที่ผู้ใช้งานต้องพิจารณาคือ

- Initial Particle Centers : สำหรับใช้ในการแสดงค่าศูนย์กลางเริ่มต้นของแต่ละพาร์ทิเคิลที่ใช้ในการจัดกลุ่ม ซึ่งผลลัพธ์ในส่วนนี้โปรแกรมจะทำการสุ่มข้อมูลตามค่าพารามิเตอร์ที่ได้จากการระบุโดยผู้ใช้
- Final Particle Centers : สำหรับใช้ในการแสดงค่าศูนย์กลางสุดท้ายของพาร์ทิเคิลที่ทำให้ผลลัพธ์ในการจัดกลุ่มดีที่สุดและแสดงจำนวนข้อมูลในแต่ละกลุ่ม รวมทั้งแสดงค่า Square Error ของแต่ละกลุ่มข้อมูล
- Square Error for Entire Clustering : สำหรับใช้แสดงผลรวมของค่า Square Error ของทุกกลุ่มที่ได้จากการจัดกลุ่มข้อมูลในพาร์ทิเคิล
- Fitness Value : สำหรับแสดงค่าฟิตเนสของพาร์ทิเคิล
- Cluster Membership : สำหรับใช้แสดงผลลัพธ์ในการจัดกลุ่มของข้อมูลทั้งหมด โดยแสดงว่าข้อมูลแต่ละแถวนั้นถูกจัดกลุ่มให้อยู่ในคลัสเตอร์ใด รวมทั้งยังบอกค่า Minimum distance ซึ่งเป็นค่าที่บอกว่าข้อมูลแต่ละแถวอยู่ห่างจากศูนย์กลางกลุ่มเท่าไหร่
- Data Count : สำหรับใช้แสดงจำนวนของแถวหรือเรคอร์ดทั้งหมดในการจัดกลุ่ม
- Back : สำหรับใช้เมื่อต้องการกลับไปยังหน้าจอที่ใช้ตรวจสอบค่าพารามิเตอร์และค่าฟิตเนส
- Finish : สำหรับใช้เมื่อต้องการออกจากโปรแกรม

4.2 ข้อมูลที่ใช้ในการทดลองจัดกลุ่ม

ข้อมูลที่นำมาทดลองได้แก่ ข้อมูล Iris, Diabetes, Crude oil, Segmentation และ Ionosphere

4.2.1 ข้อมูล Iris

ข้อมูลนี้มีจำนวนข้อมูลอยู่ 150 ตัวอย่างประกอบด้วย 4 แ อตทริบิวต์ที่แตกต่างกัน ได้แก่ ค่าความยาวกลีบเลี้ยงดอกไม้ ค่าความกว้างกลีบเลี้ยงดอกไม้ ค่าความยาวของกลีบดอกไม้ และค่าความกว้างของกลีบดอกไม้ในหน่วยเซนติเมตร มีทั้งหมด 3 คลาส (ข้อมูลของคลาส 2 และคลาส 3 คาบเกี่ยวกัน) ดังแสดงในรูปที่ 4.13 โดยแต่ละคลาสมีข้อมูล 50 ตัวอย่าง ดังนั้นจำนวนกลุ่มข้อมูลของ Iris จึงเท่ากับ 3

4.2.2 ข้อมูล Diabetes

ข้อมูลนี้เป็นข้อมูลของผู้ป่วยโรคเบาหวาน ที่เป็นเพศหญิงอายุอย่างน้อย 21 ปี มีทั้งหมด 768 ข้อมูลประกอบด้วย 8 คุณลักษณะซึ่งมี 2 คลาสได้แก่ คลาส tested negative และ tested positive

iris

	A	B	C	D	E
1	sepal length	sepal width	petal length	petal width	class
2	5.1	3.5	1.4	0.2	Iris-setosa
3	4.9	3	1.4	0.2	Iris-setosa
56	6.5	2.8	4.6	1.5	Iris-versicolor
67	6.7	3.1	4.4	1.4	Iris-versicolor
102	6.3	3.3	6	2.5	Iris-virginica
151	5.9	3	5.1	1.8	Iris-virginica

iris

รูปที่ 4.13 แสดงข้อมูล Iris ที่แบ่งตามคลาส

ดังนั้นจำนวนกลุ่มของข้อมูล Diabetes จึงมีค่าเท่ากับ 2

4.2.3 ข้อมูล Crude Oil

ข้อมูลของ Crude Oil เป็นข้อมูลที่ค่อนข้างคาบเกี่ยวกัน มีทั้งหมด 56 ข้อมูล ประกอบด้วย 5 คุณลักษณะ และมีทั้งหมด 3 คลาส ดังนั้นจำนวนกลุ่มข้อมูลที่นำมาใช้จัดกลุ่มจึงมีค่าเท่ากับ 3

4.2.4 ข้อมูล Vowel

ข้อมูลนี้ประกอบด้วยการออกเสียงสระ 871 เสียงของชาว Indian Telugu ของนักพูดชาย 3 คน ในช่วงอายุ 30-35 ปี ข้อมูลประกอบด้วยการออกเสียงสระ 3 ความถี่ และ 6 คลาสที่คาบเกี่ยวกัน ดังนั้นค่าของจำนวนกลุ่มจึงเท่ากับ 6

4.2.5 ข้อมูล Ionoshere

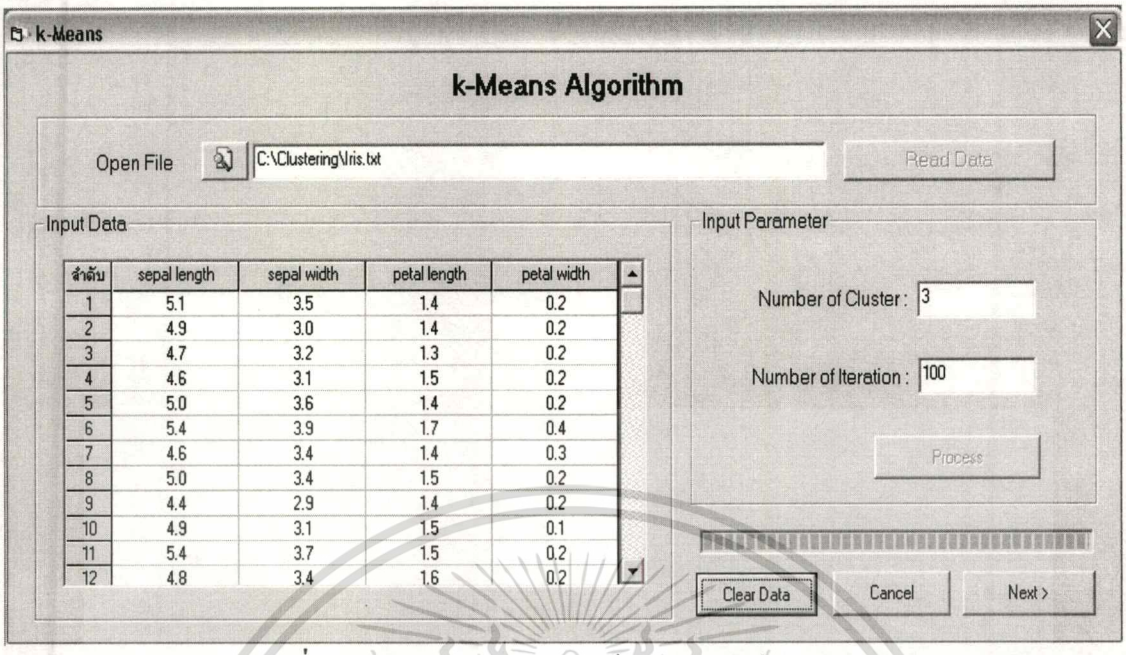
ข้อมูลนี้เป็นข้อมูลที่เกี่ยวข้องกับการส่งเรดาร์ไปที่ชั้นบรรยากาศ มีทั้งหมด 351 ข้อมูล ประกอบด้วย 34 คุณลักษณะ ซึ่งมี 2 คลาส ได้แก่ คลาส good และ bad โดยที่ good เรดาร์จะสามารถแสดงบางชนิดโครงสร้างของบรรยากาศได้ ส่วน bad เรดาร์ จะไม่สามารถแสดงได้ ดังนั้น จำนวนกลุ่มข้อมูลที่นำมาใช้จัดกลุ่มจึงมีค่าเท่ากับ 2

4.3 ผลการจัดกลุ่มข้อมูล

ในการทดลองจะทำการเปรียบเทียบผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูลระหว่างอัลกอริทึม k-Means และอัลกอริทึม Hybrid PSO กับทุกข้อมูล โดยในแต่ละอัลกอริทึมจะทำการทดลองรันเป็นจำนวน 5 ครั้ง เพื่อทำการเปรียบเทียบผลที่ได้จากทั้งสองอัลกอริทึม ซึ่งได้ผลการทดลองดังนี้

4.3.1 ผลการจัดกลุ่มข้อมูล Iris โดยใช้อัลกอริทึม k-Means

ซึ่งจะกำหนดค่าพารามิเตอร์ต่าง ๆ โดยให้จำนวนกลุ่มเท่ากับ 3 และจำนวนการทำซ้ำของอัลกอริทึมเท่ากับ 100 ดังแสดงในรูปที่ 4.14



ลำดับ	sepal length	sepal width	petal length	petal width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4
7	4.6	3.4	1.4	0.3
8	5.0	3.4	1.5	0.2
9	4.4	2.9	1.4	0.2
10	4.9	3.1	1.5	0.1
11	5.4	3.7	1.5	0.2
12	4.8	3.4	1.6	0.2

รูปที่ 4.14 การใส่ค่าพารามิเตอร์ของอัลกอริทึม k-Means

เมื่อโปรแกรมประมวลผลโดยใช้อัลกอริทึม k-Means ผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูลในรอบที่ 1 ให้ค่าความผิดพลาด (Square Error) ดังแสดงในรูปที่ 4.15

Final Cluster Centers

Square error for entire Clustering : 97.3462196941568

ลำดับ	cluster	Cluster Member	sepal length	sepal width	petal length	petal width	Square Error
1	1	61	5.883607	2.740984	4.388525	1.434426	44.597618
2	2	39	6.853846	3.076923	5.715385	2.053846	28.541989
3	3	50	5.006	3.418	1.464	0.244	24.206612

รูปที่ 4.15 แสดงค่าความผิดพลาดที่ได้จากการจัดกลุ่มโดยใช้อัลกอริทึม k-Means

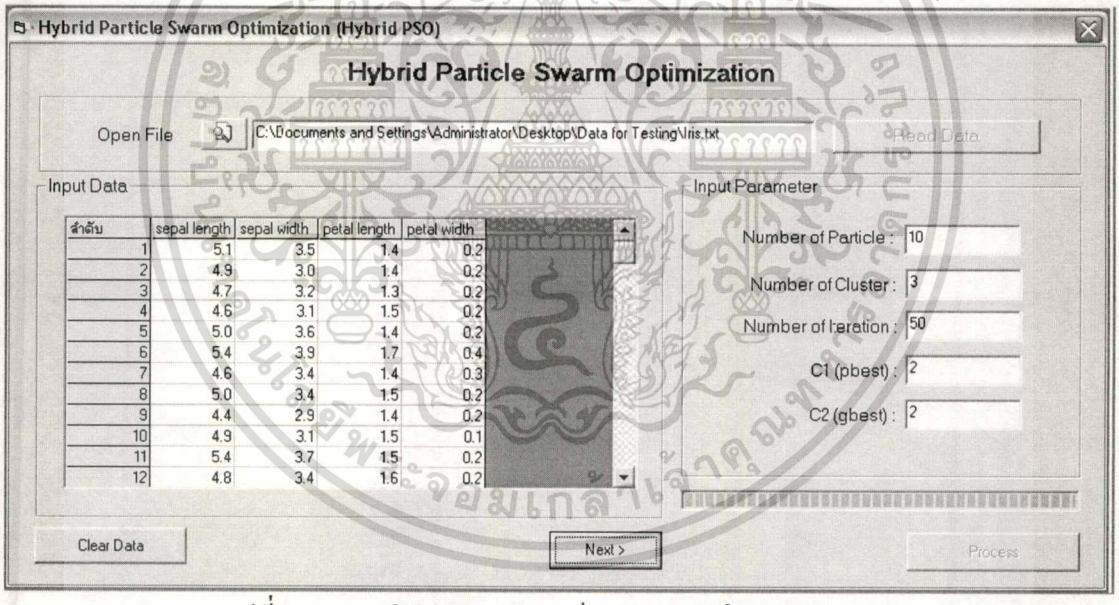
เมื่อทำการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม k-Means เป็นจำนวน 5 ครั้งในจำนวนรอบต่าง ๆ จะได้ผลลัพธ์ของค่าความผิดพลาดดังแสดงไว้ในตารางที่ 4.1

ตารางที่ 4.1 แสดงผลการจัดกลุ่มข้อมูล Iris โดยใช้อัลกอริทึม k-Means

รันครั้งที่	ค่าความผิดพลาด
1	122.27887941
2	97.3259227
3	97.3259227
4	97.3462203
5	97.3462203

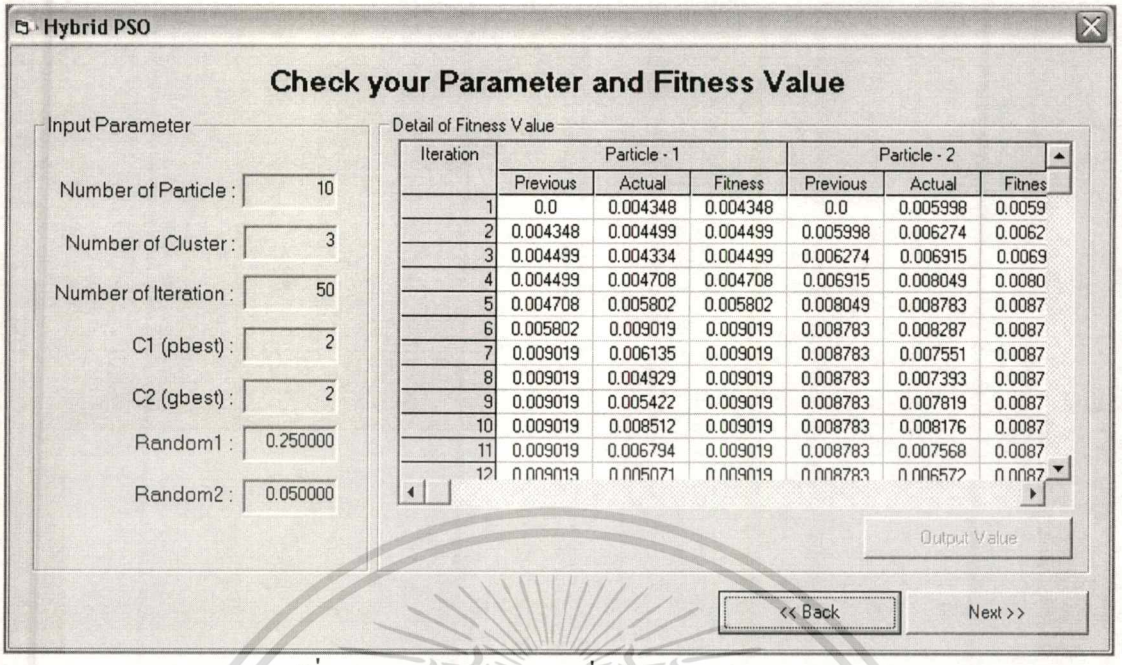
4.3.2 ผลการจัดกลุ่มข้อมูล Iris โดยใช้อัลกอริทึม Hybrid PSO

ซึ่งจะกำหนดค่าพารามิเตอร์ต่าง ๆ โดยให้จำนวนพาร์ทิเคิลเท่ากับ 5 จำนวนกลุ่มเท่ากับ 3 จำนวนการทำซ้ำและจำนวนพาร์ทิเคิลที่แตกต่างกัน และค่า C1,C2 เท่ากับ 2 ดังแสดงในรูปที่ 4.16



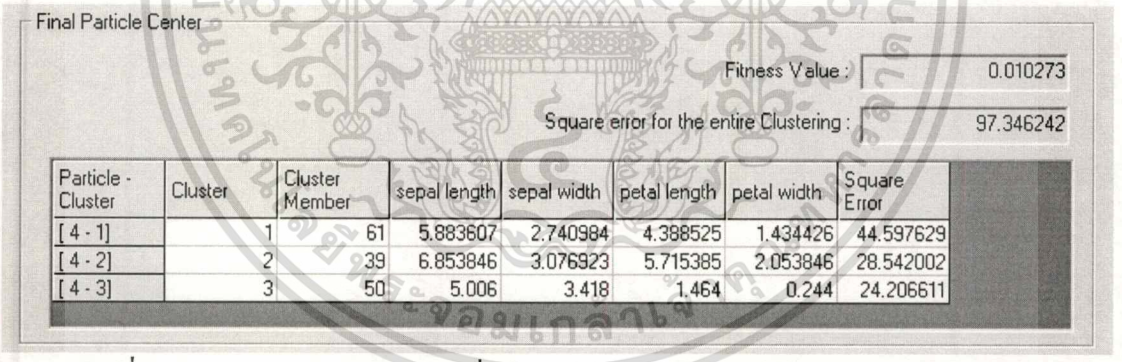
รูปที่ 4.16 การใส่ค่าพารามิเตอร์ของอัลกอริทึม Hybrid PSO

เมื่อโปรแกรมประมวลผลโดยใช้อัลกอริทึม Hybrid PSO สำเร็จแล้วสามารถดูค่าฟิตเนสที่ได้และตรวจสอบค่าพารามิเตอร์ ดังแสดงในรูปที่ 4.17



รูปที่ 4.17 แสดงค่า ฟิตเนสที่ได้จากการประมวลผล

เมื่อตรวจสอบค่าฟิตเนสเสร็จแล้ว ผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูลในรอบที่ 1 ให้ค่าความผิดพลาด (Square Error) ดังแสดงในรูปที่ 4.18



รูปที่ 4.18 แสดงค่าความผิดพลาดที่ได้จากการจัดกลุ่มโดยใช้อัลกอริทึม Hybrid PSO

เมื่อทำการจัดกลุ่มข้อมูล Iris โดยใช้อัลกอริทึม Hybrid PSO เป็นจำนวน 5 ครั้งได้ผลลัพธ์ของค่าความผิดพลาดดังแสดงไว้ในตารางที่ 4.2

ตารางที่ 4.2 แสดงผลการจัดกลุ่มข้อมูล Iris โดยใช้อัลกอริทึม Hybrid PSO

รันครั้งที่	ค่าความผิดพลาด
1	97.239834
2	97.190098
3	97.346242
4	97. 318691
5	97.339038

ผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูล Iris ในตารางที่ 4.1 และ 4.2 ค่าความผิดพลาดที่ได้จากการทำงานของอัลกอริทึม k-Means ให้ผลลัพธ์ที่ดีที่สุดเท่ากับ 97.3259227 ส่วนผลลัพธ์ที่ได้จากการจัดกลุ่มโดยใช้อัลกอริทึม Hybrid PSO ให้ผลลัพธ์ที่ดีที่สุดเท่ากับ 97.190098

4.3.3 ผลการจัดกลุ่มข้อมูล Diabetes โดยใช้อัลกอริทึม k-Means

ซึ่งจะกำหนดค่าพารามิเตอร์ต่าง ๆ โดยให้จำนวนกลุ่มเท่ากับ 2 ในจำนวนรอบต่าง ๆ จะได้ผลลัพธ์ของค่าความผิดพลาดดังแสดงไว้ในตารางที่ 4.3

ตารางที่ 4.3 แสดงผลการจัดกลุ่มข้อมูล Diabetes โดยใช้อัลกอริทึม k-Means

รันครั้งที่	ค่าความผิดพลาด
1	52072.24417
2	52072.24417
3	52072.24417
4	52072.24417
5	52072.24417

4.3.4 ผลการจัดกลุ่มข้อมูล Diabetes โดยใช้อัลกอริทึม Hybrid PSO

ซึ่งจะกำหนดค่าพารามิเตอร์ต่าง ๆ โดยให้จำนวนการทำซ้ำและจำนวนพาร์ทิเคิลที่แตกต่างกัน และค่า C1,C2 เท่ากับ 2 โดยผลการจัดกลุ่มแสดงในตารางที่ 4.4

ตารางที่ 4.4 แสดงผลการจัดกลุ่มข้อมูล Diabetes โดยใช้อัลกอริทึม Hybrid PSO

รันครั้งที่	ค่าความผิดพลาด
1	50645.935713
2	48848.563236
3	51568.279046
4	51708.006157
5	51887.428147

ผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูล Diabetes ค่าความผิดพลาดที่ได้จากการทำงานของอัลกอริทึม k-Means ให้ผลลัพธ์ที่ดีที่สุดเท่ากับ 52072.24417 ส่วนผลลัพธ์ที่ได้จากการจัดกลุ่มโดยใช้อัลกอริทึม Hybrid PSO ให้ผลลัพธ์ที่ดีที่สุดเท่ากับ 48848.563236

4.3.5 ผลการจัดกลุ่มข้อมูล Crude Oil โดยใช้อัลกอริทึม k-Means

ซึ่งจะกำหนดค่าพารามิเตอร์ต่าง ๆ โดยให้จำนวนกลุ่มเท่ากับ 3 ในจำนวนรอบต่าง ๆ จะได้ผลลัพธ์ของค่าความผิดพลาดดังแสดงไว้ในตารางที่ 4.5

ตารางที่ 4.5 แสดงผลการจัดกลุ่มข้อมูล Crude Oil โดยใช้อัลกอริทึม k-Means

รันครั้งที่	ค่าความผิดพลาด
1	279.743216
2	279.484881
3	279.743216
4	279.743216
5	279.743216

4.3.6 ผลการจัดกลุ่มข้อมูล Crude Oil โดยใช้อัลกอริทึม Hybrid PSO

ซึ่งจะกำหนดค่าพารามิเตอร์ต่าง ๆ โดยให้จำนวนการทำซ้ำและจำนวนพาร์ทิเคิลที่แตกต่างกัน และค่า C1,C2 เท่ากับ 2 โดยผลการจัดกลุ่มแสดงในตารางที่ 4.6

ตารางที่ 4.6 แสดงผลการจัดกลุ่มข้อมูล Crude Oil โดยใช้อัลกอริทึม Hybrid PSO

รันครั้งที่	ค่าความผิดพลาด
1	279.743066
2	278.895880
3	279.743066
4	279.106898
5	279.743066

ผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูล Crude Oil ค่าความผิดพลาดที่ได้จากการทำงานของอัลกอริทึม k-Means ให้ผลลัพธ์ที่ดีที่สุดเท่ากับ 279.484881 ส่วนผลลัพธ์ที่ได้จากการจัดกลุ่มโดยใช้อัลกอริทึม Hybrid PSO ให้ผลลัพธ์ที่ดีที่สุดเท่ากับ 278.895880

4.3.7 ผลการจัดกลุ่มข้อมูล Vowel โดยใช้อัลกอริทึม k-Means

ซึ่งจะกำหนดค่าพารามิเตอร์ต่าง ๆ โดยให้จำนวนกลุ่มเท่ากับ 6 ในจำนวนรอบต่าง ๆ จะได้ผลลัพธ์ของค่าความผิดพลาดดังแสดงไว้ในตารางที่ 4.7

ตารางที่ 4.7 แสดงผลการจัดกลุ่มข้อมูล Vowel โดยใช้อัลกอริทึม k-Means

รันครั้งที่	ค่าความผิดพลาด
1	157465.15548
2	149529.01545
3	151473.90351
4	151975.04736
5	151975.04736

4.3.8 ผลการจัดกลุ่มข้อมูล Vowel โดยใช้อัลกอริทึม Hybrid PSO

ซึ่งจะกำหนดค่าพารามิเตอร์ต่าง ๆ โดยให้จำนวนการทำซ้ำและจำนวนพาร์ทิเคิลที่แตกต่างกัน และค่า C1,C2 เท่ากับ 2 โดยผลการจัดกลุ่มแสดงในตารางที่ 4.8

ตารางที่ 4.8 แสดงผลการจัดกลุ่มข้อมูล Vowel โดยใช้อัลกอริทึม Hybrid PSO

รันครั้งที่	ค่าความผิดพลาด
1	151473.904105
2	151792.804207
3	149441.360274
4	157229.762251
5	151473.904105

ผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูล Vowel ค่าความผิดพลาดที่ได้จากการทำงานของอัลกอริทึม k-Means ให้ผลลัพธ์ที่ดีที่สุดเท่ากับ 149529.01545 ส่วนผลลัพธ์ที่ได้จากการจัดกลุ่มโดยใช้อัลกอริทึม Hybrid PSO ให้ผลลัพธ์ที่ดีที่สุดเท่ากับ 149441.360274

4.3.9 ผลการจัดกลุ่มข้อมูล Ionosphere โดยใช้อัลกอริทึม k-Means

ซึ่งจะกำหนดค่าพารามิเตอร์ต่าง ๆ โดยให้จำนวนกลุ่มเท่ากับ 2 ในจำนวนรอบต่าง ๆ จะได้ผลลัพธ์ของค่าความผิดพลาดดังแสดงไว้ในตารางที่ 4.9

ตารางที่ 4.9 แสดงผลการจัดกลุ่มข้อมูล Ionosphere โดยใช้อัลกอริทึม k-Means

รันครั้งที่	ค่าความผิดพลาด
1	796.3961038
2	796.5462053
3	796.3961038
4	796.5462053
5	796.5462053

4.3.10 ผลการจัดกลุ่มข้อมูล Ionosphere โดยใช้อัลกอริทึม Hybrid PSO

ซึ่งจะกำหนดค่าพารามิเตอร์ต่าง ๆ โดยให้จำนวนการทำซ้ำและจำนวนพาร์ทิเคิลที่แตกต่างกัน และค่า C1,C2 เท่ากับ 2 โดยผลการจัดกลุ่มแสดงในตารางที่ 4.10

ตารางที่ 4.10 แสดงผลการจัดกลุ่มข้อมูล Ionosphere โดยใช้อัลกอริทึม Hybrid PSO

รันครั้งที่	ค่าความผิดพลาด
1	796.285436
2	796.269140
3	796.393219
4	796.731050
5	796.285436

ผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูล Ionosphere ค่าความผิดพลาดที่ได้จากการทำงานของอัลกอริทึม k-Means ให้ผลลัพธ์ที่ดีที่สุดเท่ากับ 796.3961038 ส่วนผลลัพธ์ที่ได้จากการจัดกลุ่มโดยใช้อัลกอริทึม Hybrid PSO ให้ผลลัพธ์ที่ดีที่สุดเท่ากับ 796.269140

บทที่ 5
สรุปผลการศึกษาและข้อเสนอแนะ

จากการศึกษาการจัดกลุ่มข้อมูลโดยใช้ไฮบริดพาร์ทิเคิลสวอมออปติไมเซชั่นรวมถึงอัลกอริทึม k-Means เพื่อที่จะสร้างโครงการศึกษาพัฒนาระบบนี้ สามารถสรุปผลการศึกษา และประโยชน์ที่ได้รับ รวมทั้งข้อเสอแนะ ดังต่อไปนี้

5.1 สรุปผลการศึกษา

จากที่ได้ศึกษาการทำงานและทฤษฎีของอัลกอริทึมต่าง ๆ ในการทำดาต้าไมนิ่ง ทำให้ทราบว่า ดาต้าไมนิ่งเป็นกระบวนการที่ใช้วิเคราะห์และค้นหาข้อมูลที่ซ่อนเร้นอยู่ภายในฐานข้อมูล จึงได้มีการพัฒนาโครงการโดยนำเสนอถึงการจัดกลุ่มข้อมูล โดยใช้อัลกอริทึม k-Means และ Hybrid PSO ซึ่งเป็นอัลกอริทึมในการทำ Clustering ที่สามารถนำข้อมูลที่เป็นตัวเลข (Numeric) มาวิเคราะห์ได้ ดังนั้นจึงได้วางแนวทางในการพัฒนาระบบด้วยโปรแกรม Microsoft Visual Basic 6.0

อัลกอริทึม PSO เป็นอัลกอริทึมสำหรับค้นหาค่าที่เหมาะสมที่สุด โดยจะทำการลอกเลียนแบบพฤติกรรมทางธรรมชาติ โดยอัลกอริทึมจะมีการคำนวณค่าความเหมาะสมเพื่อให้ได้ตำแหน่งของพาร์ทิเคิลที่ดีที่สุดออกมา สำหรับเซตข้อมูลขนาดใหญ่ก็นิยมใช้อัลกอริทึม PSO เพราะสามารถนำไปสู่การค้นหาการจัดกลุ่มที่ดีที่สุดแต่ยังคงต้องการจำนวนการทำซ้ำที่มากและมีการคำนวณที่มากกว่าของ k-Means สำหรับ k-Means จะเป็นการทำงานที่ง่ายและเร็วกว่า PSO จึงได้มีการผสมผสานกันของอัลกอริทึมในการจัดกลุ่มระหว่าง PSO และ k-Means ซึ่งเป็นการรวมข้อดีและหลีกเลี่ยงข้อเสียของทั้งสองอัลกอริทึม นั่นคือวิธีการค้นหาด้วยตนเองที่มีประสิทธิภาพและได้วิธีการที่ดีที่สุดหรือวิธีการที่ใกล้เคียงที่สุดในการค้นหาคำตอบที่ซับซ้อนรวมถึงการทำงานที่รวดเร็ว โดยผลลัพธ์ของอัลกอริทึม PSO จะถูกใช้เป็นศูนย์กลางเริ่มต้นในการทำงานของอัลกอริทึม k-Means ซึ่งสุดท้ายผลลัพธ์ที่ดีที่สุดจะถูกค้นพบ

จากการศึกษาได้ทำการเปรียบเทียบประสิทธิภาพในการจัดกลุ่มข้อมูลโดยวัดจากค่าความผิดพลาดที่เกิดขึ้นระหว่างข้อมูลสมาชิกกับศูนย์กลางของกลุ่มในแต่ละคลัสเตอร์ จากข้อมูลทั้งหมดที่นำมาทดลองพบว่า การจัดกลุ่มข้อมูลโดยใช้ไฮบริดพาร์ทิเคิลสวอมออปติไมเซชั่นให้ผลค่าความผิดพลาดที่น้อยกว่าอัลกอริทึม k-Means แสดงให้เห็นว่า การจัดกลุ่มข้อมูลโดยใช้ Hybrid PSO มาค้นหาศูนย์กลางกลุ่มข้อมูลนั้นมีประสิทธิภาพในการจัดกลุ่มข้อมูลมากกว่าการจัดกลุ่มข้อมูลด้วยวิธีของ k-Means

ในการจัดกลุ่มข้อมูลแต่ละครั้งถึงแม้เราจะกำหนดค่าพารามิเตอร์นำเข้าทุกอย่างให้เหมือนกันหมด แต่ประสิทธิภาพที่ได้จากการจัดกลุ่มแต่ละครั้งอาจจะได้ผลลัพธ์ไม่เหมือนกัน

เนื่องมาจากการสุ่มค่าเริ่มต้นของโปรแกรมของแต่ละอัลกอริทึม สำหรับเรื่องเวลาในการทำงานจะแปรผันโดยตรงกับจำนวนกลุ่มที่ต้องการ จำนวนรอบในการทำซ้ำ จำนวนพาร์ทิเคิลที่เราต้องการ และค่าจำนวนรอบที่พาร์ทิเคิลเรียนรู้ในการปรับตำแหน่ง รวมถึงจำนวนของข้อมูลที่ต้องการจัดกลุ่มและความสามารถของฮาร์ดแวร์ที่ใช้

5.2 ประโยชน์ที่ได้รับจากการศึกษาและพัฒนาระบบ

จากการศึกษาในโครงการพัฒนาระบบงานนี้ สามารถสรุปผลการทดลองในการออกแบบการจัดกลุ่มข้อมูลโดยใช้ไฮบริดพาร์ทิเคิลสวอมออปติไมเซชั่น รวมถึง k-Means อัลกอริทึมได้ดังนี้

1. ทำให้เข้าใจทฤษฎี และหลักการทำงานของอัลกอริทึมต่าง ๆ ที่ใช้ในการจัดกลุ่มข้อมูลได้มากยิ่งขึ้น

2. ทำให้ทราบถึงปัญหาต่าง ๆ ที่มักเกิดขึ้นในขั้นตอนต่าง ๆ ของการจัดกลุ่มข้อมูล

3. ทำให้ได้โปรแกรมที่ใช้ในการวิเคราะห์ข้อมูล ซึ่งสามารถนำไปใช้เป็นต้นแบบในการพัฒนาโปรแกรมในลักษณะอื่น ๆ ได้

5.3 ข้อเสนอแนะ

1. ในระบบที่ทำการพัฒนาขึ้นมา ทั้งอัลกอริทึม k-Means และ Hybrid PSO สามารถที่จะรับข้อมูลเข้ามาทำการวิเคราะห์ได้ แต่ข้อมูลนั้นต้องอยู่ในรูปแบบที่เป็น Numeric เท่านั้น เนื่องมาจากข้อจำกัดทางด้านอัลกอริทึมที่สามารถจะรับข้อมูลเข้าไปในแบบ Numeric ได้เพียงรูปแบบเดียว ระบบที่ทำการพัฒนาขึ้นจึงไม่รองรับข้อมูลที่เป็นแบบ Categorical

2. สำหรับ Hybrid PSO อัลกอริทึมนั้น ควรมีความสามารถมากขึ้นกว่านี้ ได้แก่ ความสามารถในการกำหนดค่าพารามิเตอร์ของผู้ใช้ ควรที่จะสามารถลดและเพิ่มความเร็วของการเคลื่อนที่ได้ในแต่ละรอบของการทำงานได้ รวมถึงให้พาร์ทิเคิลมีความสามารถในการกระจายให้ทั่วพื้นที่โดยที่เราไม่ต้องสุ่มค่าให้กับพาร์ทิเคิล

บรรณานุกรม

C. Ching-Yi, and Ye, J. **Particle Swarm Optimization Algorithm and Its Application to Clustering Analysis.** IEEE ECNSC Taipei Taiwan. 2004.

C.K.Mohan or E.Ozean. **Particle Swarm Optimization.** [Online]. Available : http://www.cis.syr.edu/~mohan/pso/.2002.

Howard, J. Hamilton. **Clustering.** [Online]. Available : http://www2.cs.uregina.ca/~hamilton/courses/831/notes/clustering/clustering.html. 2002.

Jiawei Han, and Micheline, K. **Data Mining: Concepts And Techniques.** CA : Morgan Kaufmann San Francisco. 2002.

J. Kennedy, and R. Eberhart. **Particle Swarm Optimization.** Proc. Of IEEE International Conference on Neural Networks(ICNN) PerthAustralia. 1995.

Kardi Teknomo, PhD. **k-Means Clustering Tutorial.** [Online]. Available : http://people.vevoledu.com/kardi/turorial/kMean. 2005.

Mark C Sinclair. **Particle Swarm Optimization.** [Online]. Available : http://uk.geocities.com/markcsinclair/pso.html. 2001.

V.D Merwe, and Engelbrecht. **Data Clustering Uusing Particle Swarm Optimization.** Proceedings of IEEE ComputationCanbella Australia. 2003.

Yan Wang, and Lihua, Lin. **K-Means Clustering.** [Online]. Available : http://www.cs.ucsb.edu/~cs281b/winter2002/Misc/k-Means.ppt. 2002.

Xiaohui, Hu. **Particle Swarm Optimization.** [Online]. Available : http://www.swarmintelligence.org/index.php. 2003.

ประวัติผู้เขียน

ชื่อ-นามสกุล	นางสาวศิริพร ระเมียดดี
วัน เดือน ปีเกิด	21 มิถุนายน 2525
ที่อยู่	524 หมู่บ้านธราทร ซ.4 ถ.รามคำแหง แขวงมีนบุรี เขตมีนบุรี กรุงเทพฯ 10510
วุฒิการศึกษา	วิทยาศาสตรบัณฑิต สาขาสถิติประยุกต์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ปีที่สำเร็จการศึกษา	2547