การปรับปรุงประสิทธิภาพการจัดกลุ่มข้อมูลของอัลกอริทึมเคมีน ด้วยการหาค่าเริ่มต้นโดยวิธีการตัดแบ่งกลุ่มข้อมูล

สิริชัย ดีเลิศ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต (ระบบสารสนเทศประยุกต์)
คณะสถิติประยุกต์
สถาบันบัณฑิตพัฒนบริหารศาสตร์
2550

การปรับปรุงประสิทธิภาพการจัดกลุ่มข้อมูลของอัลกอริทึมเคมีน
ด้วยการหาค่าเริ่มต้นโดยวิธีการตัดแบ่งกลุ่มข้อมูล

สิริชัย ดีเลิศ
คณะสถิติประยุกต์

คณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาแล้วเห็นสมควรอนุมัติให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต (ระบบสารสนเทศประยุกต์)

ผู้ช่วยศาสตราจารย์ .....ประธานกรรมการ
(ดร. โอม ศรนิล)

รองศาสตราจารย์ .....กรรมการและอาจารย์ที่ปรึกษาวิทยานิพนธ์
(ดร. สุรพงค์ เอื้อวัฒนามงคล)

รองศาสตราจารย์ ..... กรรมการ
(ดร. บุญธีร์ เครือตราชู)

ผู้ช่วยศาสตราจารย์ ..... กรรมการ
(ดร. ปรีชา วิจิตรธรรมรส)

รองศาสตราจารย์ ..... คณบดี
(ดร. สุรพงค์ เอื้อวัฒนามงคล)
วันที่ 26 เดือน กันยายน พ.ศ. 2550

บทคัดย่อ

ชื่อวิทยานิพนธ์	การปรับปรุงประสิทธิภาพการจัดกลุ่มข้อมูลของอัลกอริทึมเคมีนด้วยการหาค่าเริ่มต้นโดยวิธีการตัดแบ่งกลุ่มข้อมูล
ชื่อผู้เขียน	สิริชัย ดีเลิศ
ชื่อปริญญา	วิทยาศาสตรมหาบัณฑิต (ระบบสารสนเทศประยุกต์)
ปีการศึกษา	2550

งานวิจัยนี้ เป็นการปรับปรุงประสิทธิภาพการจัดกลุ่มข้อมูลของอัลกอริทึมเคมีนด้วยการหาค่าเริ่มต้นโดยวิธีการตัดแบ่งกลุ่มข้อมูล ที่ใช้ได้ผลดีกับการตัดแบ่งชั้นสี (Color Quantization) หลักการตัดแบ่งข้อมูลคือการแบ่งข้อมูลตามแกนที่มีค่าความแปรปรวนสูงสุดให้ได้จำนวนกลุ่มตามที่ต้องการจัดกลุ่มข้อมูลด้วยอัลกอริทึมเคมีน (K-means) และใช้จุดศูนย์กลางของข้อมูลที่แบ่งแล้วเป็นจุดเริ่มต้นของการจัดกลุ่มด้วยอัลกอริทึมเคมีน การใช้จุดเริ่มต้นที่ดีจะลดข้อจำกัด และข้อเสียของการใช้ค่าเริ่มต้นแบบสุ่ม ที่ให้ผลการจัดกลุ่มที่ไม่แน่นอน และกลุ่มข้อมูลบางกลุ่มอาจไม่มีจำนวนสมาชิก

การทดสอบประสิทธิภาพของอัลกอริทึมที่นำเสนอได้ทำกับข้อมูลจาก UCI และ Web Access Log โดยเปรียบเทียบผลกับอัลกอริทึมเคมีนที่มีการกำหนดค่าเริ่มต้นแบบสุ่ม อีกทั้งยังใช้การเปรียบเทียบกับอัลกอริทึมที่ใช้ในการกำหนดค่าเริ่มต้นสำหรับการจัดกลุ่มข้อมูลของอัลกอริทึมเคมีนด้วย Cluster Center Initialization Algorithm (CCIA) จากผลการทดสอบประสิทธิภาพของการจัดกลุ่มข้อมูล ถือได้ว่าอัลกอริทึมที่นำเสนอนี้มีประสิทธิภาพดีกว่าการใช้ค่าเริ่มต้นแบบสุ่ม และให้ประสิทธิภาพใกล้เคียงกับ CCIA ซึ่งมีวิธีการที่ซับซ้อนกว่า

ABSTRACT

Title of Thesis	Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance
Author	Mr. Sirichai Deelers
Degree	Master of Science (Applied Information System)
Year	2007

In this research, we propose an algorithm to compute initial cluster centers for K-means clustering. We use novel approach for color quantization that divides color spaces into small clusters or cells with intercluster distances as large as possible and intracluster distance as small as possible. In the proposed algorithm, data in a cell is partitioned using a cutting plane that divide cell in two small cells. The plane is perpendicular to the data axis with the highest variance and is designed to reduce the sum squared errors of the two cells as much as possible. Cells are partitioned one at a time until the number of cells reaches the desired number K. The centers of the K cells become the initial cluster centers for K-means.

We evaluated our method by clustering 10 UCI data sets (UCI Machine Learning Repository) and Web Access Log data set. We also present the experimental results on some datasets in comparison with CCIA algorithm. The experimental results reveal that the proposed algorithm computes initial cluster centers that help K-means converge to better clustering than the random initial cluster centers and almost guarantee every cluster has its data membership. The proposed algorithm also performs as good as CCIA algorithm which is more difficult to implement.

กิตติกรรมประกาศ

วิทยานิพนธ์เรื่องการปรับปรุงประสิทธิภาพการจัดกลุ่มข้อมูลของอัลกอริทึมเคมีน ด้วยการหาค่นริ่มต้นโดยวิธีการตัดแบ่งกลุ่มข้อมูลนี้ สำเร็จลุล่วงได้ด้วยดี เนื่องมาจากบุคคลหลายท่านที่ได้กรุณาช่วยเหลือให้ข้อมูล ข้อเสนอแนะ คำปรึกษาแนะนำ ความคิดเห็นและกำลังใจ

ผู้เขียนขอกราบขอบพระคุณ รองศาสตราจารย์ ดร. สุรพงค์ เอื้อวัฒนามงคล ที่ได้ให้คำแนะนำ ชี้แนะแนวทาง และตรวจสอบวิทยานิพนธ์ในทุกขั้นตอน และขอกราบขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร. โอม ศรนิล ผู้ช่วยศาสตราจารย์ ดร. ปรีชา วิจิตรธรรมรส และรองศาสตราจารย์ ดร.บุญธีร์ เครือตราชู ที่ได้ชี้แนะแนวทางในการศึกษา และเปรียบเทียบประสิทธิภาพของอัลกอริทึมในการจัดกลุ่มข้อมูล

ขอขอบพระคุณคณาจารย์ของคณะสถิติประยุกต์ทุกท่าน ที่ประสิทธิ์ประสาทวิชาและถ่ายทอดความรู้ให้แก่ผู้ศึกษา และขอขอบคุณเจ้าหน้าที่คณะสถิติประยุกต์ ที่ได้ให้ความช่วยเหลือในการติดต่อประสานงานเป็นอย่างดี

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.วันชัย สุทธะนันท์ คณบดี คณะวิทยาการจัดการ มหาวิทยาลัยศิลปากร ที่ได้ให้โอกาสและสนับสนุนการศึกษาในทุกด้าน

ขอขอบคุณนักศึกษาร่วมรุ่น รุ่นพี่ และรุ่นน้องในคณะสถิติประยุกต์ที่ได้ให้ความช่วยเหลือและประสานงานตลอดช่วงเวลาที่ได้ศึกษาอยู่ที่คณะสถิติประยุกต์

ท้ายสุดนี้ ขอกราบขอบพระคุณบิดา มารดา และญาติพี่น้องที่ได้ช่วยส่งเสริมสนับสนุนและเป็นกำลังใจให้แก่ผู้จัดทำวิทยานิพนธ์ตลอดมา

สิริชัย ดีเลิศ

สิงหาคม 2550

สารบัญ

สารบัญตาราง

ตารางที่ หน้า

สารบัญภาพ

ภาพที่ หน้า

สารบัญกราฟ

กราฟที่ **หน้า**

บทที่ 1

บทนำ

ในบทนี้จะกล่าวถึงความสำคัญของปัญหา วัตถุประสงค์ของการวิจัย ขอบเขตของการวิจัย กรอบแนวความคิดและประโยชน์ที่จะได้รับจากการวิจัย โดยมีรายละเอียดดังต่อไปนี้

1.1 ความสำคัญของปัญหา

การจัดกลุ่มข้อมูล (Clustering) เป็นการจำแนกประเภทข้อมูล (Classification) จากข้อมูลของประชากรกลุ่มใหญ่เพื่อให้ได้ข้อมูลกลุ่มย่อยที่มีลักษณะหรือคุณสมบัติของข้อมูลในกลุ่มย่อยที่คล้ายคลึง (Similarity) กันภายในกลุ่ม แต่มีความแตกต่างกันระหว่างกลุ่มย่อย ซึ่งได้มีการศึกษาเกี่ยวกับการจัดกลุ่มข้อมูลเพื่อนำไปใช้ประโยชน์ในหลายสาขาวิชา เช่น สาขาวิชาชีววิทยาใช้ในการจำแนกประเภทของสิ่งมีชีวิตที่ถูกค้นพบใหม่ สาขาวิชาการแพทย์ใช้ในการวิเคราะห์สาเหตุความเจ็บป่วยเพื่อประโยชน์ในการรักษาคนไข้ได้ถูกต้องตามอาการของโรคที่พบในผู้ป่วย สาขาวิชาการบริหาร- ธุรกิจใช้ในการจัดกลุ่มลูกค้า เพื่อประสิทธิภาพในการบริหารความสัมพันธ์กับลูกค้าและส่งเสริมการตลาด สาขาวิชาวิทยาการคอมพิวเตอร์ใช้ในการจัดกลุ่มข้อมูลเว็บเพจ (Webpage Clustering) เพื่อประโยชน์ในการจัดสรรทรัพยากรได้อย่างเหมาะสม และสามารถตรวจสอบความผิดปกติของการสื่อสารข้อมูลผ่านเครือข่าย เป็นต้น

การจัดกลุ่มข้อมูลเป็นเทคนิคหนึ่งในการทำเหมืองข้อมูล (Data Mining) โดยการคาดคะเนลักษณะหรือประมาณค่าที่ชัดเจนของข้อมูลที่จะเกิดขึ้น ซึ่งเป็นการใช้พื้นฐานจากข้อมูลที่ผ่านมาในอดีต และเป็นการหาแบบจำลองเพื่ออธิบายลักษณะบางอย่างของข้อมูลที่มีอยู่จากการจัดกลุ่มให้กับข้อมูล โดยความหมายของการทำเหมืองข้อมูลมีความหลากหลายตามสาขาวิชาที่ได้มีการศึกษาในเรื่องนี้ ทั้งด้านสถิติ (Statistics) ด้านการเรียนรู้ของเครื่อง (Machine Learning) ด้านการรู้จำรูปแบบ (Pattern Recognition) ด้านการใช้เทคนิคทางคณิตศาสตร์และทางสถิติในการทดสอบทฤษฎีทางเศรษฐศาสตร์ (Econometrics) ด้านระบบผู้เชี่ยวชาญ (Expert Systems) เป็นต้น

ปัจจุบันมีงานวิจัยเกี่ยวกับการจัดกลุ่มข้อมูลเป็นจำนวนมาก โดยมีการศึกษาทั้งรูปแบบของอัลกอริทึมใหม่ การปรับปรุงอัลกอริทึมเดิม หรือการรวมอัลกอริทึมที่มีอยู่ เพื่อการจัดกลุ่มข้อมูลได้อย่างเหมาะสมและมีประสิทธิภาพมากที่สุด ทั้งนี้ประสิทธิภาพในการจัดกลุ่มข้อมูลขึ้นอยู่กับปัจจัยหลายประการได้แก่ ความเร็วในการจัดกลุ่มข้อมูล การจัดกลุ่มข้อมูลได้เหมาะสมตามคุณลักษณะและประเภทของข้อมูล ความสามารถในการจัดกลุ่มข้อมูลที่มีขนาดใหญ่ได้ ความสามารถในการจัดกลุ่มข้อมูลได้เหมาะสมกับรูปร่างของข้อมูลประเภทต่างๆ และการจัดกลุ่มข้อมูลโดยไม่ขึ้นกับค่าของข้อมูลที่มีลักษณะผิดปกติ (Noisy Data) หรือค่านอกกลุ่ม (Outlier) ดังนั้นการเลือกอัลกอริทึมที่เหมาะสม จึงมีผลต่อประสิทธิภาพที่ดีในการจัดกลุ่มข้อมูลที่แตกต่างกันไป

อัลกอริทึมเคมีน (K-means Algorithm) เป็นอัลกอริทึมในการจัดกลุ่มข้อมูลแบบแบ่งส่วน (Partitional Algorithms) ที่ได้รับความนิยมสูง เนื่องจากความเร็ว ความง่ายต่อความเข้าใจ และการนำไปใช้ในการอธิบายผลลัพธ์ของกลุ่มข้อมูลได้เป็นอย่างดี แต่ด้วยข้อจำกัดในการกำหนดค่าเริ่มต้น (Initialization) เพื่อเป็นตัวแทน (Centroid) ในการจัดกลุ่มข้อมูลตามจำนวนกลุ่มที่ต้องเป็นแบบสุ่ม ทำให้ได้ผลลัพธ์จากการจัดกลุ่มข้อมูลแต่ละครั้งแตกต่างกันหรือเรียกว่าค่าเหมาะสมเฉพาะที่ (Local Optimal Point) และค่านอกกลุ่มมีผลกระทบต่อผลลัพธ์ในการจัดกลุ่มข้อมูลที่ไม่เหมาะสม รวมถึงในกลุ่มข้อมูลที่ได้จากการจัดกลุ่มที่มีการกำหนดค่าเริ่มต้นแบบสุ่มอาจไม่มีข้อมูลภายในกลุ่มได้

ด้วยเหตุนี้ การกำหนดค่าเริ่มต้นที่เหมาะสมในการจัดข้อมูลให้กับอัลกอริทึมเคมีน จึงเป็นปัญหาที่ควรศึกษาและได้รับการปรับปรุงแก้ไข เพราะเป็นการเพิ่มประสิทธิภาพในการจัดกลุ่มข้อมูลที่มีข้อดีอยู่แล้วให้มีประสิทธิภาพดียิ่งขึ้น ทำให้มั่นใจได้ว่าการจัดกลุ่มข้อมูลด้วยอัลกอริทึมที่นำเสนอนี้ มีประสิทธิภาพในการจัดกลุ่มได้ดีกว่าการจัดกลุ่มข้อมูลที่ใช้การกำหนดค่าเริ่มต้นแบบสุ่ม และสามารถนำไปใช้ประโยชน์ในสาขาวิชาที่เกี่ยวข้องกับการจัดกลุ่มข้อมูลต่อไป

1.2 วัตถุประสงค์ของการวิจัย

ศึกษาและปรับปรุงประสิทธิภาพ การจัดกลุ่มข้อมูลของอัลกอริทึมเคมีนโดยการกำหนดค่าเริ่มต้นด้วยวิธีการตัดแบ่งกลุ่มข้อมูล เปรียบเทียบกับประสิทธิภาพการจัดกลุ่มข้อมูลของอัลกอริทึมเคมีนที่มีการกำหนดค่าเริ่มต้นแบบสุ่ม รวมถึงศึกษาเปรียบเทียบประสิทธิภาพการจัดกลุ่มข้อมูลกับอัลกอริทึม CCIA ในการกำหนดค่าเริ่มต้นสำหรับอัลกอริทึมเคมีน ที่ได้มีผู้ศึกษาวิจัยไว้แล้ว

1.3 ขอบเขตของการวิจัย

1) พัฒนา และออกแบบอัลกอริทึมในการกำหนดค่าเริ่มต้นของการจัดกลุ่มข้อมูลของอัลกอริทึมเคมีนโดยวิธีการตัดแบ่งกลุ่มข้อมูล

2) พัฒนาโปรแกรมเพื่อทดสอบอัลกอริทึมในการกำหนดค่าเริ่มต้นตามข้อ 1

3) ข้อมูลที่ใช้ในการทดลอง ได้แก่

(1) ข้อมูล UCI Repository ในการจัดกลุ่มข้อมูลแบบการเรียนรู้ของเครื่องจากเว็บไซต์ http://www.ics.uci.edu/~mlearn/MLRepository.html จำนวน 10 ชุดข้อมูล ได้แก่ Wine, Iris, Letter, Segmentation, Pendigits, Optdigits, Pima, Glass, Ionosphere, และ Vehicle

(2) ข้อมูล web access log จากเว็บไซต์ http://www.cs.washington.edu/ai/adaptive-data/ ซึ่งเป็นข้อมูลจากการใช้งานของเว็บไซต์ http://machines.hyperreal.org ในช่วง 02/12/97 ถึง 4/30/99

4) สรุปผลการทดสอบ และเปรียบเทียบผลการจัดกลุ่มข้อมูลด้วยวิธีการกำหนดค่าเริ่มต้นแบบสุ่ม รวมทั้งเปรียบเทียบผลการทดลองกับผลการทดลองด้วยวิธีการกำหนดค่าเริ่มต้นด้วยอัลกอริทึม CCIA

1.4 ขั้นตอนการวิจัย

ขั้นตอนการทำวิจัยแบ่งออกเป็น 5 ขั้นตอนดังนี้

1) ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง ได้แก่

(1) ศึกษางานวิจัยต่าง ๆ ที่เกี่ยวข้องกับการจัดกลุ่มข้อมูลทั่วไป

(2) ศึกษางานวิจัยต่าง ๆ ที่เกี่ยวข้องกับการจัดกลุ่มข้อมูลเว็บเพจ และการประเมินประสิทธิภาพของอัลกอริทึมในการจัดกลุ่มข้อมูลเว็บเพจ

2) ศึกษาและวิเคราะห์ เพื่อหาแนวคิดต่าง ๆ ที่เกี่ยวข้องกับการจัดกลุ่มข้อมูลรวมถึงรวบรวมข้อมูลที่ใช้ในทดสอบ และการประเมินประสิทธิภาพของอัลกอริทึม

3) ออกแบบ และพัฒนาอัลกอริทึมสำหรับการตัดแบ่งกลุ่มข้อมูลเพื่อหาค่าเริ่มต้นในการจัดกลุ่มข้อมูลของอัลกอริทึมเคมีน

4) ทดสอบข้อมูลและประเมินประสิทธิภาพของอัลกอริทึมสำหรับการจัดกลุ่มข้อมูลที่นำเสนอ เปรียบเทียบกับอัลกอริทึมเดิมที่มีการกำหนดค่าเริ่มต้นแบบสุ่ม

5) ทดสอบข้อมูล และประเมินประสิทธิภาพของอัลกอริทึมการหาค่าเริ่มต้นที่นำเสนอ และเปรียบเทียบผลกับอัลกอริทึมการกำหนดค่าเริ่มต้นที่มีการนำเสนองานวิจัยไว้แล้ว

6) วิเคราะห์ สรุปผลการวิจัย และข้อเสนอแนะ

1.5 ประโยชน์ที่จะได้รับจากการวิจัย

สามารถเพิ่มประสิทธิภาพในการจัดกลุ่มข้อมูลของอัลกอริทึมเคมีน โดยการกำหนดค่าเริ่มต้นที่เหมาะสม จากเดิมที่มีการกำหนดค่าเริ่มต้นแบบสุ่มซึ่งทำให้ประสิทธิภาพในการจัดกลุ่มด้วยอัลกอริทึมเคมีนมีความไม่แน่นอน เมื่อมีการกำหนดค่าเริ่มต้นที่ดีจึงเป็นการสร้างความเชื่อมั่นในประสิทธิภาพการจัดกลุ่มข้อมูล จากผลการทดลองการจัดกลุ่มข้อมูลโดยวิธีเคมีนและใช้ค่าเริ่มต้นจากอัลกอริทึมที่นำเสนอนี้ จะทำให้ทุกกลุ่มข้อมูลที่ได้มีสมาชิกของกลุ่มอย่างแน่นอน

ในบทต่อไปจะกล่าวถึง ทฤษฎีและงานวิจัยที่เกี่ยวข้องกับการวิจัยนี้

บทที่ 2

การทบทวนวรรณกรรม

ในบทนี้จะกล่าวถึงการศึกษางานวิจัยที่เกี่ยวข้อง ซึ่งประกอบด้วยงานวิจัยและทฤษฎีที่เกี่ยวข้องในการจัดกลุ่มข้อมูล การวัดประสิทธิภาพในการจัดกลุ่มข้อมูล ซึ่งมีรายละเอียดดังต่อไปนี้

2.1 การจัดกลุ่มข้อมูล (Clustering)

การวิเคราะห์การจัดกลุ่มข้อมูล เป็นกระบวนการในการจำแนกวัตถุหรือข้อมูลจากระดับประชากร ให้อยู่ในรูปแบบของข้อมูลประชากรย่อยหรือข้อมูลระดับตัวอย่างที่มีความหมายเพื่อประสิทธิภาพในการอธิบายรูปแบบปัญหา และลักษณะเฉพาะของแต่ละกลุ่มย่อยของข้อมูลได้

ลักษณะทั่วไปในการจัดกลุ่มข้อมูล (Jain and Dubes, 1998) เป็นการจำแนกลักษณะของวัตถุด้วยความสัมพันธ์ระหว่างวัตถุที่แทนด้วยแมตริกซ์ใกล้เคียง (Proximity Matrix) ตามแถวและคอลัมน์ในรูปแบบมิติของข้อมูล (D-dimensional) หรือคุณลักษณะของข้อมูล (Attribute) ซึ่งเป็นการวัดความใกล้เคียงหรือความเหมือนกัน (Similarity) ของวัตถุ โดยวัดระยะห่างของวัตถุตามความเหมือนหรือความแตกต่าง (Dissimilarity) เช่น การวัดระยะห่างแบบ Euclidean การวัดระยะห่างแบบ Hamming และการวัดระยะห่างแบบ Mahalanobis เป็นต้น

2.1.1 ประเภทการแบ่งกลุ่มข้อมูล

การจัดประเภทของการแบ่งกลุ่มข้อมูล (Jain and Dubes, 1998: 56-57) มีดังนี้

2.1.1.1 Exclusive versus Nonexclusive เป็นการจำแนกข้อมูลแบบผูกขาดกับการจัดกลุ่มของข้อมูล (Exclusive) แต่ละวัตถุจะอยู่ในกลุ่มย่อยเดียวเท่านั้น สำหรับการจำแนกข้อมูลแบบไม่ ผูกขาด (Non-exclusive) มีการซ้อนทับกันระหว่างกลุ่มของวัตถุหรืออาจกล่าวได้ว่าวัตถุสามารถอยู่ได้หลายกลุ่ม เช่น การจัดกลุ่มคนโดยอายุ หรือเพศ เป็นแบบ Exclusive ในขณะที่การจัดกลุ่มผู้ป่วยเป็นแบบ Nonexclusive เพราะคนหนึ่งคนอาจป่วยมากกว่าหนึ่งโรคได้ ตัวอย่างอัลกอริทึมที่ใช้ในการจัดกลุ่มข้อมูลแบบ Nonexclusive เช่น Fuzzy clustering เป็นต้น

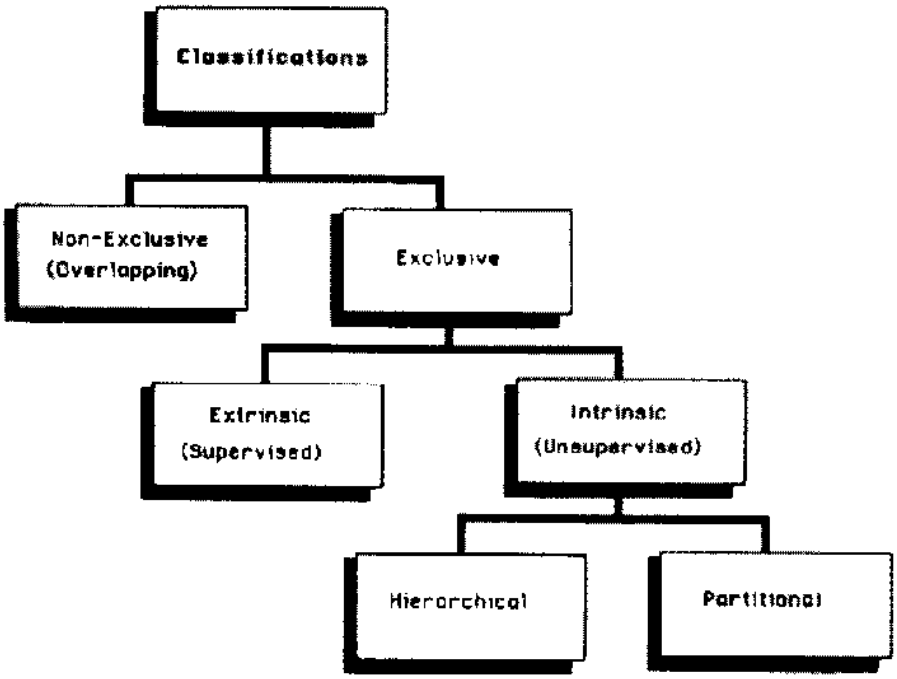
2.1.1.2 Intrinsic versus Extrinsic การจำแนกแบบ Intrinsic ใช้เฉพาะเมตริกซ์ความใกล้เคียง (Proximity Matrix) ในการจัดกลุ่มข้อมูล ที่เรียกว่า "Unsupervised Learning" ในการจดจำรูปแบบ เพราะ ไม่มีการกำหนดค่าของกลุ่มข้อมูลให้กับแต่ละข้อมูลล่วงหน้า แต่สำหรับการจำแนกข้อมูลแบบ Extrinsic เป็นการให้ค่าจากการจัดกลุ่มข้อมูลแก่วัตถุโดยเมตริกซ์ความใกล้เคียง เมื่อจำแนกตามปัญหาแล้ว จะกำหนดหรือจำแนกวัตถุอื่นตามลักษณะของข้อมูลที่ให้ค่าไว้แล้ว การจัดกลุ่มแบบ Extrinsic ยึดหลักข้อมูลที่มีการสอนก่อนที่เรียกว่า "Teacher" หรือ "Supervised Learning" การจำแนกการจัดข้อมูลแบบ Intrinsic หรือ Extrinsic นั้น ได้จากรูปแบบการให้ค่าของวัตถุมีการกำหนดค่าของกลุ่มข้อมูลระหว่างการจัดกลุ่มข้อมูลหรือไม่ หรือมีการจัดคู่หรือให้ค่าความสำคัญของข้อมูลหรือไม่ เช่น การวัดสุขภาพโดยการรวบรวมข้อมูลจากผู้ที่สูบบุหรี่ และไม่สูบบุหรี่ การจำแนกแบบ Intrinsic คือการจัดกลุ่มโดยวัดความคล้ายคลึงกันระหว่างเครื่องชี้วัดสุขภาพของข้อมูลที่มี แล้วอธิบายลักษณะของกลุ่มข้อมูลจากปัจจัยเกี่ยวกับการสูบบุหรี่ และแนวโน้มในการเจ็บป่วยอื่น ๆ ในกลุ่มข้อมูล ขณะที่การจำแนกแบบ Extrinsic เป็นการศึกษาเพื่อวินิจฉัย ผู้สูบบุหรี่จากเครื่องชี้วัดสุขภาพของบุคคลนั้นๆ

2.1.1.3 Hierarchical versus Partitional เป็นการจัดกลุ่มข้อมูลแบบลำดับชั้นกับการจัดกลุ่มข้อมูลแบบแบ่งส่วน ซึ่งเป็นส่วนย่อยของ Intrinsic โดยอัลกอริทึม Hierarchical ทั่วไปได้แก่

1) การรวมกลุ่ม (Agglomerative) เป็นการจำแนกแบบแบ่งลำดับชั้น ที่แต่ละวัตถุหรือข้อมูลเริ่มจากการมีกลุ่มของตัวเอง และค่อย ๆ รวมกับแต่ละกลุ่มของวัตถุเป็นลำดับชั้นเพื่อให้ได้กลุ่มที่ใหญ่ขึ้น จนกระทั่งทุกข้อมูลอยู่ในกลุ่มเดียวกัน

2) การแบ่งแยก (Divisive) เป็นการจำแนกแบบแบ่งลำดับชั้นที่มีขั้นตอนการทำแบบย้อนกลับโดยเริ่มจากทุกวัตถุอยู่ในกลุ่มเดียวกัน และมีการแบ่งย่อยจากกลุ่มใหญ่สู่กลุ่มที่เล็กลง

ปัจจุบันมีวิธีการจัดกลุ่มข้อมูลหลากหลายรูปแบบ ได้แก่ การจัดกลุ่มแบบลำดับชั้น (Hierarchical clustering), การจัดกลุ่มแบบแบ่งส่วน (Partitional clustering), การจัดกลุ่มแบบความน่าจะเป็น (Probabilistic clustering), การจัดกลุ่มแบบกราฟ (Graph based clustering), การจัดกลุ่มแบบฟัซซี่ (Fuzzy clustering), การจัดกลุ่มแบบโครงข่ายประสาทเทียม (Neural Network based clustering) และการจัดกลุ่มในรูปแบบผสม (Hybrid) ซึ่งแต่ละประเภทมีข้อดีและข้อเสียต่างกัน ทั้งด้านความเร็วในการจัดกลุ่มข้อมูล รูปร่างข้อมูล ประเภทของข้อมูล รวมทั้งขนาดของข้อมูล แต่ในงานวิจัยนี้ศึกษาเฉพาะการจัดกลุ่มข้อมูลแบบแบ่งส่วนเท่านั้น



ภาพที่ 2.1 แสดงการแบ่งประเภทข้อมูล (Classification Types)

แหล่งที่มา: Jain and Dubes, 1998: 56.

2.1.2 อัลกอริทึมในการจัดกลุ่มข้อมูลที่ดีควรมีคุณสมบัติดังนี้ (Han and Kamber, 2001: 337)

2.1.2.1 Scalability: มีหลายอัลกอริทึมในการจัดกลุ่มข้อมูลสามารถทำงานได้ดีในข้อมูลน้อยๆ ประมาณ 200 ข้อมูลเป็นต้น แต่สำหรับฐานข้อมูลขนาดใหญ่ที่มีหลายล้านข้อมูล การจัดกลุ่มข้อมูลที่มีจำนวนข้อมูลมาก ๆ อาจทำให้ผลลัพธ์ที่ได้มีความคลาดเคลื่อน อัลกอริทึมที่สามารถจัดกลุ่มข้อมูลที่มีขนาดใหญ่ได้จึงเป็นที่ต้องการสำหรับการจัดกลุ่มข้อมูลที่มีประสิทธิภาพ

2.1.2.2 Ability to Deal with Different Types of Attribute: ความสามารถในการจัดกลุ่มข้อมูลที่มีความหลากหลายของลักษณะข้อมูล อัลกอริทึมส่วนใหญ่ออกแบบสำหรับการจัดกลุ่มข้อมูลที่สามารถนับได้ อย่างไรก็ดีได้มีการประยุกต์ในการจัดกลุ่มข้อมูลประเภทอื่น ๆ อีก เช่น ข้อมูลนามบัญญัติ ข้อมูลแบบกลุ่ม ข้อมูลที่มีการเรียงลำดับ หรือประเภทของข้อมูลที่มีความหลากหลายประเภท

2.1.2.3 Discovery of Cluster with Arbitrary Shape: การค้นหากลุ่มข้อมูลกับรูปร่างของข้อมูลที่หลากหลาย โดยมีหลายอัลกอริทึมที่ใช้การวัดระยะห่างด้วย Euclidean หรือ

Manhattan ซึ่งอัลกอริทึมที่ใช้การวัดระยะห่างจะขึ้นอยู่กับรูปร่างข้อมูลแบบทรงกลม (Spherical) ด้วยขนาดและความหนาแน่นของข้อมูล อย่างไรก็ดี การจัดกลุ่มข้อมูลควรทำได้กับข้อมูลหลายรูปแบบ จึงมีความสำคัญในการพัฒนาอัลกอริทึมที่สามารถจัดกลุ่มข้อมูลที่มีรูปร่างข้อมูลหลายรูปแบบได้

2.1.2.4 Minimal Requirement for Domain Knowledge to Determine Input Parameter: ความต้องการในการกำหนด ขอบเขตค่าพารามิเตอร์ของการป้อนข้อมูลเข้าควรมีน้อยที่สุด

2.1.2.5 Ability to Deal with Noisy Data: ความสามารถในการจัดการเกี่ยวกับข้อมูลที่ผิดปกติ (Noisy Data) ซึ่งในฐานข้อมูลมีค่านอกกลุ่มของข้อมูล ค่าข้อมูลสูญหาย ข้อมูลที่ไม่ทราบ-ค่า ข้อมูลที่มีการป้อนค่าผิดพลาด บางอัลกอริทึมประสิทธิภาพในการจัดกลุ่มลดลงจากค่าเหล่านี้ทำให้คุณภาพในการจัดกลุ่มข้อมูลไม่มีประสิทธิภาพ

2.1.2.6 Insensitivity to the Order of Input Record: การจัดกลุ่มข้อมูลไม่ควรขึ้นอยู่กับลำดับในการนำข้อมูลเข้าประมวลผลในการจัดกลุ่มข้อมูล

2.1.2.7 High Dimensionality: ในฐานข้อมูล หรือ คลังข้อมูลมีจำนวน คุณลักษณะของข้อมูลจำนวนมาก อัลกอริทึมส่วนใหญ่สามารถจัดกลุ่มได้ดีกับจำนวนมิติขนาดเล็ก จึงเป็นความท้าทายของอัลกอริทึมในการจัดกลุ่มข้อมูลขนาดใหญ่ ที่มีมิติของข้อมูลมาก โดยเฉพาะข้อมูลที่มีลักษณะเบาบาง (Sparse) หรือข้อมูลที่มีลักษณะบิดเบี้ยวมาก (Highly Skewed)

2.1.2.8 Interpretability and Usability: ความสามารถในการแปลผล และการนำไปใช้ประโยชน์ได้อย่างครอบคลุมและเหมาะสมกับขอบเขตของข้อมูลที่ศึกษา

2.2 อัลกอริทึมการจัดกลุ่มข้อมูล (Clustering Algorithm) ที่ใช้ในงานวิจัย

2.2.1 อัลกอริทึมเคมีน (K-means Algorithm)

วิธีการของ K-Means Algorithm (Han and Kamber, 2001: 349) เริ่มจากกำหนดจุดที่ต้องการแบ่งข้อมูลแบบสุ่ม กำหนดขอบเขตของการแบ่งข้อมูล จากนั้นทำการเลื่อนจุดให้ไปอยู่กึ่งกลางที่สุดของแต่ละ Cluster เมื่อได้จุดใหม่แล้วให้ทำการลากเส้นกำหนดขอบเขตของ Cluster ใหม่ ซึ่งจะได้ขอบเขตของ Cluster ที่ดีที่สุด โดยมีขั้นตอนดังนี้

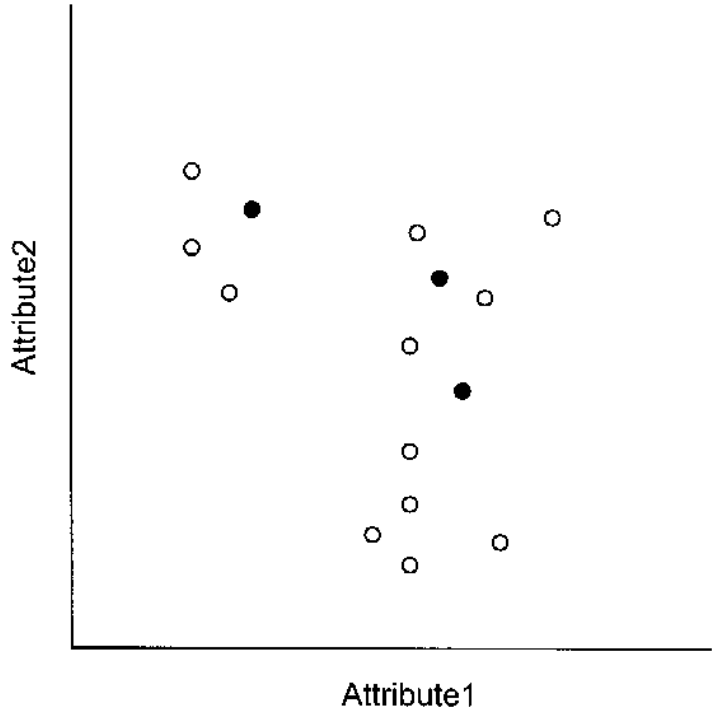
1) เลือกตำแหน่งอย่างสุ่ม จากภาพที่ 2.2 มีการเลือก 3 จุดอย่างสุ่มคือ เลือกที่จะจัดกลุ่มจำนวน 3 กลุ่ม (K=3)

2) สร้างขอบเขตโดยการเชื่อมโยงแต่ละจุดที่เลือกอย่างสุ่ม และแบ่งขอบเขตข้อมูลด้วยเส้นประ ดังภาพที่ 2.3

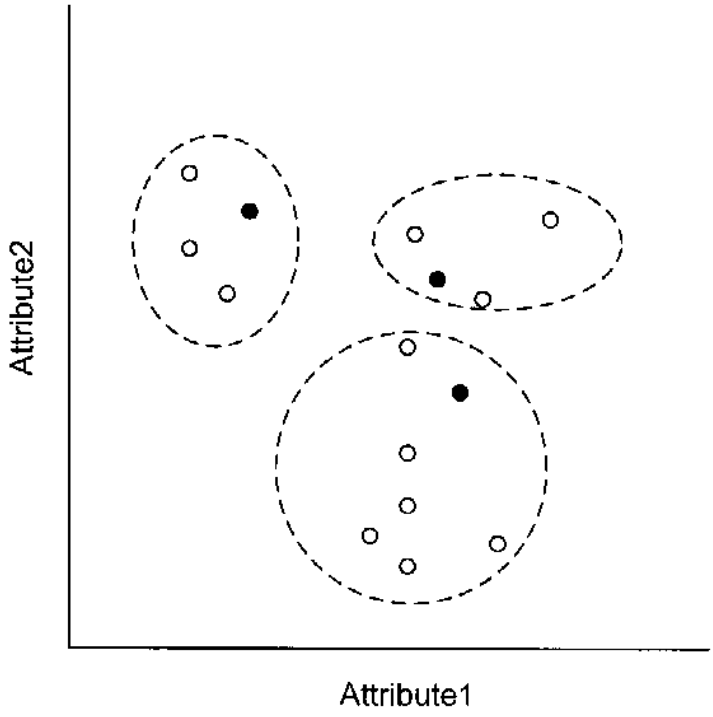
3) เปลี่ยนจุดตัวแทนของกลุ่มตามค่าเฉลี่ยของค่าข้อมูลในแต่ละกลุ่ม ดังรูปภาพที่ 2.4

4) สร้างเส้นแบ่งขอบเขตของข้อมูลตามจำนวนกลุ่มที่กำหนดในข้อ 1

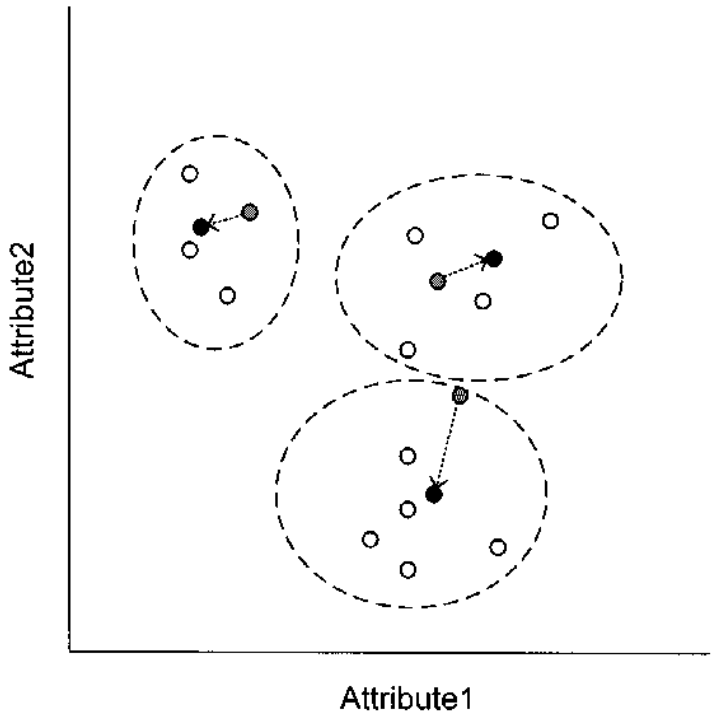
5) ทำซ้ำจนกระทั่งไม่มีการเปลี่ยนแปลงค่าจุดกึ่งกลางของแต่ละกลุ่ม หรือตามจำนวนรอบที่กำหนด



ภาพที่ 2.2 แสดงการเลือกค่าเริ่มต้นสำหรับการแบ่งกลุ่มข้อมูล 3 กลุ่ม (จุดสีดำทึบ)



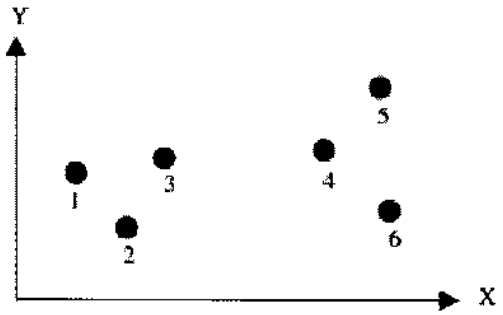
ภาพที่ 2.3 แสดงขอบเขตของข้อมูลในแต่ละกลุ่มตามความใกล้เคียงกับจุดตัวแทนกลุ่ม



ภาพที่ 2.4 แสดงการเปลี่ยนจุดกึ่งกลางหลังจากมีการจัดกลุ่มข้อมูลและขอบเขตใหม่ของกลุ่มข้อมูล

2.2.2 อัลกอริทึมในการจัดแบ่งชั้นสี

เยาวเรศ ศิริสถิตย์กุล (2546:35-42) ได้นำเสนออัลกอริทึมการแบ่งชั้นสีโดยใช้ระยะทางระหว่างสีที่ติดกันตามแกนสีที่มีความแปรปรวนสูงสุด โดยใช้หลักการแบ่งกลุ่มข้อมูลครั้งละสองกลุ่มย่อยที่ทำให้ระยะทางภายในกลุ่มมีค่าน้อยที่สุดและระยะทางระหว่างกลุ่มมีค่ามากที่สุด



ภาพที่ 2.5 แสดงจุดข้อมูล 6 จุด ใน 2 มิติ และตัวเลขที่อยู่ใต้จุดข้อมูลเป็นค่าที่ใช้แสดงลำดับของการจัดเรียง

แหล่งที่มา: เยาวเรศ ศิริสถิตย์กุล, 2546: 38.

ตัวอย่างการจัดกลุ่มจุดข้อมูลสีใน 2 มิติ กำหนดให้มีจุดข้อมูลทั้งหมด 6 จุด และแต่ละจุดถูกจัดเรียงตามค่าในแกน X ของจุด โดยที่ค่าของข้อมูลตามแกน X จะมีค่าความแปรปรวนมากที่สุด เพราะเป็นแกนที่มีแนวโน้มที่ระยะทางระหว่างกลุ่มทั้งสองมีค่ามากที่สุด และยังเป็นการลดความแปรปรวนรวมเมื่อมีการแบ่งกลุ่มตามแกนนั้นๆ อีกด้วย ดังแสดงในภาพที่ 2.5

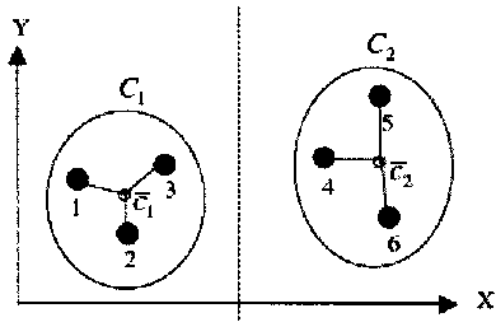
การจัดกลุ่มข้อมูล โดยการแบ่งข้อมูลออกเป็น 2 กลุ่มย่อยที่แยกจากกัน คือ C_1 และ C_2 โดยใช้ระนาบ (Plane) ที่ตั้งฉากกับแกน X เป็นตัวแบ่ง ซึ่ง Plane ที่ใช้ในการแบ่งข้อมูลนี้จะต้องทำให้ผลรวมความผิดพลาดรวมในการจัดข้อมูลของทั้งสองกลุ่มมีค่าน้อยที่สุด กำหนดให้ $\bar{c}_1$ และ $\bar{c}_2$ เป็นจุดกึ่งกลาง (Centroid) ของกลุ่มที่ 1 และกลุ่มที่ 2 ตามลำดับ ผลรวมความผิดพลาดในการแบ่งข้อมูลของกลุ่มที่ 1 สามารถวัดได้จาก

$$\sum_{c_i \in C_1} d(c_i, \bar{c}_1) \tag{2.1}$$

และผลรวมความผิดพลาดในการแบ่งกลุ่มข้อมูลกลุ่มที่ 2 สามารถวัดได้จาก

$$\sum_{c_i \in C_j} d(c_i, \overline{c_2}) \qquad (2.2)$$

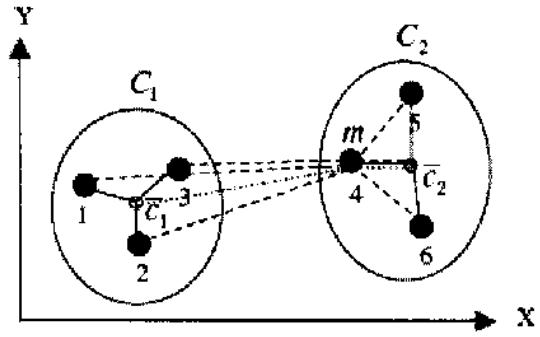
โดยที่ c_i เป็นจุดของข้อมูลลำดับที่ i ของกลุ่ม ดังนั้นการแบ่งด้วย Plane ตามแกน X จะต้องทำให้ผลรวมความผิดพลาดในการแบ่งกลุ่มข้อมูลทั้งสองกลุ่มรวมกัน (2.1 + 2.2) มีค่าน้อยที่สุด แสดงดังภาพที่ 2.6



ภาพที่ 2.6 แสดงการจัดกลุ่มจุดข้อมูล 6 จุดโดยที่เส้นทึบแสดงระยะทางระหว่างข้อมูลกับจุดกึ่งกลางของกลุ่ม

แหล่งที่มา: เยาวเรศ ศิริสถิตย์กุล, 2546: 39.

การแบ่งกลุ่มข้อมูลดังกล่าว สามารถทำได้โดยการเลือก Plane ที่คิดผ่านจุดข้อมูล m ซึ่งเป็นจุดอ้างอิงสำหรับการแบ่งกลุ่มข้อมูล เนื่องจาก $d(c_i,\overline{c_1}) \le d(c_i,c_m)+d(\overline{c_1},c_m)$ และ $d(c_i,\overline{c_2}) \le d(c_i,c_m)+d(\overline{c_2},c_m)$ ดังแสดงในภาพที่ 2.7



ภาพที่ 2.7 แสดงการจัดกลุ่มจุดข้อมูล 6 จุด โดยเลือกจุด m เป็นจุดอ้างอิง โดยที่เส้นทึบแสดงระยะทางระหว่างจุดกึ่งกลางกลุ่ม เส้นประแสดงระยะทางระหว่างจุดอ้างอิง m และจุดภายในกลุ่ม และเส้นประจุด แสดงระยะทางระหว่างจุดอ้างอิง m และจุดกึ่งกลางกลุ่ม

แหล่งที่มา: เยาวเรศ ศิริสถิตย์กุล, 2546: 39.

ดังนั้น

$$\sum_{c_i \in C_1} d(c_i, \bar{c_1}) \le \sum_{c_i \in C_1} d(c_i, c_m) + d(\bar{c_1}, c_m) \cdot |C_1|$$

และ

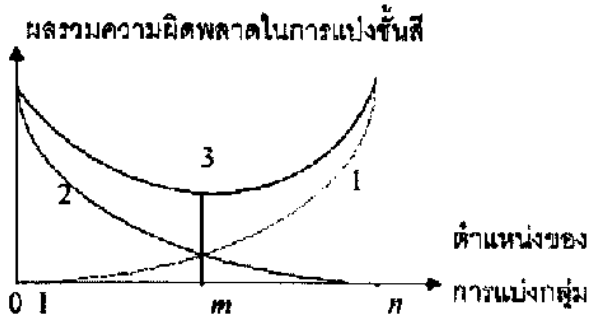
$$\sum_{c_i \in C_2} d(c_i, \bar{c_2}) \le \sum_{c_i \in C_2} d(c_i, c_m) + d(\bar{c_2}, c_m) \cdot |C_2| \qquad (2.3)$$

โดยที่ c_i เป็นจุดของข้อมูลในแต่ละกลุ่ม และ m เป็นจุดอ้างอิงสำหรับแบ่งกลุ่มข้อมูล $|C_1|$ และ $|C_2|$ เป็นจำนวนข้อมูลในกลุ่มที่ 1 และ 2 ตามลำดับ

ดังนั้น เมื่อใช้จุด m เป็นจุดอ้างอิง ผลรวมความผิดพลาดในการแบ่งข้อมูลกลุ่มที่ 1 จะถูกจำกัดด้วยผลรวมความผิดพลาดของข้อมูลในกลุ่มที่ 1 ไปยังจุด m หรือ $\sum_{c_i \in C_1} d(c_i, c_m)$ และผลรวมความผิดพลาดในการแบ่งกลุ่มข้อมูลกลุ่มที่สอง ก็สามารถถูกจำกัดด้วยผลรวมความผิดพลาดของกลุ่มข้อมูลกลุ่มที่ 2 ไปยังจุด m หรือ $\sum_{c_i \in C_2} d(c_i, c_m)$

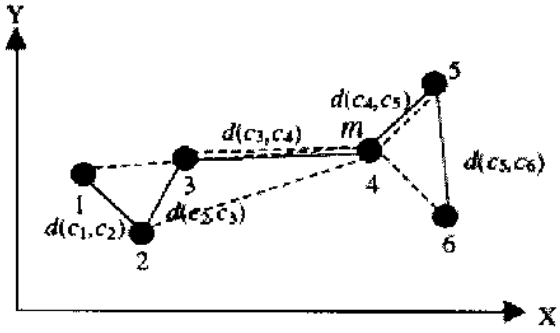
ด้วยเหตุนี้ความพยายามที่จะทำให้ผลรวมความผิดพลาดในการแบ่งข้อมูลทั้งสองกลุ่ม มีค่าน้อยที่สุด สามารถทำได้โดยการพยายามทำให้ผลรวมค่าความผิดพลาดของข้อมูลทั้งหมดไปยังจุดแบ่ง m มีค่าน้อยที่สุด

จากเหตุผลดังกล่าว เราสามารถแสดงความสัมพันธ์ของผลรวมความผิดพลาดในการแบ่งข้อมูล เมื่อเทียบกับจุดแบ่ง m ตามภาพที่ 2.8 กำหนดให้แกน X แทนตำแหน่งของการแบ่งกลุ่ม และแกน Y แทนผลรวมความผิดพลาดในการแบ่งกลุ่มข้อมูล จากกราฟ เส้นที่ 1 แสดงผลรวมความผิดพลาดในการแบ่งข้อมูลในกลุ่มที่ 1 ซึ่งจะเห็นได้ว่า เมื่อ m=0 ผลรวมความผิดพลาดในการแบ่งกลุ่มที่ 2 จะมีค่าเท่ากับผลรวมความผิดพลาดในการแบ่งข้อมูลทั้งหมด ในขณะที่ผลรวมความผิดพลาดในการแบ่งข้อมูลในกลุ่มที่ 1 เท่ากับศูนย์ เส้นที่ 2 แสดงผลรวมความผิดพลาดในการแบ่งข้อมูลของกลุ่มที่ 2 ซึ่งจะเห็นได้ว่า เมื่อ m=n ผลรวมความผิดพลาดในการแบ่งข้อมูลของกลุ่มที่ 1 จะมีค่าเท่ากับผลรวมความผิดพลาดทั้งหมด ในขณะที่ผลรวมความผิดพลาดในการแบ่งกลุ่มที่ 2 เท่ากับศูนย์ เส้นที่ 3 ซึ่งเป็นเส้นโค้งคล้ายพาราโบลา แสดงผลรวมความผิดพลาดในการแบ่งข้อมูลของกลุ่มที่ 1 และกลุ่มที่ 2



ภาพที่ 2.8 กราฟแสดงความสัมพันธ์ระหว่างผลรวมความผิดพลาดในการแบ่งกลุ่ม
แหล่งที่มา: เยาวเรศ ศิริสถิตย์กุล, 2546: 40.

จะเห็นได้ว่า จุดที่ต่ำที่สุดของเส้นโค้งพาราโบลา คือ จุดที่เหมาะสมที่สุด (m) ในการแบ่งกลุ่ม ซึ่ง ณ ตำแหน่งนี้ผลรวมความผิดพลาดในการแบ่งกลุ่มเมื่อเทียบกับจุด m ของกลุ่มที่ 1 และกลุ่มที่ 2 มีค่าเท่ากัน

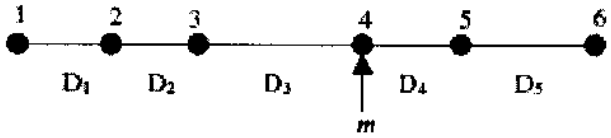


ภาพที่ 2.9 แสดงจุดข้อมูล 6 จุด โดยที่เส้นทึบแสดงระยะทางของจุดที่ติดกัน $d(c_j, c_{j+1})$ และ เส้นประแสดงระยะทางระหว่างจุดอ้างอิง **m** และจุดข้อมูลใดๆ
แหล่งที่มา: เยาวเรศ ศิริสถิตย์กุล, 2546: 41.

การหาจุดแบ่งที่ดีที่สุด ที่ทำให้ผลรวมความผิดพลาดในการแบ่งข้อมูล เมื่อเทียบกับจุด m ของทั้งสองกลุ่มเท่ากัน จะต้องใช้เวลาคำนวณ $O(n^2)$ เพื่อเป็นการลดเวลาในการคำนวณ จึงได้นำระยะทางของจุดที่อยู่ติดกันตามแกน X มาใช้ในการหาจุดแบ่งดังกล่าว โดยมีหลักเกณฑ์ดังนี้

กำหนดให้ $D_j = d(c_j, c_{j+1})^2$ คือ Squared Euclidean Distance ระหว่างจุดที่ติดกันตามแกน X ถ้า i เป็นจุดในกลุ่มที่ 1 $d(c_m, c_i) \le \sum_{j=1}^{m-1} D_j$ ในทำนองเดียวกัน ถ้า c_i เป็นจุดในกลุ่มที่ 2 $d(c_m, c_i) \le \sum_{j=m}^{i} D_j$ ดังแสดงในภาพที่ 2.9

จากสมการ การแบ่งจุด m บน 2 มิติ สามารถลดรูปการหาจุดแบ่ง m บน 1 มิติ ดังแสดงในภาพที่ 2.10



ภาพที่ 2.10 แสดงจุดข้อมูล 6 จุดบนเส้นตรงใน 1 มิติ และส่วนของเส้นตรงที่เชื่อมจุดที่ติดกัน
แหล่งที่มา: เยาวเรศ ศิริสถิตย์กุล, 2546: 41.

ดังนั้นจุดแบ่ง m ที่ดีที่สุด ได้แก่จุด Centroid ของกลุ่มข้อมูลในภาพที่ 2.10 ซึ่งให้ค่า $\sum_{i=1}^{m-1} d(c_m, c_i) \approx \sum_{i=m}^{n} d(c_m, c_i)$ กำหนดให้ $dsum_i = \sum_{j=1}^{i} D_j$ และ $centroidDist$ คือ Centroid ของข้อมูลบนเส้นตรงใน 1 มิติ สามารถคำนวณได้จาก

$$centroidDist = \frac{\sum_{i=1}^{n} dsum_i}{n} \tag{2.4}$$

ดังที่ได้กล่าวมาในตอนต้นแล้วว่า หลักการเลือกแกนเพื่อใช้ในการแบ่งกลุ่มจะเลือกแกนข้อมูลที่มีความแปรปรวนสูงสุด เนื่องจากข้อมูลตามแกนนี้จะมีการกระจายตัวสูงสุด ทำให้เมื่อแบ่งกลุ่มข้อมูลจะมีแนวโน้มที่จะให้ค่า Inter Cluster Distance ระหว่างข้อมูล 2 กลุ่ม สูงกว่าการใช้แกนอื่น

2.3 งานวิจัยที่เกี่ยวข้องกับการกำหนดค่าเริ่มต้นในการจัดกลุ่มข้อมูล

ได้มีการศึกษา และปรับปรุงการกำหนดค่าเริ่มต้นในการจัดกลุ่มข้อมูลของอัลกอริทึมเคมีน เพื่อลดข้อเสียจากการกำหนดค่าเริ่มต้นแบบสุ่ม โดยมีการพัฒนาวิธีการและอัลกอริทึมการหาค่าเริ่มต้นในการจัดกลุ่มข้อมูลอย่างต่อเนื่อง ตั้งแต่ปี ค.ศ. 1965 จนถึงปัจจุบัน ซึ่งพอสรุปได้ดังนี้

Forgy (1965) ได้นำเสนอการเลือกค่าเริ่มต้น K กลุ่มอย่างง่าย โดยเลือกค่าจุดเริ่มต้นแบบสุ่มในฐานข้อมูล (อ้างถึงใน Anderberg, 1973) ถ้าค่าที่เลือกใกล้กับค่ากึ่งกลางตามความหนาแน่นของข้อมูลในแต่ละกลุ่มแล้ว ผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูลจะมีประสิทธิภาพ แต่การเลือกค่า

เริ่มต้นด้วยวิธีการนี้ไม่สามารถรับประกันได้ว่า ค่าเริ่มต้นที่ได้มีความเหมาะสมหรือใกล้ค่ากึ่งกลางของข้อมูลจริง ถึงแม้ว่าใช้การเลือกค่าซ้ำหลาย ๆ รอบ

McQueen (1967) ได้นำเสนอวิธีการที่คล้ายกับ Forgy โดยเริ่มจากเลือกค่า K จุดแบบสุ่ม และคำนวณค่าสมาชิกของกลุ่มข้อมูลที่ใกล้กับจุดค่าเริ่มต้นที่เลือก คำนวณตัวแทนของกลุ่มใหม่เรียกว่า Centroid โดยการคำนวณค่าเฉลี่ยของข้อมูลในแต่ละกลุ่ม จากนั้นหาสมาชิกของแต่ละกลุ่มกับค่า Centroids ที่ได้ และคำนวณค่า Centroids ใหม่ ทำซ้ำจนกระทั่งสมาชิกทุกตัวไม่มีการเปลี่ยนแปลงกลุ่มข้อมูล ซึ่งในฐานข้อมูลขนาดใหญ่วิธีนี้จะมีความยุ่งยากมากในการนำไปใช้งานและใช้เวลาในการคำนวณค่าทุกครั้งที่มีการเปลี่ยนกลุ่มข้อมูล

Tou และ Gonzales (1974) ได้นำเสนอวิธีการ Simple Cluster Seeking (SCS) โดยกำหนดค่าเริ่มต้นค่าแรกในฐานข้อมูล คำนวณระยะห่างระหว่างจุดแรกและจุดถัดไป ซึ่งถ้าค่าที่ได้มากกว่าขอบเขตที่กำหนด (Threshold) จะเลือกจุดนั้นเป็นจุดเริ่มต้นที่สอง แต่ถ้าค่าที่ได้น้อยกว่าขอบเขตที่กำหนด จะเลือกจุดอื่นๆต่อไป เมื่อได้จุดเริ่มต้นสองจุดแล้ว เลือกค่าจุดที่สามโดยคำนวณระยะห่างของจุดที่เป็นค่าเริ่มต้นสองจุดแรกกับจุดใด ๆ ที่มีระยะห่างมากกว่าค่าขอบเขตที่กำหนดเป็นจุดที่สามในการกำหนดค่าเริ่มต้น ทำเช่นนี้จนกระทั่งได้จุดค่าเริ่มต้นในการจัดกลุ่มข้อมูลที่ต้องการ (K) ซึ่งวิธีการนี้ขึ้นอยู่กับความเหมาะสมของการเลือกลำดับของจุดในฐานข้อมูล และค่าขอบเขตที่กำหนด

Linde, Buzo และ Gray (1980) ได้นำเสนอวิธีการ Binary Splitting (BS) โดยใช้การกำหนด Vector Quantiser Codebooks เรียกว่า อัลกอริทึมเหมือนแบบลำดับชั้น (Hierarchical Use of The K-means Algorithm) ในลำดับแรก $K=1$ คำนวณค่าตัวแทนกลุ่ม (Centroid) จากนั้นแบ่งข้อมูลเป็นสองกลุ่มตาม $c+\epsilon$ และ $c-\epsilon$ โดยที่ ϵ คือค่าเวกเตอร์แบบสุ่มขนาดเล็ก (Some Small Random Vector) จากนั้นทำการแบ่งกลุ่มข้อมูลเช่นนี้จนกระทั่งได้จำนวนกลุ่มข้อมูลที่ต้องการ วิธีการนี้ใช้การคำนวณค่อนข้างซับซ้อน อีกทั้งการแบ่งกลุ่มข้อมูลขึ้นอยู่กับการกำหนดค่า ϵ ที่เหมาะสมด้วย

Kaufman และ Rousseeuw (1990) ได้เสนอวิธีการเลือกจุดเริ่มต้น ด้วยจุดแรกที่เป็นศูนย์กลางของข้อมูลมากที่สุด และเลือกจุดที่สองจากจุดใด ๆ ที่ทำให้ผลรวมระยะห่างกำลังสอง (Distortion) ของข้อมูลกับ จุดแรกและจุดใด ๆ มีค่าน้อยที่สุด และทำซ้ำตามขั้นตอนดังกล่าวจนกระทั่งได้จำนวนกลุ่มที่ต้องการ ซึ่งข้อเสียของวิธีนี้คือใช้การคำนวณมากในการหาค่าจุดเริ่มต้นแต่ละครั้ง แต่วิธีการนี้เหมาะสมสำหรับจำนวนกลุ่มข้อมูลตัวอย่างจากฐานข้อมูลที่มีจำนวนน้อย

Babu และ Murty (1993) ได้นำเสนอการเลือกจุดเริ่มต้นที่ใกล้กับการหาค่าเหมาะสมในการจัดกลุ่มข้อมูลโดยใช้ Genetic Programming โดยที่ค่าของเซตของจุดเริ่มต้นจำนวนหนึ่งซึ่งอยู่ในโดเมนของการค้นหา (Population) คำนวณค่าความสมบูรณ์ (Fitness) หรือฟังก์ชันจุดประสงค์โดย

ใช้อัลกอริทึมเคมีนจากการเลือกจุดจำนวนหนึ่งที่มีค่า Distortion ที่เหมาะสม จากนั้นจะสร้างประชากรรุ่นต่อไปที่ดีขึ้น โดยการทำ Crossover และ Mutation และกระทำซ้ำไปเรื่อย ๆ จนกระทั่งได้เกณฑ์ที่ใช้ในการหยุดการกระทำซ้ำ อัลกอริทึมนี้ขึ้นอยู่กับจำนวนขนาดของประชากร และความน่าจะเป็นของ Crossover และ Mutation อีกทั้งการหาคำตอบที่เหมาะสมจากการทำซ้ำของอัลกอริทึมเคมีนในแต่ละรุ่นของประชากรอาจใช้เวลานาน ในกรณีที่ฐานข้อมูลมีขนาดใหญ่

Huang และ Harris (1993) ได้นำเสนอวิธีการ Direct Search Binary Splitting (DSBS) วิธีการนี้คล้ายกับ Binary Splitting Algorithm (Linde, Buzo and Gray, 1980) ยกเว้นการเพิ่มประสิทธิภาพในขั้นตอนการแบ่งแยกข้อมูล โดยใช้ Principle Component Analysis (PCA) ในการเลือกค่าเวกเตอร์ $\in$ ที่เป็นการปรับปรุงคุณภาพของการจัดกลุ่มข้อมูล

Katsavounidis, Kuo และ Zhen (1994) ได้นำเสนอ KKZ Algorithm โดยเริ่มจากการหาจุดเริ่มต้นจุดแรก ซึ่งควรเป็นค่าที่อยู่บริเวณริมหรือขอบของกลุ่มข้อมูล (Edge) จุดที่มีระยะห่างมากที่สุดจากจุดแรกเป็นจุดเริ่มต้นที่สอง คำนวณระยะห่างระหว่างจุดใด ๆ กับจุดเริ่มต้นทั้งสองจุด เลือกจุดที่ใกล้กับสองจุดแรกมากที่สุดแต่มีระยะห่างมากกว่าเป็นจุดเริ่มต้นที่สาม และทำซ้ำโดยการเลือกจุดที่ระยะห่างมากที่สุดจากจุดต่าง ๆ ที่ใกล้กับจุดเริ่มต้น K จุดมากที่สุดจนกระทั่งได้จำนวนค่าเริ่มต้นที่ต้องการ สำหรับข้อมูลที่มีค่านอกกลุ่ม (Outlier) มีผลกระทบต่อการเลือกค่าเริ่มต้นของอัลกอริทึมนี้ ดังนั้นการกำจัดค่านอกกลุ่มจึงทำให้การจัดกลุ่มข้อมูลมีประสิทธิภาพดีขึ้น

Daoud และ Roberts (1996) ได้นำเสนอวิธีการในการจัดกลุ่มข้อมูลโดยการแบ่งข้อมูลเป็น M disjoint subspace (S_j, $j=1, 2, 3..., M$) โดยที่ N_j คือจำนวนของจุดในแต่ละ subspace และ $K_j = K (N_j/n)$ เป็นค่าของจุดเริ่มต้นใน j^{th} subspace

Bradley และ Fayyad (1998) ได้กล่าวถึงการกำหนดค่าเริ่มต้นการจัดกลุ่มข้อมูลของอัลกอริทึมเคมีน โดยใช้ Refinement Algorithm ซึ่งเป็นการเลือก J ตัวอย่างย่อย (Sub sample) ของข้อมูลทั้งหมด คำนวณการจัดกลุ่มข้อมูลด้วยเคมีนในแต่ละกลุ่มย่อยได้ Centroids ของแต่ละกลุ่มตัวอย่างย่อย คือ CM_i , $i=1,2,3,...,J$ ใช้ค่า CM_i เป็นค่าเริ่มต้นในการจัดกลุ่มค่า Centroids ทั้งหมดด้วยเคมีน เพื่อหาค่า FM หรือ Distortion น้อยที่สุด และใช้ Centroids ที่ได้เป็นค่าเริ่มต้นในการจัดกลุ่มข้อมูลด้วยอัลกอริทึมเคมีน

Likas, Vlassis และ Verbeek (2003) ได้นำเสนออัลกอริทึม The Global K-means Clustering และใช้การหาค่าเริ่มต้นสำหรับอัลกอริทึมเคมีน ด้วย Kd-trees ในการแบ่งข้อมูลเป็น Buckets ตามระนาบที่ตั้งฉากกับแกนที่มีความแปรปรวนสูงสุด จากการใช้ Principle Component Analysis (PCA) แบ่งข้อมูลจนกระทั่งได้จำนวน Buckets ที่ต้องการ หรือ ตามจำนวนข้อมูลใน

Buckets ที่กำหนด และใช้ค่า Centroids ในแต่ละ Buckets เป็นค่าเริ่มต้นในการจัดกลุ่มข้อมูลด้วยอัลกอริทึมเคมีน

Khan และ Ahmad (2004) ได้กล่าวถึงการหาค่าเริ่มต้นในการจัดกลุ่มข้อมูลด้วยอัลกอริทึมเคมีน โดยสร้างอัลกอริทึมที่เรียกว่า Cluster Center Initialization Algorithm (CCIA) โดยใช้หลักการจัดกลุ่มข้อมูลด้วยอัลกอริทึมเคมีนแต่ละ Attribute ตามพื้นที่ใต้โค้งการแจกแจงแบบปกติให้เท่าๆกัน ด้วยเปอร์เซ็นไทล์ (Percentiles) ของจุดลำดับที่ s โดยที่ $S=1,2,...,K$ โดยการหาค่า Z_s จากพื้นที่ใต้โค้งการแจกแจงแบบปกติ $-\infty$ ถึง $\frac{2s-1}{2K}$ จากนั้นคำนวณค่าเฉลี่ย (μ_j) และส่วนเบี่ยงเบนมาตรฐาน (σ_j) ของ Attribute ที่ j คำนวณค่าเริ่มต้นในการจัดกลุ่มข้อมูลของ Attribute จาก $x_s = z_s * \sigma_j + \mu_j$ ตามจำนวนกลุ่มข้อมูลที่ต้องการ และ จัดกลุ่มข้อมูล Attribute ด้วยเคมีนจากค่าเริ่มต้นดังกล่าว กำหนดค่าหมายเลขกลุ่มให้แต่ละข้อมูลของ Attribute ตามค่าที่ได้จากการจัดข้อมูลด้วยเคมีน ดำเนินการจัดกลุ่มข้อมูลทุก Attribute จากนั้นสร้างรูปแบบของการเลือกกลุ่มข้อมูลในทุกข้อมูลตาม Attribute และรวมรูปแบบกลุ่มข้อมูล (Pattern Strings) ทุกข้อมูลที่มี รูปแบบกลุ่มข้อมูลเหมือนกัน จะได้ค่ารูปแบบกลุ่มข้อมูลจำนวน K' กลุ่ม คำนวณค่า Centroid ของ K' กลุ่มรูปแบบข้อมูล

CCIA เลือกค่า K' กลุ่มที่จัดเรียงเรียบร้อยแล้ว (รวมรูปแบบกลุ่มข้อมูลที่เหมือนกันของแต่ละข้อมูล) ถ้าค่าที่ได้มากกว่าค่า K $(K'>K)$ จะทำการรวมกลุ่มที่มีความคล้ายกัน ด้วยอัลกอริทึม Density-Based Multi Scale Data Condensation (DBMSDC) เพื่อให้ได้จำนวนกลุ่มเท่ากับ K ที่ต้องการ และคำนวณค่า Centroid ของ K กลุ่ม เพื่อใช้เป็นค่าเริ่มต้นในการจัดกลุ่มข้อมูลต่อไป

2.4 การวัดประสิทธิภาพของการจัดกลุ่มข้อมูลในงานวิจัย

ในงานวิจัยนี้ใช้การวัดประสิทธิภาพการจัดกลุ่มข้อมูลต่าง ๆ ดังนี้

2.4.1 Entropy-Based Measure of Impurity

$$Entropy = -\sum_{j=1}^{c} P_j \log P_j \qquad (2.5)$$

เมื่อ P_j เป็นสัดส่วนข้อมูลของคลาส j ในกลุ่มข้อมูลหนึ่ง สำหรับค่า Entropy ที่น้อยที่สุดเท่ากับศูนย์ แสดงว่า มีตัวข้อมูลเพียง Class เดียว หรือ Impurity น้อย ซึ่งแสดงว่ามีการจัด

กลุ่มข้อมูลที่ดีและข้อมูลในกลุ่มนั้นมีความสมบูรณ์มาก ในกรณีที่กลุ่มข้อมูลมีคลาสข้อมูลกระจายพอ ๆ กัน ค่า Entropy จะมีค่าสูงสุด ซึ่งถือว่า กลุ่มข้อมูลมี Impurity มากหรือ มีการจัดกลุ่มข้อมูลที่ไม่ดี

2.4.2 Sum Squared Error (SSE)

ใช้ในการหาผลรวมของระยะทางกำลังสองระหว่างข้อมูลแต่ละตัว กับค่าตัวแทนกลุ่ม(Centroid) ของข้อมูล เพื่อแสดงว่าผลลัพธ์ที่ได้จากการจัดกลุ่มมีการกระจายมากน้อยเพียงใด ข้อมูลที่มีการจัดกลุ่มที่ดีควรมีค่า SSE น้อย โดยมีสูตรในการคำนวณ ดังนี้

$$SSE = \sum_{j=1}^{n} dist^2(x_j, c_j) \tag{2.6}$$

โดยที่ n คือจำนวนของข้อมูล x_j และ c_j คือค่าตัวแทนกลุ่มข้อมูล (Centroid) ของข้อมูล x_j ค่า SSE ควรมีค่าน้อย เพื่อแสดงถึงความสัมพันธ์ระหว่างข้อมูลกับตัวแทนกลุ่มที่มีระยะทางใกล้กันมากหรืออาจกล่าวได้ว่ามีความสัมพันธ์กันสูง

2.4.3 การวัดความถูกต้องของการจัดกลุ่มข้อมูลเว็บเพจ

วิธีการที่ใช้ในการวัดความถูกต้องของการจัดกลุ่มข้อมูลเว็บเพจ (Perkowitz and Etzioni, 1999) สามารถใช้ค่าที่เรียกว่า Weighted Average Hits ซึ่งสามารถคำนวณได้ดังต่อไปนี้

ให้ V เป็นเซตของผู้ใช้ที่เยี่ยมชมเว็บไซต์ และ V_c เป็นซับเซต (Subset) ของ V ที่สมาชิกหรือผู้ใช้ที่เข้าถึงเว็บเพจในกลุ่มเว็บเพจที่ c อย่างน้อยหนึ่งเว็บเพจ

เซตของเว็บเพจที่ถูกเยี่ยมชมใน c คือ v และ c หรืออาจเขียนได้ว่า $(v \cap c)$ จะได้ค่าเฉลี่ยของฮิต (Hits) จากผู้ใช้ที่เข้าถึงเว็บเพจในแต่ละกลุ่ม โดยการหารด้วยจำนวนกลุ่มที่ได้จากการจัดกลุ่มข้อมูลเว็บเพจทั้งหมด คือ

$$Average\ Percentage\ Hits = \frac{\dfrac{\sum_{v \in V_c} |v \cap c|}{V_c}}{|c|} \tag{2.7}$$

แต่เนื่องจากจำนวนเว็บเพจในแต่ละกลุ่มมีผลต่อค่าเฉลี่ยฮิต ในงานวิจัยนี้จึงใช้ค่าน้ำหนัก (Weight) ของค่าเฉลี่ยร้อยละของฮิตแต่ละกลุ่มด้วยจำนวนของเว็บเพจในแต่ละกลุ่มและจำนวนเว็บเพจทั้งหมด จะได้ค่าเฉลี่ยของฮิตที่ถูกเยี่ยมชมทั้งหมดแบบถ่วงน้ำหนัก คือ

$$Weighted\ Average\ Hits = \frac{\sum_{i=1}^{c} \frac{\sum_{v \in V_{c_i}} |v \cap c_i|}{|V_{c_i}|} \cdot \frac{|w_i|}{|w|}}{|c|} \qquad (2.8)$$

โดยที่ w_i คือจำนวนเว็บเพจในแต่ละกลุ่ม, w คือจำนวนเว็บเพจทั้งหมด, $|v \cap c_i|$ คือ ผู้ใช้ที่เยี่ยมชมเว็บเพจในกลุ่มที่ i , $|V_{c_i}|$ คือจำนวนผู้ใช้ทั้งหมดที่เข้าถึงเว็บเพจในกลุ่มที่ i อย่างน้อยหนึ่งเว็บเพจ และ c คือจำนวนกลุ่มข้อมูลเว็บเพจทั้งหมด

ซึ่งค่าเฉลี่ยฮิตแบบถ่วงน้ำหนัก (Weighted Average Hits) ที่ได้ควรมีค่าสูง แสดงว่าการจัดกลุ่มข้อมูลเว็บเพจได้อย่างถูกต้องและมีประสิทธิภาพตามการเยี่ยมชมเว็บเพจของผู้ใช้

2.4.4 ร้อยละของค่าความผิดพลาดจากการจัดกลุ่มข้อมูล

ในงานวิจัยนี้มีการเปรียบเทียบประสิทธิภาพการจัดกลุ่มข้อมูลของอัลกอริทึมที่นำเสนอกับอัลกอริทึมที่มีการทำวิจัยไว้แล้ว โดยเปรียบเทียบประสิทธิภาพการจัดกลุ่มข้อมูลที่มีการกำหนดกลุ่มไว้แล้ว ตามงานวิจัยดังกล่าว (Khan และ Ahmad, 2004) ซึ่งใช้การคำนวณประสิทธิภาพโดยการหาร้อยละของข้อมูลที่มีการจัดกลุ่มผิดพลาด โดยใช้สูตรคือ

$$Error(Percentage) = \left(\frac{NumberOfMissclassifiedPattern}{TotalNumberOfPatterns} \right) \times 100 \qquad (2.9)$$

โดยที่ Number of Misclassified Pattern คือจำนวนข้อมูลที่มีการจำแนกไม่ถูกต้อง และ Total Number of Patterns คือจำนวนข้อมูลทั้งหมดที่ถูกจำแนก

บทที่ 3

อัลกอริทึมที่นำเสนอและการทดลอง

3.1 อัลกอริทึมที่ใช้ในงานวิจัยการจัดกลุ่มข้อมูลด้วยการกำหนดค่าเริ่มต้นโดยวิธีการตัดแบ่งกลุ่มข้อมูล

การทำงานของอัลกอริทึมการหาค่าเริ่มต้นสำหรับการจัดกลุ่มข้อมูล โดยใช้ระยะทางระหว่างข้อมูลที่ติดกันตามแกนของข้อมูลที่มีความแปรปรวนสูงสุดนั้น ใช้หลักการของการแบ่งชั้นสีตังที่ได้กล่าวมาแล้วในบทที่ 2 เนื่องด้วยแกนที่มีความแปรปรวนสูงสุดมีแนวโน้มที่ระยะทางระหว่างกลุ่มข้อมูลจะมีค่ามากเมื่อมีการแบ่งข้อมูลเป็นสองกลุ่มหรือเซลล์ เพราะมีการกระจายของข้อมูลในแกนที่ถูกเลือกสูง และเป็นการลดความคลาดเคลื่อนจากการแบ่งกลุ่มข้อมูลเพราะการเลือกเซลล์เพื่อทำการแบ่งข้อมูลนั้น จะเลือกตามผลต่างมากที่สุดของความแปรปรวนของเซลล์เดิมกับความแปรปรวนของเซลล์ย่อยสองเซลล์ (จากการแบ่งเซลล์เดิมเป็นสองเซลล์ย่อย) ดังนั้นเมื่อแบ่งเซลล์ดังกล่าวจึงเป็นการลดความคลาดเคลื่อนได้มากที่สุดจากแบ่งกลุ่มข้อมูลในแต่ละครั้ง โดยที่จำนวนกลุ่มข้อมูลจะเพิ่มครั้งละหนึ่งกลุ่มเท่านั้น จากเหตุผลที่กล่าวมานี้เป็นไปตามหลักการจัดกลุ่มข้อมูลที่ดีโดยให้มีระยะทางภายในกลุ่มข้อมูลน้อยที่สุดและระยะทางระหว่างกลุ่มข้อมูลมากที่สุด

การทำงานของอัลกอริทึมการหาค่าเริ่มต้นสำหรับการจัดกลุ่มข้อมูล โดยวิธีการตัดแบ่งข้อมูล มีรายละเอียดดังต่อไปนี้

อินพุท: ข้อมูล $c = \{(R_{ij}) \mid i=1, 2,\ldots n \text{ (Instances)}$ และ $j= 1, 2,\ldots m \text{ (Attributes)}\}$
จำนวนกลุ่มที่ต้องการคือ K และอัลกอริทึมที่ใช้ในการจัดกลุ่มข้อมูล

เอาท์พุท: จำนวนค่าเริ่มต้นในการจัดกลุ่มข้อมูล K กลุ่ม และ จำนวนกลุ่มข้อมูลที่ได้จากการจัดกลุ่มข้อมูลด้วยอัลกอริทึมเคมีน ค่า SSE และ Entropy

ขั้นตอนการทำงาน

1. กำหนดให้ c เป็น Current Cell ที่จะแบ่งข้อมูล

2. จัดเรียงข้อมูลใน c ตามค่าแต่ละ Attribute จากค่าน้อยไปมากและเชื่อมโยงข้อมูลที่เรียงตามค่าแต่ละ Attribute ด้วย Linked list
3. คำนวณความแปรปรวนของแต่ละ Attribute ใน c
4. เลือก Attribute ที่มีความแปรปรวนสูงที่สุดเป็นแกนหลักในการจัดเรียงข้อมูล
5. คำนวณค่า Squared Euclidean Distance ของข้อมูลที่ติดกันโดยใช้ Attribute ที่มีความแปรปรวนสูงสุดเป็นแกนหลัก ตามสูตร $D_i = d(c_i, c_{i+1})^2$ และคำนวณระยะทางรวมจากข้อมูลเริ่มต้นของเซลล์ที่ผ่านการจัดเรียงไปยังข้อมูลลำดับที่ i โดยใช้สูตร

$$dsum_i = \sum_{j=1}^{i} D_j$$

6. คำนวณ *centroidDist* ของกลุ่มข้อมูลจากสูตร

$$centroidDist = \frac{\sum_{i=1}^{n} dsum_i}{n}$$

7. แบ่งเซลล์ c เป็นออกเป็น 2 กลุ่มย่อยตามแกนข้อมูลหลัก ณ จุด centroidDist ทำการ Scan และแยก linked list ของแต่ละ Attribute ให้เป็น 2 ส่วนตามข้อมูลใน 2 กลุ่มย่อย
8. หาค่า Delta Variance โดย นำ Variance ก่อนแบ่งกลุ่มย่อยลบด้วย Variance ของทั้งสองกลุ่มย่อย และนำ c ใส่ใน Max Heap โดยจัดเรียงตามค่า Delta Variance (ในครั้งแรกมีเพียงข้อมูลเดียวกัน จึงไม่มีการจัดเรียงข้อมูล)
9. ทำการลบ cell ที่มี Delta Variance สูงสุดจาก Max Heap และกำหนดให้ cell นั้น เป็น Current cell (c)
10. ตรวจสอบจำนวนสมาชิกของกลุ่มย่อยที่ 1 และสมาชิกของกลุ่มย่อยที่ 2 ของ Current cell และทำการแบ่งกลุ่มย่อยที่ 1 และ 2 ถ้าจำนวนสมาชิกในกลุ่มไม่เท่ากับศูนย์ ตามขั้นตอนในข้อที่ 3 ถึง 7 คำนวณ Delta Variance ของกลุ่มย่อยที่ 1 และ 2 จากนั้นใส่กลุ่มย่อยที่มีจำนวนสมาชิก ใน Max Heap
11. ดำเนินการตามข้อ 8-9 จนกระทั่ง ได้ขนาดของ Max heap เท่ากับจำนวนกลุ่มที่ต้องการคือ K

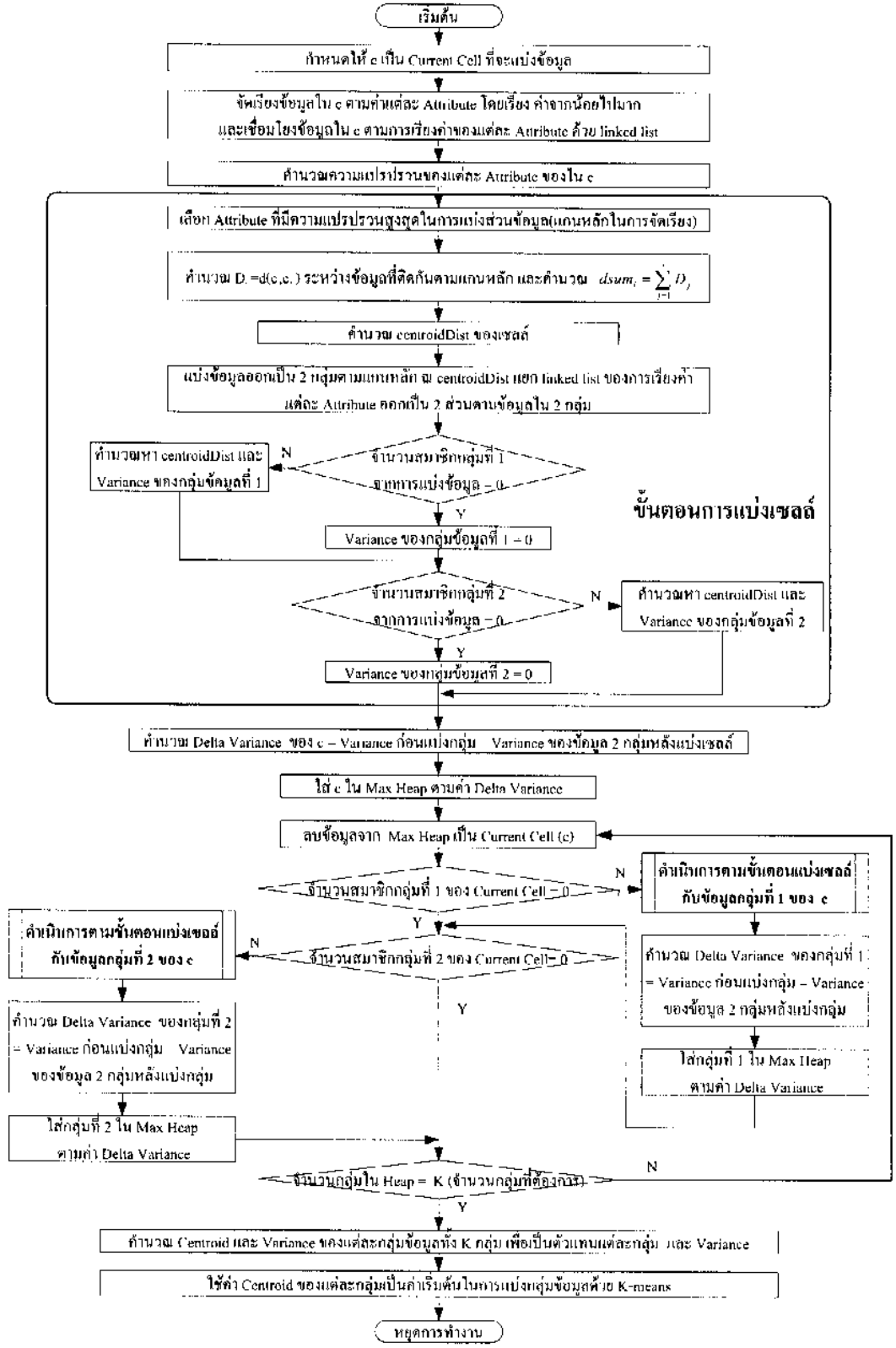
12. คำนวณ Centroid ของแต่ละกลุ่มข้อมูลใน Max heap เพื่อใช้เป็นค่าตัวแทนของกลุ่มข้อมูลเหล่านั้น
13. ใช้ค่าตัวแทนกลุ่มที่ได้ (Centroid) เป็นค่าเริ่มต้นในการจัดกลุ่มข้อมูลด้วยอัลกอริทึมเคมีน

3.2 การวิเคราะห์ความซับซ้อนด้านเวลา ของอัลกอริทึมการหาค่าเริ่มต้นโดยวิธีการตัดแบ่งข้อมูล

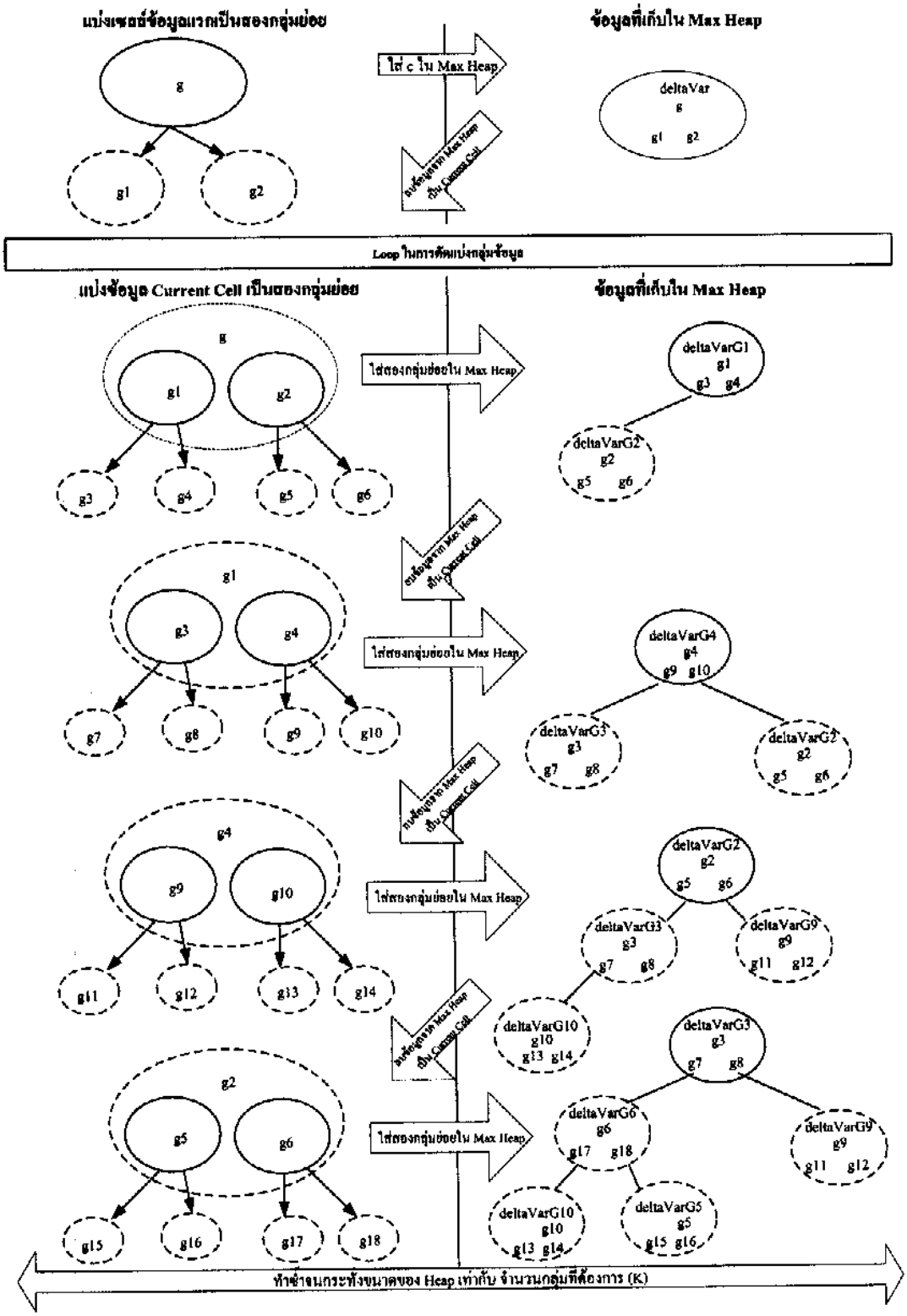
อัลกอริทึมในการหาค่าเริ่มต้นด้วยวิธีการตัดแบ่งข้อมูล โดยใช้ระยะทางระหว่างข้อมูลที่ติดกันตามแกนที่มีความแปรปรวนสูงสุดนั้น เริ่มจากการแบ่งกลุ่มข้อมูลทั้งหมดออกเป็นสองกลุ่มย่อยให้ผลรวมความผิดพลาดในการแบ่งกลุ่ม เมื่อเทียบกับจุด m ของทั้งสองกลุ่มเท่ากัน หลังจากนั้นจะใส่ข้อมูลใน Max heap และลบข้อมูลใน Max heap เพื่อทำการแบ่งกลุ่มข้อมูลสองกลุ่มย่อยเป็นอีกสองกลุ่มย่อย และใส่ข้อมูลที่แบ่งเซลล์แล้วใน Max heap แล้วลบข้อมูลที่มีค่า Delta Variance มากที่สุดมาทำการแบ่งกลุ่มจนกระทั่งข้อมูลใน Max heap ได้เท่ากับจำนวนกลุ่มข้อมูลที่ต้องการจัดกลุ่มข้อมูลด้วย K-means ซึ่งการแบ่งด้วยวิธีนี้จะให้ผลลัพธ์เป็นต้นไม้ทวิภาคแบบสมบูรณ์ (Complete Binary Tree)

ในการแบ่งกลุ่มข้อมูลแต่ละครั้ง จะต้องทำการหา Variance ของ Attribute ในแต่ละกลุ่มย่อย ซึ่งใช้เวลา $O(N)$ โดยที่ N คือจำนวน Attribute ทั้งหมด และเรียงลำดับข้อมูลจากน้อยไปมากตามแกน (Attribute) ที่มีความแปรปรวนสูงสุด ซึ่งใช้เวลา $O(NlogN)$ เพราะใช้การเรียงข้อมูลแบบฮีป (Heap Sort) จากนั้นหาระยะทางระหว่างข้อมูลที่ติดกันตามแกนที่เลือก คำนวณหาจุดแบ่งข้อมูล (Centroid) และทำการแบ่งข้อมูลโดยใช้เวลา $O(N)$ ดังนั้นเวลาในการแบ่งข้อมูลแต่ละครั้งจะใช้เวลา $O(NlogN)$ ซึ่งเวลาที่ใช้ทั้งหมดมีค่าเท่ากับ $O(KNlogN)$

สำหรับการจัดกลุ่มข้อมูลด้วยอัลกอริทึมเคมีนนั้น จะใช้เวลา $O(tkN)$ โดยที่ t คือจำนวนรอบในการทำซ้ำ k คือจำนวนกลุ่มข้อมูล และ N คือจำนวนข้อมูลทั้งหมด ซึ่งเมื่อมีการใช้ค่าเริ่มต้นด้วยอัลกอริทึมที่นำเสนอจะใช้จำนวนรอบในการทำซ้ำน้อยกว่า จำนวนรอบในการทำซ้ำด้วยการกำหนดค่าเริ่มต้นแบบสุ่มโดยเฉพาะอย่างยิ่งเมื่อข้อมูลมีจำนวนมาก เนื่องจากอัลกอริทึมที่นำเสนอจะอยู่ใกล้กลุ่มข้อมูลที่เหมาะสมอยู่แล้ว แต่สำหรับค่าเริ่มต้นแบบสุ่มนั้น จะเริ่มด้วยค่าที่ไม่แน่นอน จึงทำให้บางครั้งใช้จำนวนรอบในการทำซ้ำมาก อีกทั้งคำตอบที่ได้อาจไม่ใช่ค่าที่เหมาะสมในการจัดกลุ่มข้อมูลอีกด้วย



ภาพที่ 3.1 แสดงขั้นตอนการตัดแบ่งข้อมูลเพื่อหาค่าเริ่มต้นสำหรับอัลกอริทึมเคมีน



ภาพที่ 3.2 แบบจำลองการทำงานของวิธีการตัดแบ่งกลุ่มข้อมูลเพื่อหาค่าเริ่มต้นในการจัดกลุ่มข้อมูลด้วยอัลกอริทึมเคมีน

3.3 การทดลอง

3.3.1 การทดลองที่ 1

การหาค่าเริ่มต้นในการจัดกลุ่มข้อมูลด้วยอัลกอริทึมเคมีน ใช้ชุดข้อมูล (Datasct) จำนวน 10 ชุด จาก UCI Repository of Machine Learning Databases (Blake and Merz, 1998) ในการทดสอบประสิทธิผลของอัลกอริทึม ซึ่งมีรายละเอียดของข้อมูลพอสังเขป ดังตารางที่ 3.1

ตารางที่ 3.1 แสดงลักษณะข้อมูลทั่วไปที่ใช้ในงานวิจัย

ชื่อฐานข้อมูล (Dataset)	จำนวนข้อมูล (Instances)	จำนวนมิติ (Feature)	จำนวนกลุ่มข้อมูล (Class)	ร้อยละของ ค่า Continuous หรือ Integer	ร้อยละที่เป็นค่าแบบ Nominal	ค่าสูญหาย
optdigit	5620	64	10	100	0	ไม่มี
pendigit	10992	12	10	100	0	ไม่มี
glass	214	10	7	100	0	ไม่มี
segmentation	2310	19	7	100	0	ไม่มี
ionosphere	351	34	2	100	0	ไม่มี
iris	150	4	3	100	0	ไม่มี
letter	20000	16	26	100	0	ไม่มี
pima-diabetes	768	8	2	100	0	ไม่มี
vehicle	846	18	4	100	0	ไม่มี
wine	178	13	3	100	0	ไม่มี

แหล่งที่มา: ปรับจาก Blake and Merz, 1998.

ซึ่งแสดงลักษณะของข้อมูลทั่วไปที่ใช้ในการทดลองที่ 1 โดยวิธีการลดมิติของข้อมูลด้วย Principle Component Analysis (PCA) ในโปรแกรม Clementine 9.0 ให้แต่ละข้อมูลมีมิติเพียงสามมิติ ดังแสดงด้วยกราฟในภาคผนวก ข

3.3.2 การทดลองที่ 2

Web Access Log เป็นข้อมูลการเข้าถึงวัตถุเว็บต่างๆ ของเว็บไซต์จากผู้ใช้ โดยเก็บข้อมูลไว้ที่ Web Server ในงานวิจัยนี้ใช้แหล่งข้อมูลจากเว็บไซต์ http://www.cs.washington.edu /ai/adaptive-data/ ซึ่งเป็นข้อมูลการใช้เว็บไซต์ http://machines.hyperreal.org ในช่วงเวลา 02/12/97 ถึง 4/30/99 ข้อมูลดังกล่าวเนื้อหาบางส่วนได้ถูกเปลี่ยนแปลง แต่โครงสร้างหลักของเว็บไซต์ยังคงเหมือนเดิม การเก็บข้อมูล Web Log Data ในรูปแบบที่ไม่ระบุชื่อผู้เข้าใช้ ในส่วนของ IP address ถูกสร้างจากการร้องขอของเครื่องให้บริการและถูกเปลี่ยนเป็นตัวเลขที่ไม่มีความหมาย ข้อมูลในแต่ละไฟล์เก็บข้อมูลในแต่ละวัน โดยหมายเลขเดียวกันแต่เก็บข้อมูลต่างวัน ไม่จำเป็นต้องมาจากแหล่งข้อมูลเดียวกัน

รูปแบบ ชื่อไฟล์คือ m.YYMMDD (ซึ่งเป็นไฟล์นามสกุล .gzip) ข้อมูลในแต่ละวันในหนึ่งปีรวมอยู่ในไฟล์เดียวกัน เช่น m.1997.tar เป็นต้น

รูปแบบ Web log Access คือ O:<origin> || T:<time> || U:<url> || R:<referrer>

origin:	แหล่งที่มาของเครื่องให้บริการที่มีร้องขอข้อมูล (ไม่ระบุชื่อ)
time:	เวลาในการร้องขอข้อมูล
url:	url ที่มีการร้องขอ
referrer:	หน้าที่ถูกอ้างอิง (ที่อยู่ของ links ที่ถูกเลือก)

ตัวอย่างเช่น O:00000000000000107040 || T:1998/07/31-10:01:33 || U:/music/machines /categories/software/Mac || R:http://www.hyperreal.org/music/machines/ categories/software/

Log หลังวันที่ 7/31/ 98 รูปแบบของ log จะอยู่ในรูปแบบwww.hyperreal.org| anon0000000000000000000|GET/music/machines/guide/HTTP/1.1|text /html|301|1998/08/24-00:00:02|-|262|-|-|http://www.hyperreal.org/music/machines/|Mozilla/4.0 (compatible; MSIE 4.0; Windows 95)

โดยแต่ละส่วนจะถูกคั่นด้วยเครื่องหมาย "|" และ IP ที่ไม่ระบุที่อยู่จะขึ้นต้นด้วย "anon" และ URL "GET" คือ URL ที่ถูกอ้างอิงถึง

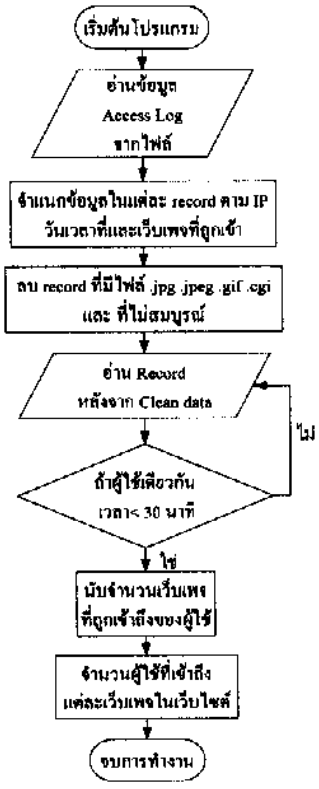
Web Access Log ดังกล่าวถูกใช้ในการงานวิจัยของ Mike Perkowitz และ Oren Etzioni เรื่อง "Adaptive Web Site: Conceptual Cluster Mining" ณ University of Washington ซึ่งงานวิจัยที่นำเสนอนี้ ใช้ข้อมูลและการประเมินประสิทธิภาพการจัดกลุ่มข้อมูลเว็บเพจจากงานวิจัยดังกล่าว

3.4 ขั้นตอนการทดลอง

3.4.1 ขั้นตอนการเตรียมข้อมูลก่อนการวิจัย (Preprocessing)

ข้อมูลสำหรับการทดลองที่ 1 จาก UCI Repository of machine learning databases จำนวน 10 ชุดข้อมูล ซึ่งเป็นข้อมูลจากแหล่งข้อมูลที่มีอยู่แล้ว (Secondary Data) มีลักษณะของข้อมูลในรูปแบบสถิติพื้นฐานทั่วไปโดยไม่มีข้อมูลสูญหายในชุดข้อมูลดังกล่าว จึงเป็นข้อมูลที่สามารถนำไปใช้ในการทดสอบได้โดยไม่ต้องมีการการตรวจสอบข้อมูลก่อน (Data Cleaning)

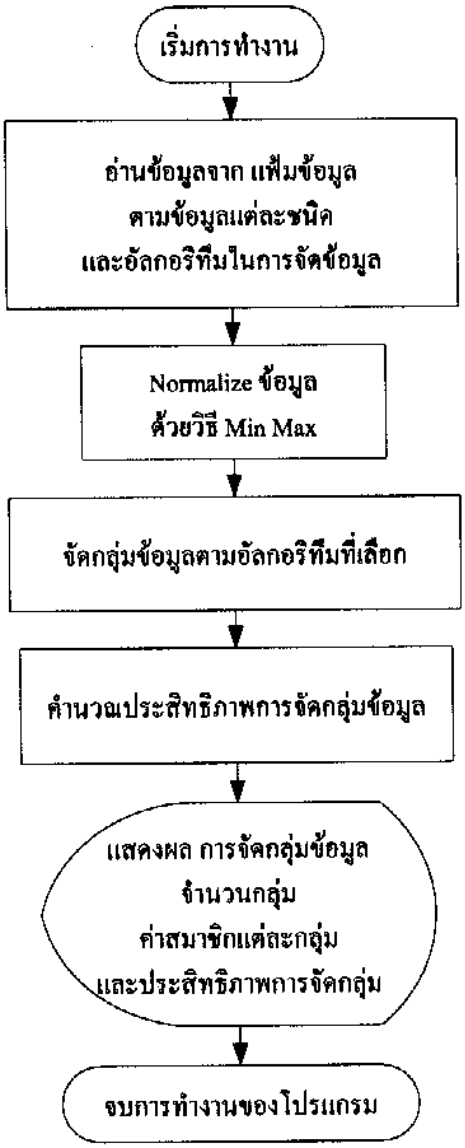
ข้อมูลสำหรับการทดสอบที่ 2 จาก Web Access Log ต้องมีการทำ Data Cleaning ก่อน เนื่องจากวัตถุประสงค์ของงานวิจัยคือทดสอบการแบ่งกลุ่มข้อมูลเว็บเพจ จึงต้องนำข้อมูลที่ไม่เกี่ยวข้องกับการจัดกลุ่มเว็บเพจ ออกจากข้อมูลที่นำไปวิเคราะห์ ได้แก่ ไฟล์ .jpg .jpeg .gif และ .cgi รวมถึงการร้องขอวัตถุเว็บที่ไม่สมบูรณ์ จากนั้นกำหนดเวลาเพื่อแยกผู้ใช้ในแต่ละ Transaction โดยงานวิจัยนี้กำหนดช่วงเวลาการใช้เว็บเพจของผู้ใช้ 30 นาที กล่าวคือหมายเลข IP เดียวกันที่เข้าถึงเว็บเพจในวันเดียวกัน ถ้าใช้เวลาเยี่ยมชมเกิน 30 นาที จะถือว่าเป็นผู้ใช้ต่างกัน จากนั้นทำการนับข้อมูลการเข้าถึงแต่ละเว็บเพจของผู้ใช้ และนำข้อมูลที่ได้ไปทำการแบ่งกลุ่มข้อมูลต่อไป ดังรูปภาพที่ 3.3



ภาพที่ 3.3: แสดงขั้นตอนการ clean data ของ Web Access Log

3.4.2 ขั้นตอนการทดสอบข้อมูลด้วยอัลกอริทึมแคมีนโดยการกำหนดค่าเริ่มต้นแบบสุ่มและการกำหนดค่าเริ่มต้นด้วยวิธีการตัดแบ่งกลุ่มข้อมูล

ขั้นตอนการทดสอบข้อมูล เป็นการนำข้อมูลที่ได้จากขั้นตอนการเตรียมข้อมูล เพื่อจัดกลุ่มข้อมูลในแต่ละอัลกอริทึม ซึ่งการทำงานของโปรแกรมการแบ่งกลุ่มข้อมูลโดยสังเขปเป็นดังนี้



ภาพที่ 3.4: แสดงขั้นตอนการทำงานของโปรแกรมการจัดกลุ่มข้อมูล

3.4.3 ขั้นตอนการประเมินประสิทธิภาพการจัดกลุ่มข้อมูล

การประเมินประสิทธิภาพในการจัดกลุ่มข้อมูลของการทดลองที่ 1 จะใช้ผลรวมความคลาดเคลื่อนกำลังสอง (SSE) ของแต่ละกลุ่มจากค่าของข้อมูลในแต่ละกลุ่มกับตัวแทนกลุ่ม ค่าที่ได้ควรมีค่าน้อย เพราะแสดงถึงระยะห่างหรือการกระจายของข้อมูลภายในกลุ่มที่มีการจัดกลุ่มข้อมูลแล้ว ซึ่งควรมีความสัมพันธ์กันมาก (ระยะห่างระหว่างข้อมูลน้อย) กับข้อมูลภายในกลุ่ม

นอกจากนี้ จะประเมินประสิทธิภาพด้วยการวัดความสมบูรณ์ของการแบ่งกลุ่ม (Entropy-Based Measure of Impurity) ค่าความสมบูรณ์ที่ได้ควรมีค่าน้อยเพื่อแสดงว่าข้อมูลในแต่ละกลุ่มมีความเป็นเนื้อเดียวกัน (Homogenous) คือไม่สามารถแบ่งกลุ่มข้อมูลได้อีก

สำหรับการทดลองที่ 2 ในการจัดกลุ่มข้อมูลเว็บเพจ ใช้การวัดประสิทธิภาพค่าเฉลี่ยฮิตแบบถ่วงน้ำหนัก (Weightcd Average Hits) ของกลุ่มข้อมูลเว็บเพจที่จัดกลุ่มแล้ว ถ้าค่าที่ได้มีค่าสูงแสดงว่าประสิทธิภาพในการจัดกลุ่มข้อมูลเว็บเพจสูง เพราะมีจำนวนผู้ใช้ที่เข้าถึงกลุ่มของเว็บเพจเป็นจำนวนมากและน้อยตามลักษณะของกลุ่มผู้ใช้ที่คล้ายคลึงกัน รายละเอียดการประเมินประสิทธิภาพแต่ละรูปแบบ ได้อธิบายไว้ในบทที่ 2

สำหรับผลการทดสอบการจัดกลุ่มข้อมูลด้วยอัลกอริทึมที่นำเสนอนี้ได้แสดงผลการเปรียบเทียบกับอัลกอริทึมเคมีนที่มีการกำหนดค่าเริ่มต้นแบบสุ่มในบทที่ 4 โดยแสดงประสิทธิภาพของการจัดกลุ่มข้อมูลในแต่ละชุดข้อมูล ทั้งในรูปแบบของตาราง และกราฟ ดังที่จะนำเสนอต่อไป

บทที่ 4

ผลการทดลอง

ผลการทดลองจัดกลุ่มข้อมูลด้วยอัลกอริทึมที่นำเสนอนี้ ใช้เครื่องคอมพิวเตอร์แบบพกพา โพรเซสเซอร์ Intel Core 2 Dual Processor T7200 (2.0 GHz) ประกอบด้วย 2 GB DDR RAM และ ฮาร์ดดิสก์ขนาด 120 GB ทำงานด้วยระบบปฏิบัติการ Microsoft Windows XP Professional และ ภาษาที่ใช้พัฒนาโปรแกรมสำหรับทดลองคือ Visual Studio 6.0 ด้วยภาษาซี

อัลกอริทึมในการจัดกลุ่มข้อมูลทั้งหมดใช้อัลกอริทึมเคมีนที่มีการกำหนดค่าเริ่มต้นแบบสุ่ม และอัลกอริทึมเคมีนที่มีการกำหนดค่าเริ่มต้นด้วยการตัดแบ่งกลุ่มข้อมูล โดยประสิทธิภาพของการจัดกลุ่มข้อมูลต่าง ๆ เป็นดังนี้

4.1 ผลการทดลองที่ 1

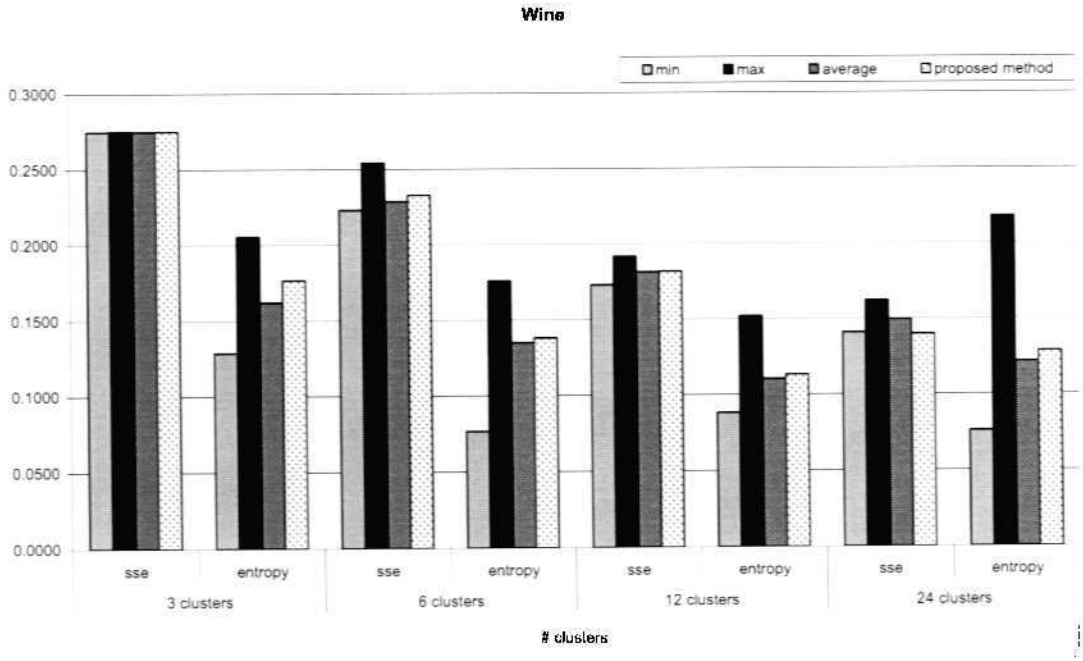
การเปรียบเทียบประสิทธิภาพการจัดกลุ่มข้อมูลในการทดลองที่ 1 ด้วยผลรวมกำลังสองของความคลาดเคลื่อน (SSE) และความสมบูรณ์ในการแบ่งกลุ่มข้อมูล (Impurity Entropy) จากค่าน้อยที่สุด ค่ามากที่สุด และค่าเฉลี่ยของการกำหนดค่าเริ่มต้นแบบสุ่มจำนวน 10 ครั้ง เปรียบเทียบกับการกำหนดค่าเริ่มต้นโดยการตัดแบ่งกลุ่มข้อมูล และเปรียบเทียบเวลาในการจัดกลุ่มข้อมูลที่มีขนาดใหญ่ 3 ชุดข้อมูล ได้แก่ ชุดข้อมูล Letter ชุดข้อมูล Pendigits และ ชุดข้อมูล Optdigits โดยชุดข้อมูลอื่น ๆ ไม่สามารถเปรียบเทียบเวลาในการจัดกลุ่มข้อมูลได้ เนื่องจากใช้เวลาในการจัดกลุ่มข้อมูลที่เร็วมาก ซึ่งได้ผลการทดสอบชุดข้อมูลต่าง ๆ ดังนี้

4.1.1 ข้อมูล Wine

การจัดกลุ่มข้อมูล Wine ด้วยการกำหนดค่าเริ่มต้นโดยการตัดแบ่งกลุ่มข้อมูลจะมีประสิทธิภาพดีกว่า เมื่อเปรียบเทียบประสิทธิภาพที่แย่ที่สุด (ค่า SSE และ Entropy มาก) ของการกำหนดค่าเริ่มต้นแบบสุ่มในทุกกลุ่มข้อมูล และมีประสิทธิภาพดีกว่าการกำหนดค่าเริ่มต้นแบบสุ่มในกรณีที่จำนวนกลุ่มข้อมูลใหญ่ ดังแสดงในตารางที่ 4.1 และกราฟที่ 4.1

ตารางที่ 4.1 แสดงประสิทธิภาพการจัดกลุ่มข้อมูล Wine ด้วย SSE และ Entropy

Data sets		**# clusters**							
		3 clusters		**6 clusters**		**12 clusters**		**24 clusters**	
Wine	**Experiment**	**sse**	**entropy**	**sse**	**entropy**	**sse**	**entropy**	**sse**	**entropy**
Trial #	1	0.2750	0.1553	0.2309	0.1242	0.1918	0.1051	0.1513	0.2178
	2	0.2751	0.1286	0.2232	0.1626	0.1772	0.0881	0.1418	0.1192
	3	0.2750	0.1553	0.2264	0.1134	0.1823	0.1177	0.1474	0.0935
	4	0.2752	0.1428	0.2229	0.1465	0.1803	0.1520	0.1496	0.0957
	5	0.2753	0.1916	0.2233	0.1758	0.1853	0.1020	0.1536	0.1499
	6	0.2751	0.1287	0.2289	0.0764	0.1824	0.1150	0.1453	0.0758
	7	0.2751	0.1672	0.2241	0.1605	0.1734	0.1048	0.1528	0.1715
	8	0.2753	0.2055	0.2543	0.1427	0.1821	0.0992	0.1620	0.0875
	9	0.2752	0.1765	0.2256	0.1537	0.1729	0.1047	0.1503	0.0866
	10	0.2751	0.1672	0.2271	0.0930	0.1860	0.1150	0.1409	0.1167
	min	**0.2750**	**0.1286**	**0.2229**	**0.0764**	**0.1729**	**0.0881**	**0.1409**	**0.0758**
	max	**0.2753**	**0.2055**	**0.2543**	**0.1758**	**0.1918**	**0.1520**	**0.1620**	**0.2178**
	average	**0.2751**	**0.1619**	**0.2287**	**0.1349**	**0.1814**	**0.1104**	**0.1495**	**0.1214**
	proposed method	**0.2752**	**0.1765**	**0.2328**	**0.1380**	**0.1817**	**0.1132**	**0.1397**	**0.1284**



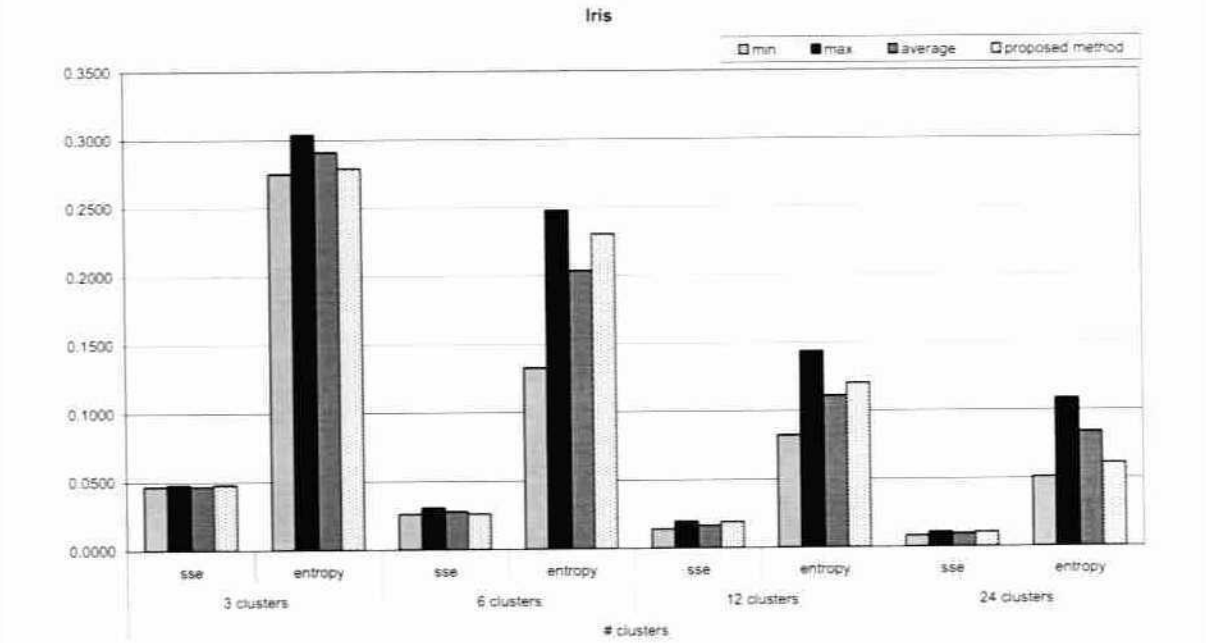
กราฟที่ 4.1 แสดงการเปรียบเทียบประสิทธิภาพการจัดกลุ่มข้อมูล Wine ด้วย SSE และ Entropy ของอัลกอริทึมที่นำเสนอ และอัลกอริทึมแบบสุ่มที่มีการประมวลผล 10 ครั้ง ด้วยค่าน้อยที่สุด มากที่สุด และค่าเฉลี่ย

4.1.2 ข้อมูล Iris

การจัดกลุ่มข้อมูล Iris ด้วยวิธีการกำหนดค่าเริ่มต้นโดยการตัดแบ่งกลุ่มข้อมูล ส่วนใหญ่มีประสิทธิภาพดีกว่าหรือใกล้เคียงกับการกำหนดค่าเริ่มต้นแบบสุ่ม ยกเว้นเมื่อเปรียบเทียบกับประสิทธิภาพที่ดีที่สุดของการเริ่มต้นแบบสุ่ม ดังแสดงใน ตารางที่ 4.2 และกราฟที่ 4.2

ตารางที่ 4.2 แสดงประสิทธิภาพการจัดกลุ่มข้อมูล Iris ด้วย SSE และ Entropy

Data sets		**# clusters**							
		3 clusters		**6 clusters**		**12 clusters**		**24 clusters**	
Iris	**Experiment**	**sse**	**entropy**	**sse**	**entropy**	**sse**	**entropy**	**sse**	**entropy**
Trial #	1	0.0467	0.2896	0.0295	0.1327	0.0196	0.1123	0.0098	0.0943
	2	0.0479	0.2751	0.0296	0.1327	0.0196	0.0910	0.0083	0.0510
	3	0.0467	0.2896	0.0261	0.2303	0.0161	0.1439	0.0096	0.0797
	4	0.0467	0.2896	0.0262	0.2141	0.0142	0.0906	0.0089	0.0734
	5	0.0467	0.2896	0.0307	0.2476	0.0142	0.0824	0.0096	0.0698
	6	0.0467	0.3040	0.0261	0.2303	0.0173	0.1436	0.0095	0.0853
	7	0.0467	0.2896	0.0296	0.1456	0.0158	0.1047	0.0098	0.0812
	8	0.0467	0.3040	0.0284	0.2413	0.0149	0.0974	0.0094	0.1086
	9	0.0467	0.2896	0.0261	0.2303	0.0164	0.1252	0.0109	0.1073
	10	0.0467	0.2896	0.0261	0.2303	0.0166	0.1199	0.0092	0.0893
	min	**0.0467**	**0.2751**	**0.0261**	**0.1327**	**0.0142**	**0.0824**	**0.0083**	**0.0510**
	max	**0.0479**	**0.3040**	**0.0307**	**0.2476**	**0.0196**	**0.1439**	**0.0109**	**0.1086**
	average	**0.0468**	**0.2910**	**0.0278**	**0.2035**	**0.0165**	**0.1111**	**0.0095**	**0.0840**
	proposed method	**0.0479**	**0.2791**	**0.0261**	**0.2303**	**0.0191**	**0.1204**	**0.0106**	**0.0612**



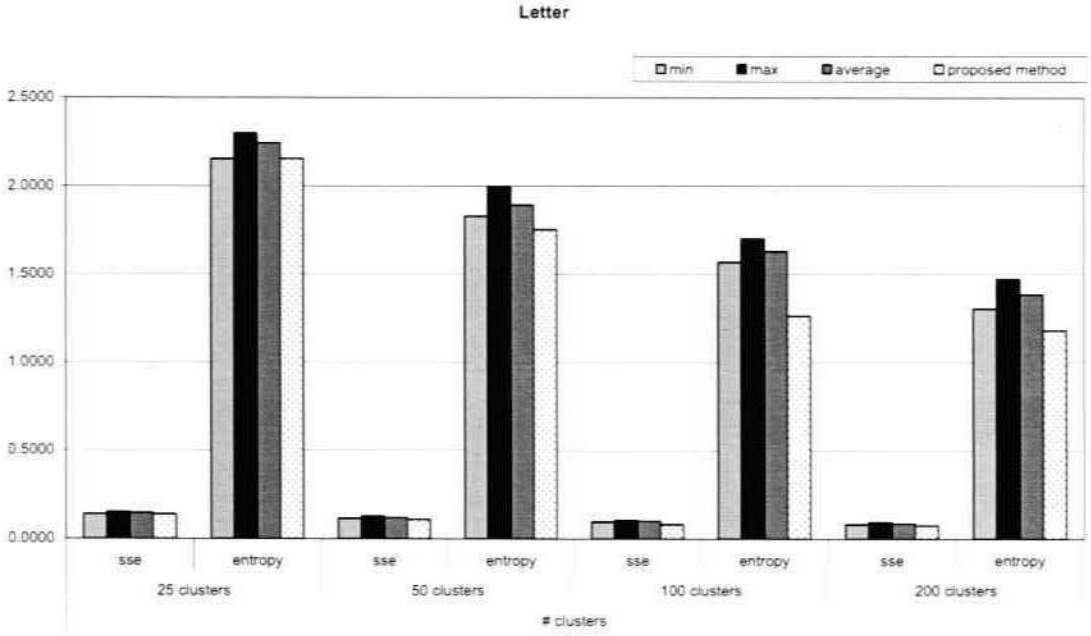
กราฟที่ 4.2 แสดงการเปรียบเทียบประสิทธิภาพการจัดกลุ่มข้อมูล Iris ด้วย SSE และ Entropy ของอัลกอริทึมที่นำเสนอ และอัลกอริทึมแบบสุ่มที่มีการประมวลผล 10 ครั้ง ด้วยค่าน้อยที่สุด มากที่สุด และค่าเฉลี่ย

4.1.3 ข้อมูล Letter

การจัดกลุ่มข้อมูล Letter ด้วยการกำหนดค่าเริ่มต้นโดยการตัดแบ่งข้อมูลจะมีประสิทธิภาพดีกว่า ในทุกกรณี ทั้งกรณีที่แย่สุด กรณีเฉลี่ย และกรณีที่ดีที่สุดของการกำหนดค่าเริ่มต้นแบบสุ่ม ดังแสดงในตารางที่ 4.3 และ กราฟที่ 4.3 อีกทั้งเมื่อเปรียบเทียบเวลาที่ใช้ในการจัดกลุ่มข้อมูลนี้ยังแสดงให้เห็นว่า อัลกอริทึมที่นำเสนอใช้เวลาในการจัดกลุ่มข้อมูลน้อยกว่าการกำหนดค่าเริ่มต้นแบบสุ่ม และมีประสิทธิภาพในการจัดกลุ่มได้ดีกว่าอีกด้วย ดังแสดงในตารางที่ 4.4 และกราฟที่ 4.4

ตารางที่ 4.3 แสดงประสิทธิภาพการจัดกลุ่มข้อมูล Letter ด้วย SSE และ Entropy

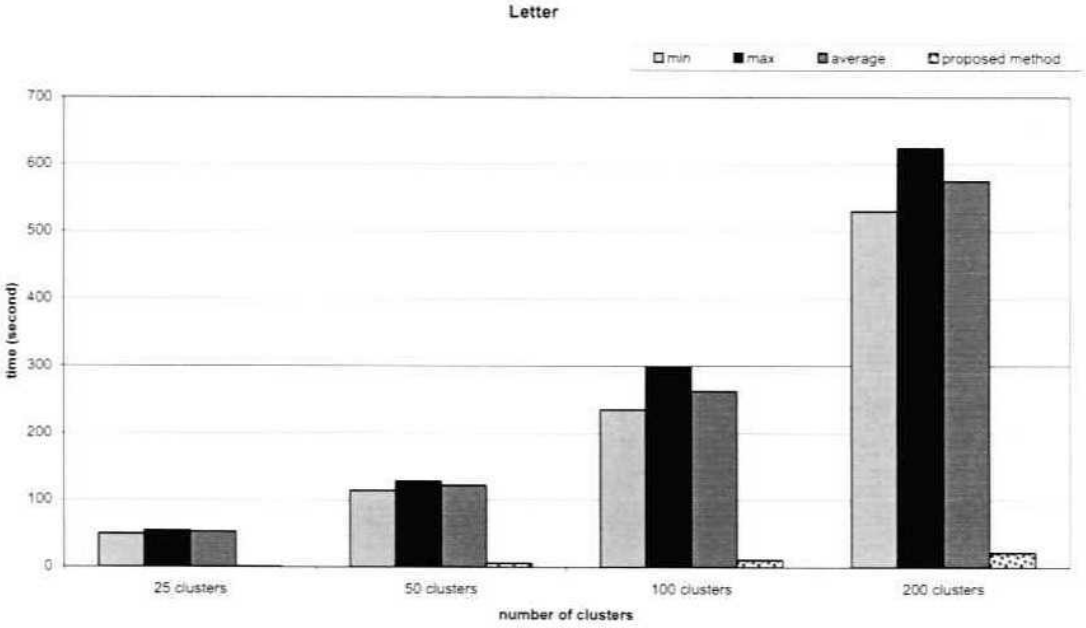
Data sets		**# clusters**							
		25 clusters		**50 clusters**		**100 clusters**		**200 clusters**	
Letter	**Experiment**	**sse**	**entropy**	**sse**	**entropy**	**sse**	**entropy**	**sse**	**entropy**
Trial #	1	0.1480	2.2569	0.1191	1.8760	0.1066	1.6876	0.0896	1.4155
	2	0.1452	2.2233	0.1168	1.8290	0.0991	1.5871	0.0890	1.3925
	3	0.1428	2.2087	0.1228	1.9359	0.1024	1.6589	0.0846	1.3691
	4	0.1445	2.2149	0.1214	1.8924	0.0995	1.6106	0.0859	1.3720
	5	0.1525	2.2991	0.1184	1.8638	0.1056	1.7009	0.0861	1.3768
	6	0.1496	2.2653	0.1288	1.9954	0.0999	1.5846	0.0839	1.3547
	7	0.1508	2.2951	0.1167	1.8498	0.1020	1.6309	0.0951	1.4723
	8	0.1406	2.1527	0.1188	1.8636	0.0961	1.5671	0.0876	1.3955
	9	0.1506	2.2785	0.1220	1.9178	0.1020	1.6518	0.0823	1.3041
	10	0.1455	2.2277	0.1206	1.9053	0.0995	1.6182	0.0858	1.3734
	min	**0.1406**	**2.1527**	**0.1167**	**1.8290**	**0.0961**	**1.5671**	**0.0823**	**1.3041**
	max	**0.1525**	**2.2991**	**0.1288**	**1.9954**	**0.1066**	**1.7009**	**0.0951**	**1.4723**
	average	**0.1470**	**2.2422**	**0.1205**	**1.8929**	**0.1013**	**1.6298**	**0.0870**	**1.3826**
	proposed method	**0.1399**	**2.1544**	**0.1105**	**1.7521**	**0.0825**	**1.2616**	**0.0764**	**1.1796**



กราฟที่ 4.3 แสดงการเปรียบเทียบประสิทธิภาพการจัดกลุ่มข้อมูล Letter ด้วย SSE และ Entropy ของอัลกอริทึมที่นำเสนอ และอัลกอริทึมแบบสุ่มที่มีการประมวลผล 10 ครั้ง ด้วยค่าน้อยที่สุด มากที่สุด และค่าเฉลี่ย

ตารางที่ 4.4 แสดงการเปรียบเทียบเวลาที่ใช้ในการจัดกลุ่มข้อมูล Letter ของอัลกอริทึมที่นำเสนอกับอัลกอริทึมแบบสุ่ม

Data sets		**time (seconds)**			
		25	**50**	**100**	**200**
Letter	**Experiment**	**clusters**	**clusters**	**clusters**	**clusters**
Trial #	1	53	122	235	563
	2	52	122	262	593
	3	53	123	262	571
	4	52	116	299	562
	5	52	122	252	576
	6	51	114	258	566
	7	49	124	267	530
	8	54	128	275	548
	9	54	122	248	618
	10	50	124	266	624
	min	**49**	**114**	**235**	**530**
	max	**54**	**128**	**299**	**624**
	average	**52**	**122**	**262**	**575**
	proposed method	**1**	**6**	**11**	**22**



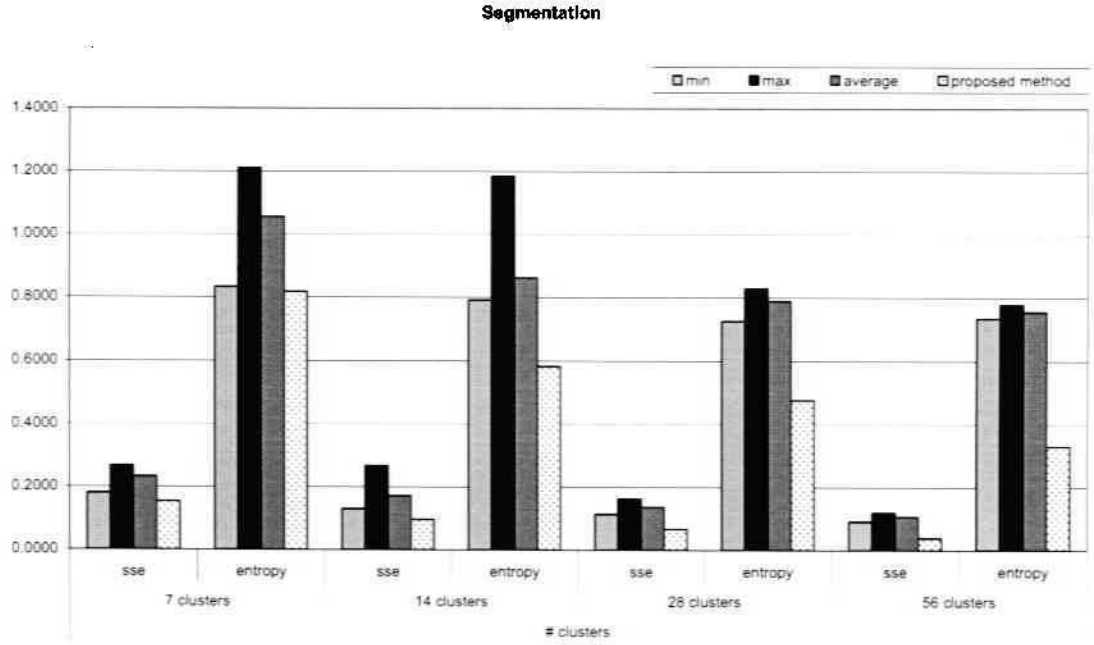
กราฟที่ 4.4 แสดงการเปรียบเทียบเวลาที่ใช้ในการจัดกลุ่มข้อมูล Letter ของอัลกอริทึมที่นำเสนอ และอัลกอริทึมแบบสุ่มที่มีการประมวลผล 10 ครั้ง ด้วยค่าน้อยที่สุด มากที่สุด และค่าเฉลี่ย

4.1.4 ข้อมูล Segmentation

การจัดกลุ่มข้อมูล Segmentation ด้วยการกำหนดค่าเริ่มต้นโดยวิธีการตัดแบ่งข้อมูลมีประสิทธิภาพดีกว่าในทุกกรณีของการกำหนดค่าเริ่มต้นแบบสุ่มที่มีการประมวลผล 10 ครั้ง ดังแสดงในตารางที่ 4.5 และกราฟที่ 4.5

ตารางที่ 4.5 แสดงประสิทธิภาพการจัดกลุ่มข้อมูล Segmentation ด้วย SSE และ Entropy

Data sets		**# clusters**							
Segmenta-		**7 clusters**		**14 clusters**		**28 clusters**		**56 clusters**	
tion	**Experiment**	**sse**	**entropy**	**sse**	**entropy**	**sse**	**entropy**	**sse**	**entropy**
Trial #	1	0.2682	1.1953	0.1811	0.8330	0.1566	0.7919	0.1083	0.7351
	2	0.1811	0.8330	0.1633	0.8536	0.1251	0.7918	0.0916	0.7428
	3	0.2681	1.1845	0.1811	0.8330	0.1495	0.8298	0.1110	0.7485
	4	0.2682	1.2099	0.1632	0.8241	0.1566	0.7919	0.1074	0.7465
	5	0.2682	1.2099	0.1307	0.7918	0.1170	0.7433	0.0986	0.7521
	6	0.2682	1.2099	0.2681	1.1846	0.1632	0.8241	0.1197	0.7780
	7	0.1811	0.8329	0.1307	0.7918	0.1576	0.8241	0.1168	0.7703
	8	0.2682	1.2099	0.1632	0.8531	0.1182	0.7918	0.1081	0.7543
	9	0.1811	0.8329	0.1811	0.8330	0.1168	0.7261	0.1092	0.7543
	10	0.1813	0.8330	0.1632	0.8241	0.1142	0.7762	0.1017	0.7775
	min	**0.1811**	**0.8329**	**0.1307**	**0.7918**	**0.1142**	**0.7261**	**0.0916**	**0.7351**
	max	**0.2682**	**1.2099**	**0.2681**	**1.1846**	**0.1632**	**0.8298**	**0.1197**	**0.7780**
	average	**0.2334**	**1.0551**	**0.1726**	**0.8622**	**0.1375**	**0.7891**	**0.1072**	**0.7559**
	proposed method	**0.1545**	**0.8178**	**0.0967**	**0.5830**	**0.0668**	**0.4764**	**0.0388**	**0.3303**



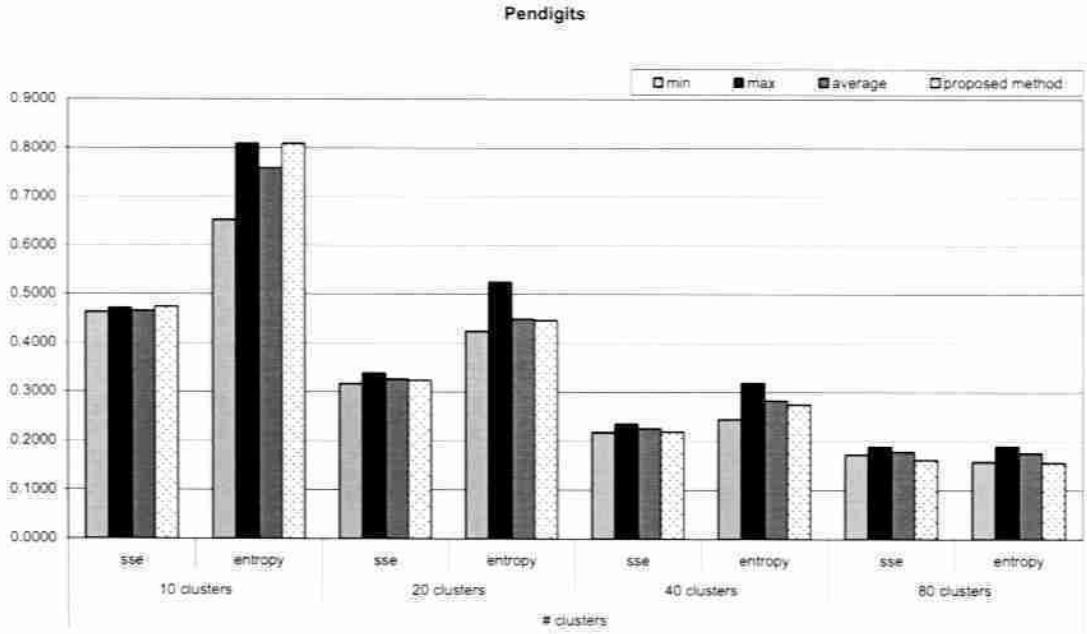
กราฟที่ 4.5 แสดงการเปรียบเทียบประสิทธิภาพการจัดกลุ่มข้อมูล Segmentation ด้วย SSE และ Entropy ของอัลกอริทึมที่นำเสนอ และอัลกอริทึมแบบสุ่มที่มีการประมวลผล 10 ครั้ง ด้วยค่าน้อยที่สุด มากที่สุด และค่าเฉลี่ย

4.1.5 ข้อมูล Pendigits

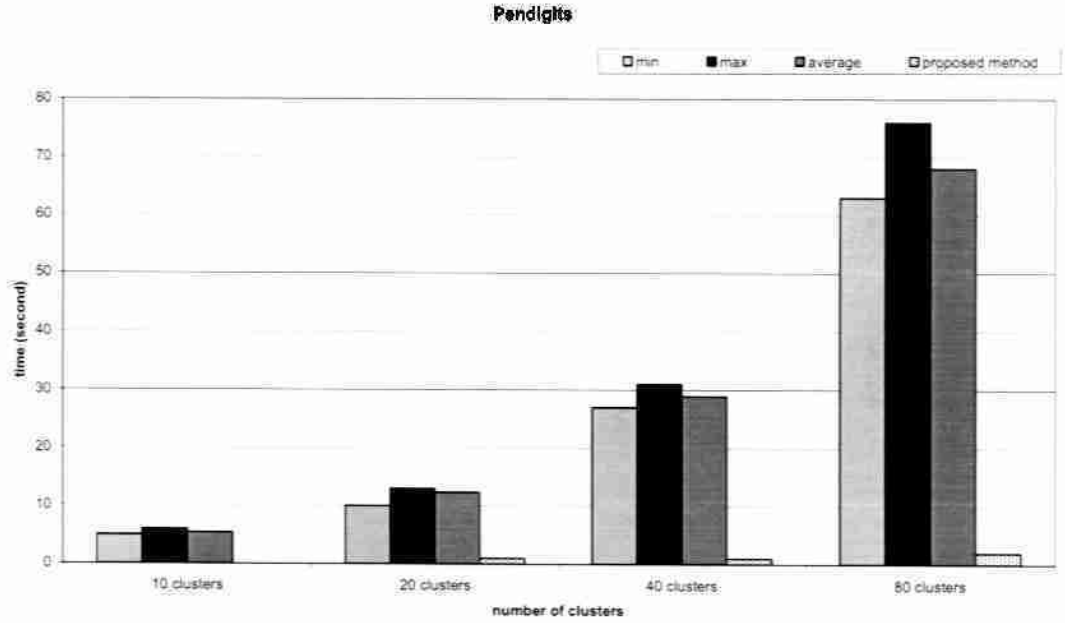
การจัดกลุ่มข้อมูล Pendigits ด้วยการกำหนดค่าเริ่มต้นโดยการตัดแบ่งข้อมูลมีประสิทธิภาพดีกว่า กรณีแย่ที่สุด และกรณีค่าเฉลี่ยของการกำหนดค่าเริ่มต้นแบบสุ่ม สำหรับการจัดกลุ่มข้อมูลขนาดใหญ่นั้น อัลกอริทึมที่นำเสนอมีประสิทธิภาพดีกว่าในทุกกรณีของการกำหนดค่าเริ่มต้นแบบสุ่ม ดังแสดงในตารางที่ 4.6 และกราฟที่ 4.6 แต่เมื่อเปรียบเทียบเวลาในการจัดกลุ่มข้อมูลแล้ว อัลกอริทึมที่นำเสนอจัดกลุ่มข้อมูลได้เร็วกว่าอัลกอริทึมการกำหนดค่าเริ่มต้นแบบสุ่ม

ตารางที่ 4.6 แสดงประสิทธิภาพการจัดกลุ่มข้อมูล Pendigits ด้วย SSE และ Entropy

Data sets		# clusters							
		10 clusters		20 clusters		40 clusters		80 clusters	
Pendigits	**Experiment**	**sse**	**entropy**	**sse**	**entropy**	**sse**	**entropy**	**sse**	**entropy**
Trial #	1	0.4661	0.7188	0.3193	0.4311	0.2243	0.2814	0.1826	0.1877
	2	0.4643	0.7953	0.3315	0.4404	0.2249	0.3060	0.1745	0.1585
	3	0.4708	0.8096	0.3165	0.4241	0.2234	0.2816	0.1891	0.1859
	4	0.4631	0.7479	0.3211	0.4481	0.2351	0.2928	0.1788	0.1726
	5	0.4635	0.7380	0.3208	0.4599	0.2297	0.2839	0.1729	0.1822
	6	0.4631	0.7493	0.3334	0.4583	0.2192	0.2446	0.1731	0.1604
	7	0.4633	0.7897	0.3253	0.4347	0.2273	0.2627	0.1808	0.1737
	8	0.4633	0.7904	0.3316	0.4345	0.2179	0.2740	0.1788	0.1897
	9	0.4644	0.8002	0.3225	0.4326	0.2231	0.2834	0.1798	0.1702
	10	0.4709	0.6528	0.3383	0.5244	0.2336	0.3193	0.1759	0.1785
	min	**0.4631**	**0.6528**	**0.3165**	**0.4241**	**0.2179**	**0.2446**	**0.1729**	**0.1585**
	max	**0.4709**	**0.8096**	**0.3383**	**0.5244**	**0.2351**	**0.3193**	**0.1891**	**0.1897**
	average	**0.4653**	**0.7592**	**0.3260**	**0.4488**	**0.2259**	**0.2830**	**0.1786**	**0.1759**
	proposed method	**0.4736**	**0.8091**	**0.3237**	**0.4464**	**0.2193**	**0.2749**	**0.1617**	**0.1564**



กราฟที่ 4.6 แสดงการเปรียบเทียบประสิทธิภาพการจัดกลุ่มข้อมูล Pendigits ด้วย SSE และ Entropy ของอัลกอริทึมที่นำเสนอ และอัลกอริทึมแบบสุ่มที่มีการประมวลผล 10 ครั้ง ด้วยค่าน้อยที่สุด มากที่สุด และค่าเฉลี่ย



กราฟที่ 4.7 แสดงการเปรียบเทียบเวลาที่ใช้ในการจัดกลุ่มข้อมูล Pendigits ของอัลกอริทึมที่นำเสนอและอัลกอริทึมแบบสุ่มที่มีการประมวลผล 10 ครั้ง ด้วยค่าน้อยที่สุด มากที่สุด และค่าเฉลี่ย

ตารางที่ 4.7 แสดงการเปรียบเทียบเวลาที่ใช้ในการจัดกลุ่มข้อมูล Pendigits ของอัลกอริทึม-ที่นำเสนอ กับอัลกอริทึมแบบสุ่ม

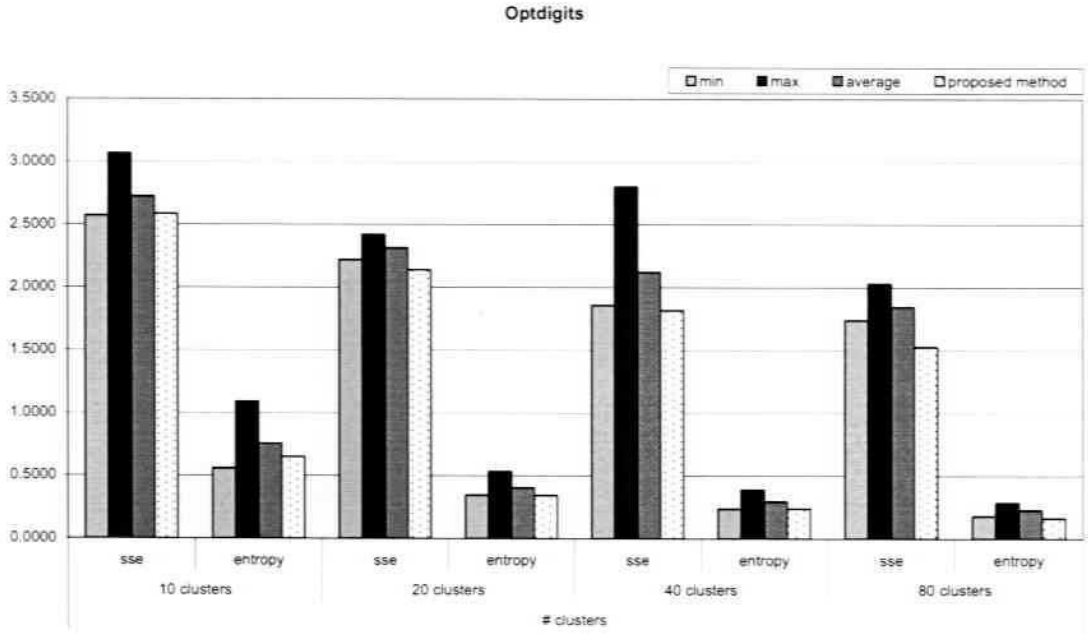
		time (seconds)			
Data sets		**10**	**20**	**40**	**80**
Pendigits	**Experiment**	**clusters**	**clusters**	**clusters**	**clusters**
Trial #	1	5	10	31	65
	2	5	12	30	68
	3	5	13	30	63
	4	6	13	30	69
	5	6	13	28	68
	6	6	13	28	67
	7	5	12	29	65
	8	5	12	27	67
	9	5	12	27	76
	10	6	13	29	73
	min	**5**	**10**	**27**	**63**
	max	**6**	**13**	**31**	**76**
	average	**5**	**12**	**29**	**68**
	proposed method	**0**	**1**	**1**	**2**

4.1.6 ข้อมูล Optdigits

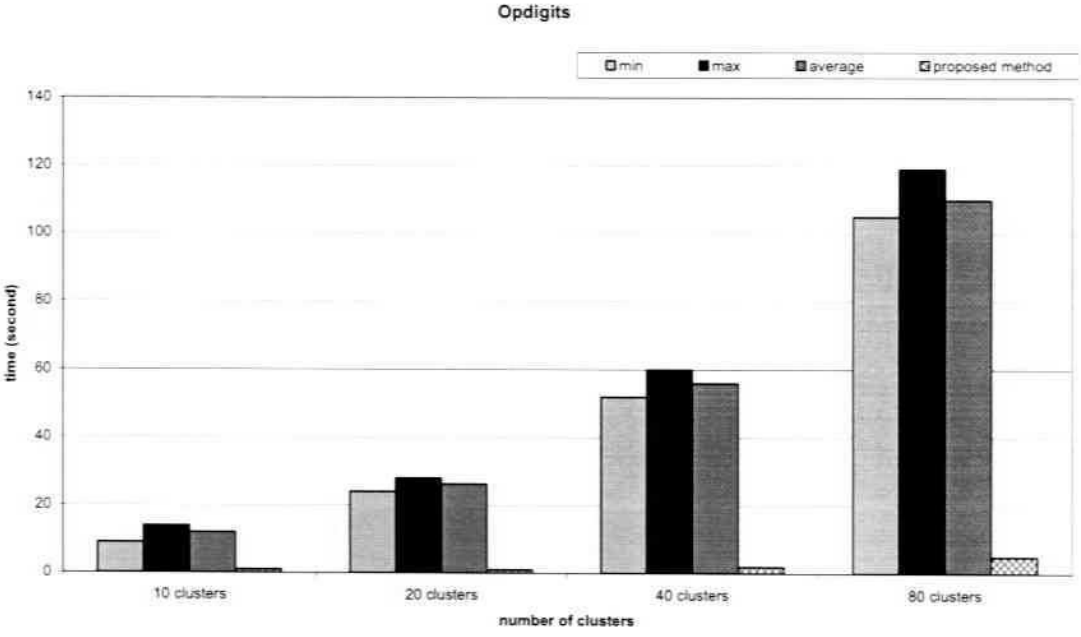
การจัดกลุ่มข้อมูล Optdigits ด้วยการกำหนดค่าเริ่มต้นโดยการตัดแบ่งข้อมูลจะมีประสิทธิภาพดีกว่าการกำหนดค่าเริ่มต้นแบบสุ่มในทุกกรณี ยกเว้นกรณีที่จำนวนกลุ่มข้อมูลขนาด 10 กลุ่ม กรณีดีที่สุดจากการกำหนดค่าเริ่มต้นแบบสุ่มมีประสิทธิภาพดีกว่า ดังแสดงในตารางที่ 4.8 และกราฟที่ 4.8 ซึ่งเมื่อเปรียบเทียบเวลาในการจัดกลุ่มข้อมูลแล้ว อัลกอริทึมที่นำเสนอมีความเร็วในการจัดกลุ่มข้อมูล มากกว่าการกำหนดค่าเริ่มต้นแบบสุ่ม ดังแสดงในตาราง ที่ 4.9 และกราฟที่ 4.9

ตารางที่ 4.8 แสดงประสิทธิภาพการจัดกลุ่มข้อมูล Optdigits ด้วย SSE และ Entropy

Data sets		# clusters							
		10 clusters		**20 clusters**		**40 clusters**		**80 clusters**	
Optdigits	**Experiment**	sse	entropy	sse	entropy	sse	entropy	sse	entropy
Trial #	1	2.6111	0.6624	2.2483	0.3608	2.8026	0.3178	1.9857	0.2872
	2	3.0661	1.0903	2.3156	0.3698	2.1049	0.3024	2.0306	0.2664
	3	2.7291	0.7524	2.3156	0.3671	2.0045	0.3017	1.8064	0.2087
	4	2.7744	0.8236	2.3992	0.4730	2.0476	0.2682	1.8182	0.2088
	5	2.6597	0.6910	2.2199	0.3481	1.9738	0.2593	1.8508	0.2338
	6	2.6598	0.6921	2.4195	0.5341	2.2838	0.3907	1.8544	0.2456
	7	2.7744	0.8276	2.2730	0.3458	2.0900	0.3162	1.7392	0.1806
	8	2.6096	0.6771	2.3705	0.3991	2.1308	0.3142	1.7464	0.2176
	9	2.7877	0.7774	2.3086	0.4744	1.8569	0.2384	1.8496	0.2266
	10	2.5713	0.5579	2.2700	0.3748	1.9161	0.2465	1.7646	0.2012
	min	**2.5713**	**0.5579**	**2.2199**	**0.3458**	**1.8569**	**0.2384**	**1.7392**	**0.1806**
	max	**3.0661**	**1.0903**	**2.4195**	**0.5341**	**2.8026**	**0.3907**	**2.0306**	**0.2872**
	average	**2.7243**	**0.7552**	**2.3140**	**0.4047**	**2.1211**	**0.2955**	**1.8446**	**0.2277**
	proposed method	**2.5844**	**0.6498**	**2.1396**	**0.3439**	**1.8173**	**0.2396**	**1.5274**	**0.1622**



กราฟที่ 4.8 แสดงการเปรียบเทียบประสิทธิภาพการจัดกลุ่มข้อมูล Optdigits ด้วย SSE และ Entropy ของอัลกอริทึมที่นำเสนอ และอัลกอริทึมแบบสุ่มที่มีการประมวลผล 10 ครั้ง ด้วยค่าน้อยที่สุด มากที่สุด และค่าเฉลี่ย



กราฟที่ 4.9 แสดงการเปรียบเทียบเวลาที่ใช้ในการจัดกลุ่มข้อมูล Opdigits ของอัลกอริทึมที่นำเสนอ และอัลกอริทึมแบบสุ่มที่มีการประมวลผล 10 ครั้ง ด้วยค่าน้อยที่สุด มากที่สุด และค่าเฉลี่ย

ตารางที่ 4.9 เปรียบเทียบเวลาที่ใช้ในการจัดกลุ่มข้อมูล Optdigits ของอัลกอริทึมที่นำเสนอ กับอัลกอริทึมแบบสุ่ม

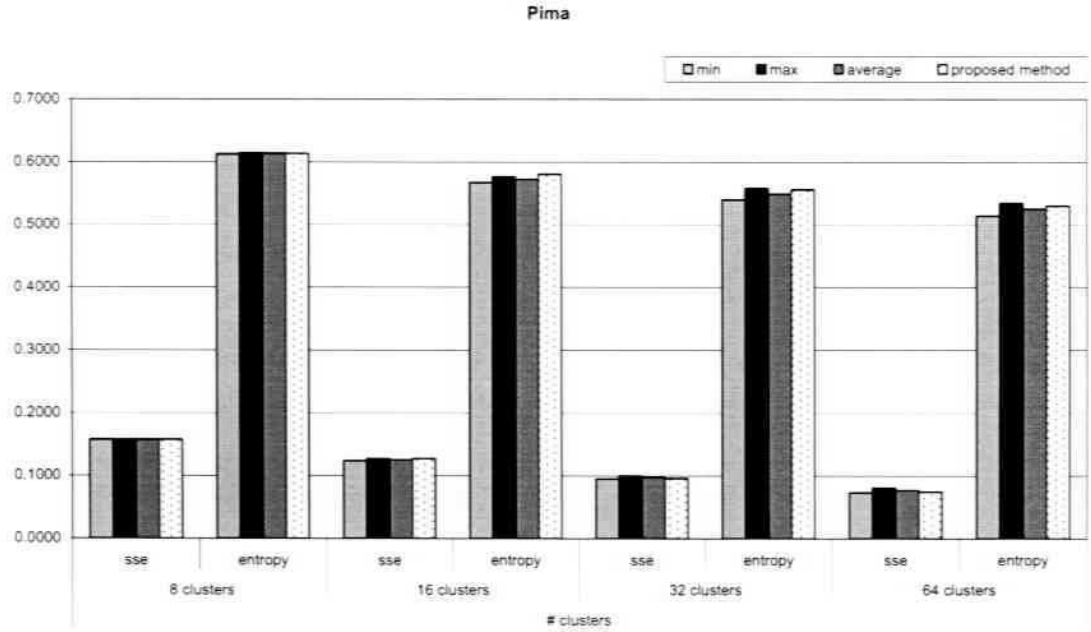
		time (seconds)			
Data sets		**10**	**20**	**40**	**80**
Optdigits	**Experiment**	**clusters**	**clusters**	**clusters**	**clusters**
Trial #	1	14	26	53	107
	2	9	24	58	105
	3	12	24	55	108
	4	12	28	57	119
	5	12	27	56	112
	6	13	27	52	111
	7	11	26	53	114
	8	13	24	57	109
	9	12	28	59	107
	10	12	28	60	107
	min	**9**	**24**	**52**	**105**
	max	**14**	**28**	**60**	**119**
	average	**12**	**26**	**56**	**110**
	proposed method	**1**	**1**	**2**	**5**

4.1.7 ข้อมูล Pima

การจัดกลุ่มข้อมูล Pima ด้วยการกำหนดค่าเริ่มต้นโดยการตัดแบ่งข้อมูลมีประสิทธิภาพดีกว่าการกำหนดค่าเริ่มต้นแบบสุ่มที่แย่ที่สุด และกลุ่มข้อมูลขนาด 4 กลุ่ม อัลกอริทึมที่นำเสนอมีประสิทธิภาพด้อยกว่าอัลกอริทึมการกำหนดค่าเริ่มต้นแบบสุ่มในทุกกรณี ซึ่งส่วนใหญ่การจัดกลุ่มข้อมูลชุดนี้ เห็นได้ว่าอัลกอริทึมที่นำเสนอมีประสิทธิภาพด้อยกว่าการกำหนดค่าเริ่มต้นแบบสุ่มที่มีประสิทธิภาพดีที่สุดของทุกกลุ่มข้อมูล ดังแสดงในตารางที่ 4.10 และกราฟที่ 4.10

ตารางที่ 4.10 แสดงประสิทธิภาพการจัดกลุ่มข้อมูล Pima ด้วย SSE และ Entropy

Data sets		# clusters							
		8 clusters		16 clusters		32 clusters		64 clusters	
Pima	**Experiment**	**sse**	**entropy**	**sse**	**entropy**	**sse**	**entropy**	**sse**	**entropy**
Trial #	1	0.1579	0.6138	0.1274	0.5727	0.0954	0.5398	0.0777	0.5285
	2	0.1579	0.6130	0.1240	0.5743	0.0954	0.5489	0.0793	0.5345
	3	0.1579	0.6148	0.1273	0.5675	0.0989	0.5511	0.0792	0.5263
	4	0.1579	0.6133	0.1241	0.5735	0.0953	0.5405	0.0757	0.5237
	5	0.1579	0.6143	0.1273	0.5688	0.1001	0.5487	0.0746	0.5220
	6	0.1579	0.6148	0.1240	0.5749	0.0999	0.5513	0.0777	0.5254
	7	0.1579	0.6145	0.1241	0.5701	0.0998	0.5451	0.0762	0.5214
	8	0.1579	0.6137	0.1241	0.5768	0.0953	0.5504	0.0758	0.5281
	9	0.1579	0.6136	0.1241	0.5746	0.0964	0.5580	0.0737	0.5141
	10	0.1579	0.6131	0.1273	0.5717	0.0997	0.5569	0.0808	0.5259
	min	**0.1579**	**0.6130**	**0.1240**	**0.5675**	**0.0953**	**0.5398**	**0.0737**	**0.5141**
	max	**0.1579**	**0.6148**	**0.1274**	**0.5768**	**0.1001**	**0.5580**	**0.0808**	**0.5345**
	average	**0.1579**	**0.6139**	**0.1254**	**0.5725**	**0.0976**	**0.5491**	**0.0771**	**0.5250**
	proposed method	**0.1579**	**0.6138**	**0.1275**	**0.5808**	**0.0959**	**0.5561**	**0.0747**	**0.5300**



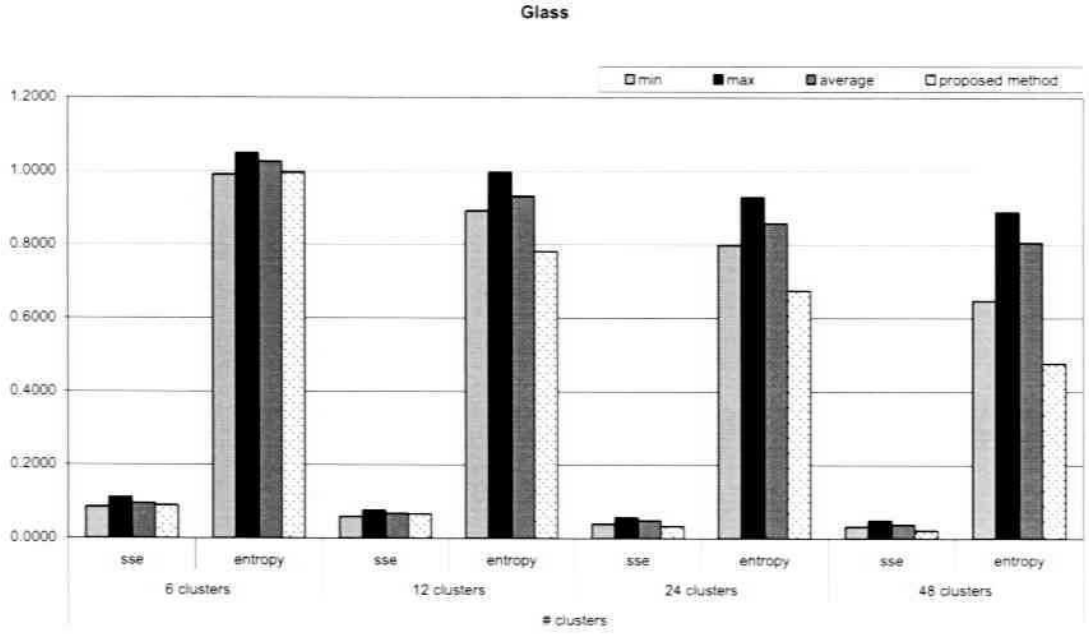
กราฟที่ 4.10 แสดงการเปรียบเทียบประสิทธิภาพการจัดกลุ่มข้อมูล Pima ด้วย SSE และ Entropy ของอัลกอริทึมที่นำเสนอ และอัลกอริทึมแบบสุ่มที่มีการประมวลผล 10 ครั้ง ด้วยค่าน้อยที่สุด มากที่สุด และค่าเฉลี่ย

4.1.8 ข้อมูล Glass

การจัดกลุ่มข้อมูล Glass ด้วยการกำหนดค่าเริ่มต้นโดยวิธีการตัดแบ่งข้อมูลมีประสิทธิภาพดีกว่าการกำหนดค่าเริ่มต้นแบบสุ่มในทุกกรณี ยกเว้นกรณีการจัดกลุ่มข้อมูลมีขนาดเล็กคือ 6 กลุ่มข้อมูล ที่ประสิทธิภาพของการกำหนดค่าเริ่มต้นแบบสุ่มกรณีค่าดีที่สุดจากการประมวลผล 10 ครั้ง มีประสิทธิภาพดีกว่า อัลกอริทึมที่นำเสนอ ดังแสดงในตารางที่ 4.11 และ กราฟที่ 4.11

ตารางที่ 4.11 แสดงประสิทธิภาพการจัดกลุ่มข้อมูล Glass ด้วย SSE และ Entropy

Data sets		# clusters							
		6 clusters		12 clusters		24 clusters		48 clusters	
Glass	Experiment	sse	entropy	sse	entropy	sse	entropy	sse	entropy
Trial #	1	0.0896	1.0439	0.0658	0.9146	0.0501	0.8367	0.0384	0.8053
	2	0.0953	1.0159	0.0687	0.9428	0.0568	0.9116	0.0384	0.8520
	3	0.0897	1.0453	0.0710	0.8915	0.0495	0.8291	0.0342	0.6471
	4	0.0955	1.0240	0.0651	0.9642	0.0440	0.8377	0.0326	0.7790
	5	0.1059	1.0014	0.0589	0.9215	0.0460	0.8573	0.0340	0.7944
	6	0.1115	1.0200	0.0640	0.9253	0.0412	0.8354	0.0401	0.8188
	7	0.0852	1.0204	0.0759	0.9960	0.0548	0.9295	0.0493	0.8886
	8	0.1018	0.9906	0.0660	0.9159	0.0544	0.8940	0.0424	0.8207
	9	0.0883	1.0472	0.0642	0.9124	0.0496	0.8462	0.0359	0.7889
	10	0.0896	1.0486	0.0702	0.9300	0.0394	0.7985	0.0378	0.8549
	min	**0.0852**	**0.9906**	**0.0589**	**0.8915**	**0.0394**	**0.7985**	**0.0326**	**0.6471**
	max	**0.1115**	**1.0486**	**0.0759**	**0.9960**	**0.0568**	**0.9295**	**0.0493**	**0.8886**
	average	**0.0952**	**1.0257**	**0.0670**	**0.9314**	**0.0486**	**0.8576**	**0.0383**	**0.8050**
	proposed method	**0.0893**	**0.9962**	**0.0656**	**0.7808**	**0.0329**	**0.6740**	**0.0219**	**0.4774**



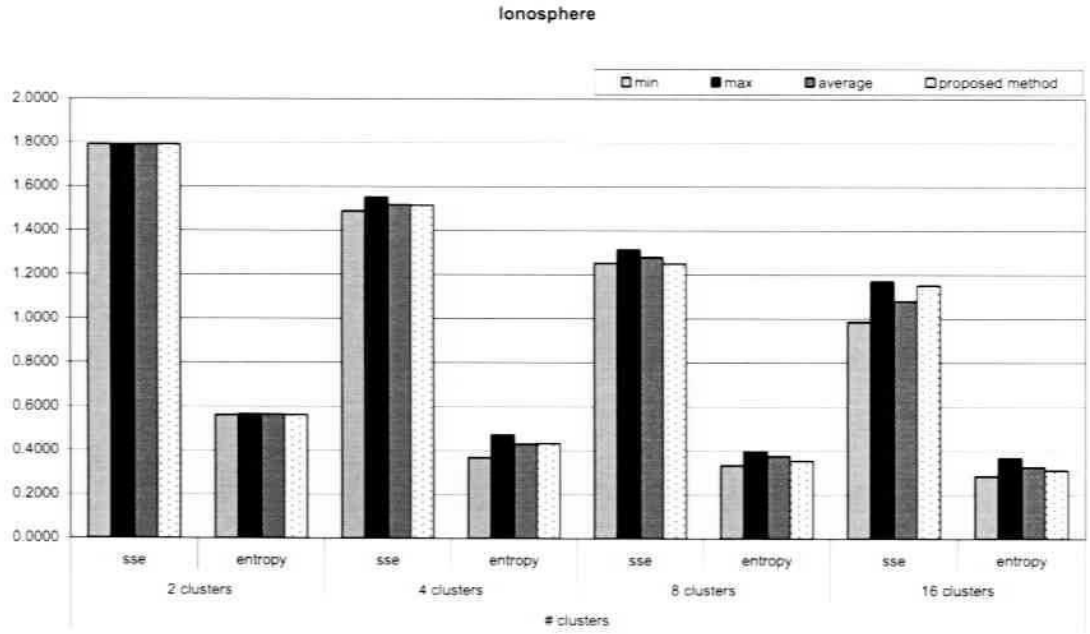
กราฟที่ 4.11 แสดงการเปรียบเทียบประสิทธิภาพการจัดกลุ่มข้อมูล Glass ด้วย SSE และ Entropy ของอัลกอริทึมที่นำเสนอ และอัลกอริทึมแบบสุ่มที่มีการประมวลผล 10 ครั้ง ด้วยค่าน้อยที่สุด มากที่สุด และค่าเฉลี่ย

4.1.9 ข้อมูล Ionosphere

การจัดกลุ่มข้อมูล Ionosphere ด้วยการกำหนดค่าเริ่มต้นโดยการตัดแบ่งข้อมูลส่วนใหญ่มีประสิทธิภาพที่ใกล้เคียงหรือดีกว่าการกำหนดค่าเริ่มต้นแบบสุ่ม ยกเว้นในบางกรณีของการกำหนด ค่าเริ่มต้นแบบสุ่มกรณีค่าดีที่สุดมีประสิทธิภาพดีกว่าอัลกอริทึมที่นำเสนอ คือการจัดกลุ่มข้อมูล 4 และ 16 กลุ่มข้อมูล ดังแสดงในตารางที่ 4.12 และ กราฟที่ 4.12

ตารางที่ 4.12 แสดงประสิทธิภาพการจัดกลุ่มข้อมูล Ionophere ด้วย SSE และ Entropy

Data sets		# clusters							
		2 clusters		4 clusters		8 clusters		16 clusters	
Ionosphere	Experiment	sse	entropy	sse	entropy	sse	entropy	sse	entropy
Trial #	1	1.7917	0.5657	1.5347	0.4696	1.2879	0.3830	1.1460	0.3364
	2	1.7917	0.5623	1.4877	0.3833	1.2542	0.3342	1.0942	0.3061
	3	1.7917	0.5657	1.5346	0.4717	1.2522	0.3957	1.1116	0.3117
	4	1.7917	0.5623	1.5353	0.4523	1.3134	0.3786	1.0578	0.3678
	5	1.7917	0.5623	1.4878	0.3676	1.2836	0.3840	1.0274	0.3301
	6	1.7917	0.5623	1.5341	0.4630	1.2764	0.3929	1.1700	0.3277
	7	1.7917	0.5657	1.5341	0.4515	1.2877	0.3705	1.0486	0.3379
	8	1.7917	0.5657	1.4874	0.4059	1.2663	0.3715	0.9858	0.2858
	9	1.7917	0.5657	1.5510	0.4071	1.3051	0.3804	1.0735	0.3211
	10	1.7917	0.5657	1.4883	0.4201	1.2627	0.3612	1.0831	0.3412
	min	**1.7917**	**0.5623**	**1.4874**	**0.3676**	**1.2522**	**0.3342**	**0.9858**	**0.2858**
	max	**1.7917**	**0.5657**	**1.5510**	**0.4717**	**1.3134**	**0.3957**	**1.1700**	**0.3678**
	average	**1.7917**	**0.5643**	**1.5175**	**0.4292**	**1.2790**	**0.3752**	**1.0798**	**0.3266**
	proposed method	**1.7917**	**0.5623**	**1.5144**	**0.4315**	**1.2497**	**0.3536**	**1.1512**	**0.3113**



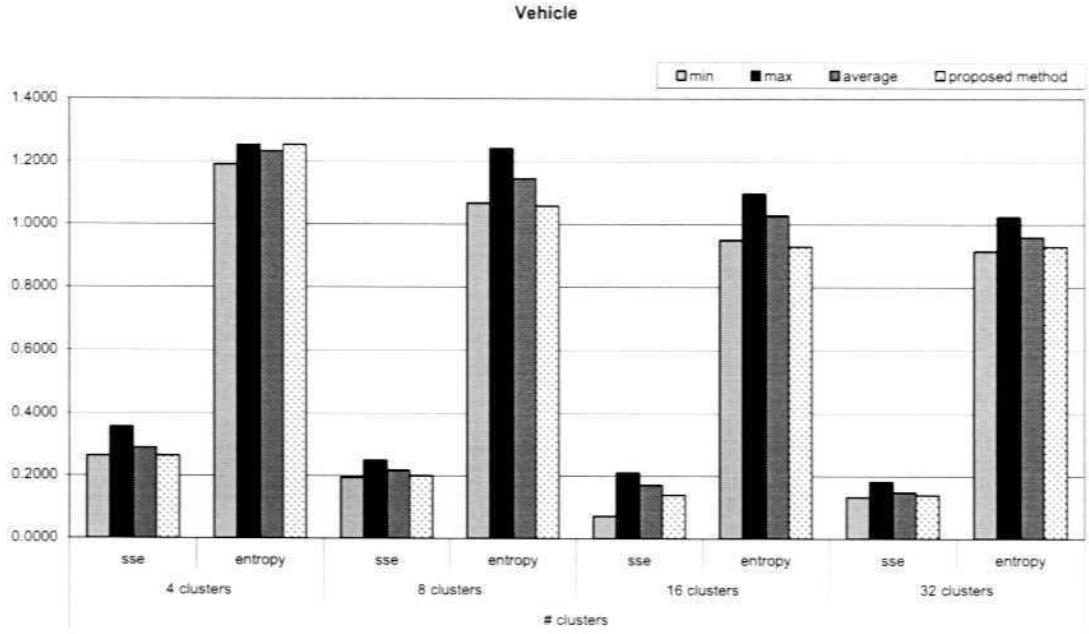
กราฟที่ 4.12 แสดงการเปรียบเทียบประสิทธิภาพการจัดกลุ่มข้อมูล Inosphere ด้วย SSE และ Entropy ของอัลกอริทึมที่นำเสนอ และอัลกอริทึมแบบสุ่มที่มีการประมวลผล 10 ครั้ง ด้วยค่าน้อยที่สุด มากที่สุด และค่าเฉลี่ย

4.1.10 ข้อมูล Vehicle

การจัดกลุ่มข้อมูล Vehicle ด้วยการกำหนดค่าเริ่มต้นโดยการตัดแบ่งข้อมูลส่วนใหญ่มีประสิทธิภาพดีกว่าการกำหนดค่าเริ่มต้นแบบสุ่ม ยกเว้นกรณีการจัดกลุ่มข้อมูลขนาด 8 และ 32 กลุ่ม เมื่อเปรียบเทียบด้วยประสิทธิภาพของ Entropy ซึ่งมีประสิทธิภาพต่างกันเพียงเล็กน้อย ดังแสดงในตารางที่ 4.12 และกราฟที่ 4.12

ตารางที่ 4.13 แสดงประสิทธิภาพการจัดกลุ่มข้อมูล Vehicle ด้วย SSE และ Entropy

Data sets		**# clusters**							
		4 clusters		**8 clusters**		**16 clusters**		**32 clusters**	
Vehicle	**Experiment**	**sse**	**entropy**	**sse**	**entropy**	**sse**	**entropy**	**sse**	**entropy**
Trial #	1	0.3570	1.2354	0.2279	1.2049	0.1942	1.0657	0.1533	0.9502
	2	0.2809	1.2400	0.2110	1.0897	0.1975	1.0962	0.1417	0.9595
	3	0.2759	1.1905	0.2499	1.2413	0.2100	1.0974	0.1554	0.9592
	4	0.2642	1.2478	0.1951	1.0676	0.1820	1.0417	0.1338	0.9691
	5	0.2807	1.2390	0.2281	1.2012	0.1715	0.9702	0.1479	0.9379
	6	0.2642	1.2475	0.2119	1.1530	0.1722	0.9960	0.1385	0.9569
	7	0.2756	1.1925	0.2283	1.1723	0.1553	0.9496	0.1331	0.9148
	8	0.2643	1.2521	0.1984	1.0954	0.1827	1.0374	0.1473	0.9447
	9	0.2810	1.2401	0.2106	1.1339	0.0719	0.9908	0.1434	0.9800
	10	0.3570	1.2354	0.2109	1.0943	0.1745	1.0384	0.1825	1.0241
	min	**0.2642**	**1.1905**	**0.1951**	**1.0676**	**0.0719**	**0.9496**	**0.1331**	**0.9148**
	max	**0.3570**	**1.2521**	**0.2499**	**1.2413**	**0.2100**	**1.0974**	**0.1825**	**1.0241**
	average	**0.2901**	**1.2320**	**0.2172**	**1.1454**	**0.1712**	**1.0283**	**0.1477**	**0.9596**
	proposed method	**0.2644**	**1.2531**	**0.1999**	**1.0577**	**0.1391**	**0.9291**	**0.1391**	**0.9291**



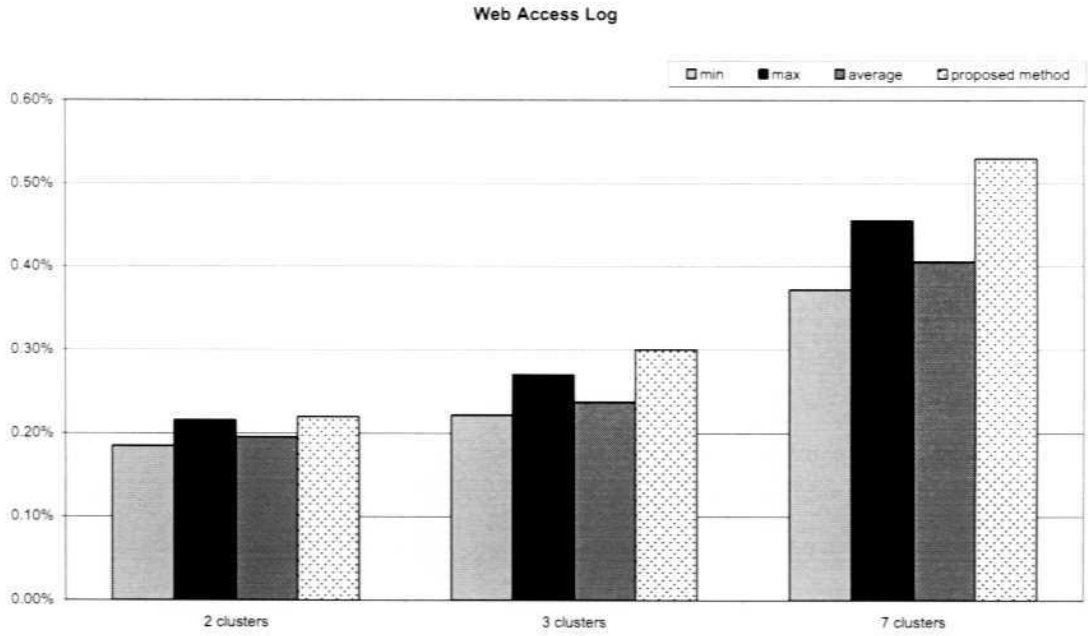
กราฟที่ 4.13 แสดงการเปรียบเทียบประสิทธิภาพการจัดกลุ่มข้อมูล Vehicle ด้วย SSE และ Entropy ของอัลกอริทึมที่นำเสนอ และอัลกอริทึมแบบสุ่มที่มีการประมวลผล 10 ครั้ง ด้วยค่าน้อยที่สุด มากที่สุด และค่าเฉลี่ย

4.2 ผลการทดลองที่ 2

การจัดกลุ่มข้อมูลเว็บเพจ ซึ่งมีขนาด 2713 Instances และ 2509 Attributes ด้วยการกำหนดค่าเริ่มต้นโดยการตัดแบ่งข้อมูล จะมีประสิทธิภาพดีกว่าการกำหนดค่าเริ่มต้นแบบสุ่ม ทั้งกรณีดีที่สุด แย่ที่สุด และค่าเฉลี่ยที่มีการประมวลผล 10 ครั้ง โดยเปรียบเทียบค่า Weighted Average Hits ที่มีการถ่วงน้ำหนักตามขนาดของกลุ่มข้อมูลเว็บเพจ กล่าวคือเมื่อกลุ่มข้อมูลมีจำนวนข้อมูลมากควรจะมีน้ำหนักต่อค่าเฉลี่ยสูงกว่ากลุ่มที่มีขนาดเล็ก ดังแสดงในตารางที่ 4.14 และกราฟที่ 4.14

ตารางที่ 4.14 แสดงประสิทธิภาพการจัดกลุ่มข้อมูล Web Access Log

Data sets		**Weighted Average Hits**		
Web access log	**Experiment**	**2 clusters**	**3 clusters**	**7 clusters**
Trial #	1	0.20%	0.23%	0.46%
	2	0.19%	0.23%	0.37%
	3	0.20%	0.24%	0.41%
	4	0.19%	0.22%	0.38%
	5	0.19%	0.24%	0.40%
	6	0.22%	0.23%	0.38%
	7	0.19%	0.27%	0.43%
	8	0.19%	0.26%	0.40%
	9	0.22%	0.23%	0.38%
	10	0.19%	0.23%	0.45%
	min	**0.19%**	**0.22%**	**0.37%**
	max	**0.22%**	**0.27%**	**0.46%**
	average	**0.20%**	**0.24%**	**0.41%**
	proposed method	**0.22%**	**0.30%**	**0.53%**



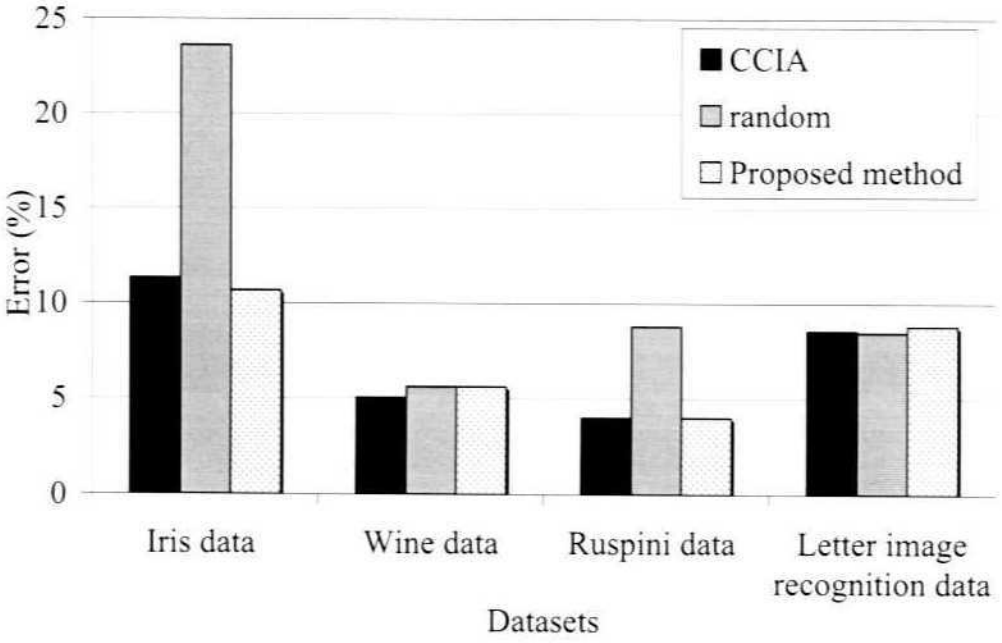
กราฟที่ 4.14 แสดงการเปรียบเทียบประสิทธิภาพการจัดกลุ่มข้อมูล Web Access Log ด้วย Weighted Average Hits ของอัลกอริทึมที่นำเสนอ และอัลกอริทึมแบบสุ่มที่มีการประมวลผล 10 ครั้ง ด้วยค่าน้อยที่สุด มากที่สุด และค่าเฉลี่ย

4.3 การเปรียบเทียบประสิทธิภาพกับอัลกอริทึม Cluster Center Initialization Algorithm (CCIA)

เมื่อเปรียบเทียบอัลกอริทึมที่นำเสนอนี้ กับอัลกอริทึม CCIA (Shehroz S. Khan และ Amir Ahmad, 2004) ในการหาจุดเริ่มต้นสำหรับการจัดกลุ่มข้อมูลด้วยเคมีน โดยวัดประสิทธิภาพการจัดกลุ่มข้อมูลด้วยร้อยละของข้อมูลที่มีจัดกลุ่มข้อมูลผิดพลาด เห็นได้ว่าประสิทธิภาพในการจัดกลุ่มข้อมูลมีค่าใกล้เคียงกัน ซึ่งบางชุดข้อมูลอัลกอริทึม CCIA มีประสิทธิภาพดีกว่า แต่ในบางกรณีอัลกอริทึมที่นำเสนอมีประสิทธิภาพดีกว่า ดังแสดงในตารางที่ 4.15 และกราฟที่ 4.15 อย่างไรก็ตาม อัลกอริทึมที่นำเสนอจะมีความซับซ้อนน้อยกว่าอัลกอริทึม CCIA มาก

ตารางที่ 4.15 แสดงประสิทธิภาพการจัดกลุ่มข้อมูลของอัลกอริทึมที่นำเสนอกับ CCIA

ชุดข้อมูล	ร้อยละของความผิดพลาดในการจัดกลุ่ม (%)		
	CCIA	**Random**	**Proposed**
Iris data	11.33	23.6	10.67
Wine data	5.05	5.61	5.62
Ruspini data	4.00	8.80	4.00
Letter image recognition data	8.55	8.47	8.8



กราฟที่ 4.15 แสดงการเปรียบเทียบประสิทธิภาพการจัดกลุ่มข้อมูลด้วยอัลกอริทึมที่นำเสนอ และอัลกอริทึม CCIA ด้วยร้อยละของความผิดพลาดในการจัดกลุ่มข้อมูล

4.4 สรุปผล

การจัดกลุ่มข้อมูลด้วยอัลกอริทึมเคมีน ที่มีการกำหนดการค่าเริ่มต้นด้วยวิธีการตัดแบ่งข้อมูลมีประสิทธิภาพดีกว่า การจัดกลุ่มข้อมูลด้วยอัลกอริทึมเคมีนที่มีการกำหนดค่าเริ่มต้นแบบสุ่ม กล่าวคือ

ข้อมูลทั่วไปที่มีคุณลักษณะของข้อมูลที่หลากหลาย ส่วนใหญ่อัลกอริทึมที่นำเสนอมีประสิทธิภาพดีกว่าการกำหนดค่าเริ่มต้นแบบสุ่ม และสามารถจัดกลุ่มข้อมูลได้ดีสำหรับข้อมูลที่มีขนาดใหญ่ เมื่อเปรียบเทียบเวลาในการจัดกลุ่มข้อมูลขนาดใหญ่แล้วเห็นได้ว่า อัลกอริทึมที่นำเสนอใช้เวลาในการจัดกลุ่มข้อมูลน้อยกว่าการกำหนดค่าเริ่มต้นแบบสุ่มมาก สำหรับการจัดกลุ่มข้อมูลขนาดเล็ก การจัดกลุ่มข้อมูลของอัลกอริทึมที่นำเสนอมีประสิทธิภาพใกล้เคียงหรือดีกว่าการกำหนดค่าเริ่มต้นแบบสุ่ม เมื่อเปรียบเทียบกับกรณีค่าเฉลี่ยของการกำหนดค่าเริ่มต้นแบบสุ่ม

ข้อมูลการเข้าถึงเว็บเพจ อัลกอริทึมที่นำเสนอยังคงมีประสิทธิภาพดีกว่าอัลกอริทึมเคมีนที่มีค่าเริ่มต้นแบบสุ่ม ถึงแม้จำนวนมิติของข้อมูลค่อนข้างมาก และข้อมูลที่มีลักษณะบางตา (Sparse)

การเปรียบเทียบกับอัลกอริทึมการกำหนดค่าเริ่มของอัลกอริทึมเคมีนที่มีผู้นำเสนอไว้แล้วนั้น เห็นได้ว่าประสิทธิภาพในการจัดกลุ่มข้อมูลมีค่าใกล้เคียงกัน ซึ่งบางชุดข้อมูลอัลกอริทึม CCIA มีประสิทธิภาพดีกว่า แต่ในบางกรณีอัลกอริทึมที่นำเสนอมีประสิทธิภาพดีกว่า ทั้งนี้อัลกอริทึมที่นำเสนอมีประสิทธิภาพดีกว่าค่าเฉลี่ยจากการจัดกลุ่มข้อมูลด้วยการกำหนดค่าเริ่มต้นแบบสุ่ม อีกทั้งยังมีความซับซ้อนน้อยกว่าอัลกอริทึม CCIA อีกด้วย

จากผลการทดลองสรุปได้ว่า การจัดกลุ่มข้อมูลด้วยอัลกอริทึมเคมีนโดยการหาค่าเริ่มต้นตามอัลกอริทึมที่นำเสนอนี้ ส่วนใหญ่มีประสิทธิภาพดีกว่าการจัดกลุ่มข้อมูลโดยกำหนดค่าเริ่มต้นแบบสุ่ม และมีประสิทธิภาพใกล้เคียงหรือด้อยกว่าในบางกรณีที่ดีที่สุดของการกำหนดค่าเริ่มต้นแบบสุ่มที่ต้องมีการประมวลผลหลายครั้ง สำหรับการทดสอบกับอัลกอริทึมการกำหนดค่าเริ่มต้นการจัดกลุ่มข้อมูลที่เคยมีการทำวิจัยไว้แล้ว อัลกอริทึมที่นำเสนอมีประสิทธิภาพดีกว่าในบางชุดข้อมูลเมื่อเปรียบเทียบกับ CCIA และมีประสิทธิภาพดีกว่าในทุกข้อมูลเมื่อเปรียบเทียบกับการกำหนดค่าเริ่มต้นแบบสุ่มโดยการประมวลผลหลายรอบ ของข้อมูลจากงานวิจัยดังกล่าว

ซึ่งจะเห็นได้ว่า อัลกอริทึมที่นำเสนอนี้สามารถจัดกลุ่มข้อมูลต่าง ๆ ได้ถูกต้องและใช้เวลาในการทำงานน้อยกว่า การกำหนดค่าเริ่มต้นแบบสุ่มโดยเฉพาะเมื่อข้อมูลมีขนาดใหญ่หรือจำนวนกลุ่มข้อมูลที่ต้องการมีค่ามาก

บทที่ 5

สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

จากผลการทดลองทั้งสองการทดลองที่ใช้ข้อมูลจาก UCI จำนวน 10 ชุด ที่มีจำนวนข้อมูลและรูปแบบของข้อมูลที่หลากหลาย และข้อมูล Web Access Log พบว่า การจัดกลุ่มข้อมูลโดยการกำหนดค่าเริ่มต้นด้วยวิธีตัดแบ่งกลุ่มนั้น ส่วนใหญ่มีประสิทธิภาพดีกว่าค่าเฉลี่ยและค่าที่แย่ที่สุดของการจัดกลุ่มข้อมูลด้วยอัลกอริทึมเคมีนที่มีการกำหนดค่าเริ่มต้นแบบสุ่มโดยการประมวลผล 10 ครั้ง และมีประสิทธิภาพใกล้เคียงหรือด้อยกว่า ประสิทธิภาพที่ดีที่สุดของการกำหนดค่าเริ่มต้นแบบสุ่ม 10 ครั้ง และจำนวนกลุ่มข้อมูลมีขนาดเล็ก แต่เมื่อกลุ่มข้อมูลมีขนาดใหญ่อัลกอริทึมที่นำเสนอส่วนใหญ่มีประสิทธิภาพดีกว่า และยังใช้เวลาน้อยมากในการจัดกลุ่มข้อมูลขนาดใหญ่อีกด้วย สำหรับการจัดกลุ่มข้อมูลเว็บเพจ อัลกอริทึมที่นำเสนอมีประสิทธิภาพดีกว่าการการกำหนดค่าเริ่มต้นแบบสุ่ม และเมื่อเปรียบเทียบกับอัลกอริทึม CCIA ที่การกำหนดค่าเริ่มต้นในการจัดกลุ่มข้อมูลนั้น อัลกอริทึมที่นำเสนอยังคงมีประสิทธิภาพดีกว่าสำหรับบางชุดข้อมูล และส่วนใหญ่มีประสิทธิภาพดีกว่าค่าเฉลี่ยของอัลกอริทึมที่มีการกำหนดค่าเริ่มต้นแบบสุ่มที่มีการประมวลผลหลายรอบจากการทดสอบประสิทธิภาพของงานวิจัยดังกล่าว

เมื่อเปรียบเทียบกับ ค่าน้อยที่สุด มากที่สุด และค่าเฉลี่ยของอัลกอริทึมการกำหนดค่าเริ่มต้นแบบสุ่มที่มีการประมวลผล 10 ครั้ง กับอัลกอริทึมที่นำเสนอในงานวิจัยนี้ เห็นได้อย่างชัดเจนว่าอัลกอริทึมที่นำเสนอ ส่วนใหญ่มีประสิทธิภาพในการจัดกลุ่มข้อมูลได้ดีกว่าการกำหนดค่าเริ่มต้นแบบสุ่ม โดยสามารถลดข้อเสียและข้อจำกัด ของอัลกอริทึมแบบสุ่มได้ดังนี้

- ข้อมูลที่มีความกระจายตัวสูงจะมีผลทำให้การจัดกลุ่มข้อมูลโดยใช้ค่าเริ่มต้นแบบสุ่มมีประสิทธิภาพลดลงและมีความไม่แน่นอน ทำให้ได้ผลลัพธ์ที่กลุ่มข้อมูลจะมีความแปรปรวนแตกต่างกันมาก แต่สำหรับอัลกอริทึมการตัดแบ่งกลุ่มข้อมูลไม่มีผลกระทบจากค่าดังกล่าวเนื่องจากอัลกอริทึมที่นำเสนอจะเลือกการตัดแบ่งกลุ่มข้อมูลตามค่า Delta Variance (ค่าความแปรปรวนก่อนแบ่งเซลล์ลบด้วยค่าความแปรปรวนของกลุ่มย่อยทั้งสองกลุ่มในเซลล์) จึงทำให้ค่าความแปรปรวนของเซลล์ย่อยมีค่าเท่า ๆ กัน

- ผลลัพธ์จากการจัดกลุ่มข้อมูลด้วยอัลกอริทึมเคมีนแบบสุ่ม ในบางครั้งไม่มีจำนวนสมาชิกจากผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูล แต่สำหรับอัลกอริทึมที่นำเสนอนี้มีการกำหนดค่าเริ่มต้นจากตัวแทนกลุ่มที่ได้จากการตัดแบ่งข้อมูลซึ่งเป็นค่าที่อยู่ในตำแหน่งใกล้เคียงกับข้อมูลที่มีอยู่จริง จึงทำให้ทุกกลุ่มข้อมูลที่ได้จากการแบ่งกลุ่มมีสมาชิกในแต่ละกลุ่มอย่างแน่นอน

- การจัดกลุ่มข้อมูลด้วยวิธีการตัดแบ่งข้อมูล เป็นการลดเวลาในการหาค่าเหมาะสมในการจัดกลุ่มข้อมูลด้วยเคมีนแบบสุ่ม เนื่องจากค่าตัวแทนกลุ่ม (Centroid) ที่ใช้เป็นค่าเริ่มต้นในการจัดกลุ่มข้อมูลด้วยเคมีนนั้นมีระยะทางใกล้เคียงกับข้อมูลที่มีอยู่ หรืออาจกล่าวได้ว่าค่าตัวแทนกลุ่มที่ได้มีความใกล้เคียงกับค่าเหมาะสม (Optimal Solution) ในการจัดกลุ่มข้อมูลอยู่แล้ว จึงใช้จำนวนรอบในการหาสมาชิกในแต่ละกลุ่มข้อมูลน้อยกว่าอัลกอริทึมที่มีการกำหนดค่าเริ่มต้นแบบสุ่ม จะเห็นได้จากการเปรียบเทียบเวลาที่ใช้ในการจัดกลุ่มข้อมูลขนาดใหญ่ นอกจากประสิทธิภาพของอัลกอริทึมที่นำเสนอจะดีกว่าแล้ว เวลาที่ใช้ในการจัดกลุ่มข้อมูลยังเร็วกว่าอัลกอริทึมกำหนดค่าเริ่มต้นแบบสุ่มอีกด้วย

- การเลือกแกนที่มีความแปรปรวนมากที่สุดเป็นแกนในการแบ่งเซลล์ เนื่องจากทำให้ระยะทางระหว่างเซลล์หรือระหว่างกลุ่มข้อมูลมีแนวโน้มที่มากกว่าเซลล์อื่น อีกทั้งยังเป็นการลดความคลาดเคลื่อนทั้งหมดจากการแบ่งกลุ่มข้อมูลของทั้งสองเซลล์ จึงทำให้วิธีการที่นำแสนอนี้มีแนวโน้มได้ผลลัพธ์ที่เป็นค่าเหมาะสมของปัญหาแบบ NP-Complete จากการหาค่าที่เหมาะสมในการแบ่งกลุ่มข้อมูลด้วยอัลกอริทึมเคมีน

ทั้งนี้อัลกอริทึมที่นำแสนอยังสามารถจัดข้อมูลที่มีความหลากหลายรูปแบบทั้งข้อมูลทั่วไปและข้อมูล Web Access Log จาก Web Server ได้อย่างดีอีกด้วย

5.2 ข้อเสนอแนะ

ในการการทดสอบประสิทธิภาพของอัลกอริทึมเพื่อหาค่าเหมาะสมแบบกว้าง (Global Optimal Solution) ยังไม่สามารถดำเนินการได้ ซึ่งผลลัพธ์ที่ได้เป็นเพียงค่าเหมาะสมเฉพาะที่ (Local Optimal Solution) ซึ่งปัญหาการหาค่าเหมาะสมในการจัดกลุ่มข้อมูลเป็นปัญหา NP-Complete จึงเป็นการยากในการหาค่าเหมาะสมสำหรับการจัดกลุ่มข้อมูลที่เหมาะสมแบบกว้าง

ควรเพิ่มการทดลองกับข้อมูลอื่น เพื่อการทดสอบประสิทธิภาพของอัลกอริทึมที่มีความหลากหลายมากขึ้น และหาวิธีการปรับปรุงประสิทธิภาพของอัลกอริทึมให้ดีขึ้น รวมถึงการปรับปรุงโปรแกรมในรูปแบบ Graphic User Interface เพื่อความสะดวกในการนำไปใช้งานต่อไป

ภาคผนวก

ภาคผนวก ก

การแบ่งกลุ่มข้อมูลเว็บ (Web Clustering)

การจัดกลุ่มข้อมูลเว็บคือ การจัดกลุ่มของวัตถุเว็บ (Web Object) ซึ่งเป็นข้อมูลต่าง ๆ ที่อยู่บนเว็บเพจ โดยการแบ่งประเภทของวัตถุ (ข้อมูล) ตามความคล้ายคลึงของวัตถุประเภทเดียวกัน และตามความแตกต่างของวัตถุของข้อมูลที่แตกต่างกัน และเป็นการค้นหาการกระจายของรูปแบบและความสัมพันธ์ตามลักษณะของข้อมูล ซึ่งการจัดกลุ่มข้อมูลเป็นการดำเนินการแบบไม่มีข้อมูลฝึกสอน (Unsupervised Learning) หรือไม่มีการจัดกลุ่มข้อมูลไว้ล่วงหน้า ในการวิจัยที่นำเสนอนี้ใช้การจัดกลุ่มข้อมูลเว็บเพจตามข้อมูลผู้ใช้ (Web Usage) โดยการจัดกลุ่มข้อมูลเว็บเพจตามข้อมูลการเข้าถึงเว็บเพจต่าง ๆ ของผู้ใช้

การจัดกลุ่มข้อมูลเกี่ยวข้องกับการวัดระยะห่างหรือความคล้ายคลึงระหว่างสองหน่วยของข้อมูล เช่น การวัดระยะห่างแบบ Euclidean, Manhattan และ Cosine เป็นต้น โดยประโยชน์จากการจัดกลุ่มข้อมูลเว็บเพจ เป็นการเพิ่มจำนวนผู้ใช้ในการเข้าถึงสารสนเทศบนเว็บไซต์ เนื่องจากการวิเคราะห์ข้อมูลพฤติกรรมของผู้ใช้ทำให้สามารถออกแบบ และปรับโครงสร้างของเว็บไซต์ได้เหมาะสมกับผู้ใช้โดยทั่วไป

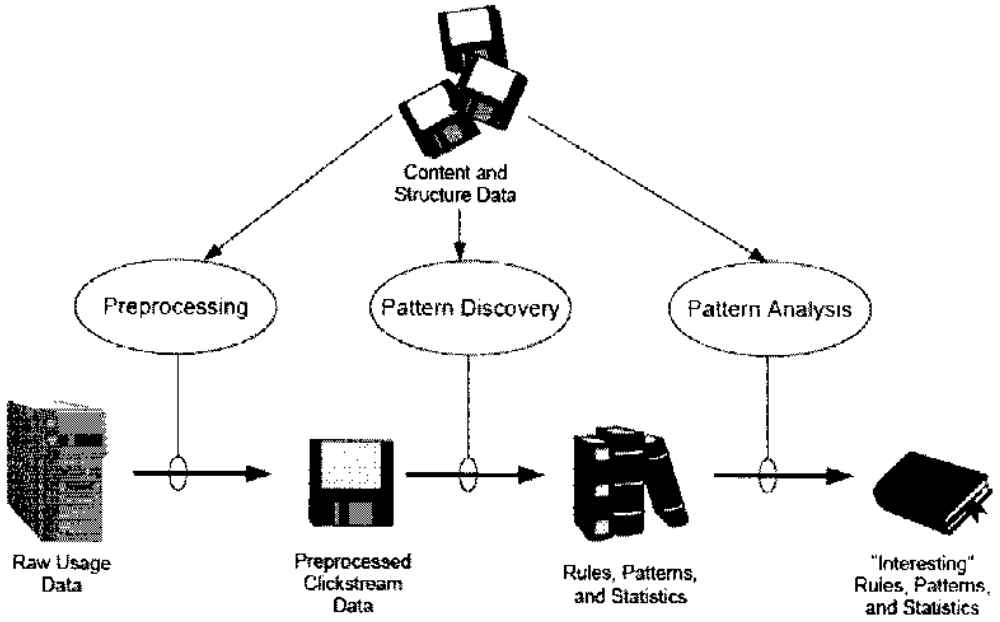
ประเภทของการแบ่งกลุ่มข้อมูลบนเว็บไซต์ มีการนำรูปแบบการจัดกลุ่มข้อมูลมาใช้ในการจัดกลุ่มข้อมูลเว็บไซต์หลายรูปแบบได้แก่ การแบ่งกลุ่มแบบลำดับชั้น (Hierarchical Clustering), การแบ่งกลุ่มแบบแบ่งส่วน (Partition Clustering), การจัดกลุ่มแบบความน่าจะเป็น (Probabilistic Clustering), การจัดกลุ่มแบบกราฟ (Graph Based Clustering), การจัดกลุ่มแบบฟัซซี่ (Fuzzy clustering), การจัดกลุ่มแบบโครงข่ายประสาทเทียม (Neural Network Based Clustering) และการจัดกลุ่มในรูปแบบผสม (Hybrid) ซึ่งแต่ละประเภทมีความเร็ว และความเหมาะสมในการจัดกลุ่มข้อมูลตามรูปร่าง ประเภท รวมถึงขนาดของข้อมูลที่แตกต่างกัน

การประยุกต์ใช้เทคนิคการทำเหมืองข้อมูล (Data Mining) เพื่อค้นหารูปแบบการเข้าถึงเว็บไซต์ของผู้ใช้นั้น แบ่งเป็น 3 ขั้นตอนคือ

1. การเตรียมข้อมูล (Preprocessing) เป็นการรวบรวมข้อมูลจาก Web Server Log File และการจำแนกข้อมูล Session ของผู้ใช้ รวมถึงการกรองข้อมูล (Data Cleaning)

2. การค้นหารูปแบบ (Pattern Discovery) เป็นการจัดการข้อมูลผู้ใช้ (User Clustering / Grouping) หรือการจัดกลุ่มเว็บเพจ (Pages Clustering/Grouping)

3. การวิเคราะห์รูปแบบ (Pattern Analysis) เป็นการวิเคราะห์รูปแบบที่ได้ว่ามีความเหมาะและสามารถนำไปใช้งานหรือไม่



ภาพแสดงขั้นตอนการทำ Web Usage Mining ในระดับสูง

แหล่งที่มา: Cooley, Mobasher, and Srivastava., 1999

ลักษณะข้อมูลเว็บเพจ เป็นการแสดงเอกสารเว็บ (Web Document) โดยที่แต่ละเว็บเพจประกอบด้วยข้อมูลต่าง ๆ ทั้งรูปแบบข้อความ ภาพ และเสียง ซึ่งสร้างเป็นไฟล์ HTML, ไฟล์ XML, ข้อมูลภาพ, ข้อมูลมัลติมีเดีย เป็นต้น

Web Server Log File เป็นการรวบรวมข้อมูลเกี่ยวกับการเข้าถึงเว็บเพจต่าง ๆ ในเว็บไซต์ โดยมีการบันทึกการร้องขอ (Request) เว็บเพจตามแหล่งที่มาของผู้ใช้ (IP Address) หรือบันทึกตามการเป็นสมาชิกของผู้ใช้ จากกรอกชื่อและรหัสผ่านเพื่อเข้าสู่ระบบ

```
www.hyperreal.org|anon0000000000000000148946|GET
/music/machines/manufacturers/Korg/Poly-800.EX-800/info/poly-800.fans
HTTP/1.1|text/plain|200|1999/01/01-00:00:05|-|15803|-|-
|http://www.hyperreal.org/music/machines/manufacturers/Korg/Poly-800.EX-800/|Mozilla/4.0
(compatible; MSIE 4.0; Windows 95)
www.hyperreal.org|anon0000000000000000148946|GET
/music/machines/manufacturers/Korg/Poly-800.EX-800/ HTTP/1.1|text/html|200|1999/01/01-
00:00:17|-|-|-|-|http://www.hyperreal.org/music/machines/manufacturers/Korg/|Mozilla/4.0
 (compatible; MSIE 4.0; Windows 95)www.hyperreal.org|
```

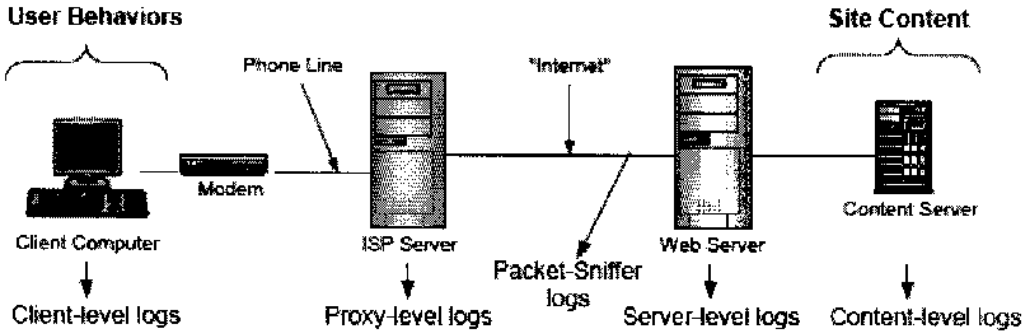
ภาพแสดงตัวอย่าง Web Access Log จากเว็บไซต์ www.hyperreal.org

Web Data Source (Cooley, 2000): ข้อมูลการใช้งานเว็บไซต์แบ่งเป็นระดับดังนี้

Client-level logs เป็นรูปแบบทั่วไปที่ผู้ใช้คนเดียวต่อหลายๆเว็บไซต์

Server-level logs เป็นรูปแบบที่หลายๆผู้ใช้เข้าใช้เว็บไซต์เดียว

Proxy-level logs เป็นรูปแบบที่หลายๆผู้ใช้เข้าใช้หลายๆเว็บไซต์



ภาพแสดงระดับการเข้าถึงเว็บไซต์ของผู้ใช้

แหล่งที่มา: Cooley, 2000.

การวิเคราะห์ข้อมูลผู้ใช้แบบ Offline จะใช้ข้อมูลจาก Server Log แต่มีปัญหาจากการใช้ข้อมูลดังกล่าวคือ การรวมรวมข้อมูลการใช้เว็บไซต์อาจจะไม่มีความน่าเชื่อถือเนื่องจากข้อมูลที่ผ่าน Cache และ Proxy

Web Server Log เป็นแหล่งข้อมูลที่สำคัญสำหรับประสิทธิภาพในการทำ Web Usage Mining เพราะเป็นแหล่งเก็บข้อมูล พฤติกรรมการค้นหาข้อมูลภายในเว็บไซต์จากผู้เยี่ยมชมเว็บไซต์ ข้อมูลที่บันทึกใน Server Log แสดงถึงการเข้าถึงเว็บไซต์โดยผู้ใช้หลาย ๆ คน โดย Log File สามารถเก็บในรูปแบบ Common Log Format (CLF) หรือ Extended Common Log Format (ECLF) ประกอบด้วย ข้อมูลฟิลด์พื้นฐาน เป็น IP Address ของ Client, User ID, วัน/เวลา (Date/Time), การร้องขอ (Request), สถานะ (Status), จำนวนไบท์ (Bite), การอ้างอิง (Referrer)

Client IP address เป็นที่อยู่ของอินเทอร์เน็ตของเครื่องที่ร้องขอข้อมูล ซึ่งอาจจะไม่แสดงที่อยู่ของ proxy-server

User ID เป็นส่วนที่มีเฉพาะผู้ที่ log in เข้าระบบเท่านั้น (Authentication)

Time/Date Stamp แทนเวลาที่ร้องขอข้อมูลและได้รับข้อมูลจาก Web Server และรายการวิธีการร้องขอ

วิธีการในการร้องขอข้อมูลมี 2 รูปแบบ คือ

GET เป็นการร้องขอ Object จาก Web Server

POST เป็นการส่งข้อมูลให้กับ Web Server

Head Request จะเป็น HTTP Header สำหรับ Object

URI อาจจะเป็น Static File ใน Local File System หรือ ชื่อของโปรแกรมที่ Executable ที่จะถูกเรียกในการตอบสนองการร้องขอ

Status Field เป็นกลุ่มของ Web Server และอธิบายลักษณะการตอบสนองจากการร้องขอ (400-499: คือการเกิด Error ในส่วนของการร้องขอ, 500-599: คือปัญหาที่เกิดจาก Server)

Size File Logs เป็นจำนวน Bytes ในการตอบกลับผลของการร้องขอ

URIs เป็นชนิดในการเข้าถึงหรือการเข้าถึงผ่าน Bookmark

Agent Field เป็นข้อความที่สามารถอธิบายถึงระบบปฏิบัติการ (Operating System) และซอฟท์แวร์ Browser ของผู้ใช้ที่เข้าถึงเว็บไซต์

ตาราง แสดงรายละเอียดข้อมูลของ Web Access Log

ชื่อฟิลด์	คำอธิบาย
User Address	ตัวเลข IP address หรือโดเมนของผู้เยี่ยมชมเว็บไซต์
Date/Time	วันและเวลาที่เยี่ยมชมเว็บไซต์
GMT offset	จำนวนชั่วโมงแสดงระยะห่างจากเวลาสากล GMT (ถ้ามีค่า -0000 หมายถึง ล็อกไฟล์ในเวลาสากล)
Action	วิธีการดำเนินการ(GET หรือ POST) ของฮิต(อยู่ในเครื่องหมายคำพูด)
URL Stem	ชื่อไฟล์ที่ถูกกระทำ
Protocol Version	เวอร์ชันของโปรโตคอล http ที่ใช้
Return Code	โค้ดแสดงผลการตอบสนองคำร้องขอ
Server to Client bytes	จำนวนไซต์ที่ส่งไปยังไคลเอนท์
Referrer	ตำแหน่งที่ผู้ชมใช้ลิงค์มายังไซต์
Browser/Platform	เว็บเบราเซอร์และแพลตฟอร์มที่ใช้ในการเยี่ยมชมเว็บไซต์

ขั้นตอนการเตรียมข้อมูล (Preprocessing)

ขั้นตอนการเตรียมข้อมูลจาก Server Log ผลลัพธ์ที่ได้จากขั้นตอนนี้คือ User Session File, Transaction File, Site Topology และการจำแนกเว็บเพจ (Page Classifications) ซึ่งความน่าเชื่อถือของ User Session File ลดลงจากการค้นหาข้อมูลและการใช้ Proxy Caching แต่ปัจจุบันวิธีในการรวบรวมข้อมูลเกี่ยวกับการอ้างอิงแคชได้รวมการใช้ Cookies และ Cache Busting ซึ่งเป็นวิธีการในการป้องกันการค้นหาข้อมูลจากการใช้ข้อมูลเวอร์ชันของเว็บเพจที่เก็บไว้ในภายในองค์กร และจะดาวโหลดเว็บเพจใหม่จาก Sever ทุกครั้งที่มีการเรียกดูเว็บเพจ ในส่วนของ Cookie สามารถถูกลบได้โดยผู้ใช้ อีกทั้ง Cache Busting มีข้อได้เปรียบในด้านความเร็วที่ Caching สร้างและดำเนินการเกี่ยวกับการดึงหน้าเว็บเพจจากเว็บไซต์ต่าง ๆ อีกหนึ่งวิธีในการกำหนดผู้ใช้คือการลงทะเบียนผู้ใช้ซึ่งมีข้อได้เปรียบ คือสามารถรวบรวมข้อมูลเกี่ยวกับผู้ใช้โดยเก็บข้อมูลแบบอัตโนมัติไว้ที่ Server ทำให้ง่ายในการจำแนกผู้ใช้จาก Session แต่เนื่องด้วยข้อจำกัดด้านความเป็นส่วนตัวของผู้ใช้ทำให้ผู้ใช้หลาย ๆ คน ไม่มีการค้นหาข้อมูลภายในเว็บไซต์ที่มีการลงทะเบียนหรือต้องกรอกข้อมูล Login ทำให้ข้อมูลที่ได้มีความผิดพลาด

การกรองข้อมูล (Data Cleaning)

เวลาที่ใช้ในการเยี่ยมชมเว็บเพจ เป็นการแสดงถึงความสนใจของผู้ใช้ในเว็บเพจนั้น ๆ โดยช่วงเวลาในการเยี่ยมชมเว็บเพจเป็นความแตกต่างระหว่างการร้องขอเว็บเพจที่ต่อเนื่องกัน แต่ข้อเสียคือผู้ใช้บางคนใช้ช่วงเวลาสั้นในการเยี่ยมชมเว็บเพจทำให้ไม่สามารถศึกษาพฤติกรรมของผู้ใช้ได้

เทคนิคการ Clean Server Log คือการกำจัดสิ่งที่ไม่เกี่ยวข้องกับการวิเคราะห์ หรือการทำเหมืองข้อมูลเว็บไซต์ ซึ่งมีความสำคัญในการวิเคราะห์ Web Log โพรโตคอล HTTP โดยแยกการติดต่อสำหรับทุกไฟล์ที่เป็นการร้องขอจาก Web Server การร้องขอเว็บเพจของผู้ใช้ให้ผลลัพธ์เป็น Log จำนวนมาก ทั้งรูปแบบกราฟฟิก และ Script ที่ถูก download ในรูปแบบไฟล์ HTML โดยทั่วไป การบันทึก Log ของไฟล์ HTML ที่ถูกร้องขอถูกเก็บไว้ใน session รวมถึงรูปภาพเป็นสิ่งที่มาพร้อมกับไฟล์ HTML ซึ่งไม่ใช่สิ่งที่ผู้ใช้ร้องขอ ดังนั้นทุกชื่อไฟล์ที่ลงท้าย URL ด้วย .cgi, .php, .asp, .jsp, .js, .dll, .cfm, .exe, .pl, .class และสัญลักษณ์ ?, =, & และ comma จะทำการตัด URL นั้นออกไป นอกจากนี้ยังตรวจสอบรหัสสถานะของการดำเนินการ ซึ่งต้องเป็นการดำเนินการที่สำเร็จ เช่น รหัสสถานะ 200 เป็นต้น ถ้าพบว่ารหัสสถานนะดำเนินการไม่สำเร็จก็จะไม่นำ URL นั้นมาวิเคราะห์เช่นกัน

การจำแนกผู้ใช้ (User Identification)

การจำแนกผู้ใช้นั้นค่อนข้างซับซ้อน ถ้า IP address เดียวกัน แต่ Agent Log หรือระบบปฏิบัติการมีการเปลี่ยนแปลง แสดงว่ามาจาก Agent ที่ต่างกัน สำหรับ IP address ก็จะแตกต่างกันด้วย กระบวนการที่ทำเป็นการระบุผู้ใช้ของข้อมูลในกลุ่มเดียวกัน ถึงแม้ว่าการจัดกลุ่มข้อมูลในขั้นตอนข้างต้นจะได้ข้อมูลที่มีหมายเลขไอพีเดียวกันอยู่รวมเป็นกลุ่มเดียวกันแล้วก็ตาม แต่ก็ยังไม่สามารถจำแนกได้ว่าข้อมูลที่เกิดขึ้นนั้นเป็นการเข้าใช้งานของผู้ใช้คนหนึ่ง ๆ เพราะเครื่องคอมพิวเตอร์เครื่องเดียวอาจจะมีผู้ใช้งานได้หลายคน ดังนั้นจึงจำเป็นต้องทำการระบุให้ได้ว่า การร้องขอที่เกิดขึ้นจากเครื่องคอมพิวเตอร์เครื่องหนึ่ง ๆ นั้น เป็นการร้องขอของผู้ใช้รายบุคคล ซึ่งจะใช้ช่วงของระยะเวลาเป็นวิธีการแบ่งข้อมูลเพื่อระบุความแตกต่าง ซึ่งในวิทยานิพนธ์นี้ใช้ช่วงเวลาในการแบ่งผู้ใช้เท่ากับ 30 นาที โดยพิจารณาจากการใช้งานของผู้ใช้เครื่องแต่ละคนที่มีการใช้งานเครื่องคอมพิวเตอร์อย่างต่อเนื่องในระยะเวลาหนึ่งของการเข้าถึงข้อมูล เพราะฉะนั้นหากเวลาที่บันทึกของข้อมูลที่ต่อเนื่องกันมีช่วงระยะเวลามากกว่า 30 นาที แสดงว่าอาจมีการเปลี่ยนผู้ใช้ในการเข้าถึงข้อมูลต่าง ๆ ของเว็บไซต์

การจำแนกเซสชัน (Session Identification)

Web Access Log ที่มีการใช้เวลาในการเยี่ยมชมเว็บไซต์ โดยจุดมุ่งหมายในการจำแนกเซสชันเพื่อจำแนกผู้ใช้แบบหนึ่งเดียว ซึ่งผู้ใช้คนเดียวกันจะมีเลข IP address เหมือนกัน เซสชันใหม่ถูกสร้างจาก IP address ใหม่ แต่ถ้ามีการเยี่ยมชมเว็บไซต์เกินกว่าเวลาที่กำหนด ให้ถือว่าเป็นเซสชันใหม่

การจัดกลุ่มข้อมูล (Data Grouping)

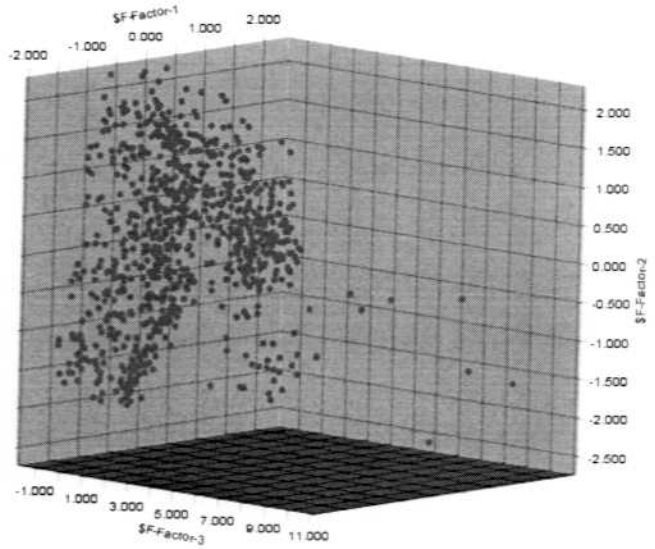
แฟ้มข้อมูลบันทึกการใช้บางประเภท จะมีการบันทึกข้อมูลเฉพาะที่สามารถระบุเครื่องคอมพิวเตอร์ เช่น ชื่อของเครื่องคอมพิวเตอร์ หรือหมายเลขไอพี และโปรแกรมตัวแทนการทำงานของผู้ใช้ (User Agent) เป็นต้น หรือบางเว็บไซต์ที่มีการสมัครใช้งาน หรือลงชื่อเพื่อใช้งาน บันทึกการเข้าใช้งานจะบันทึกชื่อการเขาใช้งาน (User Login) ซึ่งในกรณีข้างต้นสามารถใช้ข้อมูลชื่อผู้เข้าใช้สำหรับระบุผู้ใช้เพื่อจัดกลุ่มข้อมูล แต่ถ้าไม่มีการให้ผู้ใช้ลงชื่อเข้าใช้งาน อาจจะใช้ IP ในการจัดกลุ่ม โดยข้อมูลที่มีหมายเลข IP เดียวกันจะถูกบันทึกลงในแฟ้มข้อมูลเดียวกัน

ภาคผนวก ข

ลักษณะของชุดข้อมูลที่ใช้ในการทดลองเมื่อลดมิติของชุดข้อมูล โดยใช้ Principle Component Analysis (PCA) ให้เหลือเพียง 3 มิติ

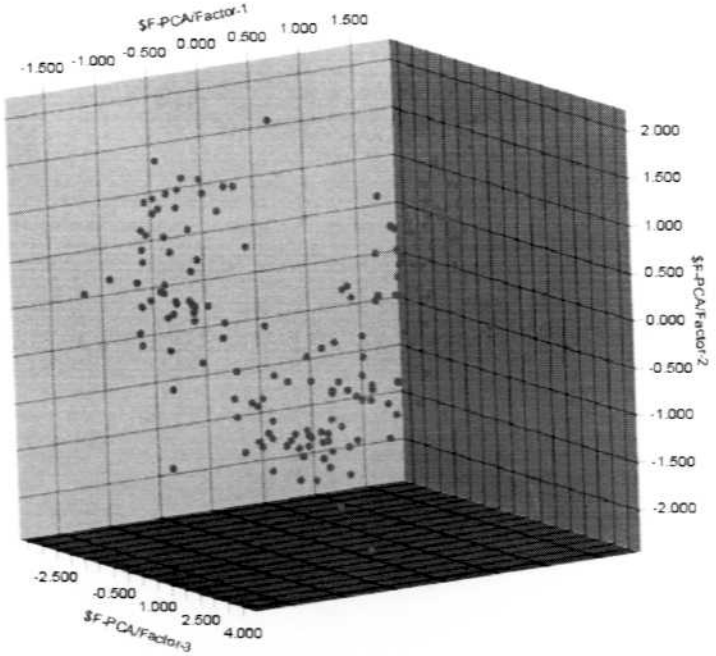
1. ชุดข้อมูล Vehicle (826x18 มิติ)

Extraction Sums of Squared Loading (Cumulative): 79.720 %

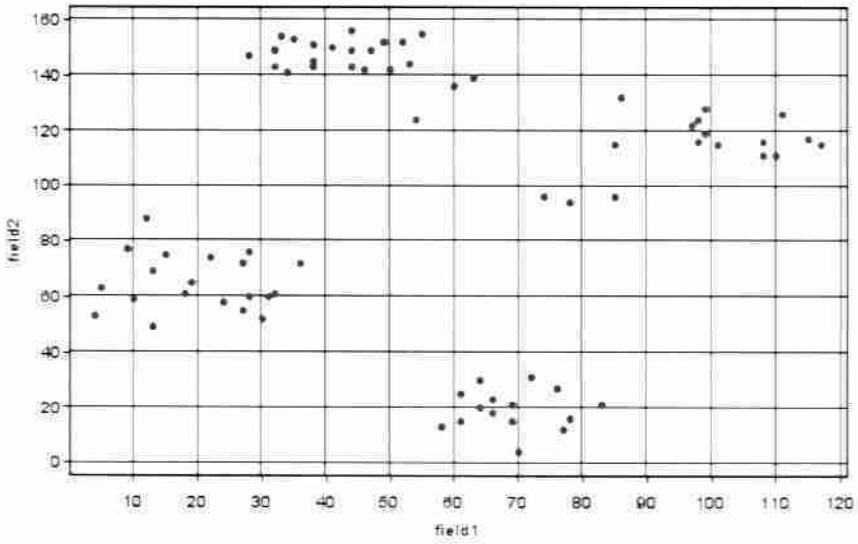


2. ชุดข้อมูล Wine (178x13 มิติ)

Extraction Sums of Squared Loading (Cumulative): 66.530 %

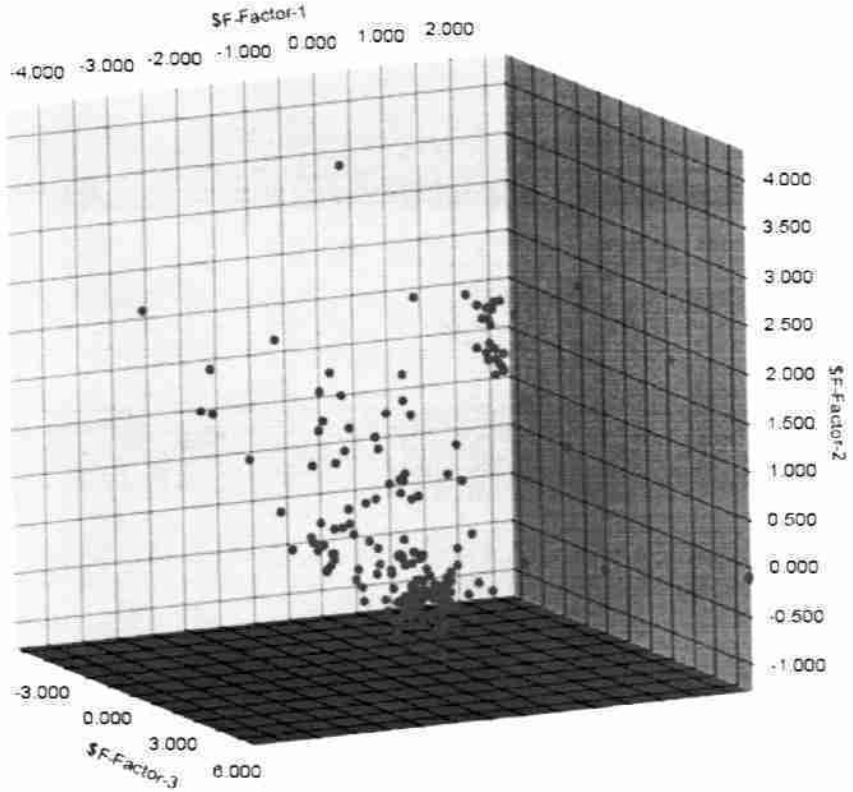


3. ชุดข้อมูล Runpini (จากการทดสอบกับอัลกอริทึม CCIA จำนวนมิติ 75x2 มิติ)



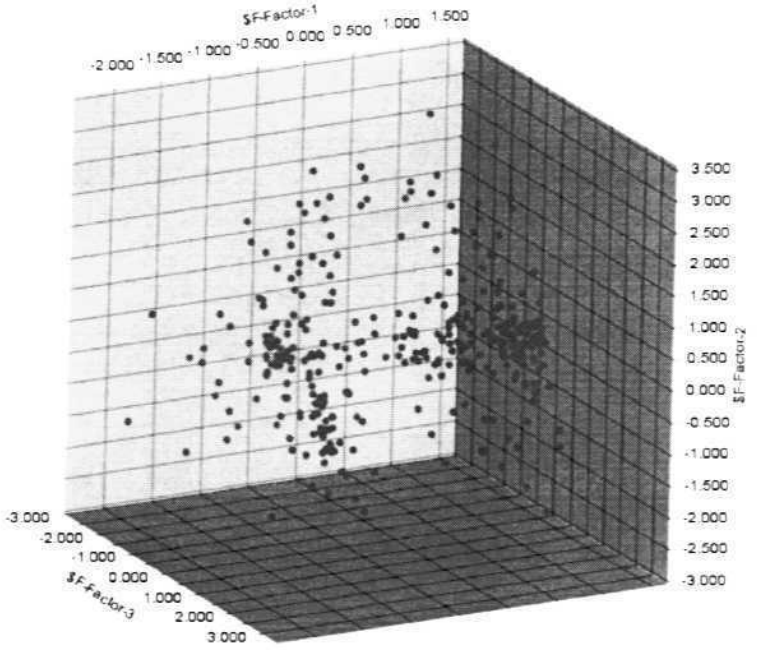
4. ชุดข้อมูล Glass (214x10 มิติ)

Extraction Sums of Squared Loading (Cumulative): 66.290 %



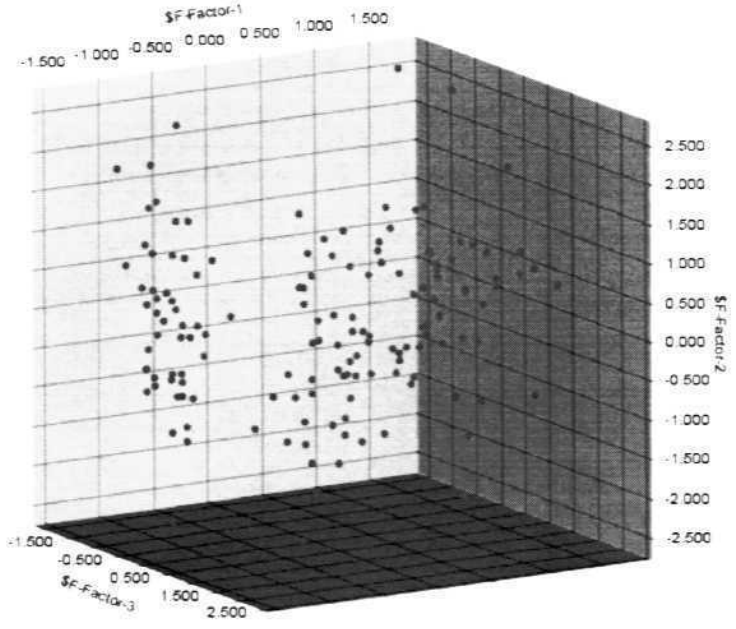
5. **ชุดข้อมูล Ionosphere (351x34 มิติ)**

Extraction Sums of Squared Loading (Cumulative): 48.008 %



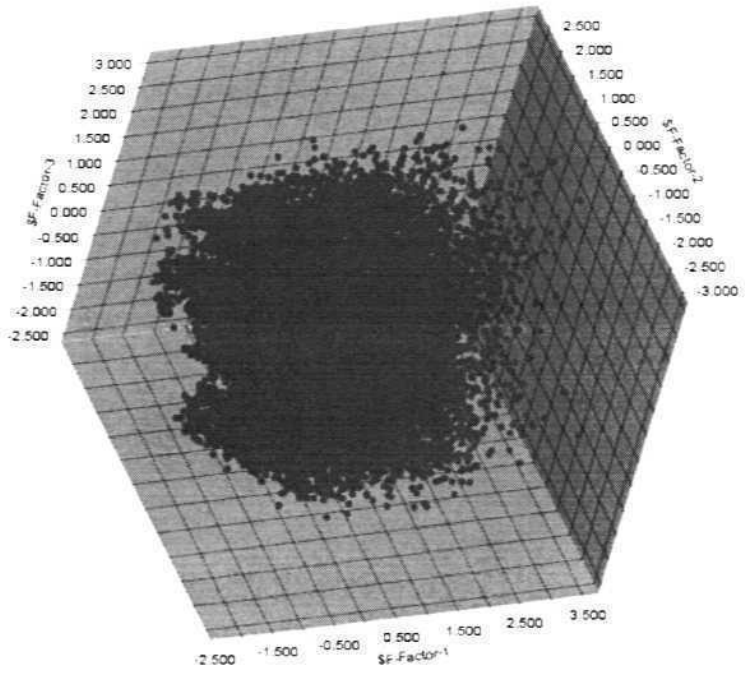
6. **ชุดข้อมูล Iris (150x4 มิติ)**

Extraction Sums of Squared Loading (Cumulative): 99.485 %



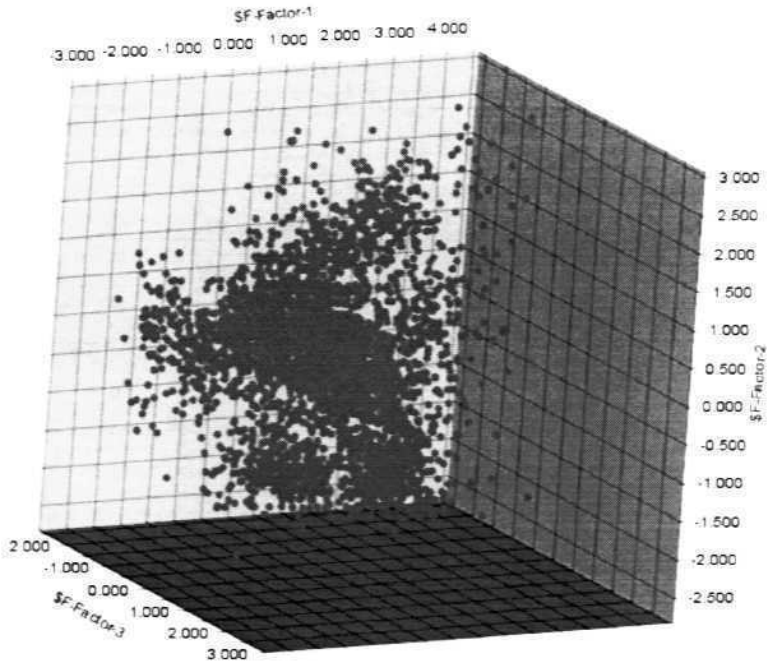
7. **ชุดข้อมูล Letter (20000x16 มิติ)**

Extraction Sums of Squared Loading (Cumulative): 54.012 %



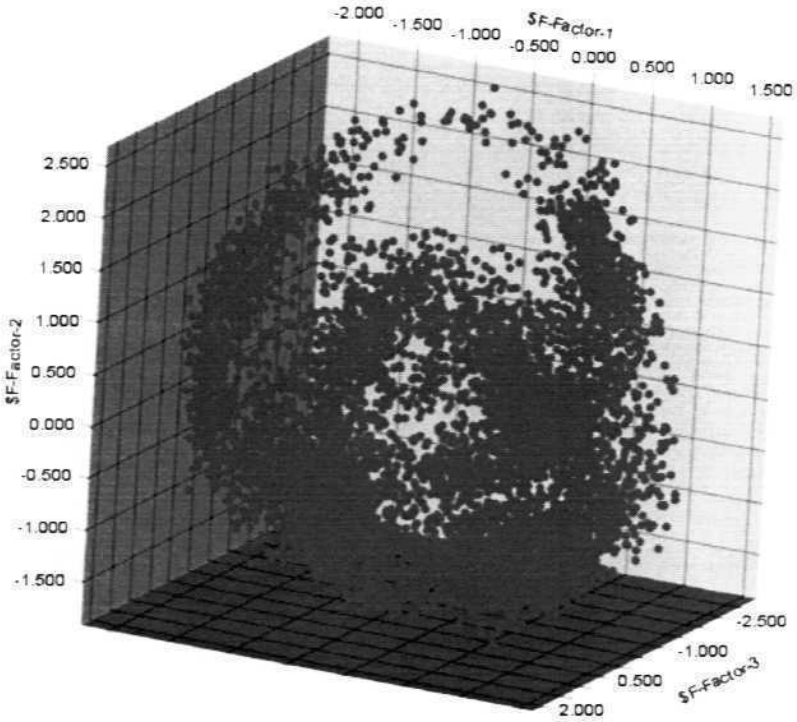
8. **ชุดข้อมูล Opdigits (5620x64 มิติ)**

Extraction Sums of Squared Loading (Cumulative): 29.466 %



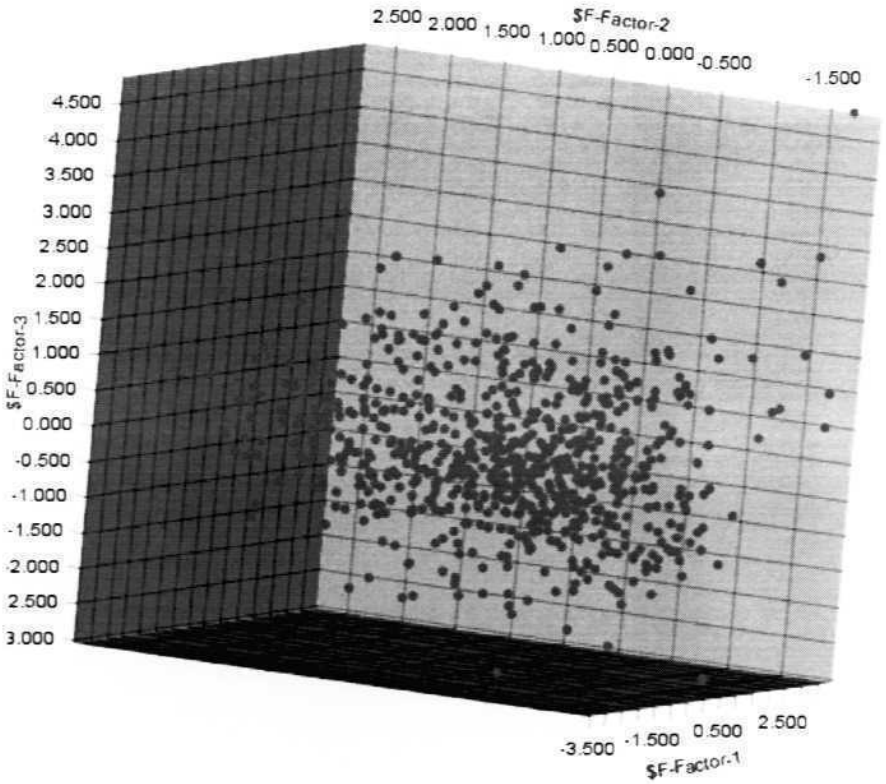
9. ชุดข้อมูล Pendigits (10992x12 มิติ)

Extraction Sums of Squared Loading (Cumulative): 65.685 %



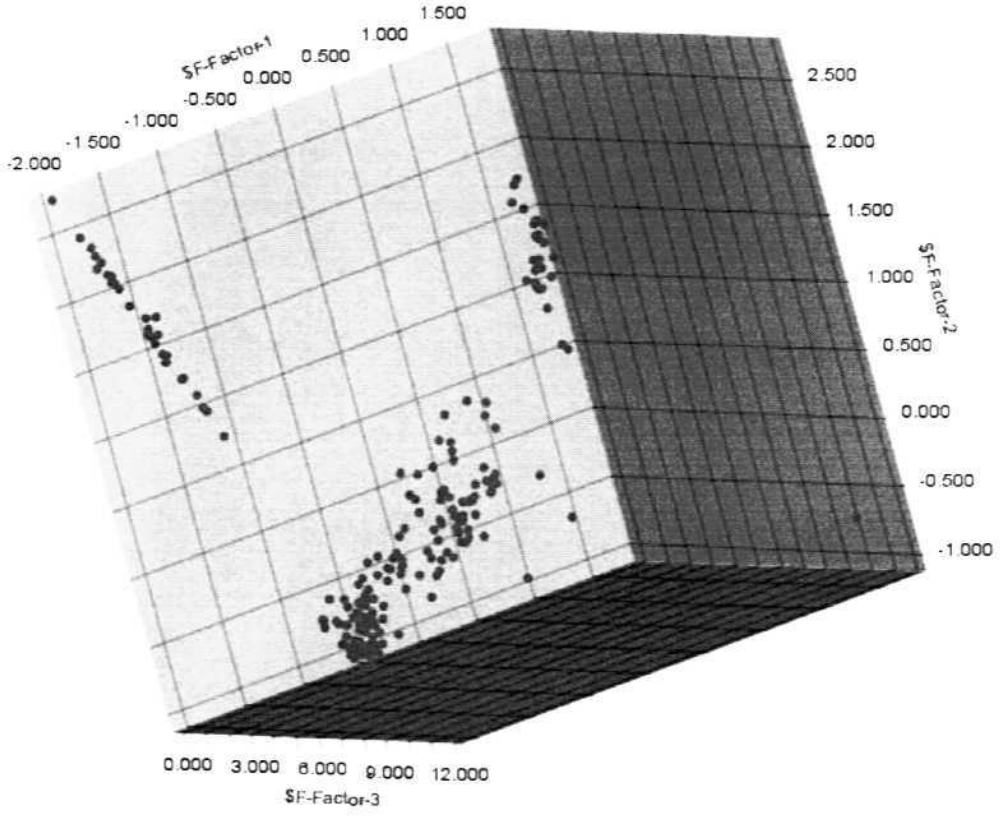
10. ชุดข้อมูล Pima (768x8 มิติ)

Extraction Sums of Squared Loading (Cumulative): 60.690 %



11. ชุดข้อมูล Segmentation (2310 x 19 มิติ)

Extraction Sums of Squared Loading (Cumulative): 74.835 %



บรรณานุกรม

จารุทรรศน์ พัฒนพันธ์ชัย. 2549. **แม่แบบการปรับปรุงประสิทธิภาพของระบบเว็บแคชชิงด้วยการทำเหมืองข้อมูลบันทึกการใช้งานเว็บ: กรณีศึกษาเครือข่ายคอมพิวเตอร์มหาวิทยาลัยสงขลานครินทร์**. วิทยานิพนธ์ปริญญามหาบัณฑิต มหาวิทยาลัยสงขลานครินทร์

เยาวเรศ ศิริสถิตย์กุล. 2546. **การแบ่งชั้นสีโดยใช้ระยะทางระหว่างสีที่ติดกันตามแกนสีที่มีความแปรปรวนสูงสุด**. วิทยานิพนธ์ปริญญามหาบัณฑิต สถาบันบัณฑิตพัฒนบริหารศาสตร์

Anderberg, M.R. 1973. **Cluster Analysis for Applications**. New York: Academic press.

Arotaritei, D. and Mitra, S. 2004. Web Mining: a Survey in the Fuzzy Framework. **Fuzzy sets and Systems**. 148: 5-19.

Babu, G.P. and Murty, M.N. 1993. A Near-Optimal Initial Seed Value Selection in K-means Algorithm Using a Genetic Algorithm. **Pattern Recognition Lett**. 14 (10): 763-769.

Berkhin, P. 2006. Survey of Clustering Data Mining Techniques. In **Grouping Multidimensional Data: Recent Advances in Clustering**. Charles Nicholas and Marc Teboulle, eds. Berlin: Springer.

Blake, C.L. and Merz, C.J.. 1998. **UCI Repository of Machine Learning Databases**. Retrieved January 31, 2007 from http://www.ics.uci.edu/~mlearn/MLRepository.html.

Bradley, P.S. and Fayyad, U.M. 1998. Refining Initial Points for K-means Clustering. In **Proceeding of The Fifteenth International Conference on Machine Learning**. San Francisco: Morgan Kaufmann. Pp. 91-99.

Cooley, R. 2000. **Web Usage Mining: Discovery and Application of Interesting Pattern from Web Data**. Doctoral dissertation, University of Minnesota.

Cooley, R., Mobasher, B. and Srivastava, J. 1999. Data Preparation for Mining World Wide Web Browsing Patterns. **Knowledge and Information System**. 1(1): 5-32.

Daoud, M.B.A. and Roberts, S.A. 1996. New Methods for the Initialization of Clusters. **Pattern Recognition Lett**. 17 (5): 451-455.

Dunham, Margaret H. 2003. **Data Mining: Introductory and Advanced Topics**. New Jersey: Prentice Hall.

Han, J. and Kamber, M. 2001. **Data Mining: Concepts and Techniques**. San Diego: Morgan Kaufmann Publishers.

Huang, C. and Harris, R., 1993. A Comparison of Several Codebook Generation Approaches. **IEEE Trans. Image Process**. 2(1), 108-122.

Jain, A. and Dubes, R. 1998. **Algorithms for Clustering Data** . Englewood Cliffs, NJ: Prentice-Hall.

Jain, A. K., Murty, M. N. and Flynn, P. J. 1999. Data Clustering: Review. **ACM Computing Survey**. 31: 264-323.

Katsavounidis, I., Kuo, C.C.J. and Zhen, Z. 1994. A New Initialization Technique for Generalized Lloyd Iteration. **Signal Process. Lett. IEEE** 1 (10): 144–146.

Kaufman, L. and Rousseeuw, P.J., 1990. **Finding Groups in Data: An Introduction to Cluster Analysis**. Canada: Wiley.

Khan, S.S. and Ahmad, A. 2004. Cluster Center Initialization for K-mean Clustering. **Pattern Recognition Letters**. 25 (11):1293-1302.

Larose, Daniel T. 2005. **Discovering Knowledge in Data: An Introduction to Data Mining.** New Jersey: John Wiley&Sons.

Likas, A., Vlassis, N. and Verbeek, J.J. 2003. The Global K-means Clustering Algorithm. **Pattern Recognition**. 36 (2): 451-461.

Linde, Y., Buzo, A. and Gray, R.M. 1980. An Algorithm for Vector Quantizer Design. **IEEE Trans. Commun**. 28: 84-95.

McQueen, J.B. 1976. Some Methods for Classification and Analysis of Multivariate Observation. In **Symposium on Mathematical Statistics and Probability**. L.M. Le Cam and J., Neyman, eds. Berkeley: University of California Press, Pp. 281-297.

Mirkin, B. G. 2005. **Clustering for Data Mining: A Data Recovery Approach**. Boca Raton: Chapman & Hall/CRC.

Mitra, P., Murthy, C.A. and Pal, S.K. 2002. Density Based Multiscale Data Condensation. **IEEE Trans. Pattern Anal. Machine Intell**. 24 (6): 734–747.

Perkowitz, M. and Etzioni, O. 1999. Adaptive Web Sites: Conceptual Cluster Mining. In **Sixteenth International Joint Conference on Artificial Intelligence**. Stockholm: Morgan Kaufmann. Pp. 264-269.

Romero, C. and Ventura, S. 2006. Educational Data Mining: A Survey from 1995 to 2005. **Expert Systems with Applications: An International Journal**. 33 (1): 135-146.

Tou, J. and Gonzales, R. 1974. **Pattern Recognition Principle**. Reading, MA: Addison Wesley.

ประวัติผู้เขียน

ชื่อ นามสกุล	นายสิริชัย ดีเลิศ
ประวัติการศึกษา	วิทยาศาสตรบัณฑิต (สถิติ) มหาวิทยาลัยศิลปากร ปีการศึกษา 2541
สถานที่ทำงาน	คณะวิทยาการจัดการ มหาวิทยาลัยศิลปากร วิทยาเขตสารสนเทศเพชรบุรี