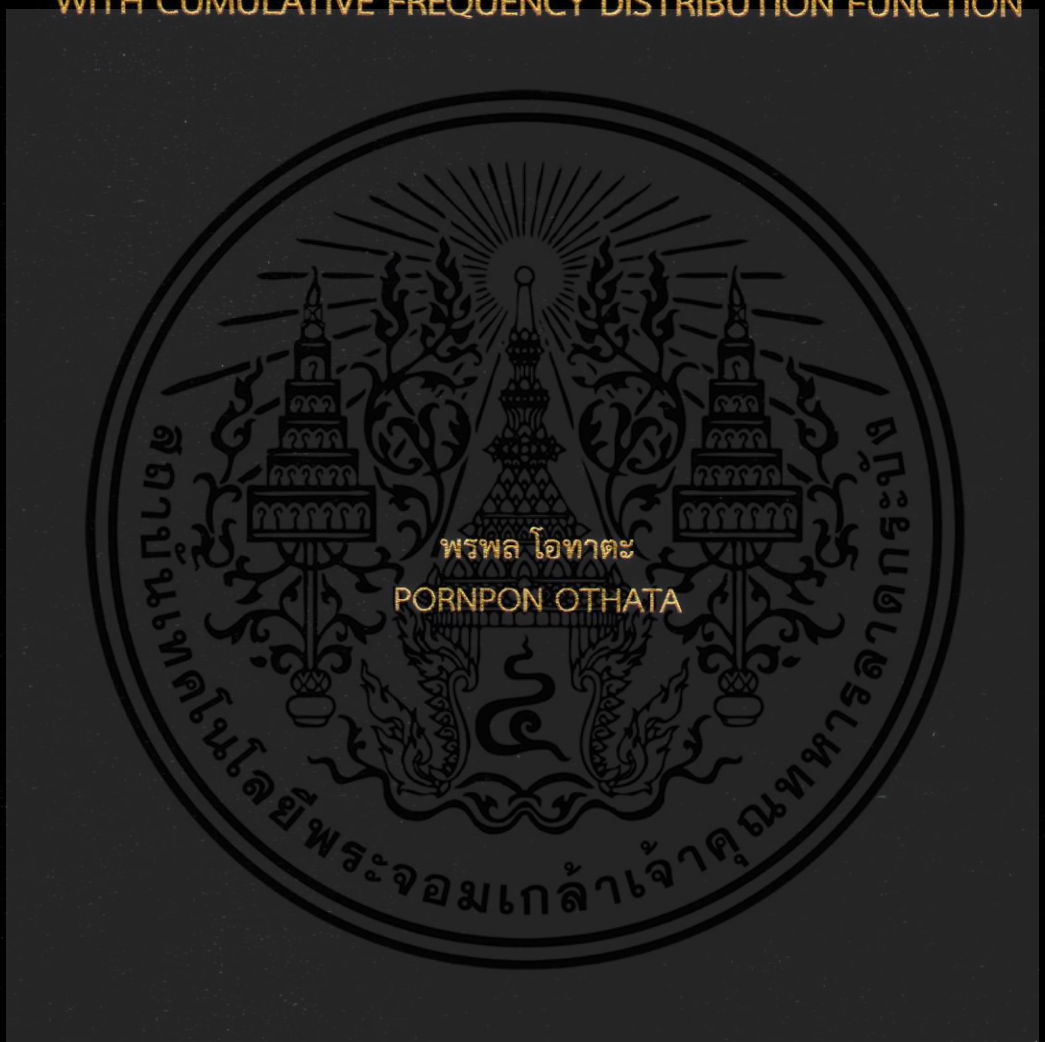
การประมาณจำนวนกลุ่มสำหรับการแบ่งกลุ่มแบบเคมีนโดยฟังก์ชันการแจกแจงความถี่สะสม

APPROXIMATING NUMBER OF CLUSTERS FOR K-MEAN CLUSTERING WITH CUMULATIVE FREQUENCY DISTRIBUTION FUNCTION



พรพล โอทาตะ

PORNPON OTHATA

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาคณิตศาสตร์ประยุกต์
ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2560
KMITL-2017-SC-M-001-026

การประมาณจำนวนกลุ่มสำหรับการแบ่งกลุ่มแบบเคมีนโดยฟังก์ชันการแจกแจงความถี่สะสม

APPROXIMATING NUMBER OF CLUSTERS FOR K-MEAN CLUSTERING WITH CUMULATIVE FREQUENCY DISTRIBUTION FUNCTION

พรพล โอทาตะ

PORNPON OTHATA

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาคณิตศาสตร์ประยุกต์
ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2560

KMITL-2017-SC-M-001-026

APPROXIMATING NUMBER OF CLUSTERS FOR K-MEAN CLUSTERING WITH CUMULATIVE FREQUENCY DISTRIBUTION FUNCTION

PORNPON OTHATA

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN APPLIED MATHEMATICS

DEPARTMENT OF MATHEMATICS
FACULTY OF SCIENCE
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2017
KMITL-2017-SC-M-001-026

คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ใบรับรองวิทยานิพนธ์

หัวข้อวิทยานิพนธ์ "การประมาณจำนวนกลุ่มสำหรับการแบ่งกลุ่มแบบเคมีนโดยฟังก์ชันแจกแจงความถี่สะสม"
(APPROXIMATING NUMBER OF CLUSTERS FOR K-MEAN CLUSTERING WITH CUMULATIVE FREQUENCY DISTRIBUTION FUNCTION)

ชื่อนักศึกษา นายพรพล โอทาตะ
รหัสประจำตัว 58605021
ปริญญา วิทยาศาสตรมหาบัณฑิต (สาขาคณิตศาสตร์ประยุกต์)
ภาควิชา คณิตศาสตร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์ รศ.ไพรบูลย์ พันธรักษ์พงษ์

คณะกรรมการสอบวิทยานิพนธ์	ลายมือชื่อ
ผศ.ดร.ภัทราวุธ จันทร์เสงี่ยม ประธานกรรมการ	
ดร.วรรณพร สรรประเสริฐ อาจารย์บัณฑิตประจำ (ในสาขาวิชาที่เกี่ยวข้อง)	
ผศ.ดร.กลศ พัฒนะรพีเลิศ ผู้ทรงคุณวุฒิจากภายนอกสถาบันฯ	
รศ.ไพรบูลย์ พันธรักษ์พงษ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์	

วัน/ เดือน/ ปี ที่สอบ 4 กรกฎาคม พ.ศ. 2560 เวลา 09.00 – 12.00 น.
สถานที่สอบ ณ ห้อง 306 อาคารพระจอมเกล้า

คณะวิทยาศาสตร์รับรองแล้ว

(รองศาสตราจารย์ ดร.ดุษณี ธนะบริพัฒน์)
คณบดีคณะวิทยาศาสตร์
วันที่ 17 **เดือน** ก.ค. **พ.ศ.** 60

หัวข้อวิทยานิพนธ์	การประมาณจำนวนกลุ่มสำหรับวิธีแบ่งกลุ่มแบบเคมีนโดยฟังก์ชันการแจกแจงความถี่สะสม
ชื่อนักศึกษา	นายพรพล โอทาตะ
รหัสประจำตัว	58605021
ปริญญา	วิทยาศาสตรมหาบัณฑิต (คณิตศาสตร์ประยุกต์)
ภาควิชา	คณิตศาสตร์
พ.ศ.	2560
อาจารย์ที่ปรึกษาวิทยานิพนธ์	รศ.ไพรบูลย์ พันธรักษ์พงษ์

บทคัดย่อ

ในงานวิจัยนี้นำเสนอ วิธีการประมาณจำนวนกลุ่มของวิธีการแบ่งกลุ่มข้อมูลแบบเคมีนสำหรับฐานข้อมูลใน $\mathbb{R}^2$ โดยขั้นตอนวิธีสำหรับการประมาณจำนวนกลุ่มของข้อมูล n ตัวแบ่งออกเป็น 2 ขั้นตอนดังนี้ (1) คำนวณการแจกแจงความถี่สะสม (CFD) ของคุณลักษณะแรก (2) หาจุดแบ่งข้อมูลของคุณลักษณะแรกโดยการพิจารณาอัตราการเปลี่ยนแปลงของ CFD (3) คำนวณการแจกแจงความถี่สะสม (CFD) ของคุณลักษณะที่สองเป็นช่วงๆ โดยใช้จุดแบ่งที่ได้จาก (2) และพิจารณาอัตราการเปลี่ยนแปลงของ CFD ของแต่ละช่วงเพื่อหาจำนวนกลุ่มของฐานข้อมูล โดยขั้นตอนวิธีนี้จะใช้เวลาในการคำนวณ $O(n)$

คำสำคัญ: การแจกแจงความถี่สะสม, วิธีแบ่งกลุ่มข้อมูลแบบเคมีน

Thesis Title	Approximating Number of Clusters for K-Mean Clustering with Cumulative Frequency Distribution Function
Student Name	Pornpon Othata
Student ID	58605021
Degree	Master of Science (Applied Mathematics)
Department	Mathematics
Year	2017
Thesis Advisor	Assoc. Prof. Praiboon Pantaragphong

Abstract

In this research will present new algorithm for calculate number of clusters for k-mean clustering with data set on $\mathbb{R}^2$. The algorithm consist of there steps for n data as following. (1) Compute the cumulative frequency distribution (CFD) of first attribute. (2) Finding breakpoint by consider rate of CFD change on first attribute. (3) Compute the CFD of second attribute with breakpoint form (2) and consider the number of cluster in each subgroup for finding the number of cluster. The new algorithm has time complexity $O(n)$.

Keywords: cumulative frequency distribution, k-mean clustering

กิตติกรรมประกาศ

วิทยานิพนธ์นี้จะไม่สามารถสำเร็จลุล่วงได้หากไม่ได้รับความช่วยเหลือและคำปรึกษาจาก

1. รองศาสตราจารย์ ไพรบูลย์ พันธรักษ์พงษ์ อาจารย์ที่ปรึกษา ที่ให้คำแนะนำและแก้ไขปัญหาต่างๆที่เกิดขึ้นเพื่อนำไปปรับใช้ ในการทำวิทยานิพนธ์นี้เป็นอย่างดี

2. ผศ.ดร.ภัทราวุธ จันทร์เสงี่ยม, ดร.วรรณพร สรรประเสริฐ และ ผศ.ดร.กลศ พัฒนระพีเลิศ กรรมการวิทยานิพนธ์ที่ให้คำแนะนำเพื่อนำไปปรับปรุงในการทำวิทยานิพนธ์นี้

3. คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ในการสนับสนุนทุนอุดหนุนการศึกษาตลอดการศึกษาระดับปริญญาโท

นาย พรพล โอทาตะ

สารบัญ

สารบัญ(ต่อ)

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

สารบัญรูปภาพ(ต่อ)

หน้า

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญ

การแบ่งกลุ่ม (clustering) เป็นวิธีการวิเคราะห์ข้อมูลเพื่อจัดกลุ่มข้อมูล โดยจะพิจารณาข้อมูลที่มีความคล้ายหรือเหมือนกันจัดไว้ในกลุ่มเดียวกัน ซึ่งในปัจจุบันนี้มีวิธีการและงานวิจัยที่เกี่ยวข้องกับการจัดกลุ่มข้อมูลอยู่มากมาย แต่มีวิธีหนึ่งที่ได้รับความนิยมอย่างแพร่หลายนั่นคือวิธีแบ่งกลุ่มแบบเคมีน

วิธีแบ่งกลุ่มแบบเคมีน (k-mean clustering) [1] เป็นวิธีที่ใช้สำหรับการจัดกลุ่มข้อมูลที่ได้รับความนิยมมากวิธีหนึ่งด้วยการจัดกลุ่มข้อมูลออกเป็น k กลุ่ม ซึ่งจะต้องระบุค่า k ก่อนเริ่มวิเคราะห์จัดกลุ่ม ค่า k ของวิธีแบ่งกลุ่มแบบเคมีน นั้นสามารถหาได้จากการพิจารณาค่า k ที่เป็นไปได้ทั้งหมดและการจะได้ค่า k ที่เหมาะสมจะใช้เวลาในการคำนวณค่อนข้างมาก รายละเอียดของวิธีแบ่งกลุ่มแบบเคมีน สามารถดูได้ในบทที่ 2 หัวข้อ 2.2

ในปัจจุบันนี้ได้มีงานวิจัยเกี่ยวกับวิธีการเลือกค่า k ที่เหมาะสมมากมาย เช่น วิธีเอลโบ (elbow method) [2] โดยวิธีนี้จะเลือกค่า k จากการวิเคราะห์ค่าคาดเคลื่อน (sum of square error) ของทุกค่า k ที่เป็นไปได้ ดังนั้นจึงทำให้มีรอบการทำงานมาก เนื่องจากต้องผ่านวิธีแบ่งกลุ่มแบบเคมีนหลายรอบ

ดังนั้นในงานวิจัยนี้จะนำเสนอขั้นตอนวิธีสำหรับประมาณค่า k ของ วิธีแบ่งกลุ่มแบบเคมีน เพื่อช่วยลดเวลาในการพิจารณาค่า k ที่ไม่จำเป็นต้องพิจารณาค่า k ที่เป็นไปได้ทั้งหมด โดยพิจารณาความหนาแน่นของข้อมูลในแต่ละคุณลักษณะ (attribute) ซึ่งพิจารณาจากอัตราการเปลี่ยนแปลงของการแจกแจงความถี่สะสม (rate of cumulative frequency distribution change) โดยเรียกวิธีนี้ว่าวิธี RCFDC โดยรายละเอียดของขั้นวิธีสามารถดูได้ในบทที่ 3

1.2 วัตถุประสงค์ของการศึกษา

1) เพื่อศึกษาเกี่ยวกับการหาจำนวนกลุ่มของวิธีการแบ่งกลุ่มข้อมูลแบบเคมีน
2) เพื่อออกแบบขั้นตอนวิธีสำหรับประมาณจำนวนกลุ่มของวิธีการแบ่งกลุ่มข้อมูลแบบเคมีน

1.3 ประโยชน์ของการศึกษา

สามารถประมาณจำนวนกลุ่มสำหรับวิธีแบ่งกลุ่มแบบเคมีนซึ่งได้ค่าที่ใกล้เคียงค่าที่เหมาะสม โดยไม่จำเป็นต้องพิจารณาจำนวนกลุ่มจากทุกค่าที่เป็นไปได้ทำให้ลดเวลาในการพิจารณาค่า k ของวิธีการแบ่งกลุ่มแบบเคมีน

1.4 ของเขตของการศึกษา

1) หาจำนวนกลุ่มของวิธีการแบ่งกลุ่มข้อมูลแบบเคมีน สำหรับข้อมูลใน $\mathbb{R}^2$

2) ใช้ฐานข้อมูล Iris setosa, ฐานข้อมูล air passengers และ ฐานข้อมูล tooth growth ในการศึกษา

1.5 ขั้นตอนในการศึกษาและดำเนินงาน

1) ศึกษาขั้นตอนวิธีของวิธีการแบ่งกลุ่มข้อมูลแบบเคมีน

2) รวบรวมและศึกษางานวิจัยและหนังสือที่เกี่ยวข้องกับการประมาณจำนวนกลุ่มของการแบ่งกลุ่มข้อมูลแบบเคมีน

3) ศึกษาทฤษฎีเกี่ยวกับการแจกแจงความถี่สะสมสะสมของตัวแปรสุ่ม

4) ออกแบบและเขียนขั้นตอนวิธีประมาณจำนวนกลุ่มของวิธีการแบ่งกลุ่มข้อมูลแบบเคมีน

5) ปรับปรุง และแก้ไขในข้อผิดพลาดที่เกิดขึ้น

6) สรุปผลการดำเนินงาน และทำรูปเล่มวิทยานิพนธ์

สามารถแสดงกิจกรรมดำเนินงานได้ ดังตารางที่ 1.1

ตารางที่ 1.1 แผนการดำเนินงานวิจัย

กิจกรรมดำเนินงาน	เดือน / ปี (2559 - 2560)									
	ส.ค.	ก.ย.	ต.ค.	พ.ย.	ธ.ค.	ม.ค.	ก.พ.	มี.ค.	เม.ษ.	พ.ค.
1. ศึกษาขั้นตอนวิธีของวิธีการแบ่งกลุ่มข้อมูลแบบเคมีน										
2. รวบรวม และศึกษางานวิจัยและหนังสือที่เกี่ยวข้องกับการประมาณจำนวนกลุ่มของการแบ่งกลุ่มข้อมูลแบบเคมีน										
3. ศึกษาทฤษฎีเกี่ยวกับฟังก์ชันสะสมของตัวแปรสุ่ม										
4. ออกแบบและเขียนขั้นตอนวิธีประมาณจำนวนกลุ่มของวิธีการแบ่งกลุ่มข้อมูลแบบเคมีน										
5. เปรียบเทียบผลลัพธ์ของการประมาณค่ากับค่าจริงที่ได้จากวิธีการแบ่งกลุ่มข้อมูลแบบเคมีน										
6. ปรับปรุง และแก้ไขในข้อผิดพลาดที่เกิดขึ้น										
7. สรุปผลการดำเนินงาน และทำรูปเล่มวิทยานิพนธ์										

บทที่ 2

เอกสารและงานวิจัยที่เกี่ยวข้อง

2.1 การแบ่งกลุ่ม (Clustering)

การแบ่งกลุ่ม (Clustering) [1] เป็นวิธีการวิเคราะห์ข้อมูล ซึ่งใช้ในการทำเหมืองข้อมูล (Data mining) โดยจะแบ่งชุดข้อมูลออกเป็นกลุ่ม (cluster) โดยนำข้อมูลที่มีคุณลักษณะเหมือนกันหรือคล้ายกันจัดไว้ในกลุ่มเดียวกัน ขั้นตอนวิธีที่ใช้ในการแบ่งกลุ่มจะอาศัยความเหมือน (similarity) หรือความใกล้ชิด (proximity) โดยคำนวณจากการวัดระยะระหว่างข้อมูล โดยใช้การวัดระยะแบบต่าง ๆ เช่น การวัดระยะแบบยูคลิด (Euclidean distance) การวัดระยะแบบแมนฮัตตัน (Manhattan distance) การวัดระยะแบบเชบิเชฟ (Chebychev distance)

โดยการแบ่งกลุ่มในปัจจุบันนั้นมีหลายวิธี เช่น การแบ่งกลุ่มแบบเคมีน (K-means Clustering), การแบ่งกลุ่มเป็นลำดับขั้น (Hierarchical Clustering), การแบ่งกลุ่มสเปกตรัมและกราฟ (Spectral and Graph Clustering) เป็นต้น ซึ่งวิธีการแบ่งกลุ่มที่ได้รับความนิยมอย่างแพร่หลายคือ การแบ่งกลุ่มแบบเคมีน

2.2 ขั้นตอนการแบ่งกลุ่มแบบเคมีน (k-mean clustering algorithm)

วิธีแบ่งกลุ่มแบบเคมีน (k-mean clustering) [1] เป็นวิธีสำหรับแบ่งกลุ่ม (cluster analysis) สำหรับเซตข้อมูล D ในการทำเหมืองข้อมูล วิธีการแบ่งกลุ่มแบบเคมีนเป็นการแบ่งข้อมูล n สิ่งออกเป็น k กลุ่ม โดยวิธีแบ่งกลุ่มแบบเคมีนมีลายละเอียดดังต่อไปนี้

ให้ $C = \{C_1, C_2, ..., C_k\}$ เป็นเซตของกลุ่มที่ได้จากวิธีแบ่งกลุ่มแบบเคมีน โดยการพิจารณาว่าเป็นการจัดกลุ่มที่ดีหรือไม่ จะพิจารณาจากค่าผลรวมค่าคลาดเคลื่อนกำลังสอง (sum of squared errors) กำหนดให้ฟังก์ชันค่าคลาดเคลื่อนกำลังสองเป็นดังนี้

$$SSE(C) = \sum_{i=1}^{k} \sum_{x_j \in C_i} \left\| x_j - \mu_i \right\|^2 \qquad (1)$$

ขั้นตอนแรกของวิธีแบ่งกลุ่มแบบเคมีนจะทำการการสุ่มสร้างจุดในเซตข้อมูล k จุด โดยจุดที่สุ่มขึ้นจะเรียกว่า เซนทรอยซ์ (Centroid) ซึ่งจะสุ่มสร้างจุดกระจายอย่างสม่ำเสมอในเซตข้อมูล โดยในแต่ละรอบของวิธีเคมีน ประกอบไปด้วย 2 ขั้นตอน

1) การกำหนดกลุ่ม (cluster assignment)

กำหนดค่าเฉลี่ยของ k กลุ่ม นั่นคือแต่ละจุด $x_j \in D$ จะอยู่ในบริเวณใกล้เคียงกับค่าเฉลี่ยของแต่ละกลุ่ม จึงทำให้สามารถจัดกลุ่มโดยแต่ละกลุ่ม C_i จะประกอบไปด้วยจุดที่อยู่ใกล้กับค่าเฉลี่ย μ_i ของกลุ่มนั้นมากกว่าค่าเฉลี่ยในกลุ่มอื่นๆ นั่นคือแต่ละจุด x_j จะอยู่ในกลุ่ม C_{j^*} เมื่อ

$$j^* = \arg\min_{i=1}^{k} \left\{ \left\| x_j - \mu_i \right\|^2 \right\} \qquad (2)$$

2) ปรับปรุงเซนทรอยด์ (Centroid Update)

กำหนดเซตของกลุ่ม $C_i,\ i = 1, 2, ..., k$ และคำนวณค่าเฉลี่ยใหม่ของแต่ละกลุ่มใน C_i เพื่อเปลี่ยนตำแหน่งของเซนทรอยด์ไปยังค่านั้น การกำหนดกลุ่มและขั้นตอนของการปรับปรุงเซนทรอยด์ จะดำเนินการซ้ำไปเรื่อยๆ จนกว่าค่าเฉลี่ยจะลู่เข้าค่าๆหนึ่งที่ไม่มีการเปลี่ยนแปลงอีก หรือเซนทรอยด์ไม่มีการเปลี่ยนแปลงนั่นเอง ซึ่งเราสามารถหยุดได้ถ้า

$$\sum_{i=1}^{k} \left\| \mu_i^t - \mu_i^{t-1} \right\|^2 \le \varepsilon, \qquad (3)$$

เมื่อ $\varepsilon > 0$ คือค่าการลู่เข้า, t คือรอบปัจจุบันของขั้นตอน และ μ_i^t คือค่าเฉลี่ยของกลุ่ม C_i ในรอบที่ t ซึ่งคำนวณได้จาก

$$\mu_i^t = \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j \qquad (4)$$

โดยขั้นตอนวิธีของวิธีการแบ่งกลุ่มแบบเคมีน สามารถแสดงได้ดังขั้นตอนวิธี 2.1 [1]

ในแง่ของความซับซ้อนในการคำนวณ เราจะเห็นว่าขั้นตอนกำหนดกลุ่มใช้เวลา $O(nkd)$ เพราะว่าในแต่ละ n จุด มีการคำนวณระยะห่างกับแต่ละ k กลุ่ม ซึ่งมี d การดำเนินการใน d มิติ และขั้นตอนการคำนวณค่าเซนทรอยด์ใหม่ใช้เวลา $O(nd)$ เพราะว่ามี n จุดใน d มิติ สมมติว่ามีการดำเนินการ t รอบ ดังนั้นเวลารวมของวิธีเคมีนคือ $O(tnkd)$

ขั้นตอนวิธีที่ 2.1 การแบ่งกลุ่มแบบเคมีน

K-Means (D,k,ε) :
$t=0$, Random initialize k centroids: $\mu_1^t,\mu_2^t,...,\mu_k^t \in \mathbb{R}^d$
repeat $t \leftarrow t+1$
  $C_j \leftarrow \varnothing$ for all $j=1,...,k$
  // Cluster Assignment Step
  foreach $x_j \in D$ **do**
    $j^* \leftarrow \arg\min_i \left\{ \left\| x_j - \mu_i^t \right\|^2 \right\}$
    //Assign x_j to closest centroid
    $C_{j^*} \leftarrow C_j \cup \{x_j\}$
  // Centroid Update Step
  foreach $i=1$ **to** k **do**
    $\mu_i^t \leftarrow \frac{1}{|C_i|} \Sigma_{x_j \in C_i} x_j$
until $\Sigma_{i=1}^k \left\| \mu_i^t - \mu_i^{t-1} \right\|^2 \le \varepsilon$

ตัวอย่างที่ 1 จากฐานข้อมูล iris setosa (รายละเอียดดูได้ในหัวข้อ 2.5) โดยใช้ความยาวกลีบดอกและความยาวกลีบเลี้ยงเป็นข้อมูล 2 มิติ มีจำนวนข้อมูล $n=150$ และต้องการแบ่งกลุ่มเป็น 3 กลุ่ม ซึ่งสอดคล้องกับ 3 ประเภทของ iris setosa โดยใช้วิธีเคมีน เริ่มต้นโดยการสุ่มค่าเฉลี่ยของกลุ่ม

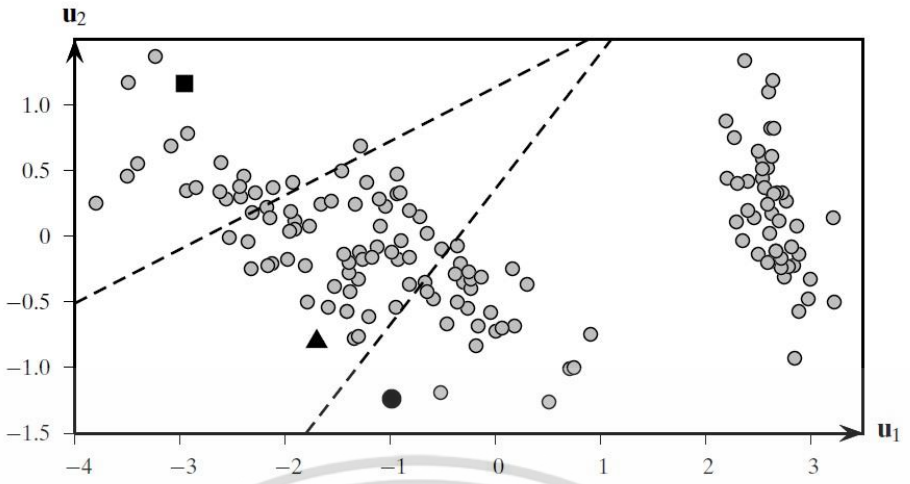
$$\mu_1 = (-0.98,-1.24)^T \qquad \mu_2 = (-2.96,1.16)^T \qquad \mu_3 = (-1.69,-0.80)^T$$

โดยรูปของการแบ่งกลุ่มขั้นแรก แสดงดังรูปที่ 2.1 และค่าเฉลี่ยหลังจากการดำเนินการรอบแรกจะแสดงดังรูปที่ 2.2

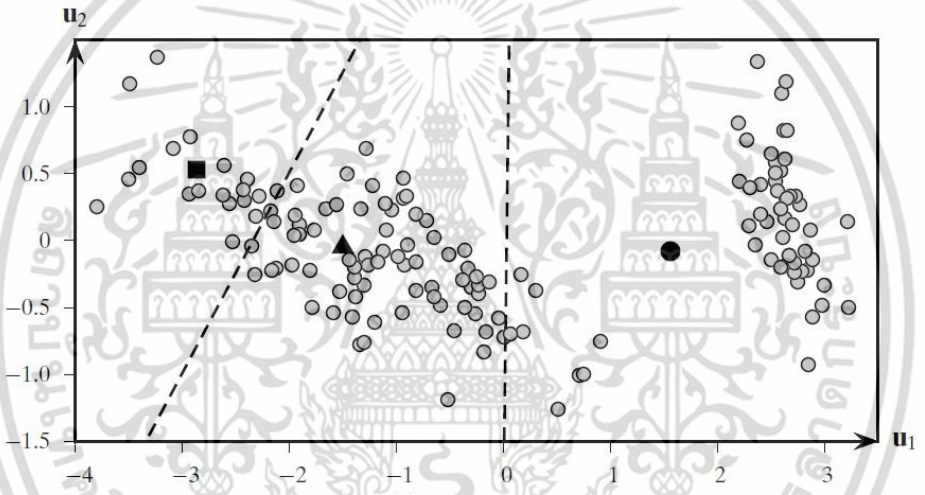
$$\mu_1 = (1.56,-0.08)^T \qquad \mu_2 = (-2.86,0.53)^T \qquad \mu_3 = (-1.50,-0.05)^T$$

และสุดท้าย กลุ่มที่ลู่เข้าแล้วและค่าเฉลี่ยสุดท้าย หลังจากผ่านการดำเนินการทั้งหมด 8 รอบ แสดงดังรูปที่ 2.3

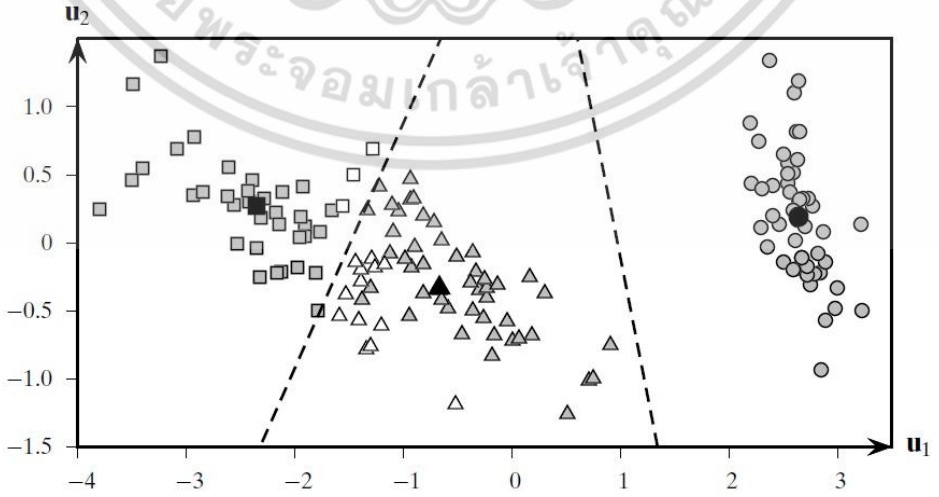
$$\mu_1 = (2.64,0.19)^T \qquad \mu_2 = (-2.35,0.27)^T \qquad \mu_3 = (-0.66,-0.33)^T$$



รูปที่ 2.1 สุ่มจุดเซนทรอยซ์ของแต่ละกลุ่ม



รูปที่ 2.2 ปรับปรุงเซนทรอยซ์รอบที่ 1



รูปที่ 2.3 ปรับปรุงเซนทรอยซ์รอบที่ 8 (ลู่เข้า)

2.3 การแจกแจงความถี่สะสม (Cumulative Frequency Distribution)

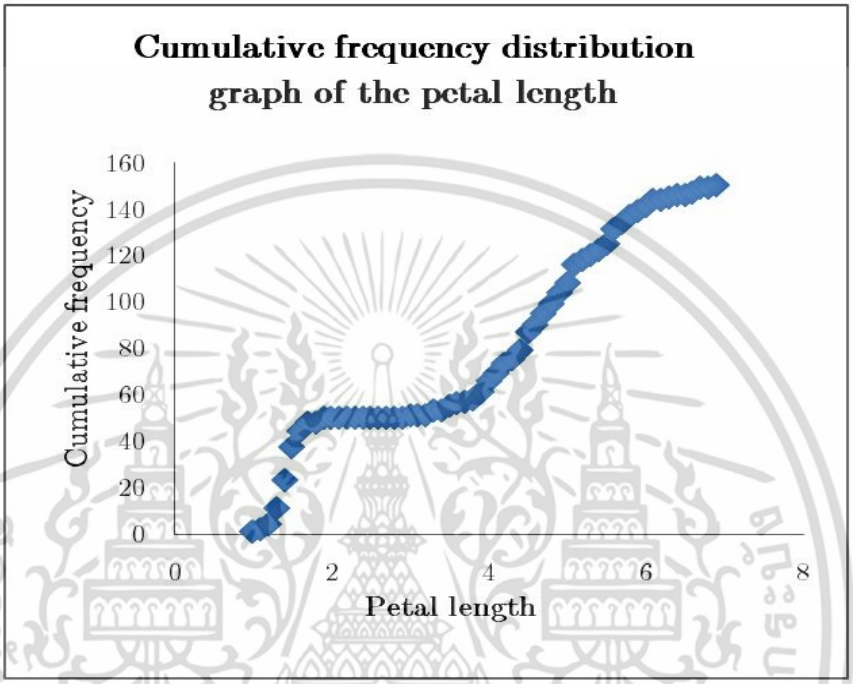
การแจกแจงความถี่สะสม (CFD) [2] ของค่าที่เป็นไปได้ค่าใดหรืออันตรภาคชั้นใด หมายถึงผลรวมของความถี่ของค่านั้นหรืออันตรภาคชั้นนั้น กับความถี่ของค่าหรือของอันตรภาคชั้นที่มีช่วงต่ำกว่าทั้งหมด หรือสูงกว่าทั้งหมดอย่างใดอย่างหนึ่ง โดยจะนับความถี่จากค่าที่มีค่ามากกว่าขอบเขตล่างแต่ไม่เกินขอบเขตบนของอันตรภาคชั้น

โดยตัวอย่างของการแจกแจงความถี่สะสมของความยาวกลีบดอกในฐานข้อมูล iris setosa โดยใช้ความกว้างอันตรภาคชั้นเป็น 0.3 แสดงดังตารางแจกแจงความถี่สะสมต่อไปนี้

ตารางที่ 2.1 การแจกแจงความถี่สะสมของความยาวกลีบดอกในฐานข้อมูล iris setosa

Petal length	Frequency	CFD
0.7-1.0	1	1
1.0-1.3	10	11
1.3-1.6	33	44
1.6-1.9	6	50
1.9-2.2	0	50
2.2-2.5	0	50
2.5-2.8	0	50
2.8-3.1	1	51
3.1-3.4	2	53
3.4-3.7	4	57
3.7-4.0	9	66
4.0-4.3	9	75
4.3-4.6	15	90
4.6-4.9	14	104
4.9-5.2	14	118
5.2-5.5	7	125
5.5-5.8	12	137
5.8-6.1	7	144
6.1-6.4	2	146
6.4-6.7	3	149
6.7-7.0	1	150

โดยกราฟของการแจกแจงความถี่สะสมของความยาวกลีบดอกในฐานข้อมูล iris setosa จะแสดงดังรูปที่ 2.4

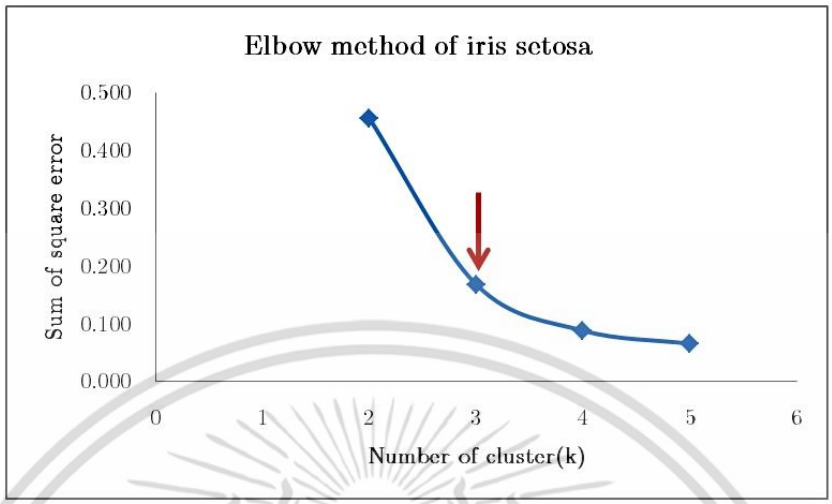


รูปที่ 2.4 กราฟของฟังก์ชันการแจกแจงสะสมสำหรับตัวแปรสุ่มแบบไม่ต่อเนื่องของความยาวกลีบดอกจากฐานข้อมูล iris setosa

2.4 วิธีเอลโบ (Elbow Method)

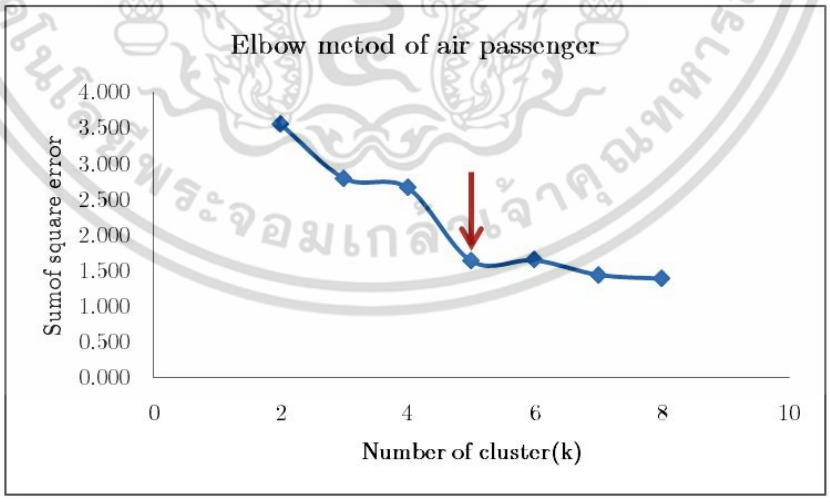
วิธีเอลโบ (Elbow Method) [3] เป็นวิธีสำหรับตีความและตรวจสอบความคล้ายคลึงกันในการวิเคราะห์แบบคลัสกลุ่ม ที่ออกแบบมาเพื่อช่วยในการหาจำนวนกลุ่มที่เหมาะสมสำหรับชุดข้อมูล โดยวิธีนี้จะใช้ในการตรวจสอบจำนวนกลุ่มที่ได้จากวิธี RCFDC ว่ามีความแม่นยำมากน้อยเพียงใด วิธีนี้จะดูเปอร์เซ็นต์ของความคาดเคลื่อน (sum square error) กับจำนวนกลุ่ม โดยควรเลือกใช้หลายกลุ่มในการพิจารณาเพื่อการวิเคราะห์จำนวนกลุ่มที่ดีขึ้น เมื่อสังเกตจากการพล๊อตกราฟของเปอร์เซ็นต์ของความแปรปรวนหรือความคาดเคลื่อนกับจำนวนกลุ่มหากมีการแปลงเปอร์เซ็นต์ของค่าความแปรปรวนหรือค่าคาดเคลื่อนมีค่าลดลงมาก ซึ่งแสดงลักษณะเป็นมุมปรากฏในกราฟ เราจะเลือกจำนวนกลุ่มที่จุดนี้ ซึ่งเรียกการเลือกจุดแบบนี้ว่า “กฎเกณฑ์เอลโบ” ตัวอย่างของการใช้วิธีเอลโบใน

การหาจำนวนกลุ่มที่เหมาะสมสำหรับฐานข้อมูล iris setosa และ air passenger (รายละเอียดสามารถดูได้ในหัวข้อ 2.5) จะแสดงดังกราฟในรูปที่ 2.5 และ 2.6 ตามลำดับ



รูปที่ 2.5 วิธีเอลโบของฐานข้อมูล iris setosa

จากรูปที่ 2.5 จะเห็นว่าในจุด $k=3$ อัตราการลดลงของค่าคาดเคลื่อนมีค่ามาก จากกฎเกณฑ์เอลโบ เราจะได้ว่าค่า k ที่เหมาะสมคือ 3



รูปที่ 2.6 วิธีเอลโบของฐานข้อมูล air passenger

จากรูปที่ 2.6 ในทำนองเดียวกันเราจะได้ว่าค่า k ที่เหมาะสมคือ 5

2.5 ฐานข้อมูลที่ใช้ในการทดลอง

2.5.1 ฐานข้อมูล Iris setosa

ฐานข้อมูล Iris setosa [4] เป็นฐานข้อมูลเกี่ยวกับการจำแนกชนิดของดอก iris ซึ่งประกอบไปด้วยคุณลักษณะภายนอก 4 ลักษณะของดอก iris คือ ความยาวกลีบเลี้ยง (sepal length), ความกว้างกลีบเลี้ยง (sepal width), ความยาวกลีบดอก (petal length) และ ความกว้างกลีบดอก (petal width) โดยในฐานข้อมูลนี้ได้จำแนกชนิดของดอก iris ออกเป็น 3 ชนิด ได้แก่ Iris-setosa, Iris-versicolor และ Iris-virginica ซึ่งลักษณะของดอก iris ทั้ง 3 ชนิดจะมีลักษณะที่คล้ายคลึงกันมากซึ่งแสดงดังรูปที่ 2.7 (ที่มา http://dataaspirant.com/2017/01/25/svm-classifier-implemenation-python-scikit-learn/)

รูปที่ 2.7 ลักษณะของดอก iris ทั้ง 3 ชนิด

2.5.2 ฐานข้อมูล Air passengers

ฐานข้อมูล Air passengers [5] เป็นฐานข้อมูลเกี่ยวกับจำนวนผู้โดยสารสายการบินในแต่ละเดือนในปี 1949-1960 ซึ่งในฐานข้อมูลจะประกอบไปด้วย จำนวนผู้โดยสาร (Air Passengers) และปี ค.ศ ที่เดินทาง (Time)

2.5.3 ฐานข้อมูล Tooth Growth

ฐานข้อมูล Tooth Growth [6] เป็นฐานข้อมูลเกี่ยวกับผลของวิตามินซีกับการเจริญเติบโตของฟัน ซึ่งในฐานข้อมูลจะประกอบไปด้วย ความยาวของฟัน (len), ประเภทอาหารเสริม (VC หรือ OJ) และปริมาณเป็นมิลลิกรัมต่อ 1 โดส (dose)

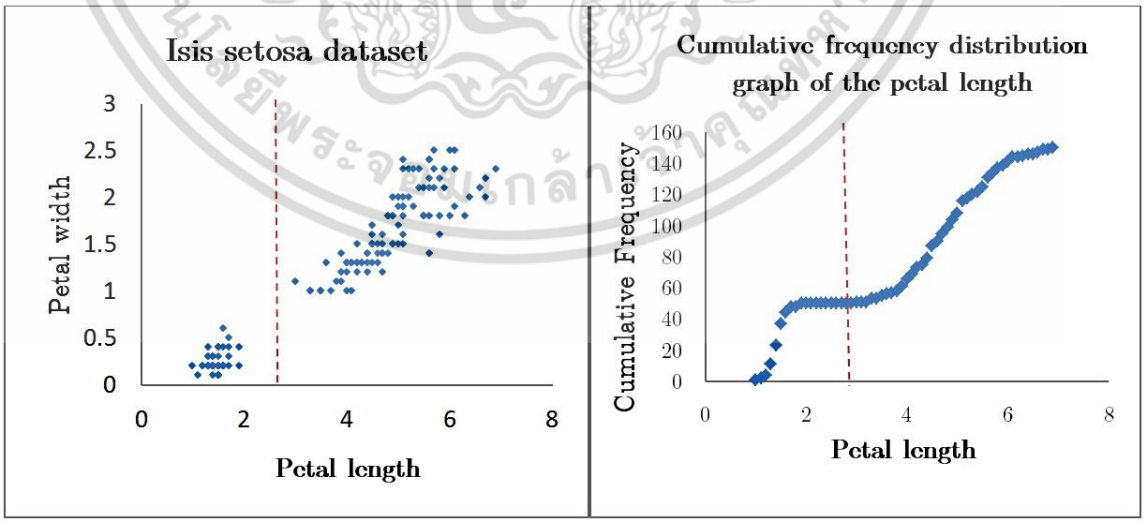
บทที่ 3

วิธี RCFDC

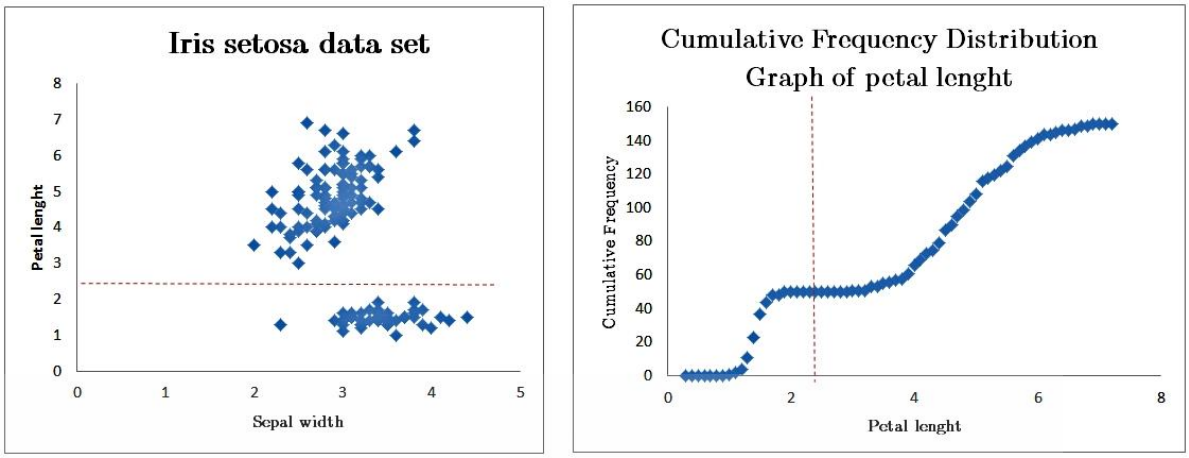
วิธี RCFDC เป็นวิธีที่ใช้คำนวณจำนวนกลุ่ม (k) ที่เหมาะสมสำหรับวิธีการแบ่งกลุ่มแบบเคมีน สำหรับแต่ละจุดข้อมูล x_i ในฐานข้อมูล D สำหรับฐานข้อมูลใน $\mathbb{R}^2$ โดยวิธีนี้สามารถคำนวณจำนวนกลุ่มได้โดยการคำนวณเพียงครั้งเดียว

3.1 อัตราการเปลี่ยนแปลงของ CFD

จำนวนกลุ่มที่คำนวณของวิธี RCFDC จะได้มาจากการพิจารณาความหนาแน่นของข้อมูลในฐานข้อมูล ซึ่งจะพิจารณาความความหนาแน่นของข้อมูลในรูปของ CFD ของฐานข้อมูล โดยเราจะสังเกตได้จากการพิจารณาอัตราการเปลี่ยนแปลงของ CFD โดยถ้าอัตราการเปลี่ยนแปลงของ CFD มีค่ามาก นั่นแสดงให้เห็นว่าบริเวณนั้นของฐานข้อมูลมีความหนาแน่นของข้อมูลมาก ในทางกลับกันถ้าอัตราการเปลี่ยนแปลงของ CFD มีค่าน้อย นั่นแสดงให้เห็นว่าบริเวณนั้นของฐานข้อมูลมีความหนาแน่นน้อย โดยตัวอย่างสำหรับฐานข้อมูล iris setosa ในแต่ละคุณลักษณะแสดงดังรูปที่ 3.1 และ 3.2



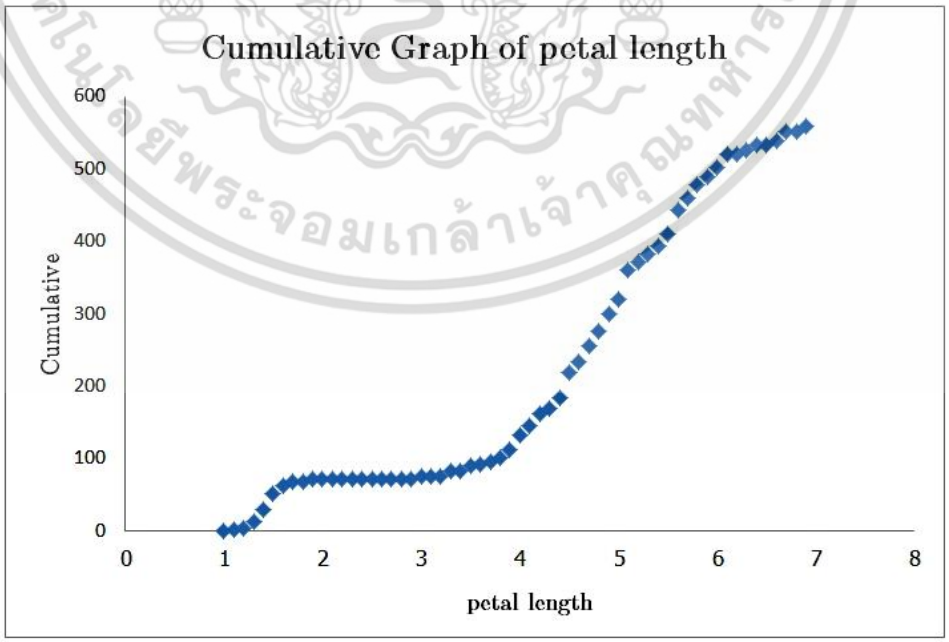
รูปที่ 3.1 รูปภาพแสดงความสัมพันธ์ระหว่างความหนาแน่นของข้อมูลความยาวกลีบดอกและความกว้างกลีบดอกของฐานข้อมูล iris setosa กับอัตราการเปลี่ยนแปลงของ CFD



รูปที่ 3.2 ความสัมพันธ์ระหว่างความหนาแน่นของข้อมูลความยาวกลีบดอกและกว้างกลีบเลี้ยงของฐานข้อมูล iris setosa กับอัตราการเปลี่ยนแปลงของ CFD

จากรูปที่ 3.1 และ 3.2 เราสามารถสังเกตได้ว่า ถ้าอัตราการเพิ่มขึ้นของ CFD มีค่าน้อยหรือไม่เพิ่มเลยติดต่อกันเป็นช่วงระยะหนึ่ง นั่นแสดงว่าบริเวณนั้นมีข้อมูลบางเบาหรือไม่มีข้อมูลอยู่เลย เราจึงนำคุณลักษณะนี้มาใช้ในการออกแบบขั้นตอนวิธีสำหรับวิธี RCFDC

นอกจากการพิจารณาอัตราการเปลี่ยนแปลงของ CFD แล้ว ยังสามารถพิจารณาอัตราการเปลี่ยนแปลงของฟังก์ชันสะสมได้เช่นกัน แต่ฟังก์ชันสะสมจะทำให้สเกลของกราฟมีค่าสูงมากซึ่งแสดงดังรูปที่ 3.3



รูปที่ 3.3 กราฟฟังก์ชันสะสมของความยาวกลีบดอกของฐานข้อมูล iris setosa

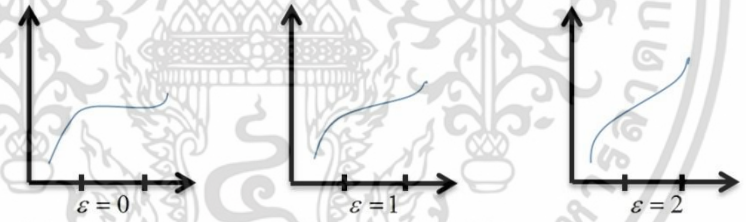
3.2 ขั้นตอนวิธีของ RCFDC

ขั้นตอนวิธีของวิธี RCFDC ได้ออกแบบโดยใช้หลักการพิจารณาอัตราการเปลี่ยนแปลงของ CFD ของข้อมูล ซึ่งจะทำการแบ่งช่วงของข้อมูลเพื่อใช้ในการคำนวณจำนวนกลุ่มโดยพิจารณาว่าอัตราการเปลี่ยนแปลงของ CFD มีค่าน้อยกว่าหรือเท่ากับค่าเปรียบเทียบอัตราการเปลี่ยนแปลง(ε) และมีความต่อเนื่องของช่วงที่อัตราการเปลี่ยนแปลงของ CFD มีค่าน้อยว่า ε (δ)

การหาจำนวนกลุ่มของข้อมูล n ตัวและมีคุณสมบัติแรกคือ x_i และคุณสมบัติที่สองคือ y_i โดยเลือกค่า ε และค่า δ ที่เหมาะสม (หลักการเลือกค่า ε และค่า δ ที่เหมาะสมสามารถดูได้ใน 3.4.2) แบ่งออกเป็น 3 ขั้นตอน ซึ่งมีรายละเอียดดังนี้

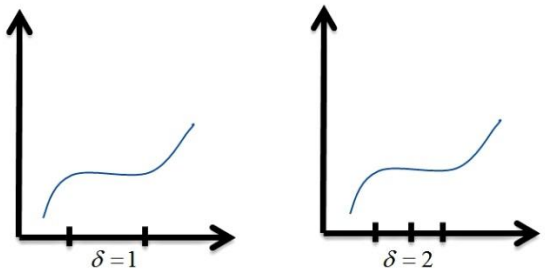
(1) คำนวณ CFD ของคุณลักษณะแรกของข้อมูล โดยเลือกค่าอันตรภาคชั้นที่เหมาะสม (หลักการเลือกค่าอันตรภาคชั้นที่เหมาะสมสามารถดูได้ใน 3.4.1)

(2) พิจารณาอัตราการเปลี่ยนแปลงของ CFD ที่ได้จาก (1) ถ้าอัตราการเปลี่ยนแปลงของ CFD มีค่าน้อยกว่าหรือเท่ากับ ε และช่วงที่อัตราการเปลี่ยนแปลงของ CFD มีค่าน้อยว่าหรือเท่ากับ ε มีความต่อเนื่องเท่ากับ δ เราจะได้ว่าจุด x_i นั้น เป็นจุดแบ่งข้อมูล เพื่อใช้ในการคำนวณค่า CFD สำหรับคุณสมบัติที่สองของข้อมูล โดยการพิจารณาค่า ε และ δ จะแสดงดังรูปที่ 3.4 และ 3.5 ตามลำดับ



รูปที่ 3.4 การพิจารณาค่า ε

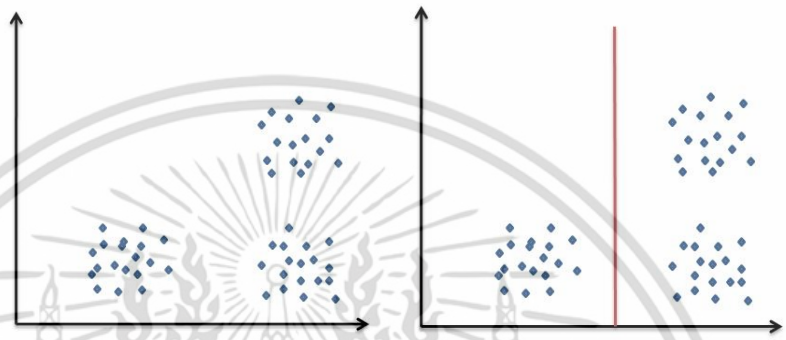
จากรูปที่ 3.4 การเลือก $\varepsilon = 2$ หมายความว่าผลต่างระหว่างช่วง x_i กับ x_{i+1} หรือ y_i กับ y_{i+1} มีค่าไม่เกิน 2 เช่น $x_i = 4$ และ $x_{i+1} = 6$ เป็นต้น



รูปที่ 3.5 การพิจารณาค่า δ ($\varepsilon = 0$)

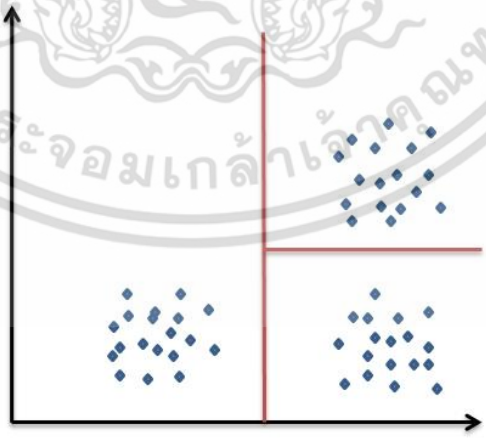
จากรูปที่ 3.5 การเลือก $\delta = 2$ มีช่วงที่ผลต่างระหว่างช่วง x_i กับ x_{i+1} หรือ y_i กับ y_{i+1} มีค่าไม่เกิน ε ต่อเนื่องกันอย่างน้อย 2 ช่วง เช่น $x_i = 6$, $x_{i+1} = 6$ และ $x_{i+2} = 6$ ($\varepsilon = 0$) เป็นต้น

โดยเมื่อเสร็จขั้นตอนนี้แล้วจะแบ่งช่วงของข้อมูลสมมติได้ดังรูปที่ 3.6



รูปที่ 3.6 การแบ่งช่วงข้อมูลหลังเสร็จขั้นตอนที่ 2

(3) คำนวณ CFD ของข้อมูลเป็นช่วงๆ ที่แบ่งโดยจุดแบ่งที่ได้จาก (2) และเปรียบเทียบอัตราการเปลี่ยนแปลงของ CFD ในแต่ละช่วง ถ้าอัตราการเปลี่ยนแปลงของ CFD มีค่าน้อยกว่าหรือเท่ากับ ε และช่วงที่อัตราการเปลี่ยนแปลงของ CFD มีค่าน้อยว่าหรือเท่ากับ ε มีความต่อเนื่องเท่ากับ δ เราจะได้ว่า ณ บริเวณนั้นจะสามารถแบ่งกลุ่มได้ โดยเมื่อเสร็จขั้นตอนนี้แล้วจะแบ่งช่วงของข้อมูลสมมติได้ดังรูปที่ 3.7



รูปที่ 3.7 การแบ่งช่วงข้อมูลหลังเสร็จขั้นตอนที่ 3

จากทั้งสามขั้นตอนข้างต้น เราสามารถเขียนเป็นขั้นตอนวิธีได้ดังขั้นตอนวิธีที่ 3.1

ขั้นตอนวิธี 3.1 วิธี RCFDC

```
1   Input ε, δ, n, x_i, y_i
2   c = 0, numbercluster = 0, j = 0
3   CFx_i ← compute CFD of first attribute
4   for each i = 1 to n do
5       ΔCFx = CFx_{i+1} − CFx_i
6       if ΔCFx ≤ ε
7           c = c + 1
8           if c = δ
9               j = j + 1, n_j = n_j + 1
10              s_j = x_i
11          else
12      else
13          c = 0
14  for each k = 1 to j do
15      CFy_i ← compute CFD of second attribute at point less than s_k
16      numbercluster ← numbercluster+1
17      for each i = 1 to n_k do
18          ΔCFy = CFy_{i+1} − CFy_i
19          if ΔCFy ≤ ε
20              c = c + 1
21              if c = δ
22                  numbercluster ← numbercluster+1
23              else
24          else
25              c = 0
```

จากขั้นตอนวิธีที่ 3.1 บรรทัดที่ 1 และสองจะเป็นการใส่ค่าและกำหนดค่าต่างๆที่จำเป็น บรรทัดที่ 3-13 เป็นการคำนวณค่า CFD และพิจารณาอัตราการเปลี่ยนแปลงของ CFD ของคุณสมบัติแรก เพื่อหาจุดแบ่ง บรรทัดที่ 14-25 จะเป็นการคำนวณค่า CFD และพิจารณาอัตราการเปลี่ยนแปลงของ CFD ของคุณสมบัติที่สองในแต่ละช่วง เพื่อหาจำนวนกลุ่มสำหรับวิธีการแบ่งกลุ่มแบบเคมีน

3.3 Big O Notation ของวิธี RCFDC

ในแง่ความซับซ้อนของวิธี RCFDC สามารถวิเคราะห์ในรูปของ Big O Notation ในแต่ละขั้นตอนของการคำนวณจำนวนกลุ่มของฐานข้อมูลที่มีข้อมูล n ตัว ได้ดังนี้

- การคำนวณ CFD ของคุณสมบัติแรกใช้เวลา $O(n)$ เนื่องจากข้อมูลแต่ละตัวถูกพิจารณา 1 ครั้ง ดังนั้นจึงมีการพิจารณาทั้งหมด n ครั้ง (บรรทัดที่ 3 ของขั้นตอนวิธี 3.1)
- การพิจารณาอัตราการเปลี่ยนแปลงของ CDF ของคุณสมบัติแรก เพื่อหาจุดแบ่งข้อมูลใช้เวลา $O(n)$ เนื่องจากจะมีการจับคู่ค่า CFD เพื่อหาอัตราการเปลี่ยนแปลง ทั้งหมด $n-1$ คู่ (บรรทัดที่ 4-13 ของขั้นตอนวิธี 3.1)
- การคำนวณ CFD ของแต่ละช่วงในคุณสมบัติที่สองใช้เวลารวม $O(n)$ เนื่องจากในแต่ละช่วงจะพิจารณาข้อมูล n_i ครั้ง แต่การพิจารณารวมของทุกช่วงมีค่า $n_1+n_2+...+n_m=n$ และการพิจารณาจำนวนกลุ่มของแต่ละช่วง CFD ในคุณสมบัติที่สองใช้เวลา $O(n)$ ในทำนองเดียวกันกับการพิจารณาอัตราการเปลี่ยนแปลงของ CDF ของคุณสมบัติแรก (บรรทัดที่ 14-25 ของขั้นตอนวิธี 3.1)

ดังนั้น การทำงานรวมในการคำนวณจำนวนกลุ่มของฐานข้อมูลที่มีข้อมูล n ตัว มีค่าเท่ากับ $4n-2$ และมี Big O Notation เท่ากับ $O(n)$

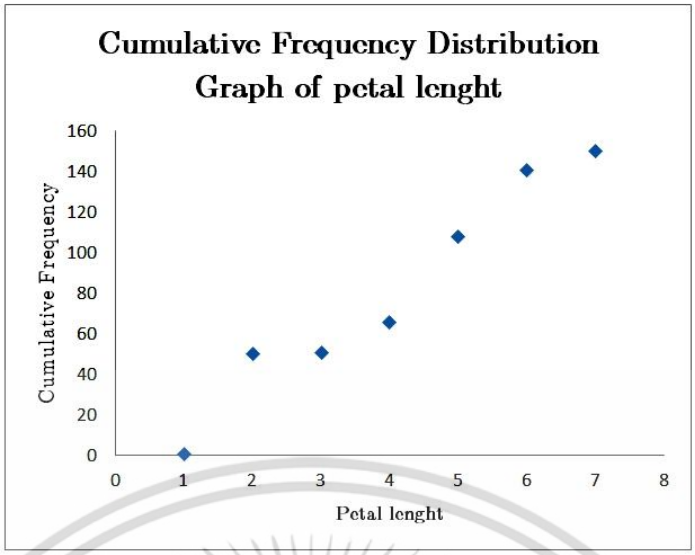
3.4 ปัจจัยที่มีผลต่อวิธี RCFDC

3.4.1 ความกว้างอันตรภาคชั้นของ CFD

การเลือกความกว้างของแต่ละอันตรภาคชั้นของ CFD นับเป็นปัจจัยที่มีความสำคัญเป็นอย่างมาก เนื่องจากถ้าเราเลือกความกว้างอันตรภาคชั้นที่กว้างเกินไป อาจจะทำให้เราไม่สามารถพิจารณาอัตราการเปลี่ยนแปลง CFD ของข้อมูลได้ เพราะว่าอันตรภาคชั้นที่กว้างจะทำให้ข้ามการนับความถี่ในช่วงที่ข้อมูลเบาบางไป ทำให้เราเห็นว่าอัตราการเปลี่ยนแปลงเพิ่มขึ้นอยู่ตลอดไม่มีบริเวณที่ไม่มีการเพิ่มหรือเพิ่มขึ้นเพียงเล็กน้อยเลย ซึ่งแสดงตัวอย่างของกราฟที่ได้การใช้ CFD บนความยาวกลีบดอกของฐานข้อมูล iris setosa ในค่าความกว้างของอันตรภาคชั้นที่ต่างกัน โดยเมื่อเลือกความกว้างของอันตรภาคชั้นเป็น 1 จะได้ตาราง CFD ดังตารางที่ 3.1 และจะได้กราฟของ CFD ดังรูปที่ 3.8 ในทำนองเดียวกันเมื่อเลือกความกว้างของอันตรภาคชั้นเป็น 0.5 และ 0.1 จะได้กราฟของ CFD ดังรูปที่ 3.9 และ 3.10 ตามลำดับ

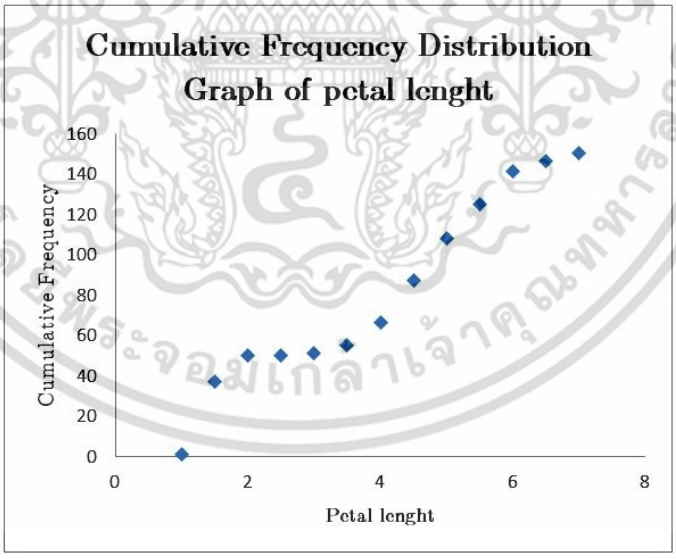
ตารางที่ 3.1 การแจกแจงความถี่สะสมของความยาวกลีบดอกในอันตรภาคชั้นเท่ากับ 1

Petal length	Frequency	CFD
0-1	1	1
1-2	40	50
2-3	1	51
3-4	15	66
4-5	42	108
5-6	33	141
6-7	19	150



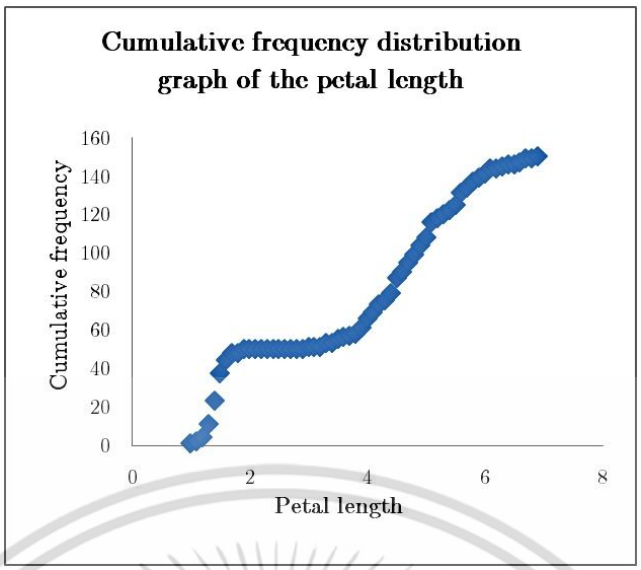
รูปที่ 3.8 กราฟ CFD บนความยาวกลีบดอกของฐานข้อมูล iris setosa ค่าอันตรภาคชั้น 1

จากรูปที่ 3.8 จะเห็นว่าความต่อเนื่องของช่วงที่มีอัตราการเปลี่ยนแปลงของ CFD คงที่ น้อยเกินไปทำให้ไม่สามารถแบ่งช่วงข้อมูลได้



รูปที่ 3.9 กราฟ CFD บนความยาวกลีบดอกของฐานข้อมูล iris setosa ค่าอันตรภาคชั้น 0.5

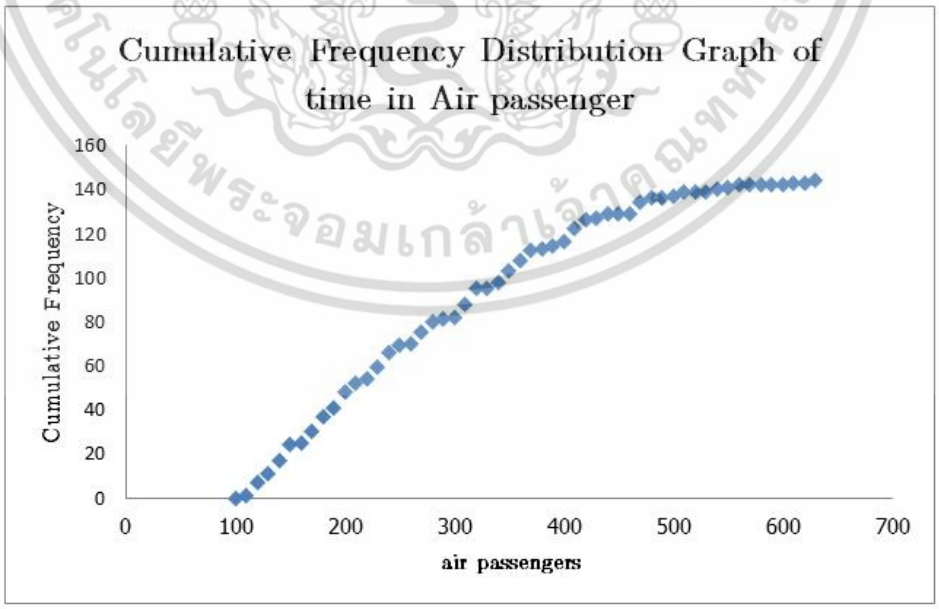
จากรูปที่ 3.9 จะเห็นว่าความต่อเนื่องของช่วงที่มีอัตราการเปลี่ยนแปลงของ CFD คงที่ มากขึ้นกว่ารูปที่ 3.8 ทำให้สามารถแบ่งช่วงข้อมูลได้ดีขึ้น



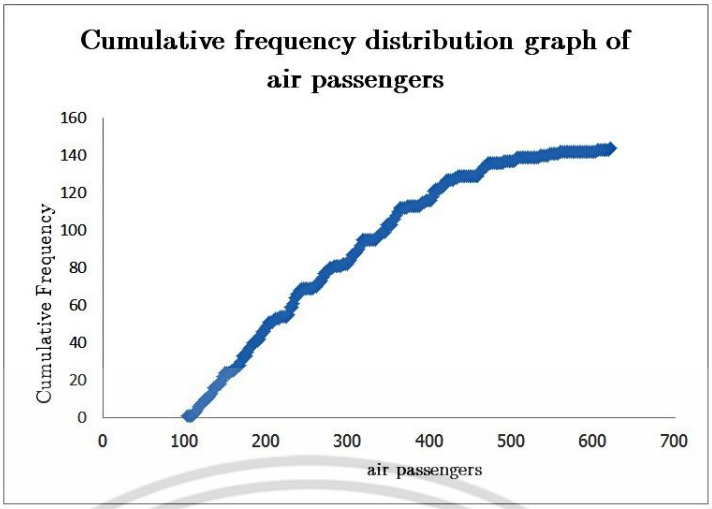
รูปที่ 3.10 กราฟ CFD บนความยาวกลีบดอกของฐานข้อมูล iris setosa ค่าอันตรภาคชั้น 0.1

จากรูปที่ 3.10 จะเห็นว่าความต่อเนื่องของช่วงที่มีอัตราการเปลี่ยนแปลงของ CFD คงที่ มากขึ้นกว่ารูปที่ 3.9 ทำให้สามารถแบ่งช่วงข้อมูลได้ดีขึ้น

ในทำนองเดียวกันตัวอย่างของกราฟที่ได้การใช้ CFD บนเวลาที่เดินทางของฐานข้อมูล air passengers ในค่าอันตรภาคชั้นที่ต่างกัน ดังรูปที่ 3.11 และ 3.12



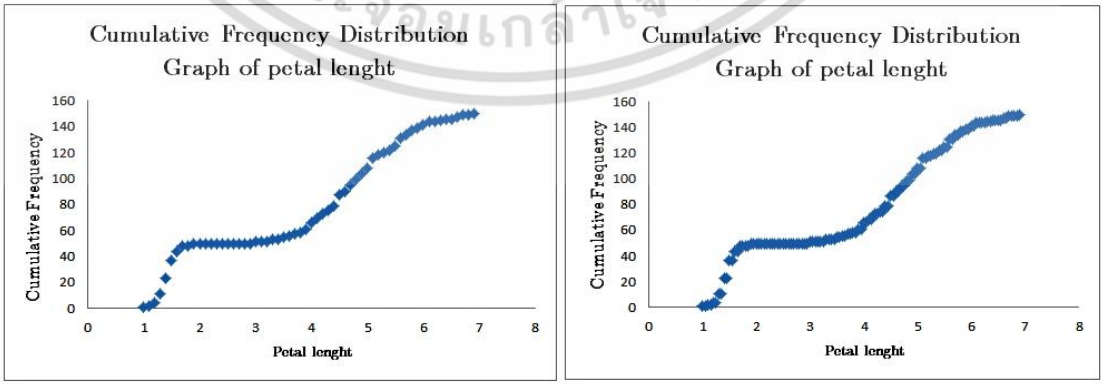
รูปที่ 3.11 กราฟ CFD บนจำนวนผู้โดยสารของฐานข้อมูล air passenger ค่าอันตรภาคชั้น 10



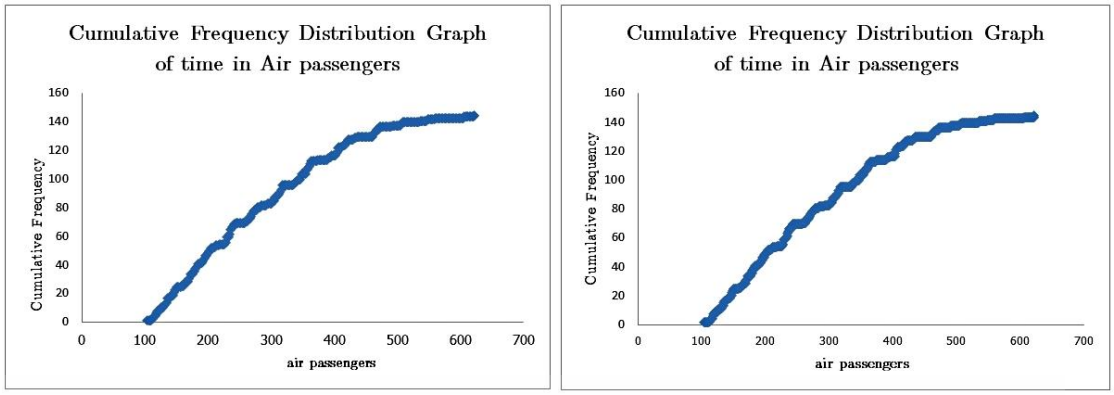
รูปที่ 3.12 กราฟ CFD บนจำนวนผู้โดยสารของฐานข้อมูล air passenger ค่าอันตรภาคชั้น 2

ในทางกลับกันถ้าเลือกใช้อันตรภาคชั้นที่น้อยมากๆ ถึงแม้ว่าจะทำให้ค่าที่คำนวณได้มีความแม่นยำ แต่ก็จะทำให้ใช้เวลาในการทำ CFD มากขึ้นเนื่องจากต้องพิจารณาหลายอันตรภาคชั้น ซึ่งมากเกินความจำเป็นเนื่องจากค่าที่ได้จากการเลือกค่าอันตรภาคชั้นที่พอเหมาะก็มีความแม่นยำเช่นกัน และอาจทำให้เกิดการแบ่งกลุ่มที่มากกว่าความเป็นจริงได้

ซึ่งโดยปกติแล้วเราจะเลือกอันตรภาคชั้นที่น้อยที่สุดไม่เกินอัตราการเพิ่มของค่าของข้อมูลที่น้อยที่สุด เช่นจากตัวอย่างของความยาวกลีบดอกของฐานข้อมูล iris setosa เราจะเลือกค่าอันตรภาคชั้นเท่ากับ 0.1 โดยตัวอย่างของการเลือกค่าอันตรภาคชั้นที่น้อยเกินความจำเป็นจะแสดงดังรูปที่ 3.13 และ 3.14



รูปที่ 3.13 กราฟ CFD บนความยาวกลีบดอกของฐานข้อมูล iris setosa ค่าอันตรภาคชั้น 0.1 และ 0.05



รูปที่ 3.14 กราฟ CFD บนเวลาที่เดินทางของฐานข้อมูล air passengers ค่าอันตรภาคชั้นเท่ากับ 2 และ 1

จากรูปที่ 3.13 และ 3.14 จะเห็นได้ว่าอัตราการเปลี่ยนแปลงของกราฟ CFD ของค่าอันตรภาคชั้นทั้งสองแทบไม่มีความแตกต่างกัน ดังนั้นค่าจำนวนกลุ่มที่คำนวณได้จึงมีค่าไม่แตกต่างกัน

3.4.2 การเลือกค่า ε และ δ

การเลือกค่า ε และ δ นับเป็นสิ่งสำคัญในการหาจำนวนกลุ่มของฐานข้อมูล โดยใช้วิธี RCFDC เนื่องจากการใช้ค่า ε และ δ ที่แตกต่างกัน จะทำให้ได้ผลลัพธ์ที่แตกต่างกัน ดังนั้นเราจึงจำเป็นต้องเลือกค่า ε และ δ ให้เหมาะสม โดยขึ้นอยู่กับความต้องการของผู้ใช้ว่าต้องการความละเอียดมากน้อยเพียงใด และความหนาแน่นของฐานข้อมูล

โดยการเลือกค่า ε น้อยและ δ มากจะเหมาะกับฐานข้อมูลที่มีลักษณะเบาบางและมีช่วงที่ไม่มีข้อมูลหรือมีช่วงที่มีจำนวนข้อมูลน้อย การเลือกค่า ε น้อยและ δ น้อยจะเหมาะกับฐานข้อมูลที่มีลักษณะเบาบาง แต่ไม่มีช่วงไม่มีข้อมูลหรือมีช่วงที่มีจำนวนข้อมูลน้อย การเลือกค่า ε มากและ δ มากจะเหมาะกับฐานข้อมูลที่มีลักษณะหนาแน่นและมีช่วงทีจำนวนข้อมูลเบาบางกว่าช่วงอื่นๆ การเลือกค่า ε มากและ δ มากจะเหมาะกับฐานข้อมูลที่มีลักษณะหนาแน่นและไม่มีช่วงทีจำนวนข้อมูลเบาบางกว่าช่วงอื่นๆ

ซึ่งตัวอย่างของการทดลองใช้ค่า ε และ δ ที่แตกต่างกันเพื่อเปรียบเทียบผลที่ได้จากวิธี RCFDC บนฐานข้อมูล iris setosa แสดงดังตารางที่ 3.1

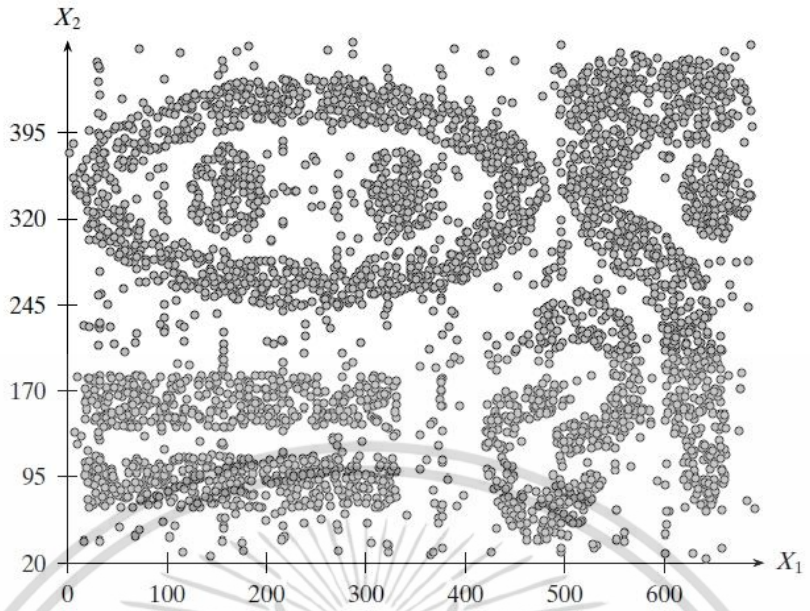
ตารางที่ 3.1 การเปรียบเทียบผลลัพธ์ของวิธี RCFDC ที่ใช้ค่า ε และ δ ที่แตกต่างกันบนฐานข้อมูล iris setosa

ε value	δ value	Number of clusters	Sum of square error
0	1	3	0.169
	2	2	0.457
	3	2	0.457
1	1	5	0.066
	2	3	0.169
	3	2	0.457
2	1	8	0.037
	2	5	0.066
	3	3	0.169

จากตารางที่ 3.1 จะเห็นว่าเมื่อเลือกค่า ε มากขึ้นจะต้องเลือกค่า δ ที่มากขึ้นด้วย เพื่อให้ได้ผลลัพธ์เป็น 3 ตามชนิดของดอก iris และถ้าเลือกค่า δ น้อยเกินไปจะส่งผลให้เกิดการแบ่งกลุ่มที่มากเกินความจำเป็น และถ้าเลือกค่า δ มาก จะทำให้เกิดการแบ่งกลุ่มที่น้อยหรือไม่เกิดการแบ่งกลุ่มเลย

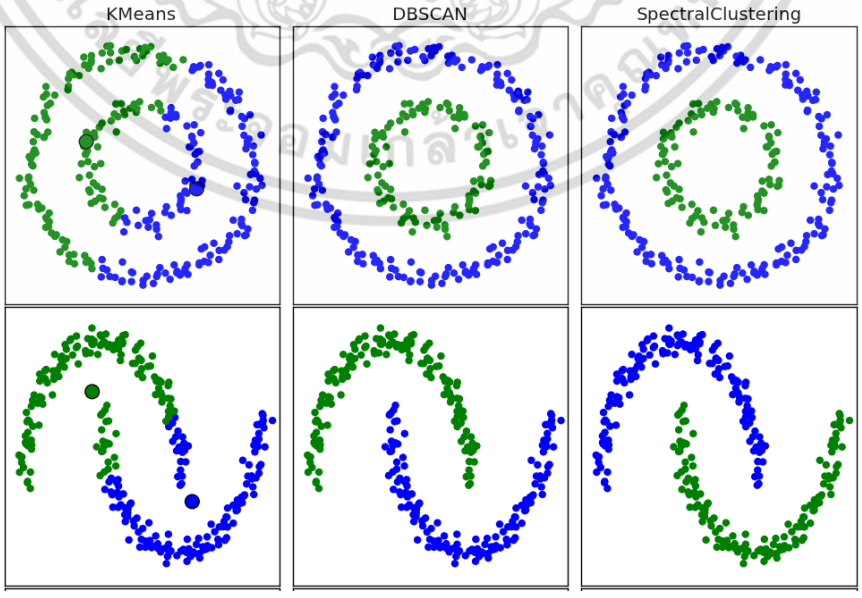
3.4.3 ลักษณะของฐานข้อมูล

ลักษณะของฐานข้อมูลนับเป็นอีกหนึ่งปัจจัยที่มีผลต่อผลลัพธ์ของวิธี RCFDC โดยลักษณะของฐานข้อมูลที่เหมาะสมกับวิธี RCFDC จะขึ้นอยู่กับลักษณะข้อมูลที่เหมาะสมกับวิธีการแบ่งกลุ่มแบบเคมีน คือ ฐานข้อมูลแบบเซตนูน (convex set) เช่นฐานข้อมูล iris setosa และ air passengers ซึ่งตัวอย่างแสดงดังรูปที่ 4.1 และ 4.6 ตามลำดับ ส่วนลักษณะของฐานข้อมูลที่ไม่เหมาะสมกับวิธี RCFDC และการแบ่งกลุ่มแบบเคมีน คือ ฐานข้อมูลแบบไม่ใช่เซตนูน (non-convex set) เช่น ฐานข้อมูล tooth growth ดังรูปที่ 4.10 และลักษณะฐานข้อมูลที่ไม่ใช่เซตนูนแบบอื่นๆ แสดงดังรูปที่ 3.15 [1]



รูปที่ 3.15 ฐานข้อมูลที่ไม่ใช่เซตนูน

โดยฐานข้อมูลแบบไม่ใช่เซตนูนเมื่อใช้การแบ่งกลุ่มแบบเคมีนในการแบ่งกลุ่มแล้วจะทำได้ผลลัพธ์ที่ได้จากการแบ่งกลุ่มไม่ตรงกับความเป็นจริง ซึ่งตัวอย่างการใช้การแบ่งกลุ่มแบบเคมีนฐานข้อมูลแบบ non-convex เปรียบเทียบกับการแบ่งกลุ่มแบบอื่นๆ แสดงดังรูปที่ 3.16 (ที่มา http://ogrisel.github.io/scikit-learn.org/sklearn-tutorial/modules/clustering.html)



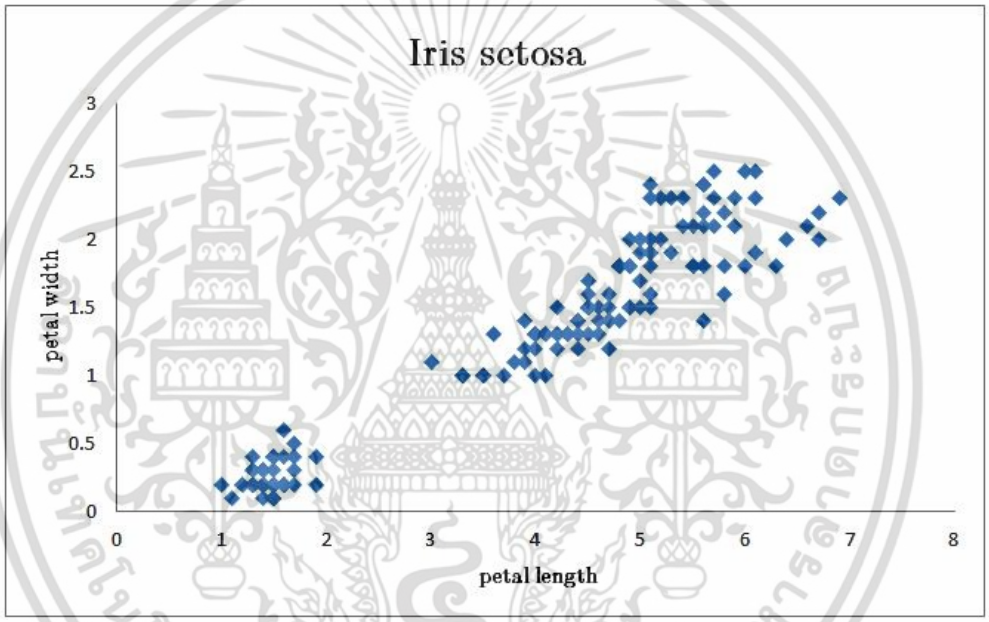
รูปที่ 3.16 การแบ่งกลุ่มแบบเคมีนฐานข้อมูลที่ไม่ใช่เซตนูนเปรียบเทียบกับการแบบอื่นๆ

บทที่ 4

ผลการทดลองของวิธี RCFDC

4.1 ผลการทดลองของวิธี RCFDC กับฐานข้อมูล Iris setosa

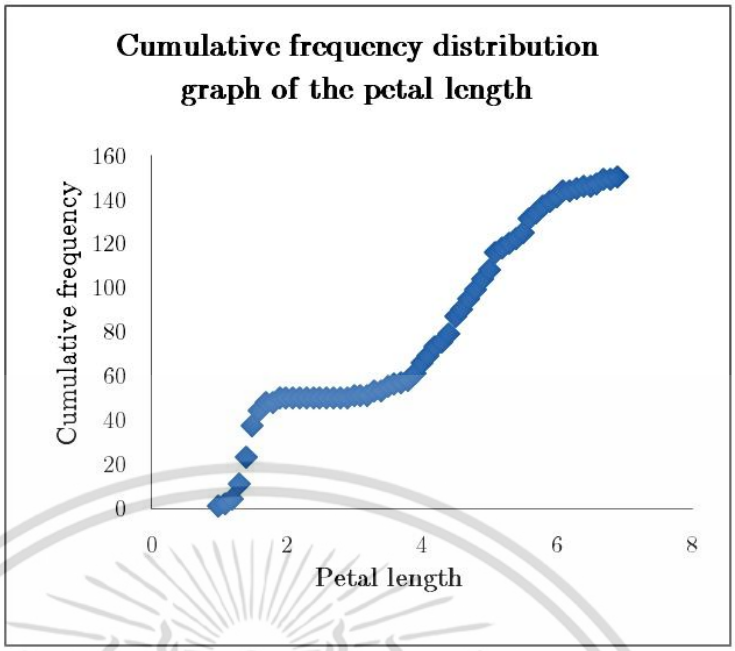
จากการใช้วิธี RCFDC กับฐานข้อมูล iris setosa โดยเลือก ความยาวกลีบดอกกับความกว้างกลีบดอก เป็น 2 คุณลักษณะเพื่อใช้ในการหาจำนวนกลุ่ม ซึ่งมีลักษณะของฐานข้อมูลดังรูปที่ 4.1



รูปที่ 4.1 ฐานข้อมูล Iris setosa

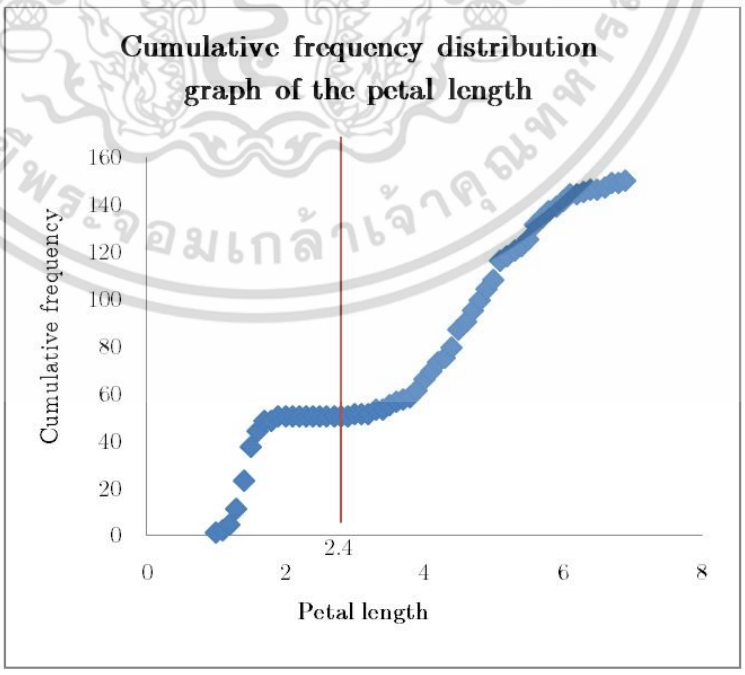
ซึ่งมีลำดับการดำเนินการดังนี้ โดยเลือก $\varepsilon = 1$ และ $\delta = 2$

1) คำนวณ CFD บนความยาวกลีบดอกของฐานข้อมูล iris setosa โดยเลือกอันตรภาคชั้นเป็น 0.1 จะได้กราฟ CFD ดังรูปภาพที่ 4.2



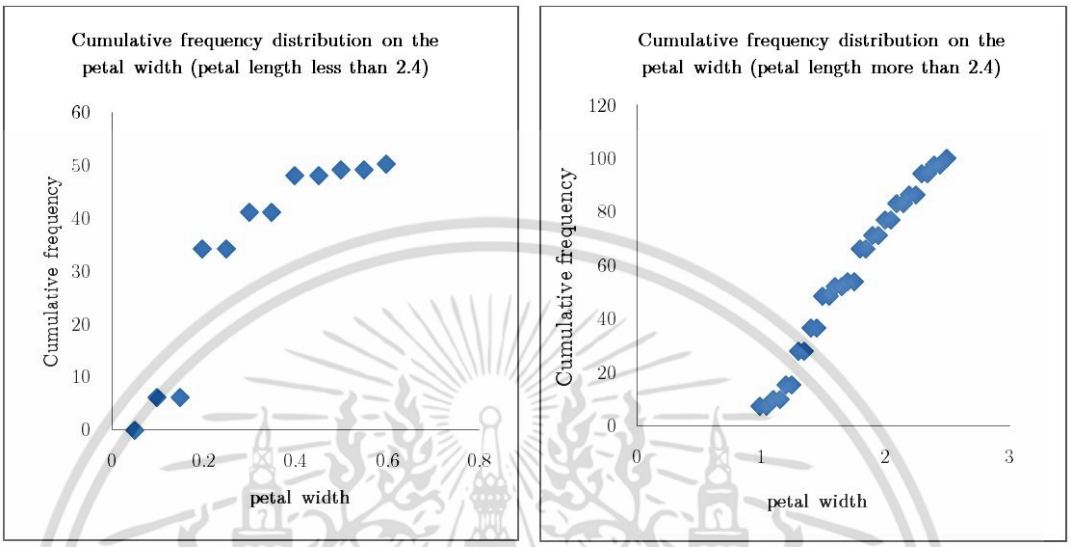
รูปที่ 4.2 กราฟ CFD บนความยาวกลีบดอกของฐานข้อมูล iris setosa ค่าอันตรภาคชั้น 0.1

2) พิจารณาอัตราการเปลี่ยนแปลงของ CFD ของความยาวกลีบดอกเพื่อหาจุดแบ่งข้อมูล จะได้จุดแบ่งคือจุด 2.4 แสดงดังรูปที่ 4.3

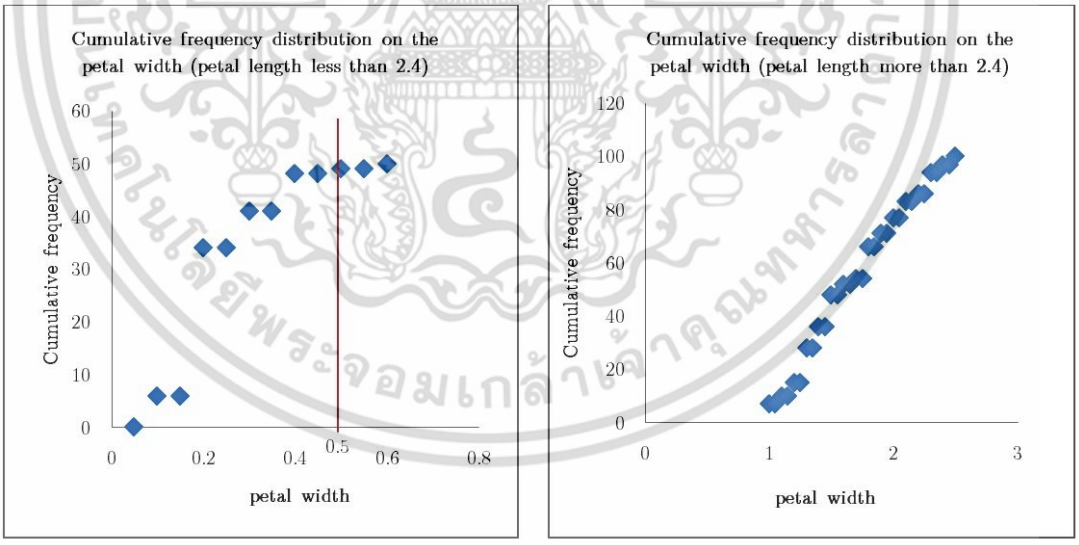


รูปที่ 4.3 จุดแบ่งข้อมูลของความยาวกลีบดอก

3) คำนวณ CFD ของความกว้างกลีบดอกในแต่ละช่วงที่แบ่งโดยจุดแบ่งที่ได้จากขั้นตอนก่อนหน้า โดยเลือกความกว้างของอันตรภาคชั้นเป็น 0.05 จะได้กราฟของ CFD ดังรูปที่ 4.4 และผลของการพิจารณาค่าอัตราการเปลี่ยนแปลงของ CFD ดังรูปที่ 4.5



รูปที่ 4.4 กราฟ CFD ของแต่ละช่วงของความกว้างกลีบดอก

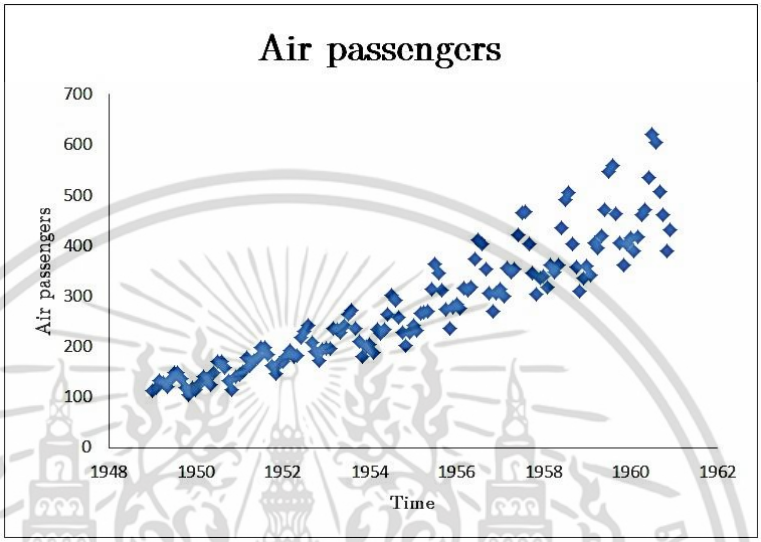


รูปที่ 4.5 ผลการแบ่งกลุ่มในแต่ละช่วงของความกว้างกลีบดอก

จากรูปที่ 4.5 จะได้ว่า ช่วงแรกของฐานข้อมูลสามารถแบ่งกลุ่มได้สองกลุ่ม ส่วนช่วงที่สองไม่สามารถแบ่งกลุ่มได้ ดังนั้นเราจะได้จำนวนกลุ่มของฐานข้อมูล iris setosa คือ 3

4.2 ผลการทดลองของวิธี RCFDC กับฐานข้อมูล Air passengers

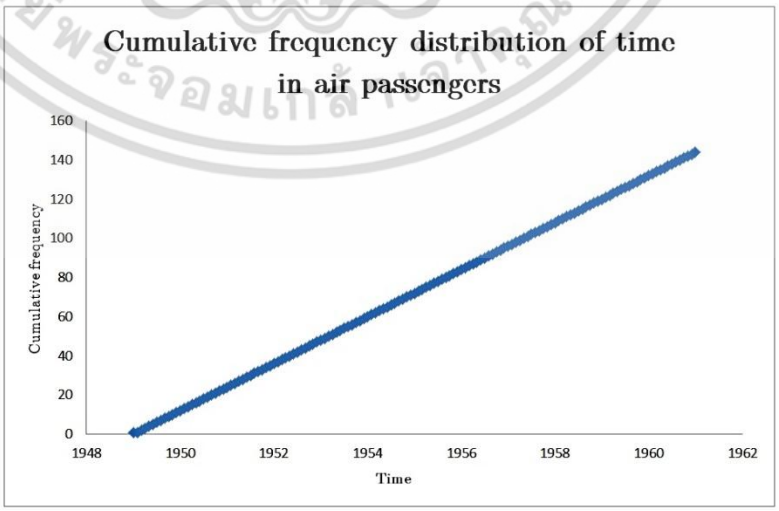
จากการใช้วิธี RCFDC กับฐานข้อมูล air passengers โดยเลือกจำนวนผู้โดยสารกับเวลาที่เดินทางเป็น 2 คุณลักษณะเพื่อใช้ในการหาจำนวนกลุ่ม ซึ่งมีลักษณะของฐานข้อมูลดังรูปที่ 4.6



รูปที่ 4.6 ฐานข้อมูล Air passengers

ซึ่งมีลำดับการดำเนินการดังนี้ โดยเลือก $\varepsilon = 0$ และ $\delta = 4$

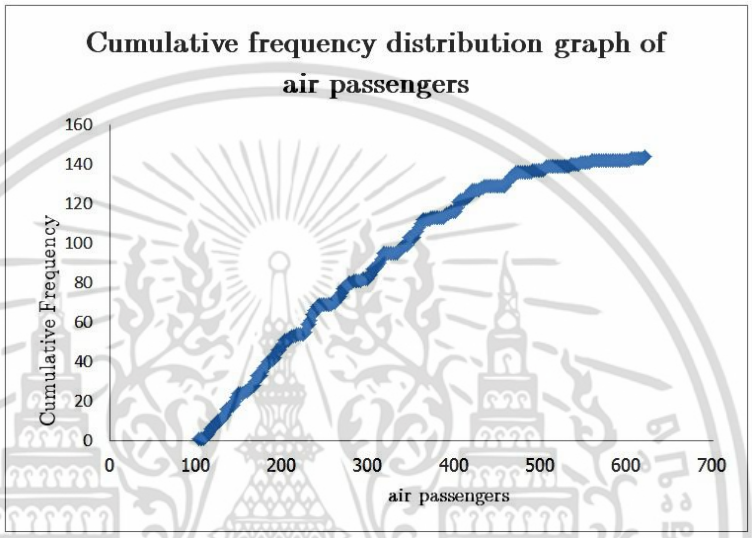
1) คำนวณ CFD บน time ของฐานข้อมูล air passenger โดยเลือกอันตรภาคชั้นเป็น 0.083 จะได้กราฟ CFD ดังรูปภาพที่ 4.7



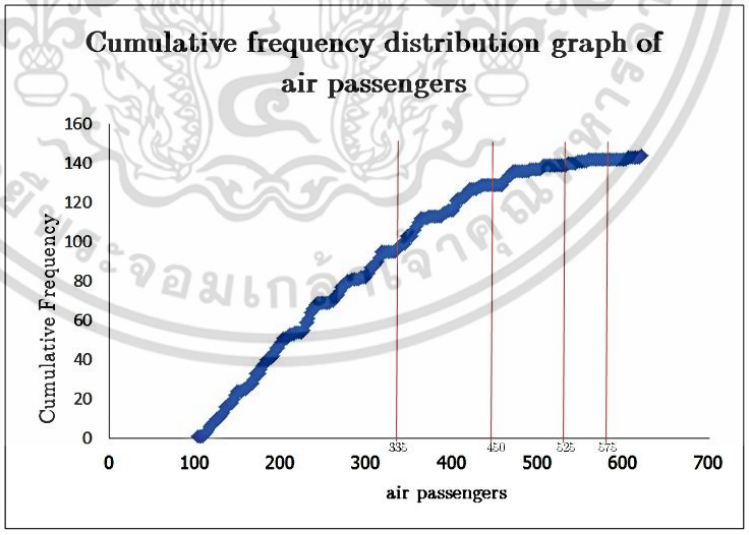
รูปที่ 4.7 กราฟ CFD บน time ของฐานข้อมูล air passenger ค่าอันตรภาคชั้น 0.083

2) พิจารณาอัตราการเปลี่ยนแปลงของ CFD ของ time เพื่อหาจุดแบ่งข้อมูล จะได้ว่าไม่สามารถหาจุดแบ่งได้

3) คำนวณ CFD ของ air passengers ในแต่ละช่วงที่แบ่งโดยจุดแบ่งที่ได้จากขั้นตอนก่อนหน้า โดยเลือกความกว้างของอันตรภาคชั้นเป็น 2 จะได้กราฟของ air passengers และผลของการพิจารณาค่าอัตราการเปลี่ยนแปลงของ CFD ดังรูปที่ 4.8 และ 4.9 ตามลำดับ



รูปที่ 4.8 กราฟ CFD ของแต่ละช่วงของ air passengers ในฐานข้อมูล Air passengers

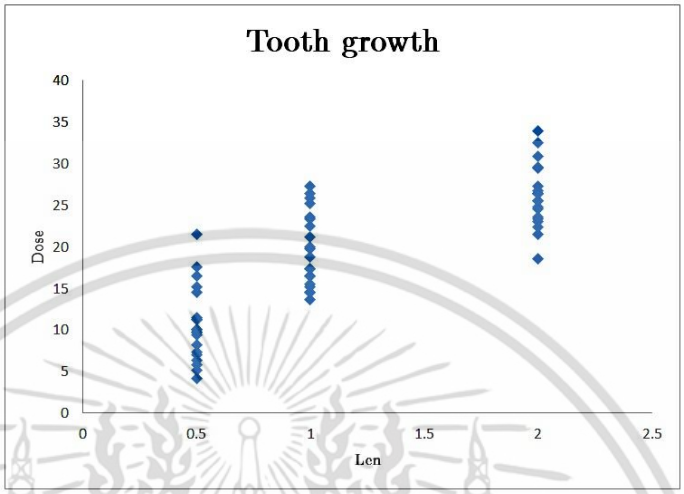


รูปที่ 4.9 ผลการแบ่งกลุ่มในแต่ละช่วงของ air passengers ในฐานข้อมูล Air passengers

จากรูปที่ 4.9 จะได้ว่ามีจุดแบ่งทั้งหมด 4 จุด ดังนั้นจะได้ว่าจำนวนกลุ่มของฐานข้อมูล air passenger คือ 5

4.3 ผลการทดลองของวิธี RCFDC กับฐานข้อมูล Tooth Growth

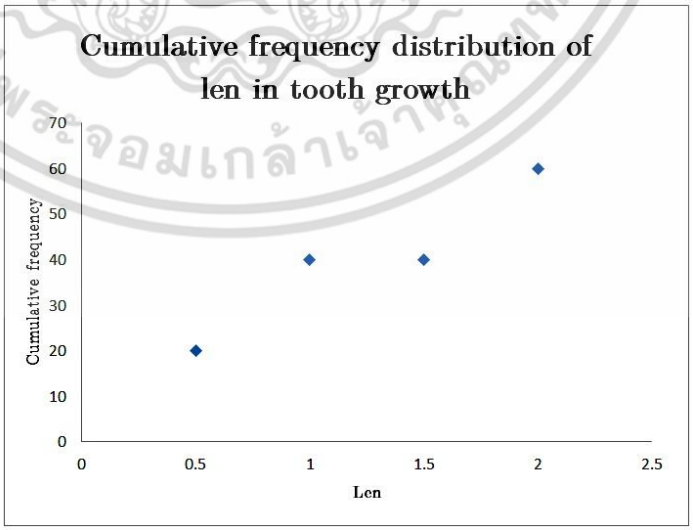
จากการใช้วิธี RCFDC กับฐานข้อมูล tooth growth โดยเลือกความยาวของฟันกับโดส เป็น 2 คุณลักษณะเพื่อใช้ในการหาจำนวนกลุ่ม ซึ่งมีลักษณะของฐานข้อมูลดังรูปที่ 4.10



รูปที่ 4.10 ฐานข้อมูล tooth growth

ซึ่งมีลำดับการดำเนินการดังนี้ โดยเลือก $\varepsilon = 0$ และ $\delta = 7$ และ

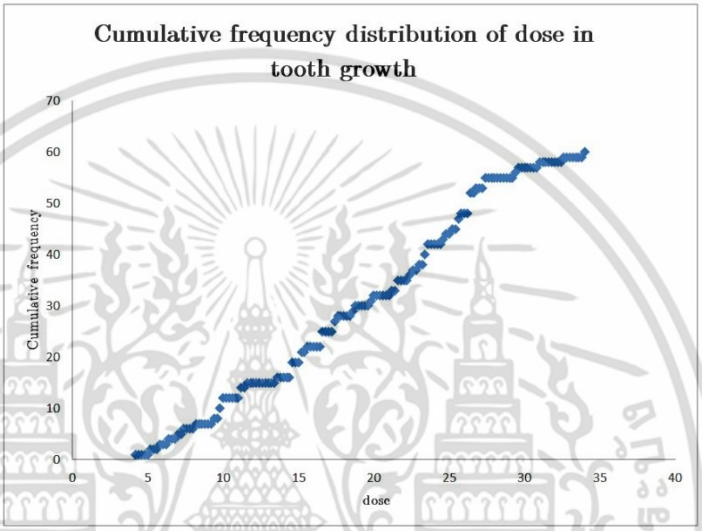
1) คำนวณ CFD บน len ของฐานข้อมูล tooth growth โดยเลือกอันตรภาคชั้นเป็น 0.5 จะได้กราฟ CFD ดังรูปภาพที่ 4.11



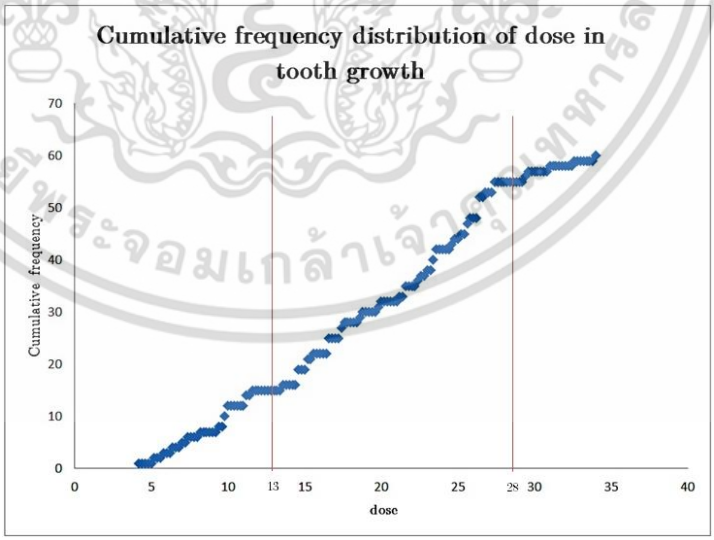
รูปที่ 4.11 กราฟ CFD บนความยาวของฐานข้อมูล tooth growth ค่าอันตรภาคชั้น 0.5

2) พิจารณาอัตราการเปลี่ยนแปลงของ CFD ของความยาวเพื่อหาจุดแบ่งข้อมูล จะได้ว่าไม่สามารถหาจุดแบ่งได้

3) คำนวณ CFD ของโดสในแต่ละช่วงที่แบ่งโดยจุดแบ่งที่ได้จากขั้นตอนก่อนหน้า โดยเลือกความกว้างของอันตรภาคชั้นเป็น 0.2 จะได้กราฟของโดสและผลของการพิจารณาค่าอัตราการเปลี่ยนแปลงของ CFD ดังรูปที่ 4.12 และ 4.13 ตามลำดับ



รูปที่ 4.12 กราฟ CFD ของแต่ละช่วงของโดส



รูปที่ 4.13 ผลการแบ่งกลุ่มในแต่ละช่วงของโดส

จากรูปที่ 4.13 จะได้ว่าจำนวนกลุ่มของฐานข้อมูล tooth growth คือ 3

บทที่ 5

สรุปผลของวิธี RCFDC และข้อเสนอแนะ

5.1 สรุปผลของวิธี RCFDC

จากการใช้วิธี RCFDC ในการหาจำนวนกลุ่มของฐานข้อมูลบน R^2 พบว่าประสิทธิภาพของผลลัพธ์จะขึ้นอยู่กับการเลือกค่า ε และ δ ที่เหมาะสมสำหรับแต่ละฐานข้อมูล ซึ่งขึ้นอยู่กับความหนาแน่นและการกระจายตัวของข้อมูล ถ้าฐานข้อมูลมีความหนาแน่นและการกระจายตัวของข้อมูลที่มาก เราควรเลือก ε ที่มีค่ามากและ δ ที่มีค่าน้อย ในทางกลับกันถ้าฐานข้อมูลมีข้อมูลที่เบาบางและกระจายตัวน้อย เราควรเลือก ε ที่มีค่าน้อยและ δ ที่มีค่ามาก

ในการหาผลลัพธ์ของวิธี RCFDC เหมาะกับฐานข้อมูลแบบเซตนูน ซึ่งตัวอย่างแสดงในบทที่ 4 และไม่เหมาะกับฐานข้อมูลที่ไม่ใช่เซตนูน ซึ่งตัวอย่างของ ฐานข้อมูลที่ไม่ใช่แสดงในบทที่ 4 เช่นกัน

ผลลัพธ์ที่ได้จากวิธี RCFDC ของฐานข้อมูล iris setosa และ air passengers เมื่อเทียบกับวิธี Elbow จะได้ผลดังตารางที่ 5.1

ตารางที่ 5.1 การเปรียบเทียบผลลัพธ์ที่ได้จากวิธี RCFDC กับวิธี Elbow

Data set	Elbow method	RCFDC	
	k value	k value	Remark
Iris setosa	3	3	$\varepsilon = 1, \delta = 2$
Air passenger	5	5	$\varepsilon = 0, \delta = 4$

ในแง่ของ Big O Notation สามารถสรุปได้ว่า การคำนวณ CFD ของคุณสมบัติแรกใช้เวลา $O(n)$, การพิจารณาอัตราการเปลี่ยนแปลงของ CDF ของคุณสมบัติแรก เพื่อหาจุดแบ่งข้อมูลใช้เวลา $O(n)$, การคำนวณ CFD ของแต่ละช่วงในคุณสมบัติที่สองใช้เวลารวม $O(n)$ และการพิจารณาจำนวนกลุ่มของแต่ละช่วง CFD ในคุณสมบัติที่สองใช้เวลา $O(n)$ ดังนั้น Big O Notation รวมในการคำนวณจำนวนกลุ่มของฐานข้อมูลที่มีข้อมูล n ตัว มีค่า $O(n)$

5.2 ข้อเสนอแนะ

1) วิธี RCFDC นอกจากจะสามารถหาผลลัพธ์ของจำนวนกลุ่มของฐานข้อมูลบน R^2 ได้แล้ว ในทำนองเดียวกันสามารถนำไปพัฒนาเพื่อหาผลลัพธ์ของจำนวนกลุ่มของฐานข้อมูลบน R^3 ได้

2) หาวิธีในการคำนวณหาค่า ε และ δ ที่เหมาะสม ซึ่งจะทำให้สามารถใช้วิธี RCFDC ได้อย่างมีประสิทธิภาพมากยิ่งขึ้น

3) เปรียบเทียบการพิจารณาอัตราการเปลี่ยนแปลงของ CFD กับฟังก์ชันความหนาแน่นหรือการแจงแจงแบบอื่นๆ เพื่อเปรียบเทียบประสิทธิภาพในการแบ่งกลุ่ม

4) สำหรับปัญหาฐานข้อมูลที่มีจุดแบ่งในแนวทแยงดังรูปที่ 5.1 ซึ่งการพิจารณาอัตราการเปลี่ยนแปลงของ CFD บนแกน x, y ไม่สามารถแบ่งกลุ่มได้ จึงควรหาฟังก์ชันการฉายภาพ (projection) หรือพัฒนาขั้นตอนวิถีเพื่อใช้ในการแก้ปัญหานี้

รูปที่ 5.1 ฐานข้อมูลที่มีจุดแบ่งในแนวทแยง

เอกสารอ้างอิง

[1] M. J. Zaki and W. Meira JR. 2014. **Data mining and analysis.** New York. Cambridge University

[2] J. Isotalo. 2014. **Basics of statistics**. Create Space Independent. Publishing Platform.

[3] T. M. Kodinariya and P. R. Makwana. 2013. Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies.* Issue 6(1): 90-95.

[4] Vincent Arel-Bundock. 2017. **Edgar Anderson's Iris Data**. [Online]. Available : https://vincentarelbundock.github.io/Rdatasets/datasets.html

[5] Vincent Arel-Bundock. 2017. **Monthly Airline Passenger Numbers 1949-1960.** [Online]. Available: https://vincentarelbundock.github.io/Rdatasets/datasets.html

[6] Vincent Arel-Bundock. 2017. **The Effect of Vitamin C on Tooth Growth in Guinea Pigs.** [Online]. Available: https://vincentarelbundock.github.io/Rdatasets/datasets.html

ภาคผนวก

ภาคผนวก ก

ฐานข้อมูลที่ใช้ในงานวิจัย

1. ฐานข้อมูล Iris setosa

sepal length	sepal width	petal length	petal width	iris
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5	3.6	1.4	0.2	Iris-setosa
5.4	3.9	1.7	0.4	Iris-setosa
4.6	3.4	1.4	0.3	Iris-setosa
5	3.4	1.5	0.2	Iris-setosa
4.4	2.9	1.4	0.2	Iris-setosa
4.9	3.1	1.5	0.1	Iris-setosa
5.4	3.7	1.5	0.2	Iris-setosa
4.8	3.4	1.6	0.2	Iris-setosa
4.8	3	1.4	0.1	Iris-setosa
4.3	3	1.1	0.1	Iris-setosa
5.8	4	1.2	0.2	Iris-setosa
5.7	4.4	1.5	0.4	Iris-setosa
5.4	3.9	1.3	0.4	Iris-setosa
5.1	3.5	1.4	0.3	Iris-setosa
5.7	3.8	1.7	0.3	Iris-setosa
5.1	3.8	1.5	0.3	Iris-setosa
5.4	3.4	1.7	0.2	Iris-setosa
5.1	3.7	1.5	0.4	Iris-setosa
4.6	3.6	1	0.2	Iris-setosa
5.1	3.3	1.7	0.5	Iris-setosa
4.8	3.4	1.9	0.2	Iris-setosa
5	3	1.6	0.2	Iris-setosa
5	3.4	1.6	0.4	Iris-setosa
5.2	3.5	1.5	0.2	Iris-setosa
5.2	3.4	1.4	0.2	Iris-setosa
4.7	3.2	1.6	0.2	Iris-setosa

4.8	3.1	1.6	0.2	Iris-setosa
5.4	3.4	1.5	0.4	Iris-setosa
5.2	4.1	1.5	0.1	Iris-setosa
5.5	4.2	1.4	0.2	Iris-setosa
4.9	3.1	1.5	0.1	Iris-setosa
5	3.2	1.2	0.2	Iris-setosa
5.5	3.5	1.3	0.2	Iris-setosa
4.9	3.1	1.5	0.1	Iris-setosa
4.4	3	1.3	0.2	Iris-setosa
5.1	3.4	1.5	0.2	Iris-setosa
5	3.5	1.3	0.3	Iris-setosa
4.5	2.3	1.3	0.3	Iris-setosa
4.4	3.2	1.3	0.2	Iris-setosa
5	3.5	1.6	0.6	Iris-setosa
5.1	3.8	1.9	0.4	Iris-setosa
4.8	3	1.4	0.3	Iris-setosa
5.1	3.8	1.6	0.2	Iris-setosa
4.6	3.2	1.4	0.2	Iris-setosa
5.3	3.7	1.5	0.2	Iris-setosa
5	3.3	1.4	0.2	Iris-setosa
7	3.2	4.7	1.4	Iris-versicolor
6.4	3.2	4.5	1.5	Iris-versicolor
6.9	3.1	4.9	1.5	Iris-versicolor
5.5	2.3	4	1.3	Iris-versicolor
6.5	2.8	4.6	1.5	Iris-versicolor
5.7	2.8	4.5	1.3	Iris-versicolor
6.3	3.3	4.7	1.6	Iris-versicolor
4.9	2.4	3.3	1	Iris-versicolor
6.6	2.9	4.6	1.3	Iris-versicolor
5.2	2.7	3.9	1.4	Iris-versicolor
5	2	3.5	1	Iris-versicolor
5.9	3	4.2	1.5	Iris-versicolor
6	2.2	4	1	Iris-versicolor
6.1	2.9	4.7	1.4	Iris-versicolor
5.6	2.9	3.6	1.3	Iris-versicolor
6.7	3.1	4.4	1.4	Iris-versicolor
5.6	3	4.5	1.5	Iris-versicolor

5.8	2.7	4.1	1	Iris-versicolor
6.2	2.2	4.5	1.5	Iris-versicolor
5.6	2.5	3.9	1.1	Iris-versicolor
5.9	3.2	4.8	1.8	Iris-versicolor
6.1	2.8	4	1.3	Iris-versicolor
6.3	2.5	4.9	1.5	Iris-versicolor
6.1	2.8	4.7	1.2	Iris-versicolor
6.4	2.9	4.3	1.3	Iris-versicolor
6.6	3	4.4	1.4	Iris-versicolor
6.8	2.8	4.8	1.4	Iris-versicolor
6.7	3	5	1.7	Iris-versicolor
6	2.9	4.5	1.5	Iris-versicolor
5.7	2.6	3.5	1	Iris-versicolor
5.5	2.4	3.8	1.1	Iris-versicolor
5.5	2.4	3.7	1	Iris-versicolor
5.8	2.7	3.9	1.2	Iris-versicolor
6	2.7	5.1	1.6	Iris-versicolor
5.4	3	4.5	1.5	Iris-versicolor
6	3.4	4.5	1.6	Iris-versicolor
6.7	3.1	4.7	1.5	Iris-versicolor
6.3	2.3	4.4	1.3	Iris-versicolor
5.6	3	4.1	1.3	Iris-versicolor
5.5	2.5	4	1.3	Iris-versicolor
5.5	2.6	4.4	1.2	Iris-versicolor
6.1	3	4.6	1.4	Iris-versicolor
5.8	2.6	4	1.2	Iris-versicolor
5	2.3	3.3	1	Iris-versicolor
5.6	2.7	4.2	1.3	Iris-versicolor
5.7	3	4.2	1.2	Iris-versicolor
5.7	2.9	4.2	1.3	Iris-versicolor
6.2	2.9	4.3	1.3	Iris-versicolor
5.1	2.5	3	1.1	Iris-versicolor
5.7	2.8	4.1	1.3	Iris-versicolor
6.3	3.3	6	2.5	Iris-virginica
5.8	2.7	5.1	1.9	Iris-virginica
7.1	3	5.9	2.1	Iris-virginica
6.3	2.9	5.6	1.8	Iris-virginica

6.5	3	5.8	2.2	Iris-virginica
7.6	3	6.6	2.1	Iris-virginica
4.9	2.5	4.5	1.7	Iris-virginica
7.3	2.9	6.3	1.8	Iris-virginica
6.7	2.5	5.8	1.8	Iris-virginica
7.2	3.6	6.1	2.5	Iris-virginica
6.5	3.2	5.1	2	Iris-virginica
6.4	2.7	5.3	1.9	Iris-virginica
6.8	3	5.5	2.1	Iris-virginica
5.7	2.5	5	2	Iris-virginica
5.8	2.8	5.1	2.4	Iris-virginica
6.4	3.2	5.3	2.3	Iris-virginica
6.5	3	5.5	1.8	Iris-virginica
7.7	3.8	6.7	2.2	Iris-virginica
7.7	2.6	6.9	2.3	Iris-virginica
6	2.2	5	1.5	Iris-virginica
6.9	3.2	5.7	2.3	Iris-virginica
5.6	2.8	4.9	2	Iris-virginica
7.7	2.8	6.7	2	Iris-virginica
6.3	2.7	4.9	1.8	Iris-virginica
6.7	3.3	5.7	2.1	Iris-virginica
7.2	3.2	6	1.8	Iris-virginica
6.2	2.8	4.8	1.8	Iris-virginica
6.1	3	4.9	1.8	Iris-virginica
6.4	2.8	5.6	2.1	Iris-virginica
7.2	3	5.8	1.6	Iris-virginica
7.4	2.8	6.1	1.9	Iris-virginica
7.9	3.8	6.4	2	Iris-virginica
6.4	2.8	5.6	2.2	Iris-virginica
6.3	2.8	5.1	1.5	Iris-virginica
6.1	2.6	5.6	1.4	Iris-virginica
7.7	3	6.1	2.3	Iris-virginica
6.3	3.4	5.6	2.4	Iris-virginica
6.4	3.1	5.5	1.8	Iris-virginica
6	3	4.8	1.8	Iris-virginica
6.9	3.1	5.4	2.1	Iris-virginica
6.7	3.1	5.6	2.4	Iris-virginica

6.9	3.1	5.1	2.3	Iris-virginica
5.8	2.7	5.1	1.9	Iris-virginica
6.8	3.2	5.9	2.3	Iris-virginica
6.7	3.3	5.7	2.5	Iris-virginica
6.7	3	5.2	2.3	Iris-virginica
6.3	2.5	5	1.9	Iris-virginica
6.5	3	5.2	2	Iris-virginica
6.2	3.4	5.4	2.3	Iris-virginica
5.9	3	5.1	1.8	Iris-virginica

2. ฐานข้อมูล Air passengers

time	AirPassengers
1949	112
1949.083333	118
1949.166667	132
1949.25	129
1949.333333	121
1949.416667	135
1949.5	148
1949.583333	148
1949.666667	136
1949.75	119
1949.833333	104
1949.916667	118
1950	115
1950.083333	126
1950.166667	141
1950.25	135
1950.333333	125
1950.416667	149
1950.5	170

1950.583333	170
1950.666667	158
1950.75	133
1950.833333	114
1950.916667	140
1951	145
1951.083333	150
1951.166667	178
1951.25	163
1951.333333	172
1951.416667	178
1951.5	199
1951.583333	199
1951.666667	184
1951.75	162
1951.833333	146
1951.916667	166
1952	171
1952.083333	180
1952.166667	193
1952.25	181
1952.333333	183
1952.416667	218
1952.5	230
1952.583333	242
1952.666667	209
1952.75	191
1952.833333	172
1952.916667	194
1953	196
1953.083333	196
1953.166667	236

1953.25	235
1953.333333	229
1953.416667	243
1953.5	264
1953.583333	272
1953.666667	237
1953.75	211
1953.833333	180
1953.916667	201
1954	204
1954.083333	188
1954.166667	235
1954.25	227
1954.333333	234
1954.416667	264
1954.5	302
1954.583333	293
1954.666667	259
1954.75	229
1954.833333	203
1954.916667	229
1955	242
1955.083333	233
1955.166667	267
1955.25	269
1955.333333	270
1955.416667	315
1955.5	364
1955.583333	347
1955.666667	312
1955.75	274
1955.833333	237

1955.916667	278
1956	284
1956.083333	277
1956.166667	317
1956.25	313
1956.333333	318
1956.416667	374
1956.5	413
1956.583333	405
1956.666667	355
1956.75	306
1956.833333	271
1956.916667	306
1957	315
1957.083333	301
1957.166667	356
1957.25	348
1957.333333	355
1957.416667	422
1957.5	465
1957.583333	467
1957.666667	404
1957.75	347
1957.833333	305
1957.916667	336
1958	340
1958.083333	318
1958.166667	362
1958.25	348
1958.333333	363
1958.416667	435
1958.5	491

1958.583333	505
1958.666667	404
1958.75	359
1958.833333	310
1958.916667	337
1959	360
1959.083333	342
1959.166667	406
1959.25	396
1959.333333	420
1959.416667	472
1959.5	548
1959.583333	559
1959.666667	463
1959.75	407
1959.833333	362
1959.916667	405
1960	417
1960.083333	391
1960.166667	419
1960.25	461
1960.333333	472
1960.416667	535
1960.5	622
1960.583333	606
1960.666667	508
1960.75	461
1960.833333	390
1960.916667	432

3. ฐานข้อมูล Tooth growth

len	supp	dose
4.2	VC	0.5
11.5	VC	0.5
7.3	VC	0.5
5.8	VC	0.5
6.4	VC	0.5
10	VC	0.5
11.2	VC	0.5
11.2	VC	0.5
5.2	VC	0.5
7	VC	0.5
16.5	VC	1
16.5	VC	1
15.2	VC	1
17.3	VC	1
22.5	VC	1
17.3	VC	1
13.6	VC	1
14.5	VC	1
18.8	VC	1
15.5	VC	1
23.6	VC	2
18.5	VC	2
33.9	VC	2
25.5	VC	2
26.4	VC	2
32.5	VC	2
26.7	VC	2
21.5	VC	2
23.3	VC	2

29.5	VC	2
15.2	OJ	0.5
21.5	OJ	0.5
17.6	OJ	0.5
9.7	OJ	0.5
14.5	OJ	0.5
10	OJ	0.5
8.2	OJ	0.5
9.4	OJ	0.5
16.5	OJ	0.5
9.7	OJ	0.5
19.7	OJ	1
23.3	OJ	1
23.6	OJ	1
26.4	OJ	1
20	OJ	1
25.2	OJ	1
25.8	OJ	1
21.2	OJ	1
14.5	OJ	1
27.3	OJ	1
25.5	OJ	2
26.4	OJ	2
22.4	OJ	2
24.5	OJ	2
24.8	OJ	2
30.9	OJ	2
26.4	OJ	2
27.3	OJ	2
29.4	OJ	2
23	OJ	2

ภาคผนวก ข

งานวิจัยที่เผยแพร่ในงานประชุมวิชาการ

The 22nd Annual Meeting in Mathematics (AMM 2017)
Department of Mathematics, Faculty of Science
Chiang Mai University, Chiang Mai, Thailand

Number of cluster for k-means clustering by RCFDC method*

Pornpon Othata‡, and Praiboon Pantaragphong†
Department of Appiled Mathematics, Faculty of Science
King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand

Abstract

In this paper will present new algorithm for calculate number of clusters for k-mean clustering with data set on R^2. The algorithm consist of there steps as following. (1) Compute the cumulative frequency distribution (CFD) of first attribute. (2) Find number of group by consider rate of CFD change on first attribute. (3) Compute the CFD of second attribute and consider the number of cluster in each subgroup. The new algorithm has time complexity $O(n)$.

Keywords: cumulative frequency distribution, k-mean clustering.

2010 MSC: Primary 62H10; Secondary 62H30.

1 Introduction

Today big data in storage is very necessary to make decisions. The use of effective big data is required through data classification. The most popular method for grouping data set is K-means clustering [1].

The k-means clustering is a way for a clustering of n data into k clusters. Normally k values for K-means are chosen. Choosing the appropriate k for a dataset needs to be determined by several factors in the dataset, which is very difficult in a large data set. Therefore, selecting the appropriate k for each data set is a very important issue.

There are currently several ways to help for selecting the appropriate k, such as the elbow method [3]. These methods need to rely on the k-means clustering trial data for many k to be used for decision making. This method takes a high iteratively and takes a long time to choose appropriate k.

In this paper will present new algorithm for calculate number of clusters for k-mean clustering with dataset on R^2. The method used the rate of cumulative frequency distribution change of each attribute, which is called RCFDC. The RCFDC method can calculate k without considering of posible k.

*This research was financially supported by, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang

†Corresponding author.
‡Speaker.
E-mail address: pornpon.othata@gmail.com , praiboon.pa@kmitl.ac.th.

2 Preliminaries

In this research we have used the basic knowledge of k-means clustering and cumulative frequency distribution to develop the new algorithms.

2.1 K-means Clustering

Given a dataset with n points in a d-dimensional space and dataset $D = \{x_i\}_{i=1}^{n}$. K-means clustering [1] introduces the clusters implies by randomly choosing k point in data space.In each iteration of K-means comprises of two step: (1) cluster assignment, and (2) centroid update. The group was divided to be effective is determined by the sum square error. The function of the sum square error as follows

$$SSE(C) = \sum_{i=1}^{k} \sum_{x_j \in C_i} \|x_j - \mu_i\|^2 \tag{2.1}$$

Given the k cluster, in cluster assignment, compute the mean of k clusters, each point $x_j \in D$ is assigned to the closest mean, with each clusters C_i comprising points that are closer to μ_i than any other cluster mean. So each point x_j is assigned to cluster C_{j^*} where

$$j^* = \arg\min_{i=1}^{k} \left\{ \|x_j - \mu_i\|^2 \right\} \tag{2.2}$$

Given a set of groups C_i, $i = 1, 2, ..., k$ in the centroid update step, and calculate the new mean of each group from the point in C_i. The cluster assignment and centroid update steps are carried out iteratively until we reach a local minima. If the centroids do not change from one iteration to the next, we can stop if $\sum_{i=1}^{k} \|\mu_i^t - \mu_i^{t-1}\|^2 \leq \varepsilon$, where $\varepsilon > 0$ is the convergence threshold, t is the current iteration and μ_i^t is the average of the group C_i in the cycle t, which is calculated from $\mu_i^t = \frac{1}{n_i} \sum_{x_j \in C_i} x_j$ where n_i is the number of points in cluster C_i. Which steps of K-mean clustering are shown in algorithm 2.1

Algorithm 2.1 K-means clustering

K-means (D, k, ε) :
$t = 0$, Random initialize k centroids: $\mu_1^t, \mu_2^t, ..., \mu_k^t \in \mathbb{R}^d$
repeat $t \leftarrow t+1$
  $C_j \leftarrow \varnothing$ for all $j = 1, ..., k$
  // Cluster Assignment Step
  foreach $x_j \in D$ **do**
    $j^* \leftarrow \arg\min_i \left\{ \|x_j - \mu_i^t\|^2 \right\}$ //Assign x_j to closest centroid
    $C_{j^*} \leftarrow C_j \cup \{x_j\}$
  // Centroid Update Step
  foreach $i = 1$ **to** k **do**
    $\mu_i^t \leftarrow \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j$
until $\sum_{i=1}^{k} \|\mu_i^t - \mu_i^{t-1}\|^2 \leq \varepsilon$

In terms of computational complexity, we see that the grouping process takes $O(nkd)$, since for each n point there is a distance calculation with each k group, which is executed d times in the d-dimension. The new centroid takes $O(nd)$ because there are n points in the d-dimension is assumed to be t rounded, so the total time of the chemical method is $O(tnkd)$.

2.2 Cumulative Frequency Distribution

The cumulative frequency distribution (CFD) [2] of a possible value, or any class, is the sum of the frequency of that value, or that of that class, or the frequency of the value or of the class of frequencies that have a lower or lower overall score range. An example of a table and graph of the cumulative frequency distribution of the iris setosa dataset [4] is shown in Figure 2.1.

Petal length	Frequency	Cumulative frequency
1	1	1
1.5	36	37
2	13	50
2.5	0	50
3	1	51
3.5	4	55
4	11	66
4.5	21	87
5	22	108
5.5	17	125
6	16	141
6.5	5	146
7	4	150

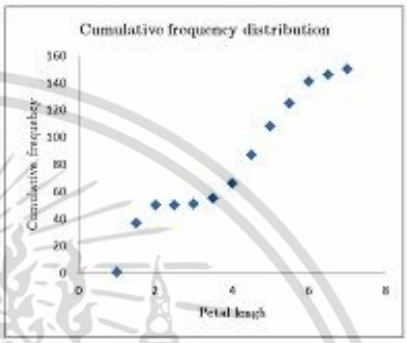


Figure 2.1 Table and graph of cumulative frequency distribution of petal length

2.3 Elbow Method

Elbow method [3] is a technique which takes a gander at the rate of fluctuation clarified as a component of the number of clusters. This strategy exists upon the possibility that one ought to pick various groups so that including another bunch doesn't give much better displaying of the data. The principal groups will include much data however sooner or later the peripheral pick up will drop significantly and gives an edge in the diagram. The appropriate k i.e. number of bunches is picked now, consequently the "elbow rule". An example of using the elbow method to determine the k value of an iris setosa dataset in the properties of petal width and petal length and air passenger dataset [5] as shown in Figure 2.2 and 2.3 respectively and we get $k = 3$ for iris setosa and $k = 5$ for air passenger.

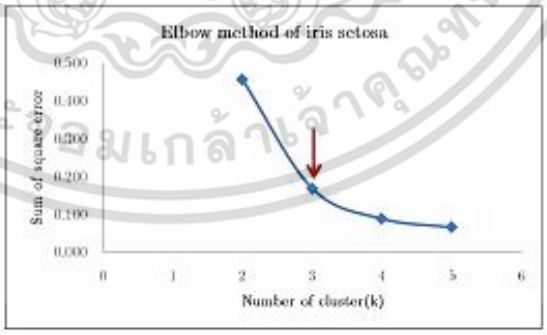


Figure 2.2 Elbow method for iris setosa

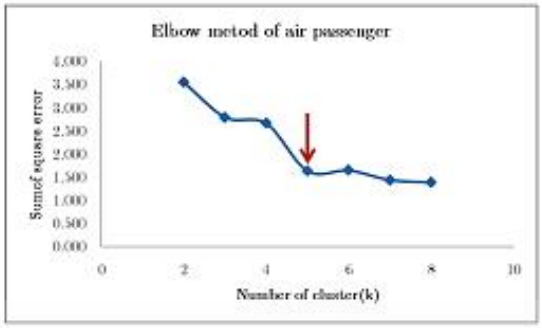


Figure 2.3 Elbow method for air passenger

3 RCFDC Method

The RCFDC method is a process of computing for get k values for k-means clustering on R^2. The RCFDC divides the data by considering the rate of cumulative frequency distribution of each attribute change. If the rate of the cumulative frequency distribution is non decrease in some interval implies that this dataset has a small amount of data at this interval. So we can divide the range of data at this interval. An example of a sparse data interval is shown in Figure 3.1.

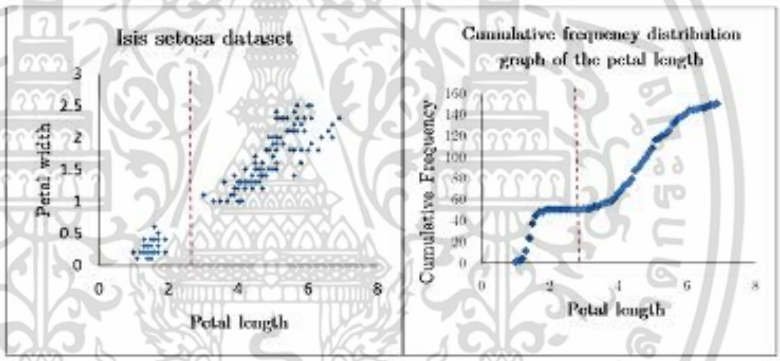


Figure 3.1 Interval of sparse data of petal lenght

The RCFDC method as shown in algorlithm 3.1 is divided into 3 steps. (1) Calculates the CFD of the first attribute of a set and selecting the appropriate scale. (2) Consider the rate of the CFD from (1) change. If the rate of change of CFD is less than ε and consecutive is equal to δ , then the point x_i is breakpoint for calculating the CFD of a second attribute. The ε and δ obtained by choose. (3) Calculates the CFD of the second attribute of the set using breakpoint x_i are given from (2). Considering the number of cluster as well as considering breakpoint in (2) for each CFD for each breakpoint.

In terms of computational complexity, we see that the CFD process on first attribute takes $O(n)$, since iris setosa dataset has n points. The time to consider the breakpoint on first attribute takes $O(n)$. The time of calculate CFD on second attribute takes $O(n)$, since number of all group is $\sum_{i=1}^{m} n_i = n$. The time of consider rate of CFD change of second attribute for finding k takes $O(n)$. So the total time of the RCFDC is $O(n) + O(n) + O(n) + O(n) = O(n)$.

Algorithm 3.1 RCFDC Method

Input ε, δ, n, x_i, y_i
$c=0$, numbercluster $=0$, $j=0$
$CFx_i \leftarrow$ compute CFD of first attribute
foreach i = 1 to n do
    $\Delta CFx = CFx_{i+1} - CFx_i$
    if $\Delta CFx < \varepsilon$
        $c = c+1$
        if $c == \delta$
            $j = j+1$, $n_j = n_j+1$
            $s_j = x_i$
        else
    else
        $c = 0$
foreach k = 1 to j do
    $CFy_i \leftarrow$ compute CFD of second attribute at point less than s_k
    numbercluster $\leftarrow$ numbercluster+1
    foreach i = 1 to n_k do
        $\Delta CFy = CFy_{i+1} - CFy_i$
        if $\Delta CFy < \varepsilon$
            $c = c+1$
            if $c == \delta$
                numberclusters $\leftarrow$ numbercluster+1
            else
        else
            $c = 0$

4 Experiment

We are using RCFDC method with the iris setosa and air passenger dataset. For the iris setosa, we chosen two features such as petal length and petal width. On the experiment, we are compute the CFD on petal length with data scale 0.1. The graph of CFD as shown in Figure 4.1.

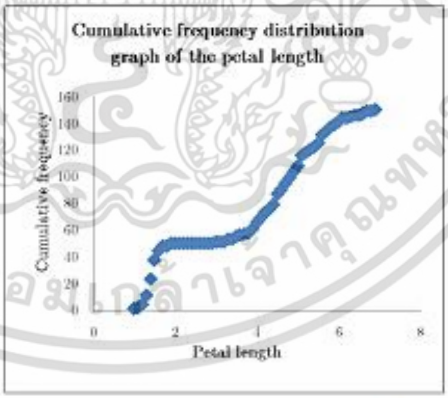


Figure 4.1 cumulative frequency distribution graph of the petal length

We consider the breakpoint for the data by rate of CFD change with $\varepsilon = 2$ and $\delta = 2$, the result as shown in Figure 4.2.

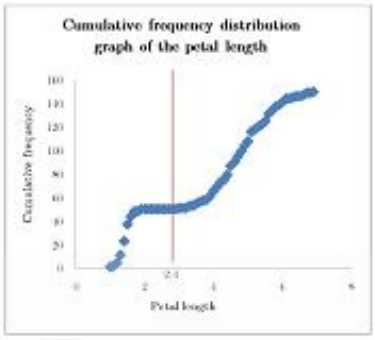


Figure 4.2 Divide point of the petal length

Then we compute the CFD of petal width with scale 0.05 with breakpoint obtained from the previous step, and find breakpoint for cluster on the petal width of the first group and second group. The result as shown in Figure 4.3 and 4.4 respectively. So we get $k = 3$ for k-means clustering.

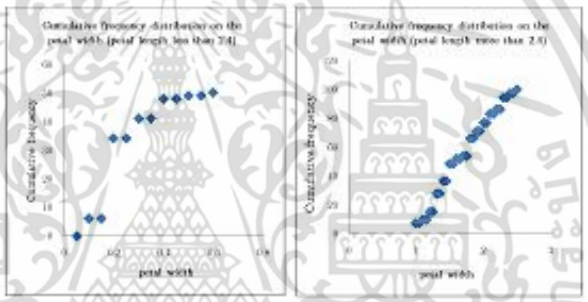


Figure 4.3 Cumulative frequency distribution on the petal width (petal length less than 2.4 and more than 2.4)

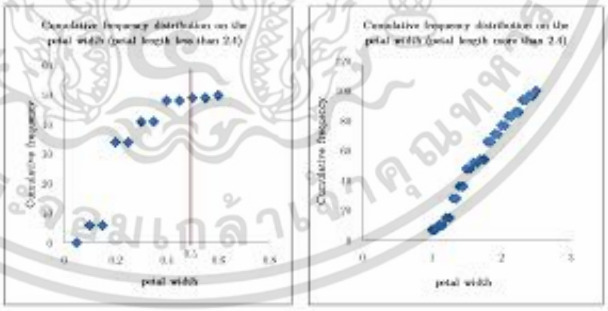


Figure 4.4 Divide point of the petal width

In the RCFDC method, the ε and δ values are very important because both values affect the breakpoint of the groups. The key to choosing the appropriate ε and δ values depends on the density of the dataset. We performed clustering using the iris sentosa dataset using various e and d values as shown in Table 4.5.

ε value	δ value	Number of clusters	Sum of square error
1	1	3	0.169
	2	2	0.457
	3	2	0.457
2	1	5	0.066
	2	3	0.169
	3	2	0.457
3	1	8	0.037
	2	5	0.066
	3	3	0.169

Table 4.5 The various of ε and δ and the number of clusters of iris setosa dataset

On experiment we find that if ε is small value, the algorithm will give a low k-value, if δ is small value, the algorithm will give a high k-value

By comparing the results of the clustering of the RCFDC with elbow method on iris setosa and air passenger data set. We found that the values were similar, as shown in Figure 4.6.

Data set	Elbow method		RCFDC		
	k value	Sum square error	k value	Sum square error	Remark
Iris setosa	3	0.169	3	0.169	$\varepsilon=2, \delta=2$
Air passenger	5	1.636	6	1.648	$\varepsilon=1, \delta=5$

Table 4.6 Comparison the results of clustering of the RCFDC with elbow method

5 Conclusions

The clustering dataset by using the RCDFC method, we found that the effective clusterings was based on the ε and δ values and selection the appropriate ε and δ values for the data set depends on the density and distribution of the dataset. If the dataset is very distributed we must choosing the ε is high and the δ is low. The clusters of datasets obtained from RCFDC selected the appropriate ε and δ that is effective.

In terms of Big(O) of RCFDC, we see that for the CFD calculation of first attribute in the dataset takes $O(n)$. The time of consider the rate of CFD change for finding breakpoint on the first attribute takes $O(n)$. The time for CFD calculation of second properties in the dataset takes $O(n)$. The time for consider the rate of CFD change for finding breakpoint on the second attribute takes $O(n)$. So the total time of RCFDC is $O(n)$.

Acknowledgment. The authors are grateful to Prof.Dr.Sorin V. Sabau for comments on the algorithm and advise program for k-mean clustering .

References

[1] M. J. Zaki and W. Meira JR., *Data mining and analysis* , Progress in Fundamental Concepts and Algorithms, Cambridge University, New York, 2014.

[2] J. Isotalo, *Basics of statistics* , Progress in Statistics, CreateSpace Independent Publishing Platform, 2014.

[3] T. M. Kodinariya1 and P. R. Makwana, *Review on determining number of Cluster in K-Means Clustering*, International Journal of Advance Research in Computer Science and Management Studies. **1** (2013), Issue 6, 90–95.

[4] Edgar Anderson's Iris Data. (n.d.). Retrieved March 15, 2017, from *https://vincentarelbundock.github.io/Rdatasets/datasets.html*

[5] Monthly Airline Passenger Numbers 1949-1960. (n.d.). Retrieved March 15, 2017, from *https://vincentarelbundock.github.io/Rdatasets/datasets.html*

ประวัติผู้เขียน

ชื่อ	นาย พรพล โอทาตะ
วัน เดือน ปีเกิด	24 สิงหาคม 2535
ที่อยู่ปัจจุบัน	บ้านเลขที่ 5/230 หมู่ที่ 3 แขวงโคกแฝด เขตหนองนอก กรุงเทพฯ 10530
ประวัติการศึกษา	2558 วิทยาศาสตรบัณฑิต สาขาคณิตศาสตร์ประยุกต์ เกรดเฉลี่ย 3.06 สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
	2560 วิทยาศาสตรมหาบัณฑิต สาขาคณิตศาสตร์ประยุกต์ เกรดเฉลี่ย 3.56 สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ทุนการศึกษาที่ได้รับ	ทุนอุดหนุนการศึกษาระดับบัณฑิตศึกษา คณะวิทยาศาสตร์ สจล.
ผลงานทางวิชาการ	1. Number of Cluster for K-Means Clustering by RCFDC Method Proceedings of AMM 2017 Page MIS-03-1 – MIS-03-8