สำนักหอสมุดกลาง พระจอมเกล้าลาดกระบัง

การใช้ Anytime Algorithm สำหรับการจัดกลุ่มด้วย K-Mean Algorithm

Using Anytime Algorithm for K-Mean Clustering



นายรวี บุญสันติสุข

นางสาวสุกัญญา พินิจ

นางสาวชนิกานต์ คันโธ

เลขหมู่
เลขทะเบียน 117236
วัน,เดือน,ปี 19 ก.ค. 2554

b 12340996
i

โครงงานพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต

สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ปีการศึกษา 2553

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

USING ANYTIME ALGORITHM FOR

K-MEAN CLUSTERING

MR. RAWEE BOONSANTISUK

MISS SUKUNYA PINIT

MISS CHANIKARN KUNTHO

A SPECIAL PROBLEM SUBMITTED IN PARTIAL FULFILMENT

OF THE REQUIRMENT FOR THE DEGREE OF BACHELOR OF SCIENCE

IN COMPUTER SCIENCE

FACULTY OF SCIENCE

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

ACADEMIC YEAR 2010

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

หัวข้อปัญหาพิเศษ	การใช้ Anytime Algorithm สำหรับการจัดกลุ่มด้วย K-Mean Algorithm	
ชื่อนักศึกษา	นายรวี	บุญสันติสุข
	นางสาวสุกัญญา	พินิจ
	นางสาวชนิกานต์	คันโช
ปริญญา	วิทยาศาสตรบัณฑิต	
สาขาวิชา	วิทยาการคอมพิวเตอร์	
ปีการศึกษา	2553	
อาจารย์ที่ปรึกษา	รศ.ดร.วีระ บุญจริง	

คณะวิทยาศาสตร์สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง อนุมัติให้ปัญหาพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร วิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ ประจำปีการศึกษา 2553

คณะกรรมการสอบ	ลายมือชื่อ
อาจารย์วีระชัย ตันยะสิทธิ์	
อาจารย์สันธนะ อู่อุดมยิ่ง	
รศ.ดร. วีระ บุญจริง	

ลิขสิทธิ์ของคณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของลิขสิทธิ์ทุกครั้งที่มีการนำไปใช้

หัวข้อปัญหาพิเศษ	การใช้ Anytime Algorithm สำหรับการจัดกลุ่มด้วย K-Mean Algorithm	
ชื่อนักศึกษา	นายรวี	บุญสันติสุข
	นางสาวสุกัญญา	พินิจ
	นางสาวชนิกานต์	คันโธ
ปริญญา	วิทยาศาสตรบัณฑิต	
สาขาวิชา	วิทยาการคอมพิวเตอร์	
ปีการศึกษา	2553	
อาจารย์ที่ปรึกษา	รศ.ดร.วีระ บุญจริง	

บทคัดย่อ

ขั้นตอนวิธีการจัดกลุ่มเป็นขั้นตอนวิธีที่ใช้ในการหากลุ่มข้อมูลโดยอาศัยการวัดความคล้ายกันของข้อมูลแบบต่างๆ ขั้นตอนวิธีการหนึ่งที่นิยมใช้ คือ ขั้นตอนวิธีการเคมีนที่สามารถใช้ได้กับงานหลายด้าน แต่ยังคงมีข้อจำกัดอยู่ ดังนั้นปัญหาปัญหาพิเศษนี้มีวัตถุประสงค์คือ มุ่งเน้นการศึกษาทดลองขั้นตอนวิธีการเอนนีไทม์เคมีน ที่มีการแบ่งข้อมูลออกเป็นส่วนๆ โดยมีจำนวนที่เท่ากับหรือแตกต่างกันในการจัดกลุ่มข้อมูล เพื่อวิเคราะห์ผลที่เกิดจากการปรับจำนวนข้อมูลแต่ละส่วนที่แบ่ง และจำนวนส่วนข้อมูลที่แบ่ง เพื่อใช้ในการจัดกลุ่มข้อมูล และเปรียบเทียบผลการทดลองที่ได้นั้นกับผลการจัดกลุ่มที่ได้จากขั้นตอนวิธีการเคมีน ซึ่งจากผลการทดลองสามารถสรุปได้ว่าขั้นตอนวิธีการเอนนีไทม์เคมีนที่มีการแบ่งข้อมูลออกเป็นส่วนๆนั้นมีประสิทธิภาพในการจัดกลุ่มมากกว่าการใช้ขั้นตอนวิธีการเคมีนในการจัดกลุ่ม

คำสำคัญ: ขั้นตอนวิธีการเคมีน, ขั้นตอนวิธีการเอนนีไทม์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Title	Anytime K-Mean Clustering Algorithm
Students	Mr. Rawee Boonsantisuk
	Miss Sukunya Pinit
	Miss Chanikarn Kuntho
Degree	Bachelor of Science
Major Program	Computer Science
Academic Year	2010
Advisor	Assoc.Prof.Dr.Veera Boonjing

ABSTRACT

A clustering algorithm attempts to find groups of data based on some similarities. One method used is the K-Mean Algorithms that can be used with multiple sides but K-Mean Clustering also has some limitations in terms of processing which must be accomplished all at once. Therefore, this research has proposed to optimization of K-Mean Algorithm by applying Anytime Algorithm to K-Mean Algorithm. Anytime K-Mean Algorithm divides the information into groups of equal or different data size to analyze the effect of adjusting, the amount of data that divides each section and the number of dimensions that divide. When the results of processing Anytime Algorithm to measure efficiency of clustering can be seen that the efficiency of Anytime K-Mean Algorithm is better.

Keyword: K-Mean Algorithm, Anytime Algorithm

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

กิตติกรรมประกาศ

ในการทำปัญหาพิเศษหัวข้อเรื่อง การใช้ Anytime Algorithm สำหรับการจัดกลุ่มด้วย K-Mean Algorithm สำเร็จลงได้ด้วยดีจากความช่วยเหลือและสนับสนุนของบุคคลหลายท่าน คณะผู้จัดทำขอขอบพระคุณบุคคลดังต่อไปนี้

ขอขอบคุณ บิดา มารดา ญาติพี่น้องทุกคนที่คอยดูแลห่วงใยเป็นกำลังใจ และให้การสนับสนุนในทุกๆด้านและให้กำลังใจเสมอมา

ขอขอบคุณ รศ.ดร.วีระ บุญจริง อาจารย์ที่ปรึกษา ที่ให้ความกรุณาให้คำแนะนำและแนะแนวทางที่เป็นประโยชน์ต่อปัญหาพิเศษ คอยดูแลให้ปัญหาพิเศษสำเร็จออกมาด้วยดี

ขอขอบคุณ อาจารย์วีระชัย ตันยะสิทธิ์ และอาจารย์สันธนะ อู่อุดมยิ่ง คณะกรรมการสอบปัญหาพิเศษ ที่กรุณาให้คำแนะนำ และช่วยตรวจสอบแก้ไขปัญหาพิเศษฉบับนี้ และขอขอบคุณคณาจารย์ในภาควิชาวิทยาการคอมพิวเตอร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบังทุกท่านที่ประสิทธ์ประสาทความรู้อันมีค่า

ขอขอบคุณพี่ๆ และเพื่อนๆ ทุกคนในภาควิชาที่คอยช่วยเหลือและให้คำปรึกษาในการทำปัญหาพิเศษฉบับนี้ให้สำเร็จลุล่วงในที่สุด และขอขมาในสิ่งที่ล่วงเกินไว้ ณ โอกาสนี้

นอกจากนี้ยังมีบุคคลท่านอื่นๆที่ไม่ได้กล่าวไว้ ณ ที่นี้จึงใคร่ขอขอบพระคุณทุกท่านที่กรุณามีส่วนร่วมในการให้ความช่วยเหลือ ให้คำปรึกษา ให้คำแนะนำ ให้การสนับสนุนตลอดจนกำลังใจในการทำปัญหาพิเศษฉบับนี้

นายรวี บุญสันติสุข
นางสาวสุกัญญา พินิจ
นางสาวชนิกานต์ คันโธ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ

	หน้า
บทคัดย่อภาษาไทย	I
บทคัดย่อภาษาอังกฤษ	II
กิตติกรรมประกาศ	III
สารบัญ	IV
สารบัญรูป	VI
สารบัญตาราง	VII
บทที่ 1 บทนำ	**1**
1.1 ที่มาและความสำคัญของปัญหา	1
1.2 วัตถุประสงค์	2
1.3 ขอบเขต	2
1.4 ประโยชน์ที่คาดว่าจะได้รับ	2
บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง	**3**
2.1 ทฤษฎีที่เกี่ยวข้อง	3
2.1.1 การเรียนรู้แบบไม่มีผู้สอน	3
2.1.1.1 การจัดกลุ่ม	3
2.1.1.2 ขั้นตอนวิธีการเคมีน	4
2.1.2 ขั้นตอนวิธีการเอนนีไทม์	11
2.1.3 การประเมินประสิทธิภาพการจัดกลุ่มข้อมูล	12
2.1.3.1 Purity's Cluster	12
2.1.3.2 Normalized Mutual Information	13
2.1.3.3 The Rand Index	13
2.1.3.4 F Measure	14

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ(ต่อ)

หน้า

บทที่ 3 ขั้นตอนวิธีการเอนนีไทม์เคมีนสำหรับการจัดกลุ่ม **15**

3.1 ขั้นตอนวิธีการเอนนีไทม์เคมีน 15

3.2 ขั้นตอนวิธีการจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีการเอนนีไทม์เคมีน 16

บทที่ 4 ผลการทดลอง **27**

4.1 ชุดข้อมูลที่ใช้ 27

4.1.1 ชุดข้อมูล Iris Plants 27

4.1.2 ชุดข้อมูล Balance Scale 28

4.1.3 ชุดข้อมูล Haberman's Survival 28

4.1.4 ชุดข้อมูล Hayes-Roth 29

4.2 เครื่องมือที่ใช้ในการทดลอง 29

4.3 วิธีการประเมินประสิทธิภาพการทดลอง 29

4.4 ตัววัดประสิทธิภาพ 30

4.5 ผลการทดลอง 30

4.5.1 ผลการทดลองกับชุดข้อมูล Iris Plants 30

4.5.2 ผลการทดลองกับชุดข้อมูล Balance Scale 34

4.5.3 ผลการทดลองกับชุดข้อมูล Haberman's Survival 38

4.5.4 ผลการทดลองกับชุดข้อมูล Hayes-Roth 42

บทที่ 5 สรุปและข้อเสนอแนะ 46

5.1 สรุป 46

5.2 ข้อเสนอแนะ 46

เอกสารอ้างอิง **47**

ภาคผนวก **48**

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป

รูปที่		หน้า
4.1	ผลการวัดประสิทธิภาพการจัดกลุ่มของชุดข้อมูล Iris ด้วย Purity	30
4.2	ผลการวัดประสิทธิภาพการจัดกลุ่มของชุดข้อมูล Iris ด้วย NMI	31
4.3	ผลการวัดประสิทธิภาพการจัดกลุ่มของชุดข้อมูล Iris ด้วย RI	32
4.4	ผลการวัดประสิทธิภาพการจัดกลุ่มของชุดข้อมูล Iris ด้วย F	33
4.5	ผลการวัดประสิทธิภาพการจัดกลุ่มของชุดข้อมูล Balance Scale ด้วย Purity	34
4.6	ผลการวัดประสิทธิภาพการจัดกลุ่มของชุดข้อมูล Balance Scale ด้วย NMI	35
4.7	ผลการวัดประสิทธิภาพการจัดกลุ่มของชุดข้อมูล Balance Scale ด้วย RI	36
4.8	ผลการวัดประสิทธิภาพการจัดกลุ่มของชุดข้อมูล Balance Scale ด้วย F	37
4.9	ผลการวัดประสิทธิภาพการจัดกลุ่มของชุดข้อมูล Haberman ด้วย Purity	38
4.10	ผลการวัดประสิทธิภาพการจัดกลุ่มของชุดข้อมูล Haberman ด้วย NMI	39
4.11	ผลการวัดประสิทธิภาพการจัดกลุ่มของชุดข้อมูล Haberman ด้วย RI	40
4.12	ผลการวัดประสิทธิภาพการจัดกลุ่มของชุดข้อมูล Haberman ด้วย F	41
4.13	ผลการวัดประสิทธิภาพการจัดกลุ่มของชุดข้อมูล Hayes-Roth ด้วย Purity	42
4.14	ผลการวัดประสิทธิภาพการจัดกลุ่มของชุดข้อมูล Hayes-Roth ด้วย NMI	43
4.15	ผลการวัดประสิทธิภาพการจัดกลุ่มของชุดข้อมูล Hayes-Roth ด้วย RI	44
4.16	ผลการวัดประสิทธิภาพการจัดกลุ่มของชุดข้อมูล Hayes-Roth ด้วย F	45

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญตาราง

ตารางที่	หน้า
2.1 แสดงคุณลักษณะของชุดข้อมูล Iris 30 ตัว	5
2.2 แสดงจุดศูนย์กลางแต่ละกลุ่ม (รอบที่ 1)	6
2.3 แสดงระยะห่างจากจุดศูนย์กลางแต่ละกลุ่มถึงข้อมูลแต่ละตัว (รอบที่ 1)	7
2.4 แสดงจุดศูนย์กลางแต่ละกลุ่ม (รอบที่ 2)	7
2.5 แสดงระยะห่างจากจุดศูนย์กลางแต่ละกลุ่มถึงข้อมูลแต่ละตัว (รอบที่ 2)	8
2.6 แสดงจุดศูนย์กลางแต่ละกลุ่ม (รอบที่ 3)	9
2.7 แสดงค่าจากการวัดประสิทธิภาพ	9
2.8 แสดงระยะห่างจากจุดศูนย์กลางแต่ละกลุ่มถึงข้อมูลแต่ละตัว (รอบที่ 3)	10
2.9 แสดงผลการจัดกลุ่มข้อมูลของชุดข้อมูล Iris 30 ตัว	10
3.1 แสดงคุณลักษณะของชุดข้อมูล Iris 15 ตัว (ส่วนที่ 1)	17
3.2 แสดงจุดศูนย์กลางแต่ละกลุ่ม (ส่วนที่ 1 รอบที่ 1)	17
3.3 แสดงระยะห่างจากจุดศูนย์กลางแต่ละกลุ่มถึงข้อมูลแต่ละตัว (ส่วนที่ 1 รอบที่ 1)	18
3.4 แสดงจุดศูนย์กลางแต่ละกลุ่ม (ส่วนที่ 1 รอบที่ 2)	19
3.5 แสดงระยะห่างจากจุดศูนย์กลางแต่ละกลุ่มถึงข้อมูลแต่ละตัว (ส่วนที่ 1 รอบที่ 2)	19
3.6 แสดงผลการจัดกลุ่มข้อมูลของชุดข้อมูล Iris 15 ตัว	20
3.7 แสดงคุณลักษณะของชุดข้อมูล Iris 13 ตัว (ส่วนที่ 2)	20
3.8 แสดงจุดศูนย์กลางแต่ละกลุ่ม (ส่วนที่ 2 รอบที่ 1)	21
3.9 แสดงระยะห่างจากจุดศูนย์กลางแต่ละกลุ่มถึงข้อมูลแต่ละตัว (ส่วนที่ 2 รอบที่ 1)	21
3.10 แสดงจุดศูนย์กลางแต่ละกลุ่ม (ส่วนที่ 2 รอบที่ 2)	22
3.11 แสดงระยะห่างจากจุดศูนย์กลางแต่ละกลุ่มถึงข้อมูลแต่ละตัว (ส่วนที่ 2 รอบที่ 2)	22
3.12 แสดงผลการจัดกลุ่มข้อมูลของชุดข้อมูล Iris 10 ตัว	23
3.13 แสดงคุณลักษณะของชุดข้อมูล Iris 8 ตัว (ส่วนที่ 3)	23
3.14 แสดงจุดศูนย์กลางแต่ละกลุ่ม (ส่วนที่ 3 รอบที่ 1)	24
3.15 แสดงระยะห่างจากจุดศูนย์กลางแต่ละกลุ่มถึงข้อมูลแต่ละตัว (ส่วนที่ 3 รอบที่ 1)	24
3.16 แสดงจุดศูนย์กลางแต่ละกลุ่ม (ส่วนที่ 3 รอบที่ 2)	25
3.17 แสดงระยะห่างจากจุดศูนย์กลางแต่ละกลุ่มถึงข้อมูลแต่ละตัว (ส่วนที่ 3 รอบที่ 2)	25
3.18 แสดงผลการจัดกลุ่มข้อมูลของชุดข้อมูล Iris 5 ตัว	26
3.19 แสดงค่าจากการวัดประสิทธิภาพ	26

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหาพิเศษ

ในปัจจุบัน การทำการแบ่งกลุ่มข้อมูลสามารถทำได้หลายวิธี โดยในแต่ละวิธีมีขั้นตอนการทำงานที่แตกต่างกัน ชุดข้อมูลแต่ละชุดที่นำมาทำการแบ่งกลุ่มก็จะใช้วิธีที่แตกต่างกันด้วย ชุดข้อมูลที่นำมาทำอาจเหมาะสมกับวิธีการแบ่งกลุ่มบางวิธีเท่านั้น เมื่อทำการแบ่งกลุ่มเสร็จสิ้นแล้ว ผลลัพธ์ที่ได้ จะต้องเห็นได้ชัดเจนว่า ข้อมูลในกลุ่มเดียวกันจะต้องมีความต่างกันน้อยมาก และข้อมูลที่ต่างกลุ่มกันจะต้องมีความต่างกันมาก แม้ว่าการแบ่งกลุ่มข้อมูลในปัจจุบันจะมีการทำงานที่สะดวกสบายและความรวดเร็ว แต่ก็ยังคงพบปัญหาในเรื่องของเวลาที่ใช้ในการแบ่งแยกประเภทข้อมูล

ขั้นตอนวิธีการเคมีนเป็นวิธีที่นิยมวิธีหนึ่งที่ใช้ในการแบ่งกลุ่มข้อมูล ซึ่งในปัจจุบันถูกพัฒนาให้มีประสิทธิภาพการทำงานที่ดีขึ้น รองรับการทำงานได้หลายรูปแบบเพื่อความสะดวกสบายและมีความรวดเร็วในการทำงาน การทำงานของขั้นตอนวิธีการเคมีนจะพิจารณาที่ศูนย์กลางข้อมูลในแต่ละกลุ่ม แล้วทำการแบ่งกลุ่มข้อมูลทีละข้อมูล ซึ่งผลลัพธ์ที่ได้จะมีความถูกต้องและเหมาะสมที่สุด แต่จะใช้ระยะเวลาในการทำงานนาน จึงจะได้ผลลัพธ์ที่สามารถนำไปใช้งานได้ แต่หากนำขั้นตอนวิธีการเอนนีไทม์ ซึ่งมีกระบวนการทำงานที่แตกต่างจากขั้นตอนวิธีการเคมีน ในเรื่องของเวลาและจำนวนข้อมูลในการประมวลผล มาประยุกต์รวมกับ ขั้นตอนวิธีการเคมีน จะสามารถหาประมวลผลข้อมูลในเวลาใดก็ได้ และประมวลผลข้อมูลจำนวนเท่าใดก็ได้ตามความต้องการ โดยผลจากการจัดกลุ่มสามารถนำไปใช้ประโยชน์ได้เลย โดยไม่รอให้ทำการประมวลผลข้อมูลครบทุกตัว

ดังนั้นการนำขั้นตอนวิธีการเคมีนและขั้นตอนวิธีการเอนนีไทม์มาประยุกต์รวมกัน จะเป็นการเพิ่มความยืดหยุ่นในด้านเวลาให้แก่ขั้นตอนวิธีการเคมีนในแบบเดิม นอกจากนี้การนำมาประยุกต์รวมกันยังส่งผลให้ประมวลในแต่ละรอบหรือในแต่ละชุดข้อมูลนั้นๆ สามารถทำได้อย่างรวดเร็วมากยิ่งขึ้น และยังมีความแม่นยำมากขึ้นอีกด้วย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

1.2 วัตถุประสงค์ของปัญหาพิเศษ

1.2.1 พัฒนางานวิจัยที่นำขั้นตอนวิธีการเคมีนมาประยุกต์เข้ากับขั้นตอนวิธีการเอนนีไทม์ ซึ่งสามารถหาผลลัพธ์จากการจัดกลุ่มชุดข้อมูลได้ เมื่อแบ่งชุดข้อมูล ออกเป็นกลุ่มย่อย

1.2.2 นำงานวิจัยที่นำขั้นตอนวิธีการเคมีน มาประยุกต์เข้ากับขั้นตอนวิธีการเอนนีไทม์ไปใช้เพิ่มความยืดหยุ่นในด้านเวลาให้กับการจัดกลุ่มข้อมูล

1.2.3 นำงานวิจัยที่นำขั้นตอนวิธีการเคมีน มาประยุกต์เข้ากับขั้นตอนวิธีการเอนนีไทม์ไปใช้เพิ่มประสิทธิภาพการจัดกลุ่มข้อมูลให้อยู่ในกลุ่มที่เหมาะสมมากขึ้น โดยไม่ต้องกระทำข้อมูลในชุดข้อมูลจนครบทุกตัว

1.3 ข้อจำกัดและขอบเขตของปัญหาพิเศษ

1.3.1 ทำการพิสูจน์ผลลัพธ์ที่ได้จากงานวิจัยนี้ซึ่งมีพื้นฐานมาจากขั้นตอนวิธีการเคมีน ว่าผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูล จะใกล้เคียงกับผลลัพธ์ที่ได้จากวิธี ขั้นตอนวิธีการเคมีน แบบปรกติ

1.3.2 งานวิจัยนี้เป็นวิธีการในการจัดกลุ่มข้อมูลเป็นกลุ่มๆ ตามจำนวนของกลุ่ม (cluster) ที่ต้องการ โดยทำการแบ่งกลุ่มข้อมูลในการประมวลผลออกเป็นกลุ่มย่อยๆ ซึ่งจะช่วยให้ได้ผลลัพธ์จากการจัดกลุ่มมีความถูกต้องและเหมาะสมมากขึ้น รวมทั้งเกิดความยืดหยุ่นของการประมวลผลในด้านเวลาให้กับการจัดกลุ่มข้อมูล

1.3.3 ทดลองหาผลลัพธ์จากการพัฒนา โดยนำชุดข้อมูลต่างๆมาผ่านขั้นตอนวิธี

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1.4.1 ได้ขั้นตอนวิธีการใหม่ที่มีการนำขั้นตอนวิธีการเคมีนมาประยุกต์เข้ากับขั้นตอนวิธีการเอนนีไทม์ ซึ่งสามารถหาผลลัพธ์ที่ถูกต้องและเหมาะสมในการจัดกลุ่มข้อมูล

1.4.2 ได้ขั้นตอนวิธีการใหม่ซึ่งเพิ่มความยืดหยุ่นในด้านเวลาให้กับการแบ่งกลุ่มข้อมูล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 การเรียนรู้แบบไม่มีผู้สอน

การเรียนรู้แบบไม่มีผู้สอนเป็นเทคนิคหนึ่งในการเรียนรู้ของเครื่อง โดยจะทำการสร้างโมเดลที่มีความเหมาะสมให้กับข้อมูล การเรียนรู้ประเภทนี้จะพิจารณาข้อมูลเป็นเซตของตัวแปรสุ่ม เพราะต้องพยายามเรียนรู้โดยอาศัยเพียงแค่ข้อมูลในชุดข้อมูลในการหาความสัมพันธ์ภายในชุดข้อมูลแล้วจึงสร้างโมเดลความหนาแน่นร่วมของชุดข้อมูลซึ่งสามารถจัดเรียงตัวเองตามลักษณะหรือจัดหมวดหมู่ของข้อมูลได้

เนื่องจากขั้นตอนการเรียนรู้แบบไม่มีผู้สอนนี้จะมีการระบุกลุ่มของข้อมูลก่อน เมื่อตัดสินใจได้แล้วว่าข้อมูลใหม่มีลักษณะที่ควรจะจัดรวมเข้ากลุ่มใด หรืออาจถือเป็นกลุ่มใหม่ก็ได้ในกรณีที่เห็นว่าไม่ควรจัดเข้ากลุ่มใด

2.1.1.1 การจัดกลุ่ม

ในการจัดกลุ่มข้อมูล กระบวนการเรียนรู้นั้นพยายามที่จะหารูปแบบในชุดข้อมูลตัวอย่าง จากนั้นจะทำการวิเคราะห์ว่าควรจะแบ่งข้อมูลออกเป็นกี่ชุด และข้อมูลตัวใดที่ควรจะอยู่ในชุดเดียวกันบ้าง แล้วทำการจัดกลุ่มให้กับข้อมูล ซึ่งการจัดกลุ่มนั้น ข้อมูลที่มีลักษณะคล้ายกันจะอยู่ในกลุ่มเดียวกันโดยไม่มีการจัดกลุ่มข้อมูลตัวอย่างไว้ล่วงหน้า ซึ่งสามารถประยุกต์ใช้กับงานได้หลากหลายไม่ว่าจะเป็นการคัดแยกข้อมูล การจำแนกข้อมูล หรือการจดจำรูปแบบ เทคนิคในการดำเนินการจัดกลุ่มของข้อมูลนั้นหลากหลายไม่ว่าจะเป็นกรรมวิธีเชิงลำดับชั้นหรือวิธีโครง ข่ายประสาทเทียมฯลฯ โดยในขอบเขตงานนี้จะกล่าวถึงการจัดกลุ่มข้อมูลตามขั้นตอนวิธีการเคมีน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.1.1.2 ขั้นตอนวิธีการจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีการเคมีน

ขั้นตอนวิธีการเคมีนเป็นวิธีการที่จะแสดงให้เห็นถึงการจัดกลุ่ม ในโปรแกรมประยุกต์ โดยขั้นตอนวิธีการเคมีน เป็นที่รู้จักกันดีในเรื่องของความง่ายของขั้นตอนวิธีในทางปฏิบัติและได้ผลดี หลักการของขั้นตอนวิธีการเคมีนจะใช้แนวคิดของฟังก์ชันสมาชิกสำหรับการระบุว่าข้อมูลหนึ่งอยู่ในกลุ่มคลัสเตอร์ไหน ด้วยหลักการของเคมีนข้อมูลหนึ่งๆ จะสามารถเป็นสมาชิกของกลุ่มใดกลุ่มหนึ่งได้กลุ่มเดียว จากจำนวน k กลุ่ม จากนั้นจะทำการคำนวณค่าระยะห่างด้วยวิธียูคลิเดียน ของแต่ละเวคเตอร์ กับกลุ่ม k กลุ่ม เพื่อหาระยะห่าง ที่น้อยที่สุดเพื่อเลือกว่า เวคเตอร์นั้นจะอยู่กลุ่มใด หลังจากเลือกเวคเตอร์เข้ากลุ่มแล้วจึงทำการคำนวณหาจุดศูนย์กลางของแต่ละกลุ่มจากเวคเตอร์ในกลุ่มนั้นๆ และจะวนทำซ้ำขั้นตอนเดิม จนกว่าจุดศูนย์กลางกลุ่มใหม่เท่ากับจุดศูนย์กลางกลุ่มเก่า แต่ขั้นตอนวิธีโดยตรงของขั้นตอนวิธีการเคมีน ต้องใช้เวลาจำนวนมากไปกับจำนวนเวคเตอร์และจำนวนกลุ่มต่อการวนลูป เป็นการคำนวณที่มีการใช้ทรัพยากรเปลืองโดยเฉพาะอย่างยิ่งสำหรับชุดข้อมูลขนาดใหญ่ นอกจากนี้ ขั้นตอนวิธีเคมีนนี้ถือได้ว่าเป็นขั้นตอนที่มีการทำซ้ำมาก และจำเป็นต้องกำหนดจำนวนกลุ่ม หรือค่า k ก่อน จึงจะสามารถทำงานได้

วิธีการจัดกลุ่มโดยใช้ ขั้นตอนวิธีการเคมีน มีขั้นตอนดังต่อไปนี้

ขั้นตอนที่ 1 : กำหนดจำนวนกลุ่มที่ต้องการจัดกลุ่ม โดยให้ k แทนจำนวนกลุ่ม

ขั้นตอนที่ 2 : แบ่งข้อมูลออกเป็นกลุ่ม โดยสุ่มให้ข้อมูลแต่ละตัวอยู่ในกลุ่ม ตามจำนวนกลุ่มที่กำหนดไว้

ขั้นตอนที่ 3 : คำนวณหาจุดศูนย์กลางแต่ละกลุ่ม

ขั้นตอนที่ 4 : คำนวณหาระยะห่างจากทุกจุดข้อมูลไปยังจุดศูนย์กลางแต่ละกลุ่ม ด้วยวิธี Euclidean distance ตามสมการ

$$d_E(j,k) = \left(\sum_{i=1}^{n} \left| x_{ji} - x_{ki} \right|^2 \right)^{1/2} \quad (1)$$

ขั้นตอนที่ 5 : ตรวจสอบเงื่อนไขผลจากการคำนวณหาระยะห่างว่าจุดข้อมูลอยู่ในตำแหน่งที่ใกล้กับจุดศูนย์กลางกลุ่มอื่นหรือไม่

ถ้า จริง; ให้ย้ายข้อมูลนั้นไปยังกลุ่มที่มีระยะห่างน้อยที่สุด และวนกลับไปทำขั้นตอนที่ 3 ใหม่อีกครั้ง

ถ้า เท็จ; ให้หยุดการวนรอบ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตัวอย่าง 2.1

จะใช้ชุดข้อมูลทั้งหมด 30 ตัว กำหนดให้ค่า k = 3

รอบที่ 1 แต่ละจุดข้อมูลถูกจัดเป็นกลุ่ม (จากการสุ่ม) โดยข้อมูลแต่ละตัวแสดงดังตาราง 2.1

ตารางที่ 2.1 แสดงคุณลักษณะของชุดข้อมูล

ข้อมูลตัวที่	Sepal Length	Sepal Width	Petal Length	Petal Width	Species
1	5	2	3.5	1	versicolor
2	6.2	2.2	4.5	1.5	versicolor
3	6	2.2	5	1.5	virginica
4	6	2.2	4	1	versicolor
5	5.8	4	1.2	0.2	setosa
6	5.2	4.1	1.5	0.1	setosa
7	5	2.3	3.3	1	versicolor
8	4.5	2.3	1.3	0.3	setosa
9	5.5	2.4	3.8	1.1	versicolor
10	5.1	3.5	1.4	0.2	setosa
11	4.9	2.4	3.3	1	versicolor
12	6.7	2.5	5.8	1.8	virginica
13	5.5	4.2	1.4	0.2	setosa
14	6.3	2.5	4.9	1.5	versicolor
15	5.7	2.5	5	2	virginica
16	4.9	3	1.4	0.2	setosa
17	4.8	3	1.4	0.3	setosa
18	4.8	3	1.4	0.1	setosa
19	4.9	2.5	4.5	1.7	virginica
20	7.7	2.6	6.9	2.3	virginica
21	6.1	2.6	5.6	1.4	virginica
22	5.8	2.6	4	1.2	versicolor
23	5.7	2.6	3.5	1	versicolor
24	5.1	3.8	1.9	0.4	setosa
25	5.1	3.8	1.6	0.2	setosa

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ข้อมูลตัวที่	Sepal Length	Sepal Width	Petal Length	Petal Width	Species
26	6.4	2.7	5.3	1.9	virginica
27	6	2.7	5.1	1.6	versicolor
28	5.8	2.7	5.1	1.9	virginica
29	5.8	2.7	5.1	1.9	virginica
30	5.8	2.7	4.1	1	versicolor

คำนวณหาค่าจุดศูนย์กลางแต่ละกลุ่ม แสดงได้ดังตาราง 2.2 และคำนวณระยะห่างจากจุดศูนย์กลางแต่ละกลุ่มถึงข้อมูลแต่ละตัว แสดงดังตาราง 2.3

ตารางที่ 2.2 แสดงจุดศูนย์กลางแต่ละกลุ่ม

จุดศูนย์กลาง	Sepal Length	Sepal Width	Petal Length	Petal Width
Setosa	5.080	3.470	1.450	0.220
Virginica	6.278	2.556	5.367	1.822
Versicolor	5.655	2.418	4.000	1.173

ทำการตรวจสอบว่า มีการย้ายกลุ่มของข้อมูลหรือไม่ ถ้ามีการย้ายกลุ่ม จะให้ทำการคำนวณหาจุดศูนย์กลางของแต่ละกลุ่มใหม่ และทำการคำนวณหาระยะห่างใหม่ในรอบที่ 2 แต่หากไม่มีข้อมูลตัวใดที่มีการย้ายกลุ่ม หมายถึง การจัดกลุ่มเสร็จสิ้นแล้ว จากการพิจารณาตารางที่ 2.3 จะเห็นได้ว่า มีข้อมูลจำนวนทั้งสิ้น 3 ตัวที่มีการย้ายกลุ่ม เพราะฉะนั้นจะทำการย้ายกลุ่มและทำการวนทำรอบที่ 2

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 2.3 แสดงระยะห่างจากจุดศูนย์กลางแต่ละกลุ่มถึงข้อมูลแต่ละตัว

ข้อมูลที่	Setosa	Virginica	Versicolor	ข้อมูลที่	Setosa	Virginica	Versicolor
1	2.642	2.470	0.940	16	0.506	4.523	2.935
2	3.716	0.994	0.838	17	0.555	4.520	2.931
3	4.608	1.274	2.050	18	0.562	4.591	2.997
4	3.094	1.658	0.444	19	3.531	1.633	1.051
5	0.929	4.723	3.363	20	6.454	2.146	3.728
6	0.654	4.633	3.230	21	4.518	0.516	1.686
7	2.325	2.578	0.981	22	2.956	1.576	0.234
8	1.317	4.699	3.066	23	2.440	2.120	0.561
9	2.760	1.899	0.264	24	0.587	4.120	2.688
10	0.065	4.544	3.031	25	0.364	4.445	2.987
11	2.282	2.621	1.044	26	4.470	0.215	1.689
12	4.998	0.608	2.176	27	4.082	0.467	1.261
13	0.844	4.656	3.302	28	4.154	0.571	1.356
14	3.996	0.570	1.158	29	4.154	0.571	1.356
15	4.135	0.709	1.301	30	2.957	1.590	0.375

รอบที่ 2 ข้อมูลถูกย้าย 3 ตัวจากรอบที่ 1 คำนวณค่าจุดศูนย์กลางแต่ละกลุ่มแสดงดังตาราง 2.4

ตารางที่ 2.4 แสดงจุดศูนย์กลางแต่ละกลุ่ม

จุด ศูนย์กลาง	Sepal Length	Sepal Width	Petal Length	Petal Width
Setosa	5.080	3.470	1.450	0.220
Virginica	6.390	2.570	5.380	1.780
Versicolor	5.480	2.390	3.850	1.150

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 2.5 แสดงระยะห่างจากจุดศูนย์กลางแต่ละกลุ่มถึงข้อมูลแต่ละตัว

ข้อมูลที่	Setosa	Virginica	Versicolor
1	2.642	2.530	0.726
2	3.716	1.013	1.049
3	4.608	1.175	2.273
4	3.094	1.674	0.593
5	0.929	4.729	3.259
6	0.654	4.651	3.103
7	2.325	2.634	0.751
8	1.317	4.741	2.862
9	2.760	1.944	0.074
10	0.065	4.568	2.878
11	2.282	2.680	0.813
12	4.998	0.527	2.393
13	0.844	4.668	3.191
14	3.996	0.567	1.382
15	4.135	0.821	1.451

ข้อมูลที่	Setosa	Virginica	Versicolor
16	0.506	4.554	2.759
17	0.555	4.555	2.749
18	0.562	4.623	2.818
19	3.531	1.734	1.036
20	6.454	2.073	3.949
21	4.518	0.527	1.885
22	2.956	1.609	0.414
23	2.440	2.149	0.487
24	0.587	4.146	2.549
25	0.364	4.468	2.846
26	4.470	0.194	1.899
27	4.082	0.529	1.460
28	4.154	0.677	1.524
29	4.154	0.677	1.524
30	2.957	1.616	0.532

ทำการตรวจสอบว่า มีการย้ายกลุ่มของข้อมูลหรือไม่ ถ้ามีการย้ายกลุ่ม จะให้ทำการคำนวณหาจุดศูนย์กลางของแต่ละกลุ่มใหม่ และทำการคำนวณหาระยะห่างใหม่ในรอบที่ 3 จากการพิจารณาตารางที่ 2.5 จะเห็นได้ว่า มีข้อมูลจำนวน 1 ตัว ที่มีการย้ายกลุ่ม เพราะฉะนั้นจะทำการย้ายกลุ่มและทำการวนทำรอบที่ 3

<u>รอบที่ 3</u> มีข้อมูลถูกย้าย 1 ตัวจากรอบที่ 2 คำนวณค่าจุดศูนย์กลางแต่ละกลุ่มแสดงดังตาราง 2.6

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 2.6 แสดงจุดศูนย์กลางแต่ละกลุ่ม

จุด ศูนย์กลาง	Sepal Length	Sepal Width	Petal Length	Petal Width
Setosa	5.080	3.470	1.450	0.220
Virginica	6.373	2.536	5.300	1.755
Versicolor	5.400	2.411	3.778	1.111

ทำการตรวจสอบว่า มีการย้ายกลุ่มของข้อมูลหรือไม่ ถ้ามีการย้ายกลุ่ม จะให้ทำการคำนวณหาจุดศูนย์กลางของแต่ละกลุ่มใหม่ และทำการคำนวณหาระยะห่างใหม่ในรอบที่ 6 หากไม่มีข้อมูลตัวใดที่มีการย้ายกลุ่มแล้วหมายถึง การจัดกลุ่มชุดข้อมูลเสร็จสิ้นแล้ว และจากการพิจารณาตาราง 2.8 จะเห็นได้ว่า ไม่มีข้อมูลตัวใด ที่มีการย้ายกลุ่ม เพราะฉะนั้นสามารถทำการสรุปผลการจัดกลุ่มข้อมูลทั้งหมด 30 ตัว ซึ่งแสดงได้ดังตาราง 2.9 และเมื่อทำการตรวจสอบความถูกต้องการจัดกลุ่ม โดยใช้ตัววัดมาตราฐาน Purity ตามสมการ (2), NMI ตามสมการ (4), RI ตามสมการ (5) และ F ตามสมการ (6) ได้ดังตาราง 2.7

ตารางที่ 2.7 แสดงค่าจากการวัดประสิทธิภาพ

Purity	NMI	RI	F (β=1)	F (β=2)	F (β=3)	F (β=4)	F (β=5)
0.8667	0.7171	0.7585	0.7647	0.7647	0.7647	0.7647	0.7647

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 2.8 แสดงระยะห่างจากจุดศูนย์กลางแต่ละกลุ่มถึงข้อมูลแต่ละตัว

ข้อมูลที่	Setosa	Virginica	Versicolor
1	2.642	2.446	0.647
2	3.716	0.921	1.165
3	4.608	1.150	2.385
4	3.094	1.585	0.683
5	0.929	4.658	3.187
6	0.654	4.582	3.017
7	2.325	2.551	0.643
8	1.317	4.656	2.760
9	2.760	1.860	0.104
10	0.065	4.492	2.786
11	2.282	2.599	0.701
12	4.998	0.600	2.502
13	0.844	4.600	3.114
14	3.996	0.481	1.493
15	4.135	0.777	1.543

ข้อมูลที่	Setosa	Virginica	Versicolor
16	0.506	4.473	2.661
17	0.555	4.474	2.649
18	0.562	4.543	2.717
19	3.531	1.677	1.061
20	6.454	2.150	4.060
21	4.518	0.542	1.982
22	2.956	1.526	0.503
23	2.440	2.065	0.464
24	0.587	4.076	2.460
25	0.364	4.396	2.755
26	4.470	0.221	2.006
27	4.082	0.479	1.559
28	4.154	0.645	1.617
29	4.154	0.645	1.617
30	2.957	1.538	0.600

ตารางที่ 2.9 แสดงผลการจัดกลุ่มข้อมูลของชุดข้อมูล ทั้งหมด 30 ตัว แสดงด้วยเลขลำดับของข้อมูล

Setosa	Virginica	Versicolor
5	2	1
6	3	4
8	12	7
10	14	9
13	15	11
16	20	19

Setosa	Virginica	Versicolor
17	21	22
18	26	23
24	27	30
25	28	-
-	29	-
-	-	-

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.1.2 ขั้นตอนวิธีการเอนนีไทม์

ขั้นตอนวิธีการเอนนีไทม์ หมายถึง กระบวนการทำงานที่สามารถหยุดได้ระหว่างการทำงาน และสามารถแสดงผลลัพธ์ได้ด้วย ขั้นตอนวิธีการเอนนีไทม์จะสามารถให้ผลลัพธ์บางส่วนได้ซึ่งความถูกต้องที่ได้จะขึ้นอยู่กับจำนวนข้อมูล และระยะเวลาที่ได้ใช้ในการประมวลผล โดยผลลัพธ์ที่ได้จากขั้นตอนวิธีการเอนนีไทม์เป็นการประมาณคำตอบจากผลลัพธ์ที่ถูกต้อง

คุณสมบัติของขั้นตอนวิธีการเอนนีไทม์

1. การวัดคุณภาพ

คุณภาพของผลลัพธ์ที่ถูกคาดการณ์สามารถถูกกำหนดได้อย่างถูกต้อง ตัวอย่างเช่น คุณภาพของผลลัพธ์จะสะท้อนให้เห็นความแตกต่างระหว่างผลลัพธ์ที่คาดการณ์กับผลลัพธ์ที่ถูกต้อง ตราบใดที่มีการกำหนดผลลัพธ์ที่ถูกต้องอยู่แล้ว

2. คุณภาพที่ยอมรับได้

คุณภาพของผลลัพธ์ที่ถูกคาดการณ์จะง่ายต่อการกำหนดขณะทำงาน และคุณภาพของผลลัพธ์จะขึ้นอยู่กับค่าความใกล้เคียงของผลลัพธ์ที่ถูกต้อง

3. ความซ้ำซากจำเจ

คุณภาพของผลลัพธ์เพิ่มขึ้นตามระยะเวลาการทำงาน และคุณภาพของข้อมูลขั้นตอนวิธีการเอนนีไทม์ สามารถลดในส่วนของกระบวนการทำงานให้น้อยลง ใช้เวลาน้อยลง แต่ผลลัพธ์ที่ได้มีค่าความใกล้เคียงกับผลลัพธ์ที่ถูกต้อง และสามารถเป็นที่ยอมรับได้

4. ความสอดคล้อง

ประสิทธิภาพของผลลัพธ์มีความสัมพันธ์กับระยะเวลาในการคำนวณและคุณภาพของข้อมูล โดยทั่วไป ขั้นตอนวิธีการ จะไม่สามารถรับรองประสิทธิภาพของผลลัพธ์ สิ่งสำคัญคือ ต้องมีค่าความแปรปรวนที่แคบเพื่อให้การคาดการณ์ด้านคุณภาพสามารถกระทำได้

5. ลดระยะเวลาการทำงาน

คุณสมบัตินี้ก่อให้เกิดผลดีต่อกระบวนการทำงานของ ขั้นตอนวิธีการ นั่นหมายถึง ถ้าแบ่งข้อมูลออกเป็นหลายๆส่วน เมื่อนำมาเข้ากระบวนการ ผลลัพธ์ที่ได้จะมีค่าใกล้เคียงกับการที่นำข้อมูลมาผ่านกระบวนการเพียงครั้งเดียว โดยไม่มีการหยุดจนกว่าจะเสร็จ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

6. ศักยภาพในการถูกรบกวน

กระบวนการทำงานสามารถถูกหยุดได้ตลอดเวลาแต่ว่าก็ยังคงสามารถกลับมาหาคำตอบต่อได้ ซึ่งนี่ถือว่าเป็นลักษณะเฉพาะของขั้นตอนวิธีการเอนนีไทม์

7. ศักยภาพก่อนกระบวนการหยุดลง

ในกระบวนการทำงานแบบ ขั้นตอนวิธีการเอนนีไทม์ เมื่อเริ่มทำงานกับข้อมูลบางส่วนแล้ว กระบวนการทำงานอาจจะถูกหยุด แต่จะสามารถนำข้อมูลที่เหลือมาทำต่อจากข้อมูลเดิมได้

2.1.3 การประเมินประสิทธิภาพการจัดกลุ่มข้อมูล

2.1.3.1 Purity's Cluster

Zhao และ Karypis (2001) ได้นิยามถึง การประเมินประสิทธิภาพการจัดกลุ่มข้อมูล โดยเรียกว่า เพียวริตี (Purity) สำหรับใช้ในการประเมินการทำงานของการจัดกลุ่มข้อมูล การคำนวณหาค่าเพียวริตีสามารถคำนวณได้จาก

$$purity(\Omega, C) = \frac{1}{N}\sum_k \max_j \left|\omega_k \cap c_j\right| \quad (2)$$

โดยที่ $\Omega = \{\omega_1, \omega_2, \dots, \omega_k\}$ คือ จำนวนข้อมูลแต่ละกลุ่มที่ได้จากการทดลอง

$C = \{c_1, c_2, \dots, c_j\}$ คือ จำนวนข้อมูลจริงในแต่ละกลุ่ม

N คือ จำนวนข้อมูลทั้งหมด

ค่าเพียวริตีจะมีค่าในช่วง 0-1 โดยค่าเพียวริตีจะสูงถึง 1 เมื่อตัวข้อมูลทุกตัวนั้นได้ถูกจัดให้อยู่ถูกต้องตามกลุ่มของมัน และจะมีค่า 0 เมื่อไม่มีข้อมูลตัวใดถูกจัดให้อยู่ตามกลุ่มอย่างถูกต้อง ค่าเพียวริตี้จะมีค่าสูงได้ง่ายเมื่อจำนวนกลุ่มของข้อมูลมีค่าสูงขึ้น ดังนั้นเราไม่สามารถใช้เพียวริตี้ในการวัดประสิทธิภาพ การจัดกลุ่มที่เทียบกับจำนวนกลุ่มของข้อมูลได้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.1.3.2 Normalized Mutual Information (NMI)

NMI เป็นตัววัดประสิทธิภาพที่สามารถทำการวัดประสิทธิภาพการจัดกลุ่ม ที่ขึ้นกับจำนวนของข้อมูลได้ การคำนวณหาค่า NMI สามารถคำนวณได้จาก

$$\mathrm{NMI}(\Omega;C) = \frac{\mathrm{I}(\Omega;C)}{[\mathrm{H}(\Omega) + \mathrm{H}(C)]/2} \tag{3}$$

โดยที่

$$I(\Omega; C) = \sum_k \sum_j \frac{|\omega_k \cap c_j|}{N} \log \frac{N|\omega_k \cap c_j|}{|\omega_k||c_j|} \tag{4}$$

$$H(\Omega) = -\sum_k \frac{|\omega_k|}{N} \log \frac{|\omega_k|}{N} \tag{5}$$

$$H(\mathrm{C}) = -\sum_j \frac{|c_j|}{N} \log \frac{|c_j|}{N} \tag{6}$$

P (ω_k), P (c_j) และ P $(\omega_k \cap c_j)$ คือ ความจะน่าเป็นของข้อมูลจากการจัดกลุ่ม, ความน่าจะเป็นของข้อมูลจริงและค่าความน่าจะเป็นของข้อมูล 2 คลาสที่เหมือนกัน ตามลำดับ ค่า NMI จะมีค่าในช่วง 0-1 นั่นคือ ในกรณีที่การแบ่งกลุ่มแย่ที่สุดจะมีค่าเป็น 0 หากการแบ่งกลุ่มดีที่สุดจะมีค่าเป็น 1

2.1.3.3 The Rand Index (RI)

RI เป็นทางเลือกอีกทางในการวัดประสิทธิภาพข้อมูลแต่ละกลุ่ม คือ การพิจารณาความคล้ายคลึงกันของข้อมูล แบ่งเป็น 4 ลักษณะ ได้แก่ ข้อมูลคล้ายกันอยู่กลุ่มเดียวกัน (TP), ข้อมูลไม่คล้ายกันอยู่คนละกลุ่มกัน (TN), ข้อมูลไม่คล้ายกันอยู่กลุ่มเดียวกัน (FP) และข้อมูลคล้ายกันอยู่คนละกลุ่ม (FN) มาใช้ในการวัดประสิทธิภาพ ดังสมการ ซึ่ง RI ถือว่าเป็นการวัดค่าความแม่นยำที่ง่าย การคำนวณหาค่า RI สามารถคำนวณได้จาก

$$RI = \frac{\mathrm{TP+TN}}{\mathrm{TP+FP+FN+TN}} \tag{7}$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

โดย ตัววัดประสิทธิภาพนี้มีการถ่วงน้ำหนัก FP และ FN เท่ากัน แต่ในบางกรณีการแบ่งกลุ่มโดยข้อมูลคล้ายกันอยู่กลุ่มเดียวกัน ผลที่ได้อาจแย่กว่าการที่ข้อมูลไม่คล้ายกันจะอยู่กลุ่มเดียวกัน ค่า RI จะมีค่าในช่วง 0-1 นั่นคือ กรณีที่การแบ่งกลุ่มแย่ที่สุดจะมีค่าเป็น 0 หากการแบ่งกลุ่มดีที่สุดจะมีค่าเป็น 1

2.1.3.4 F

F เป็นตัววัดประสิทธิภาพที่ช่วยลดข้อบกพร่องและข้อจำกัดของตัววัดประสิทธิภาพตัวอื่น โดย F เป็นการพิจารณาให้ FN มีความสำคัญมากกว่า FP โดยเลือกใช้ค่า $\beta > 1$ เป็นการถ่วงน้ำหนัก การคำนวณหาค่า F สามารถคำนวณได้จาก

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P+R} \tag{8}$$

โดยที่

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{9}$$

$$R = \frac{TP}{TP+FN} \tag{10}$$

การวัดประสิทธิภาพด้วย F เป็นการวัดประสิทธิภาพที่ดีที่สุดและมีความเหมาะสมในการใช้งานกับงานวิจัย ค่า F จะมีค่าในช่วง 0-1 นั่นคือ ในกรณีที่การแบ่งกลุ่มแย่ที่สุดจะมีค่าเป็น 0 หากการแบ่งกลุ่มดีที่สุดจะมีค่าเป็น 1

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 3
ขั้นตอนวิธีการเอนนีไทม์เคมีนสำหรับการจัดกลุ่ม

ในบทนี้ได้นำเสนอขั้นตอนวิธีการเอนนีไทม์เคมีนสำหรับการจัดกลุ่มโดยได้นำขั้นตอนวิธีการเคมีน มาปรับแต่งกับทฤษฎีเอนนีไทม์ จากนั้นทำการเปรียบเทียบผลการทดลองที่ได้จากการจัดกลุ่มโดยใช้ ขั้นตอนวิธีการเคมีนกับขั้นตอนวิธีการเอนนีไทม์เคมีนและมีการใช้ตัววัดประสิทธิภาพการจัดกลุ่มข้อมูลเพื่อวัดผลการทดลอง ซึ่งชุดข้อมูลที่ใช้ในการศึกษาทดลองคือ ชุดข้อมูล Iris Plants จำนวน 20 ตัว

3.1 ขั้นตอนวิธีการเอนนีไทม์เคมีน

ขั้นตอนวิธีการเอนนีไทม์เคมีนจะสามารถหาผลลัพธ์ตามกระบวนการทำงานของขั้นตอนวิธีการ โดยไม่จำเป็นที่จะต้องหาจากการทำงาน หรือใส่ข้อมูลที่มีเข้าไปทั้งหมด ขั้นตอนวิธีการเอนนีไทม์เคมีนจะสามารถแบ่งกลุ่มข้อมูลที่มีและนำข้อมูลทีละกลุ่มเข้ากระบวนการ ผลลัพธ์ที่ได้จาก ขั้นตอนวิธีการเอนนีไทม์เคมีนจะมีค่าใกล้เคียงกับผลลัพธ์จาก ขั้นตอนวิธีการเคมีนซึ่งตรงกับขั้นตอนวิธีการเอนนีไทม์ ที่มีคุณสมบัติตรงตามศักยภาพของกระบวนการ ก่อนกระบวนการทำงานจะหยุดลง โดยเมื่อขั้นตอนวิธีการเอนนีไทม์เริ่มทำงานกับข้อมูลบางส่วนแล้ว เมื่อกระบวนการถูกรบกวนแล้วการทำงานจะหยุดลง และสามารถนำข้อมูลที่เหลือมาทำต่อจากข้อมูลเดิมได้ และขั้นตอนวิธีการเอนนีไทม์เคมีนยังตรงกับ ขั้นตอนวิธีการเอนนีไทม์โดยที่ ขั้นตอนวิธีการเอนนีไทม์จะมีคุณสมบัติในการลดความซ้ำซากจำเจของการทำงาน โดยผลลัพธ์ที่ได้หลังจากการทำงานแต่ละกลุ่ม จะมีค่าใกล้เคียงกับผลลัพธ์สุดท้าย และยังมีคุณสมบัติของการลดระยะเวลาในการทำงานได้อีกด้วย ประสิทธิภาพของผลลัพธ์นั้นมีความสัมพันธ์กับจำนวนครั้งหรือระยะเวลาในการคำนวณ โดยถ้ามีจำนวนครั้งหรือระยะเวลาในการคำนวณมาก ผลลัพธ์ก็จะมีประสิทธิภาพมาก เทียบได้กับการทำงานของขั้นตอนวิธีการเอนนีไทม์เคมีนเมื่อเริ่มการทำงานกับแต่ละกลุ่มไปเรื่อยๆ ผลลัพธ์ที่ได้ก็จะเข้าใกล้ ผลลัพธ์ที่ถูกต้องเช่นกัน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.2 ขั้นตอนวิธีการจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีการเอนนีไทม์เคมีน

วิธีการจัดกลุ่มโดยใช้ขั้นตอนวิธีการเอนนีไทม์เคมีน มีขั้นตอนดังต่อไปนี้

ขั้นตอนที่ 1 : กำหนดจำนวนกลุ่มที่ต้องการจัดกลุ่ม โดยให้ k แทนจำนวนกลุ่ม

ขั้นตอนที่ 2 : แบ่งข้อมูลออกเป็นส่วนๆ กำหนดจำนวนข้อมูลที่ต้องการทำ และแบ่งกลุ่มข้อมูลนั้นโดยสุ่มจากข้อมูลที่ต้องการจัดกลุ่ม ตามจำนวนกลุ่มที่กำหนด

ขั้นตอนที่ 3 : คำนวณหาจุดศูนย์กลางแต่ละกลุ่ม

ขั้นตอนที่ 4 : คำนวณหาระยะห่างจากทุกจุดข้อมูลไปยังจุดศูนย์กลางแต่ละกลุ่ม ด้วยวิธี Euclidean distance ตามสมการที่ (1)

ขั้นตอนที่ 5 : ตรวจสอบเงื่อนไขผลจากการคำนวณหาระยะห่างว่าจุดข้อมูลอยู่ในตำแหน่งที่ใกล้กับจุดศูนย์กลางกลุ่มอื่นหรือไม่

ถ้า จริง; ให้ย้ายข้อมูลนั้นไปยังกลุ่มที่มีระยะห่างน้อยที่สุด และวนกลับไปทำขั้นตอนที่ 3 ใหม่อีกครั้ง

ถ้า เท็จ; ให้หยุดการวนรอบ

ขั้นตอนที่ 6 : บันทึกจุดศูนย์กลางของแต่ละกลุ่มเป็นตัวแทนของการจัดกลุ่ม เพื่อนำไปใช้ในการจัดกลุ่มกับข้อมูลส่วนต่อไป

<u>ตัวอย่าง 3.1</u>

จะใช้ชุดข้อมูล Iris มีทั้งหมด 30 ตัว แต่ละชุดแบ่งทำเป็น 3 ส่วน คือ 15, 10, 5 ตามลำดับ กำหนดให้ค่า k = 3

<u>ส่วนที่ 1</u> เริ่มต้นที่ข้อมูลจำนวน 15 ตัว ดังตารางที่ 3.1

<u>รอบที่ 1</u> แต่ละจุดข้อมูลถูกจัดเป็นกลุ่ม (จากการสุ่ม)

คำนวณหาค่าจุดศูนย์กลางแต่ละกลุ่ม แสดงได้ดังตาราง 3.2 และคำนวณระยะห่างจากจุดศูนย์กลางแต่ละกลุ่มถึงข้อมูลแต่ละตัว แสดงดังตาราง 3.3

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สำนักหอสมุดกลาง พระจอมเกล้าลาดกระบัง

ตารางที่ 3.1 แสดงคุณลักษณะของชุดข้อมูล

ข้อมูลตัวที่	Sepal Length	Sepal Width	Petal Length	Petal Width	Species
1	5	2	3.5	1	versicolor
2	6.2	2.2	4.5	1.5	versicolor
3	6	2.2	5	1.5	virginica
4	6	2.2	4	1	versicolor
5	5.8	4	1.2	0.2	setosa
6	5.2	4.1	1.5	0.1	setosa
7	5	2.3	3.3	1	versicolor
8	4.5	2.3	1.3	0.3	setosa
9	5.5	2.4	3.8	1.1	versicolor
10	5.1	3.5	1.4	0.2	setosa
11	4.9	2.4	3.3	1	versicolor
12	6.7	2.5	5.8	1.8	virginica
13	5.5	4.2	1.4	0.2	setosa
14	6.3	2.5	4.9	1.5	versicolor
15	5.7	2.5	5	2	virginica

ตารางที่ 3.2 แสดงจุดศูนย์กลางแต่ละกลุ่ม

จุดศูนย์กลาง	Sepal Length	Sepal Width	Petal Length	Petal Width
Setosa	5.220	3.620	1.360	0.200
Virginica	6.600	2.400	5.267	1.767
Versicolor	5.557	2.286	3.900	1.157

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

117236

ตารางที่ 3.3 แสดงระยะห่างจากจุดศูนย์กลางแต่ละกลุ่มถึงข้อมูลแต่ละตัว

ข้อมูลตัวที่	Setosa	Virginica	Versicolor
1	2.809	2.536	0.759
2	3.811	0.927	0.948
3	4.659	0.907	2.175
4	3.199	1.610	0.488
5	0.712	4.711	3.347
6	0.510	4.671	3.209
7	2.489	2.651	0.834
8	1.508	4.723	2.935
9	2.886	1.951	0.172
10	0.174	4.568	2.975
11	2.448	2.710	0.911
12	5.071	0.553	2.318
13	0.645	4.675	3.291
14	4.080	0.553	1.310
15	4.240	0.972	1.410

ทำการตรวจสอบว่า มีการย้ายกลุ่มของข้อมูลหรือไม่ ถ้ามีการย้ายกลุ่ม จะวนทำใหม่ในรอบที่ 2 แต่หากไม่มีการย้ายกลุ่มข้อมูลเกิดขึ้น นั่นคือ การจัดกลุ่มเสร็จสิ้นแล้ว และจากการพิจารณาตารางที่ 3.3 จะเห็นได้ว่า มีข้อมูลจำนวน 2 ตัวที่มีการย้ายกลุ่ม เพราะฉะนั้นจะทำการย้ายกลุ่มและทำการวนทำรอบที่ 2

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

<u>รอบที่ 2</u> มีข้อมูล 2 ตัวถูกย้ายจากรอบที่ 1 คำนวณค่าจุดศูนย์กลางแต่ละกลุ่มแสดงดังตาราง 3.4

ตารางที่ 3.4 แสดงจุดศูนย์กลางแต่ละกลุ่ม

จุด ศูนย์กลาง	Sepal Length	Sepal Width	Petal Length	Petal Width
Setosa	5.220	3.620	1.360	0.200
Virginica	6.460	2.380	5.040	1.660
Versicolor	5.280	2.260	3.580	1.020

ตารางที่ 3.5 แสดงระยะห่างจากจุดศูนย์กลางแต่ละกลุ่มถึงข้อมูลแต่ละตัว

ข้อมูลตัวที่	Setosa	Virginica	Versicolor
1	2.809	2.255	0.391
2	3.811	0.646	1.388
3	4.659	0.971	2.597
4	3.199	1.327	0.836
5	0.712	4.465	3.104
6	0.510	4.417	2.927
7	2.489	2.367	0.398
8	1.508	4.437	2.515
9	2.886	1.665	0.350
10	0.174	4.299	2.645
11	2.448	2.428	0.493
12	5.071	0.818	2.759
13	0.645	4.429	3.039
14	4.080	0.292	1.752
15	4.240	0.842	1.792

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ทำการตรวจสอบว่ามีการย้ายกลุ่มของข้อมูลหรือไม่ ถ้ามีการย้ายกลุ่ม จะทำวนทำใหม่ในรอบที่ 2 หากไม่มีการย้ายกลุ่ม นั่นคือ การจัดกลุ่มเสร็จสิ้นแล้ว จากการพิจารณาตาราง 3.5 จะเห็นได้ว่า ไม่มีข้อมูลใดที่มีการย้ายกลุ่ม เพราะฉะนั้นสามารถจบการจัดกลุ่มส่วนที่ 1 ได้ และจะบันทึกจุดศูนย์กลางของแต่ละกลุ่มไว้เพื่อใช้ในการคำนวณส่วนต่อไป โดยผลจากการจัดกลุ่มข้อมูลในส่วนที่ 1 แสดงดังตารางที่ 3.6

ตารางที่ 3.6 แสดงผลการจัดกลุ่มข้อมูลของชุดข้อมูล 15 ตัว ซึ่งแสดงผลด้วย ตัวเลขลำดับของข้อมูล

Setosa	Virginica	Versicolor
5	2	1
6	3	4
8	12	7
10	14	9
13	15	11

<u>ส่วนที่ 2</u> จำนวนข้อมูลทั้งหมด 13 ตัว ประกอบด้วยจุดศูนย์กลางของแต่ละกลุ่มจำนวน 3 กลุ่ม และจำนวนข้อมูล 10 ตัว ดังตาราง 3.7

<u>รอบที่ 1</u> แต่ละจุดข้อมูลถูกจัดเป็นกลุ่ม (จากการสุ่ม)

ตารางที่ 3.7 แสดงคุณลักษณะของชุดข้อมูล

ข้อมูลตัวที่	Sepal Length	Sepal Width	Petal Length	Petal Width	Species
16	4.9	3	1.4	0.2	setosa
17	4.8	3	1.4	0.3	setosa
18	4.8	3	1.4	0.1	setosa
19	4.9	2.5	4.5	1.7	virginica
20	7.7	2.6	6.9	2.3	virginica
21	6.1	2.6	5.6	1.4	virginica
22	5.8	2.6	4	1.2	versicolor
23	5.7	2.6	3.5	1	versicolor
24	5.1	3.8	1.9	0.4	setosa
25	5.1	3.8	1.6	0.2	setosa

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เมื่อคำนวณหาค่าจุดศูนย์กลางแต่ละกลุ่ม โดยนำจุดศูนย์กลางของทั้ง 3 กลุ่มมาทำการคำนวณร่วมด้วย จะได้ผลการคำนวณแสดงได้ดังตาราง 3.8 และคำนวณระยะห่างจากจุดศูนย์กลางแต่ละกลุ่มถึงข้อมูลแต่ละตัว แสดงดังตาราง 3.9

ตารางที่ 3.8 แสดงจุดศูนย์กลางแต่ละกลุ่ม

จุด ศูนย์กลาง	Sepal Length	Sepal Width	Petal Length	Petal Width
Setosa	4.987	3.370	1.510	0.233
Virginica	6.290	2.520	5.510	1.765
Versicolor	5.593	2.487	3.693	1.073

ตารางที่ 3.9 แสดงระยะห่างจากจุดศูนย์กลางแต่ละกลุ่มถึงข้อมูลแต่ละตัว

ข้อมูลตัวที่	Setosa	Virginica	Versicolor
16	0.397	4.637	2.601
17	0.434	4.636	2.598
18	0.449	4.703	2.665
19	3.443	1.720	1.235
20	6.425	2.053	4.030
21	4.463	0.429	2.003
22	2.896	1.687	0.407
23	2.377	2.232	0.259
24	0.615	4.237	2.374
25	0.455	4.560	2.667
กลุ่ม 1	0.375	4.693	2.762
กลุ่ม 2	4.201	0.530	1.709
กลุ่ม 3	2.494	2.317	0.407

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ทำการตรวจสอบว่า มีการย้ายกลุ่มของข้อมูลหรือไม่ ถ้ามีการย้ายกลุ่ม จะวนทำใหม่ในรอบที่ 2 แต่หากไม่มีการย้ายกลุ่มข้อมูลเกิดขึ้น นั่นคือ การจัดกลุ่มเสร็จสิ้นแล้ว และจากการพิจารณาตารางที่ 3.9 จะเห็นได้ว่า มีข้อมูลจำนวน 1 ตัวที่มีการย้ายกลุ่ม เพราะฉะนั้นจะทำการย้ายกลุ่มและทำการวนทำรอบที่ 2

รอบที่ 2 มีข้อมูล1 ตัวถูกย้ายจากรอบที่ 1 คำนวณค่าจุดศูนย์กลางแต่ละกลุ่มแสดงดังตาราง 3.10

ตารางที่ 3.10 แสดงจุดศูนย์กลางแต่ละกลุ่ม

จุด ศูนย์กลาง	Sepal Length	Sepal Width	Petal Length	Petal Width
Setosa	4.987	3.370	1.510	0.233
Virginica	6.753	2.527	5.847	1.787
Versicolor	5.420	2.490	3.895	1.230

ตารางที่ 3.11 แสดงระยะห่างจากจุดศูนย์กลางแต่ละกลุ่มถึงข้อมูลแต่ละตัว

ข้อมูลตัวที่	**Setosa**	**Virginica**	**Versicolor**
16	0.397	5.094	2.796
17	0.434	5.101	2.781
18	0.449	5.163	2.854
19	3.443	2.293	0.926
20	6.425	1.508	3.922
21	4.463	0.802	1.847
22	2.896	2.161	0.410
23	2.377	2.691	0.547
24	0.615	4.675	2.547
25	0.455	4.991	2.854
กลุ่ม 1	0.375	5.118	2.967
กลุ่ม 2	4.201	0.880	1.609
กลุ่ม 3	2.494	2.823	0.465

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ทำการตรวจสอบว่ามีการย้ายกลุ่มของข้อมูลหรือไม่ ถ้ามีการย้ายกลุ่ม จะทำการคำนวณหาจุดศูนย์กลางกลุ่มใหม่ นั่นคือวนทำรอบที่ 2 จากการพิจารณาตารางที่ 3.11 จะเห็นได้ว่า ไม่มีข้อมูลใดที่มีการย้ายกลุ่ม เพราะฉะนั้นสามารถนำข้อมูลในส่วนที่ 3 มาทำต่อได้ และจะบันทึกจุดศูนย์กลางของแต่ละกลุ่มไว้เพื่อใช้ในการคำนวณส่วนต่อไป โดยผลจากการจัดกลุ่มข้อมูลในส่วนที่ 2 แสดงดังตารางที่ 3.12

ตารางที่ 3.12 แสดงผลการจัดกลุ่มข้อมูลของชุดข้อมูล 10 ตัว แสดงด้วยเลขลำดับของข้อมูล

Setosa	Virginica	Versicolor
16	20	19
17	21	22
18	-	23
24	-	-
25	-	-

ส่วนที่ 3 มีข้อมูลจำนวน 8 ตัว ประกอบด้วย จุดศูนย์กลางของแต่ละกลุ่มจำนวน 3 กลุ่ม และจำนวนข้อมูล 5 ตัว ดังตารางที่ 3.13

รอบที่ 1 แต่ละจุดข้อมูลถูกจัดเป็นกลุ่ม (จากการสุ่ม)

ตารางที่ 3.13 แสดงคุณลักษณะของชุดข้อมูล

ข้อมูลตัวที่	Sepal Length	Sepal Width	Petal Length	Petal Width	Species
26	6.4	2.7	5.3	1.9	virginica
27	6	2.7	5.1	1.6	versicolor
28	5.8	2.7	5.1	1.9	virginica
29	5.8	2.7	5.1	1.9	virginica
30	5.8	2.7	4.1	1	versicolor

คำนวณหาค่าจุดศูนย์กลางแต่ละกลุ่ม แสดงได้ดังตาราง 3.14 และคำนวณระยะห่างจากจุดศูนย์กลางแต่ละกลุ่มถึงข้อมูลแต่ละตัว แสดงดังตาราง 3.15

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.14 แสดงจุดศูนย์กลางแต่ละกลุ่ม

จุด ศูนย์กลาง	Sepal Length	Sepal Width	Petal Length	Petal Width
Setosa	4.987	3.370	1.510	0.233
Virginica	6.188	2.657	5.337	1.872
Versicolor	5.740	2.630	4.365	1.277

ตารางที่ 3.15 แสดงระยะห่างจากจุดศูนย์กลางแต่ละกลุ่มถึงข้อมูลแต่ละตัว

ข้อมูลตัวที่	Setosa	Virginica	Versicolor
26	4.426	0.221	1.305
27	4.029	0.409	0.847
28	4.096	0.458	0.968
29	4.096	0.458	0.968
30	2.899	1.563	0.394
กลุ่ม 1	0.000	4.391	3.218
กลุ่ม 2	5.005	0.777	1.869
กลุ่ม 3	2.765	1.763	0.587

ทำการตรวจสอบว่า มีการย้ายกลุ่มของข้อมูลหรือไม่ ถ้ามีการย้ายกลุ่ม จะให้ทำการคำนวณหาจุดศูนย์กลางของแต่ละกลุ่มใหม่ และทำการคำนวณหาระยะห่างใหม่ในรอบที่ 2 จากการพิจารณาตารางที่ 3.15 จะเห็นได้ว่า มีข้อมูลจำนวน 1 ตัวที่มีการย้ายกลุ่ม เพราะฉะนั้นจะทำการย้ายกลุ่มและทำการวนทำรอบที่ 2

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

<u>รอบที่ 2</u> มีข้อมูล 1 ตัวถูกย้ายจากรอบที่ 1 คำนวณค่าจุดศูนย์กลางแต่ละกลุ่มแสดงดังตาราง 3.16

ตารางที่ 3.16 แสดงจุดศูนย์กลางแต่ละกลุ่ม

จุด ศูนย์กลาง	Sepal Length	Sepal Width	Petal Length	Petal Width
Setosa	4.987	3.370	1.510	0.233
Virginica	6.151	2.665	5.289	1.817
Versicolor	5.610	2.595	3.997	1.115

ตารางที่ 3.17 แสดงระยะห่างจากจุดศูนย์กลางแต่ละกลุ่มถึงข้อมูลแต่ละตัว

ข้อมูลตัวที่	Setosa	Virginica	Versicolor
26	4.426	0.265	1.717
27	4.029	0.327	1.270
28	4.096	0.408	1.371
29	4.096	0.408	1.371
30	2.899	1.486	0.266
กลุ่ม 1	0.000	4.318	2.820
กลุ่ม 2	5.005	0.833	2.276
กลุ่ม 3	2.765	1.689	0.266

ทำการตรวจสอบว่า มีการย้ายกลุ่มของข้อมูลหรือไม่ ถ้ามีการย้ายกลุ่ม จะให้ทำการคำนวณหาจุดศูนย์กลางของแต่ละกลุ่มใหม่ และทำการคำนวณหาระยะห่างใหม่ในรอบที่ 3 จากการพิจารณาตารางที่ 3.15 จะเห็นได้ว่า ไม่มีข้อมูลตัวใดที่มีการย้ายกลุ่ม เพราะฉะนั้นจะทำการสรุปผลการจัดกลุ่มข้อมูลในรอบที่ 3 แสดงดังตารางที่ 3.18

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.18 แสดงผลการจัดกลุ่มข้อมูลของชุดข้อมูล 5 ตัว ซึ่งแสดงผลด้วย ตัวเลขลำดับของข้อมูล

Setosa	Virginica	Versicolor
-	26	30
-	27	-
-	28	-
-	29	-

ตรวจสอบความถูกต้องการจัดกลุ่มทั้งหมด โดยใช้ตัววัดมาตราฐาน Purity ตามสมการ (2), NMI ตามสมการ (4), RI ตามสมการ (5) และ F ตามสมการ (6) ได้ดังตาราง 3.19

ตารางที่ 3.19 แสดงค่าจากการวัดประสิทธิภาพ

Purity	NMI	RI	F (β=1)	F (β=2)	F (β=3)	F (β=4)	F (β=5)
0.8670	0.7502	0.8261	0.7607	0.7646	0.7667	0.7676	0.7681

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 4
ผลการทดลอง

ปัญหาพิเศษนี้ได้ทำการทดลองการจัดกลุ่มข้อมูลจากชุดข้อมูลทั้งหมด 4 ชุด คือ ชุดข้อมูลที่ชื่อว่า Iris Plants, Balance Scale, Haberman's Survival และ Hayes - Roth เพื่อวิเคราะห์ผลการเปรียบเทียบค่าประสิทธิภาพในการจัดกลุ่มจากการทดลองโดยใช้ขั้นตอนวิธีการเคมีนและขั้นตอนวิธีการเอนนีไทม์เคมีน ซึ่งการทดลองนี้จะแบ่งส่วนข้อมูลออกเป็น 4 ลักษณะ คือ แบ่งเท่ากันทุกส่วน, แบ่งจากน้อยไปมาก, แบ่งจากมากไปน้อย และแบ่งแบบผสม คือ การนำ 3 ลักษณะแรกมาประยุกต์รวมกัน โดยผลการทดลองจากขั้นตอนวิธีการเอนนีไทม์เคมีน เมื่อนำมาทำการวัดประสิทธิภาพ จะใช้ค่าเฉลี่ยจากทุกลักษณะมาเปรียบเทียบกับผลการวัดประสิทธิภาพของขั้นตอนวิธีการเคมีน

4.1 ชุดข้อมูลที่ใช้

ชุดข้อมูลที่ใช้ในการทดสอบมีจำนวน 4 ชุดข้อมูล คือ ชุดข้อมูล Iris Plants, Balance Scale, Haberman's Survival และ Hayes - Roth ซึ่งคุณลักษณะเฉพาะของชุดข้อมูลชุดเหล่านี้ มีรายละเอียดดังต่อไปนี้

4.1.1 ชุดข้อมูล Iris Plants

ชุดข้อมูล Iris Plants เป็นชุดข้อมูลที่มีข้อมูลทั้งหมด 150 รายการ แบ่งเป็น 3 กลุ่ม กลุ่มละ 50 รายการ แต่ละกลุ่มคือ ประเภทของพันธุ์พืช Iris ชุดข้อมูลนี้นำมาจากเว็บไซต์ UCI Repository –http://archive.ics.uci.edu/ml/datasets/Iris คุณลักษณะของแอททริบิวต์เป็นประเภทจำนวนจริงและไม่ปรากฏข้อมูลสูญหาย รายละเอียดของข้อมูลประกอบด้วย

- ความยาวกลีบเลี้ยง (sepal length) หน่วยเป็น ซ.ม.
- ความกว้างกลีบเลี้ยง (Sepal width) หน่วยเป็น ซ.ม.
- ความยาวกลีบดอก (Petal length) หน่วยเป็น ซ.ม.
- ความกว้างกลีบดอก (Petal width) หน่วยเป็น ซ.ม.
- กลุ่ม มี 3 กลุ่ม คือ Iris Setosa, Iris Versicolour และ Iris Virginia

ผู้ให้ข้อมูลคือ Michael Marshall (MARSHALL%PLU@io.arc.na sa.gov) ให้เมื่อ ก.ค. 2531

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.1.2 ชุดข้อมูล Balance Scale

ชุดข้อมูล Balance Scale เป็นชุดข้อมูลที่มีข้อมูลทั้งหมด 650 รายการ แบ่งเป็น 3 กลุ่ม กลุ่ม 1 มี 288 รายการ กลุ่ม 2 มี 49 รายการและกลุ่ม 3 มี 288 รายการ ผลลัพธ์ของข้อมูลชุดนี้พิจารณาจากการนำเอาน้ำหนักด้านซ้ายคูณกับระยะห่างด้านซ้าย และน้ำหนักด้านขวาคูณกับระยะห่างด้านขวา มาทำการเปรียบเทียบกัน หากทางด้านซ้ายมากกว่าทางด้านขวา ผลคือ เอียงซ้าย แต่ถ้าหากว่าทางด้านขวามากกว่าทางซ้าย ผลคือ เอียงขวา หรือถ้าทั้งสองด้านเท่ากัน ผลคือข้อมูลนั้นมีความสมดุล ซึ่งข้อมูลชุดนี้คุณลักษณะของแอททริบิวต์เป็นประเภทจำนวนจริงและไม่ปรากฏข้อมูลสูญหาย รายละเอียดของข้อมูลประกอบด้วย

- น้ำหนักด้านซ้าย (Left –Weight) เป็นตัวเลขจำนวนเต็ม 1-5
- ระยะห่างด้านซ้าย (Left –Distance) เป็นตัวเลขจำนวนเต็ม 1-5
- น้ำหนักด้านขวา (Right –Weight) เป็นตัวเลขจำนวนเต็ม 1-5
- ระยะห่างด้านขวา (Right –Distance) เป็นตัวเลขจำนวนเต็ม 1-5
- กลุ่ม มี 3 กลุ่ม คือ L,B,R

ผู้ให้ข้อมูลคือ Tim Hume (hume@ics.uci.edu) ให้เมื่อ เม.ย. 2538

4.1.3 ชุดข้อมูล Haberman's Survival

ชุดข้อมูล Haberman's Survival เป็นชุดข้อมูลที่มีข้อมูลทั้งหมด 306 รายการ แบ่งเป็น 2 กลุ่ม กลุ่มหนึ่งคือ กลุ่มของผู้ป่วยที่มีชีวิตรอดมากกว่า 5 ปี มีจำนวน 225 คน และอีกกลุ่มหนึ่งคือ กลุ่มของผู้ป่วยที่เสียชีวิตภายใน 5 ปี มีจำนวน 81 คน ซึ่งข้อมูลชุดนี้คุณลักษณะของแอททริบิวต์เป็นประเภทจำนวนจริงและไม่ปรากฏข้อมูลสูญหาย รายละเอียดของข้อมูลประกอบด้วย

- อายุของผู้ป่วย เป็นตัวเลขจำนวนเต็มในช่วง [30, 83]
- ปีที่ผู้ป่วยเข้ารับการรักษา เป็นตัวเลขจำนวนเต็มอยู่ในช่วง [58, 69]
- จำนวนจุดเชื่อมของเครื่องช่วย เป็นตัวเลขจำนวนเต็มอยู่ในช่วง [0, 52]
- สถานะการอยู่รอด แบ่งเป็น 2 กลุ่ม คือ
  - ผู้ป่วยที่มีชีวิตรอดมากกว่า 5 ปี
  - ผู้ป่วยที่เสียชีวิตภายใน 5 ปี

ผู้ให้ข้อมูลคือ Tjen-Sien Lim (limt@stat.wisc.edu) ให้เมื่อ มี.ค. 2543

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.1.4 ชุดข้อมูล Hayes - Roth

ชุดข้อมูล Hayes - Roth เป็นชุดข้อมูลที่มีข้อมูลทั้งหมด 160 รายการ แบ่งเป็น 3 กลุ่ม กลุ่มละ 51 รายการ, 51 รายการ และ 30 รายการ มีแอททริบิวต์ 5 แอททริบิวต์ โดยชุดข้อมูลชุดนี้มีคุณลักษณะของแอททริบิวต์เป็นประเภทจำนวนจริงและไม่ปรากฏข้อมูลสูญหาย รายละเอียดของข้อมูลประกอบด้วย

- ชื่อ (Name) เป็นตัวเลขจำนวนเต็ม
- งานอดิเรก (Hobby) เป็นตัวเลขจำนวนเต็มเต็มที่ระบุค่า 1-3
- อายุ (Age) เป็นตัวเลขจำนวนเต็มที่ระบุค่า 1-3
- สถานภาพทางการศึกษา (Educational level) เป็นตัวเลขจำนวนเต็มที่ระบุค่า 1-4
- สถานภาพสมรส (Marital status) เป็นตัวเลขจำนวนเต็มที่ระบุค่า 1-4

ผู้ให้ข้อมูลคือ David W. Aha (aha@ics.uci.edu) ให้เมื่อ มี.ค 2532

4.2 เครื่องมือที่ใช้ในการทดลอง

ในการศึกษาทดลองขั้นตอนวิธีครั้งนี้ใช้โปรแกรม Netbean IDE 6.9.1 และ Microsoft Excel 2007 โดยนำโปรแกรมทั้งสองมาใช้ในการประมวลผลขั้นตอนวิธี การประมวลผลเชิงตัวเลข และสร้างกราฟให้สะดวกต่อการใช้ในการทำการทดลองและสรุปผลการทดลอง

4.3 วิธีการประเมินประสิทธิภาพการทดลอง

ในการวัดประสิทธิภาพของการจัดกลุ่มข้อมูล จะใช้วิธีการนำผลที่ได้จากการจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีการเคมีนและขั้นตอนวิธีการเอนนีไทม์เคมีนทั้งหมดมาทำการคำนวณหาค่าคลัสเตอร์เพียวริตี (Cluster's Purity), ค่า NMI, ค่า RI และค่า F ที่ค่า β ต่างๆ เพื่อประเมินการจัดข้อมูลว่าจัดกลุ่มได้ถูกกลุ่ม

การทดสอบจะทำการวัดประสิทธิภาพของการจัดกลุ่มข้อมูล โดยใช้ขั้นตอนวิธีการเคมีนเปรียบ-เทียบกับการจัดกลุ่มโดยใช้ขั้นตอนวิธีการเอนนีไทม์เคมีน ซึ่งขั้นตอนวิธีการเอนนีไทม์เคมีนที่ทำการแบ่งข้อมูลออกเป็นส่วนๆจะทำการทดลองหลากหลายกรณี เพื่อวัดประสิทธิภาพการจัดกลุ่มว่าการจัดกลุ่มแบบใดให้ค่าความแม่นยำมากที่สุด

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

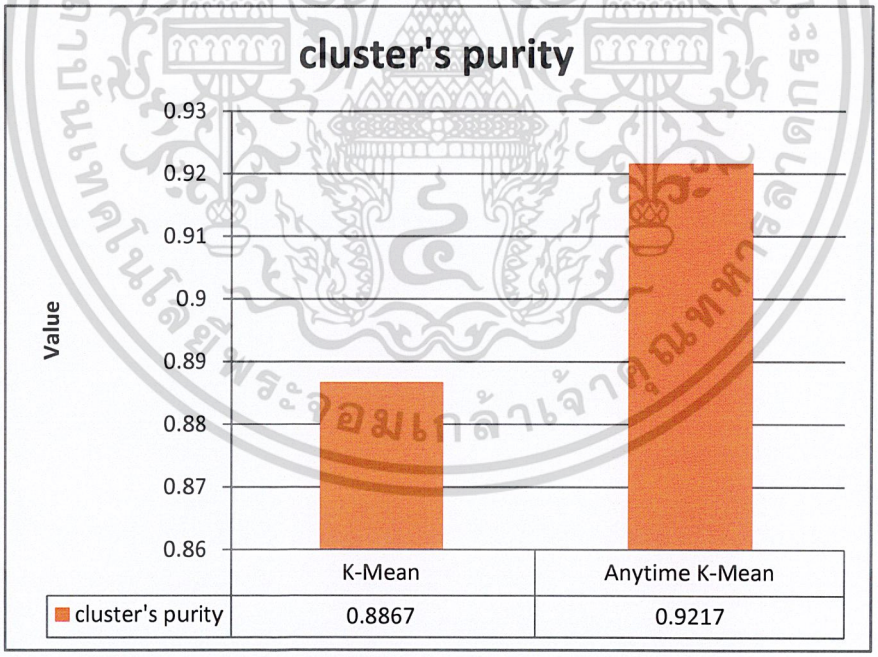
4.4 ตัววัดประสิทธิภาพ

ในการวัดประสิทธิภาพของการจัดกลุ่มข้อมูล จะใช้การคำนวณหาค่าเพียวริตี (Cluster's Purity) จากสมการที่ (2), ค่า NMI จากสมการที่ (3), ค่า RI จากสมการที่ (4) และค่า F ที่ค่า β ต่างๆ จากสมการที่ (5) ในการประเมินประสิทธิภาพการจัดกลุ่มข้อมูล

4.5 ผลการทดลอง

เมื่อทำการทดลองการจัดกลุ่มด้วยขั้นตอนวิธีการเคมีนและขั้นตอนวิธีการเอนนีไทม์เคมีน ซึ่งมีการแบ่งข้อมูลออกเป็นส่วนๆ โดยมีจำนวนการแบ่งที่แตกต่างกันแล้ว จากนั้นทำการวัดประสิทธิภาพโดยใช้ค่าคลัสเตอร์เพียวริตี โดยผลที่ได้จากการแบ่งข้อมูลออกเป็นส่วนๆ โดยมีจำนวนการแบ่งที่แตกต่างกันนั้น ได้ผลการวัดประสิทธิภาพการจัดกลุ่ม ดังต่อไปนี้

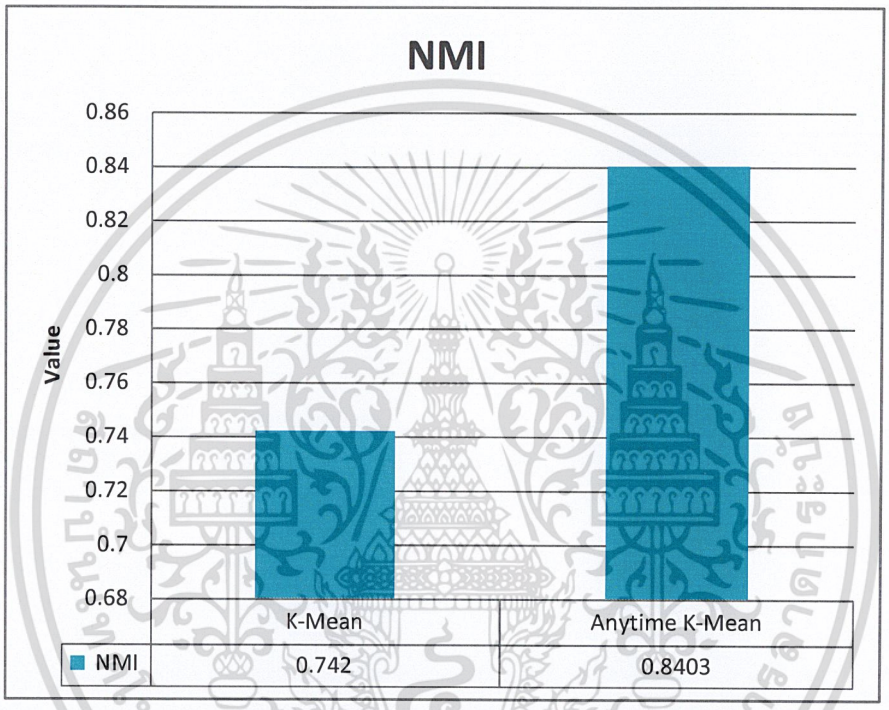
4.5.1 ผลการทดลองกับชุดข้อมูล Iris Plans



รูปที่ 4.1 กราฟแสดงผลการวัดประสิทธิภาพการจัดกลุ่มด้วย Purity โดยใช้ขั้นตอนวิธีการเคมีนกับขั้นตอนวิธีการเอนนีไทม์เคมีน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

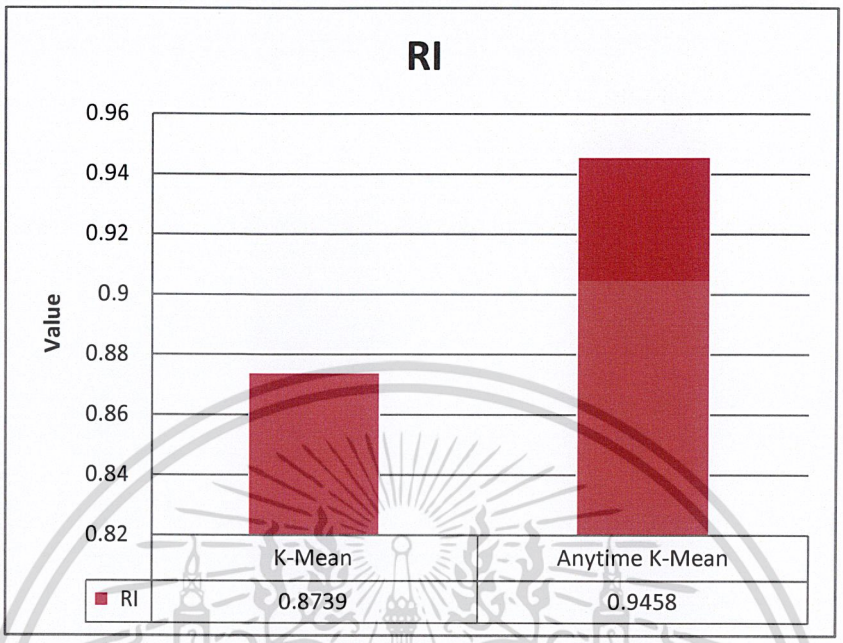
จากรูปที่ 4.1 การใช้ตัววัดประสิทธิภาพการจัดกลุ่ม โดยวัดจากค่าเพียวริตีของกลุ่มคลัสเตอร์ สำหรับข้อมูล Iris Plants นั้น ผลการจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีการเคมีนได้ค่าเพียวริตีเท่ากับ 0.8867 ซึ่งผลการจัดกลุ่มข้อมูลของขั้นตอนวิธีการเอนนีไทม์เคมีน ผลปรากฏว่า เมื่อแบ่งชุดข้อมูลออกเป็นหลายส่วน จะได้ว่าประสิทธิภาพในการจัดกลุ่มดีกว่า ซึ่งจะได้ค่าเพียวริตีโดยเฉลี่ยเท่ากับ 0.9217



รูปที่ 4.2 กราฟแสดงผลการวัดประสิทธิภาพการจัดกลุ่มด้วย NMI โดยใช้ขั้นตอนวิธีการเคมีนกับขั้นตอนวิธีการเอนนีไทม์เคมีน

จากรูปที่ 4.2 การใช้ตัววัดประสิทธิภาพการจัดกลุ่ม โดยวัดจากNMI ของกลุ่มคลัสเตอร์ สำหรับข้อมูล Iris Plants นั้น ผลการจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีการเคมีน NMI เท่ากับ 0.742 ซึ่งผลการจัดกลุ่มข้อมูลของขั้นตอนวิธีการเอนนีไทม์เคมีน ผลปรากฏว่า เมื่อแบ่งชุดข้อมูลออกเป็นหลายส่วน จะได้ว่าประสิทธิภาพในการจัดกลุ่มดีกว่า ซึ่งจะได้ NMI โดยเฉลี่ยเท่ากับ 0.8403

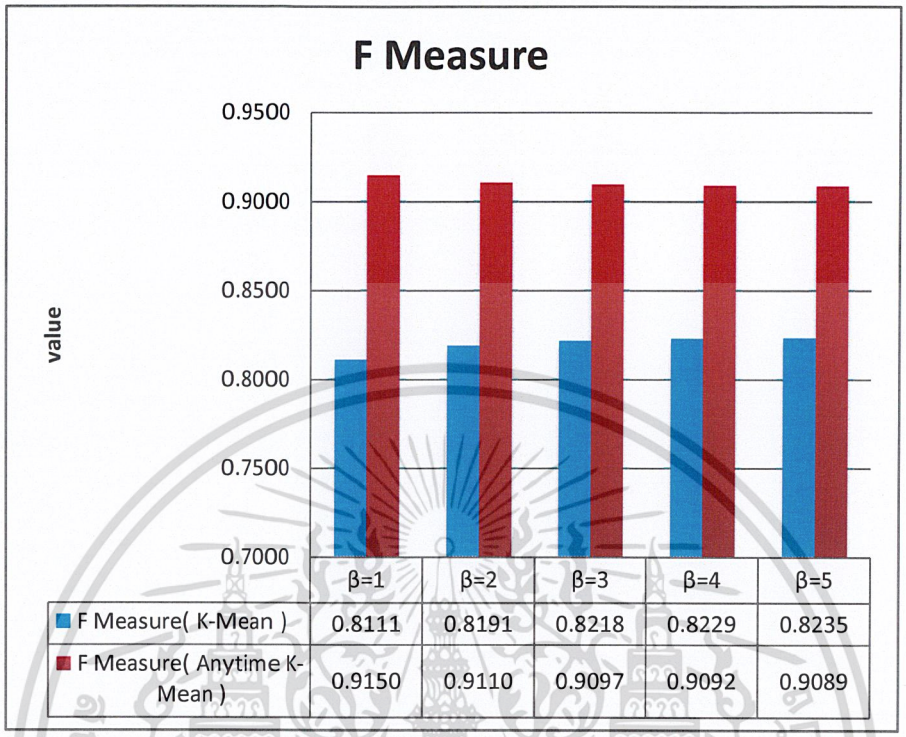
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.3 กราฟแสดงผลการวัดประสิทธิภาพการจัดกลุ่มด้วย RI โดยใช้ขั้นตอนวิธีการเคมีนกับขั้นตอนวิธีการเอนนีไทม์เคมีน

จากรูปที่ 4.3 การใช้ตัววัดประสิทธิภาพการจัดกลุ่ม โดยวัดจาก RI ของกลุ่มคลัสเตอร์สำหรับข้อมูล Iris Plants นั้น ผลการจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีการเคมีน RI เท่ากับ 0.8739 ซึ่งผลการจัดกลุ่มข้อมูลของขั้นตอนวิธีการเอนนีไทม์เคมีน ผลปรากฏว่า เมื่อแบ่งชุดข้อมูลออกเป็นหลายส่วน จะได้ว่าประสิทธิภาพในการจัดกลุ่มดีกว่า ซึ่งจะได้ RI โดยเฉลี่ยเท่ากับ 0.9458

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



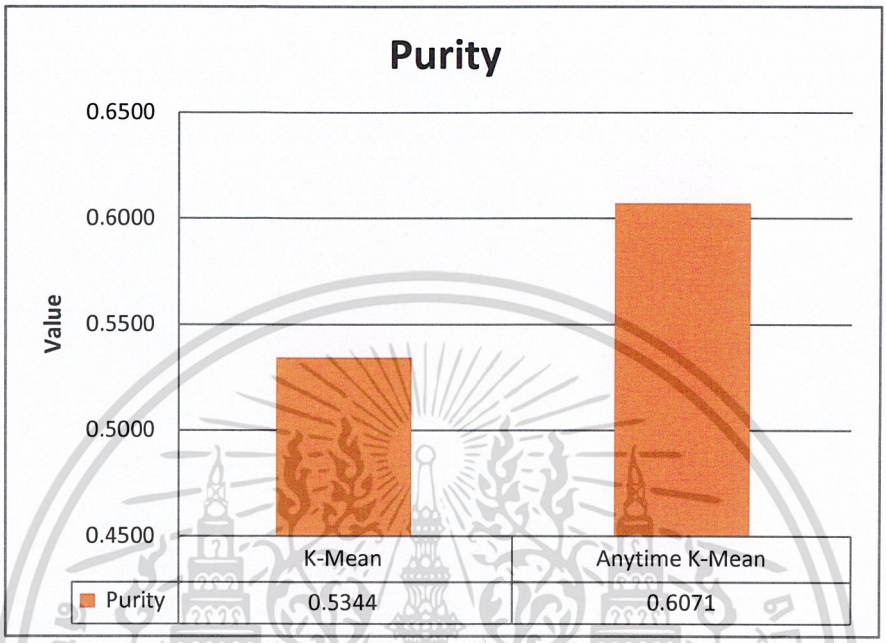
	β=1	β=2	β=3	β=4	β=5
F Measure(K-Mean)	0.8111	0.8191	0.8218	0.8229	0.8235
F Measure(Anytime K-Mean)	0.9150	0.9110	0.9097	0.9092	0.9089

รูปที่ 4.4 กราฟแสดงผลการวัดประสิทธิภาพการจัดกลุ่มด้วย F Measure โดยใช้ขั้นตอนวิธีการเคมีนกับขั้นตอนวิธีการเอนนีไทม์เคมีน

จากรูปที่ 4.4 การใช้ตัววัดประสิทธิภาพการจัดกลุ่ม โดยวัดจาก F Measure ของกลุ่มคลัสเตอร์สำหรับข้อมูล Iris Plants นั้น ผลการจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีการเคมีน เมื่อค่า β เท่ากัน ผลการจัดกลุ่มของการจัดกลุ่มด้วยขั้นตอนวิธีการเคมีนจะเห็นได้ว่า ผลการจัดกลุ่มข้อมูลของขั้นตอนวิธีการเอนนีไทม์เคมีน มีประสิทธิภาพมากกว่าการจัดกลุ่มด้วยขั้นตอนวิธีการเคมีน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

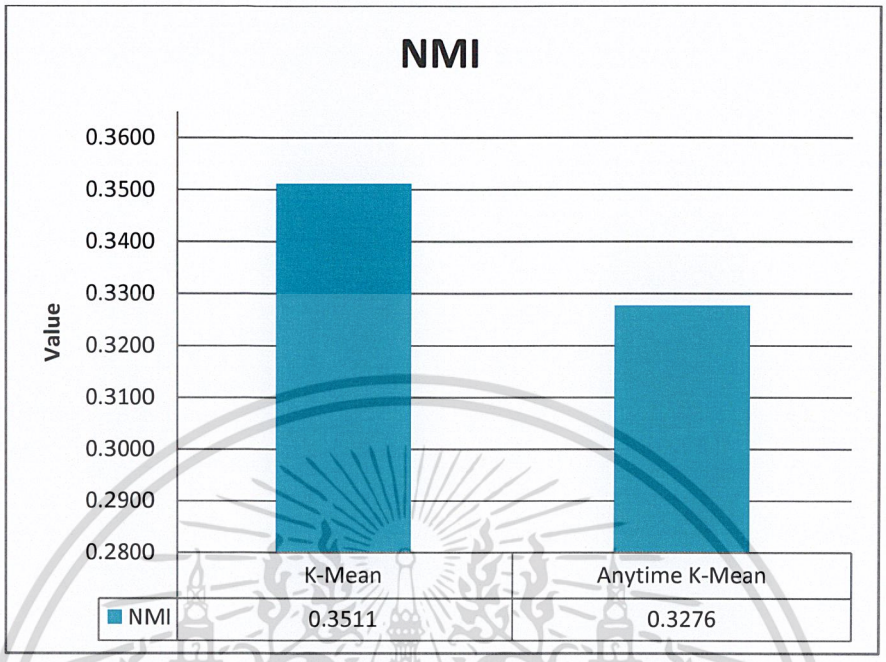
4.5.2 ผลการทดลองกับชุดข้อมูล Balance Scale



รูปที่ 4.5 กราฟแสดงผลการวัดประสิทธิภาพการจัดกลุ่มด้วย Purity โดยใช้ขั้นตอนวิธีการเคมีนกับขั้นตอนวิธีการเอนนีไทม์เคมีน

จากรูปที่ 4.5 การใช้ตัววัดประสิทธิภาพการจัดกลุ่ม โดยวัดจากค่าเพียวริตีของกลุ่มคลัสเตอร์ สำหรับข้อมูล Balance Scale นั้น ผลการจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีการเคมีนได้ค่าเพียวริตีเท่ากับ 0.5344 ซึ่งผลการจัดกลุ่มข้อมูลของขั้นตอนวิธีการเอนนีไทม์เคมีน ผลปรากฏว่า เมื่อแบ่งชุดข้อมูลออกเป็นหลายส่วน จะได้ว่าประสิทธิภาพในการจัดกลุ่มดีกว่า ซึ่งจะได้ค่าเพียวริตีโดยเฉลี่ยเท่ากับ 0.6071

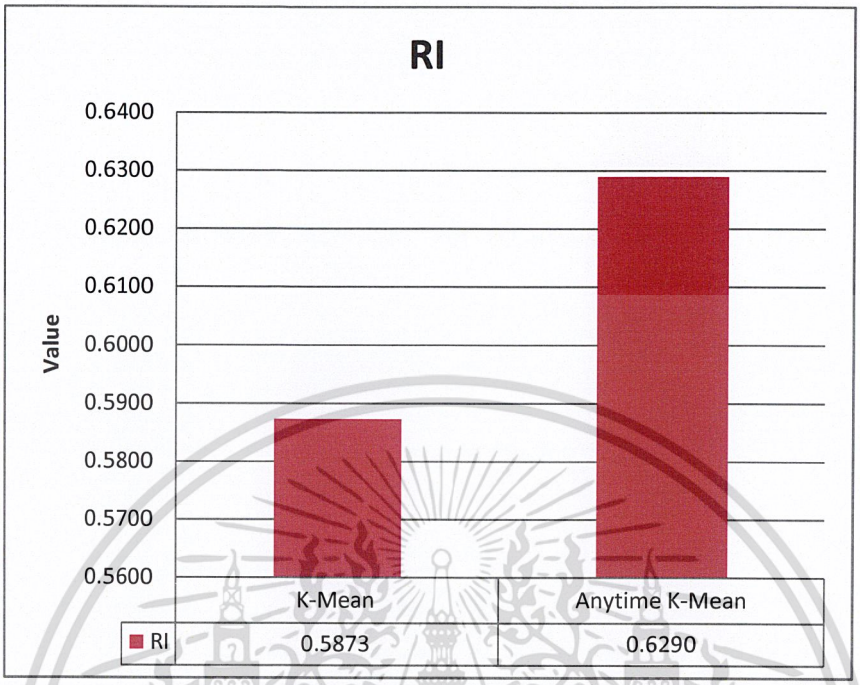
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.6 กราฟแสดงผลการวัดประสิทธิภาพการจัดกลุ่มด้วย NMI โดยใช้ขั้นตอนวิธีการเคมีนกับขั้นตอนวิธีการเอนนีไทม์เคมีน

จากรูปที่ 4.6 การใช้ตัววัดประสิทธิภาพการจัดกลุ่ม โดยวัดจาก NMI ของกลุ่มคลัสเตอร์สำหรับข้อมูล Balance Scale นั้น ผลการจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีการเคมีน NMI เท่ากับ 0.3511 ซึ่งผลการจัดกลุ่มข้อมูลของขั้นตอนวิธีการเอนนีไทม์เคมีน ได้ NMI โดยเฉลี่ยเท่ากับ 0.376 จะได้ว่าประสิทธิภาพในการจัดกลุ่มดีกว่าข้อมูลโดยใช้ขั้นตอนวิธีการเคมีนดีกว่า

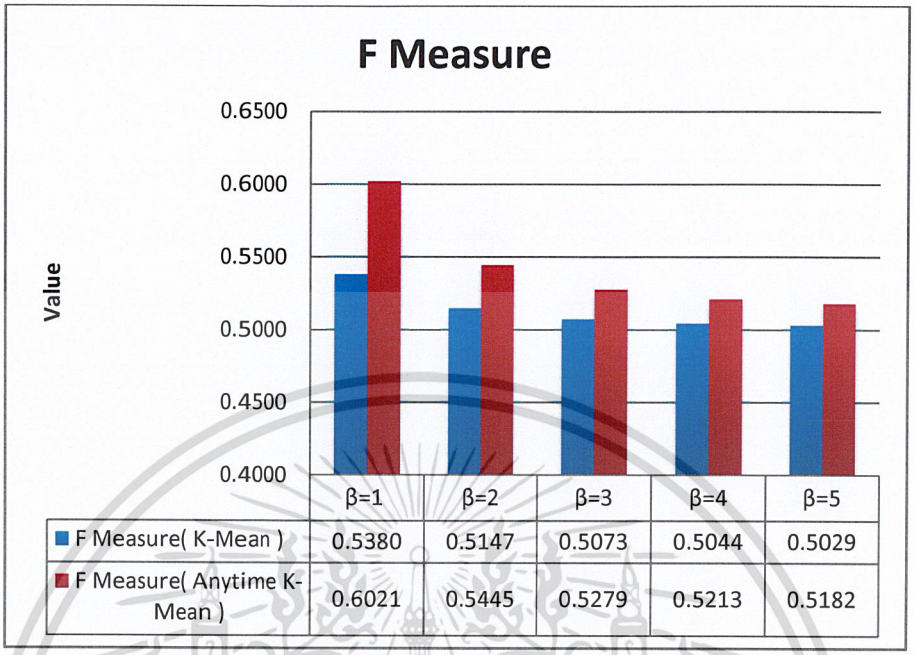
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.7 กราฟแสดงผลการวัดประสิทธิภาพการจัดกลุ่มด้วย RI โดยใช้ขั้นตอนวิธีการเคมีนกับขั้นตอนวิธีการเอนนีไทม์เคมีน

จากรูปที่ 4.7 การใช้ตัววัดประสิทธิภาพการจัดกลุ่ม โดยวัดจาก RI ของกลุ่มคลัสเตอร์สำหรับข้อมูล Balance Scale นั้น ผลการจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีการเคมีน RI เท่ากับ 0.5873 ซึ่งผลการจัดกลุ่มข้อมูลของขั้นตอนวิธีการเอนนีไทม์เคมีน ผลปรากฏว่า เมื่อแบ่งชุดข้อมูลออกเป็นหลายส่วน จะได้ว่าประสิทธิภาพในการจัดกลุ่มดีกว่า ซึ่งจะได้ RI โดยเฉลี่ยเท่ากับ 0.6290

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



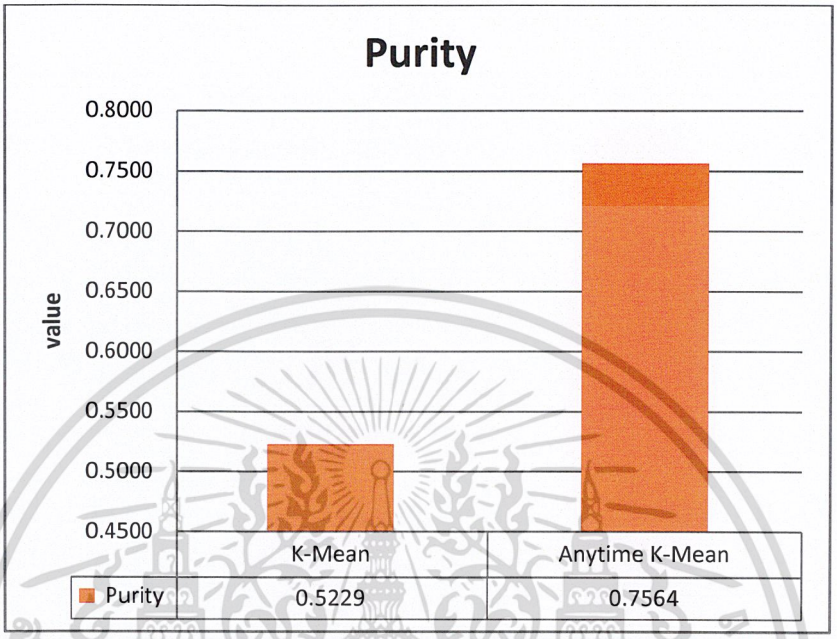
	β=1	β=2	β=3	β=4	β=5
F Measure(K-Mean)	0.5380	0.5147	0.5073	0.5044	0.5029
F Measure(Anytime K-Mean)	0.6021	0.5445	0.5279	0.5213	0.5182

รูปที่ 4.8 กราฟแสดงผลการวัดประสิทธิภาพการจัดกลุ่มด้วย F โดยใช้ขั้นตอนวิธีการเคมีนกับขั้นตอนวิธีการเอนนีไทม์เคมีน

จากรูปที่ 4.8 การใช้ตัววัดประสิทธิภาพการจัดกลุ่ม โดยวัดจาก F Measure ของกลุ่มคลัสเตอร์สำหรับข้อมูล Balance Scale นั้น ผลการจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีการเคมีน เมื่อค่า β เท่ากันผลการจัดกลุ่มของการจัดกลุ่มด้วยขั้นตอนวิธีการเคมีนจะเห็นได้ว่า ผลการจัดกลุ่มข้อมูลของขั้นตอนวิธีการเอนนีไทม์เคมีน มีประสิทธิภาพมากกว่าการจัดกลุ่มด้วยขั้นตอนวิธีการเคมีน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

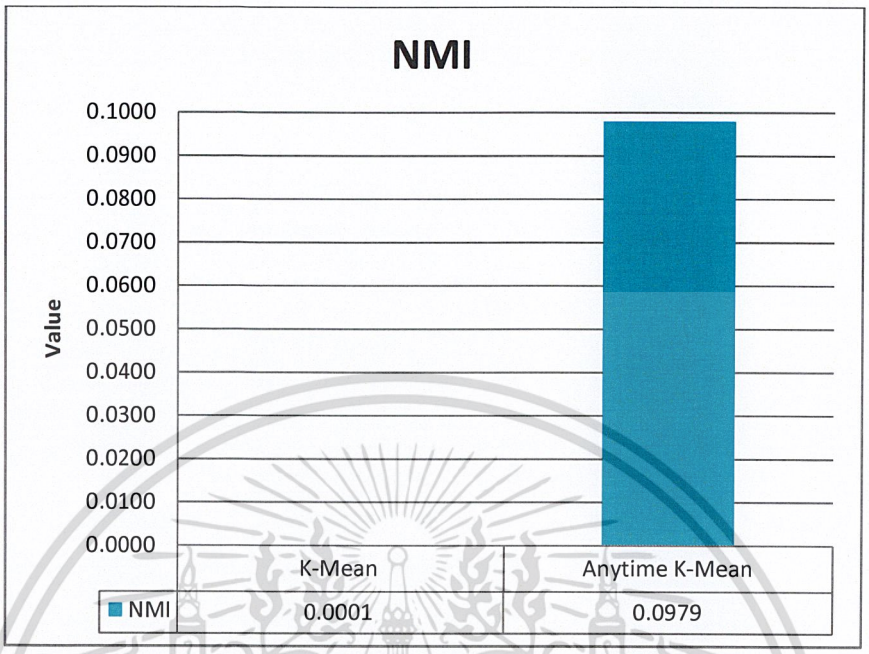
4.5.3 ผลการทดลองกับชุดข้อมูล Haberman's Survival



รูปที่ 4.9 กราฟแสดงผลการวัดประสิทธิภาพการจัดกลุ่มด้วย Purity โดยใช้ขั้นตอนวิธีการเคมีนกับขั้นตอนวิธีการเอนนีไทม์เคมีน

จากรูปที่ 4.9 การใช้ตัววัดประสิทธิภาพการจัดกลุ่ม โดยวัดจากค่าเพียวริตีของกลุ่มคลัสเตอร์สำหรับข้อมูล Haberman's Survival นั้น ผลการจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีการเคมีนได้ค่าเพียวริตีเท่ากับ 0.5229 ซึ่งผลการจัดกลุ่มข้อมูลของขั้นตอนวิธีการเอนนีไทม์เคมีน ผลปรากฏว่า เมื่อแบ่งชุดข้อมูลออกเป็นหลายส่วน จะได้ว่าประสิทธิภาพในการจัดกลุ่มดีกว่า ซึ่งจะได้ค่าเพียวริตีโดยเฉลี่ยเท่ากับ 0.7564

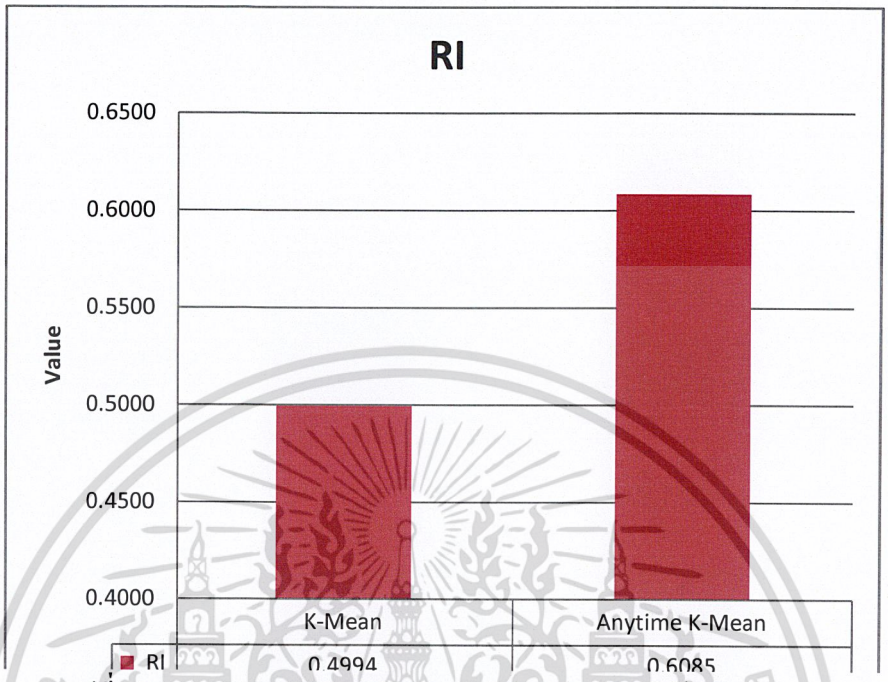
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.10 กราฟแสดงผลการวัดประสิทธิภาพการจัดกลุ่มด้วย NMI โดยใช้ขั้นตอนวิธีการเคมีนกับขั้นตอนวิธีการเอนนีไทม์เคมีน

จากรูปที่ 4.10 การใช้ตัววัดประสิทธิภาพการจัดกลุ่ม โดยวัดจาก NMI ของกลุ่มคลัสเตอร์สำหรับข้อมูล Haberman's Survival นั้น ผลการจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีการเคมีน NMI เท่ากับ 0.0001 ซึ่งผลการจัดกลุ่มข้อมูลของขั้นตอนวิธีการเอนนีไทม์เคมีน ผลปรากฏว่า เมื่อแบ่งชุดข้อมูลออกเป็นหลายส่วน จะได้ว่าประสิทธิภาพในการจัดกลุ่มดีกว่า ซึ่งจะได้ NMI โดยเฉลี่ยเท่ากับ 0.0979

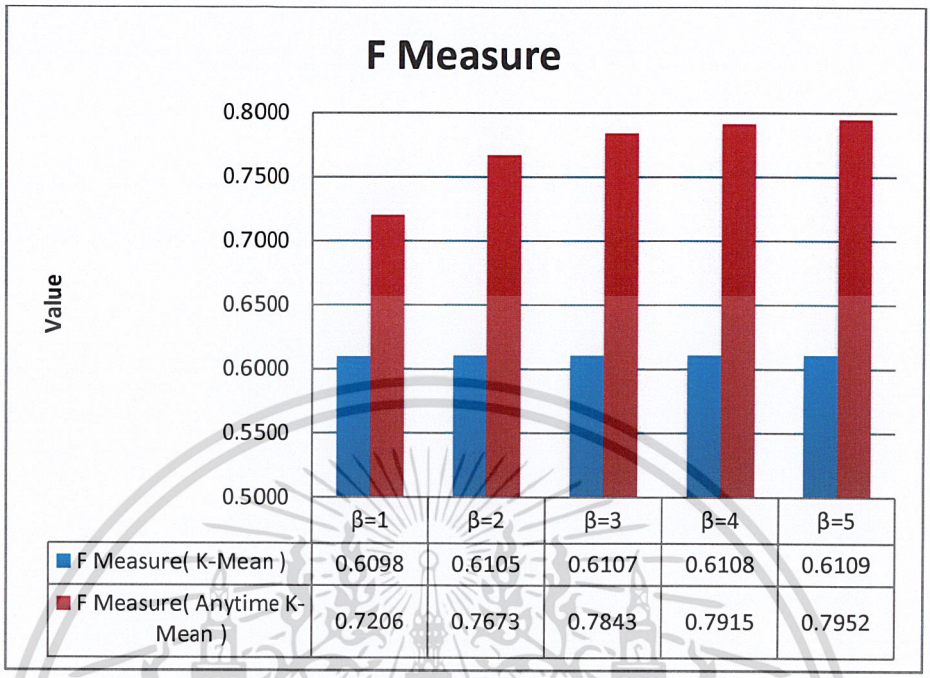
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.11 กราฟแสดงผลการวัดประสิทธิภาพการจัดกลุ่มด้วย RI โดยใช้ขั้นตอนวิธีการเคมีนกับขั้นตอนวิธีการเอนนีไทม์เคมีน

จากรูปที่ 4.11 การใช้ตัววัดประสิทธิภาพการจัดกลุ่ม โดยวัดจาก RI ของกลุ่มคลัสเตอร์ สำหรับข้อมูล Haberman's Survival นั้น ผลการจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีการเคมีน RI เท่ากับ 0.4994 ซึ่งผลการจัดกลุ่มข้อมูลของขั้นตอนวิธีการเอนนีไทม์เคมีน ผลปรากฏว่า เมื่อแบ่งชุดข้อมูลออกเป็นหลายส่วน จะได้ว่าประสิทธิภาพในการจัดกลุ่มดีกว่า ซึ่งจะได้ RI โดยเฉลี่ยเท่ากับ 0.6085

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



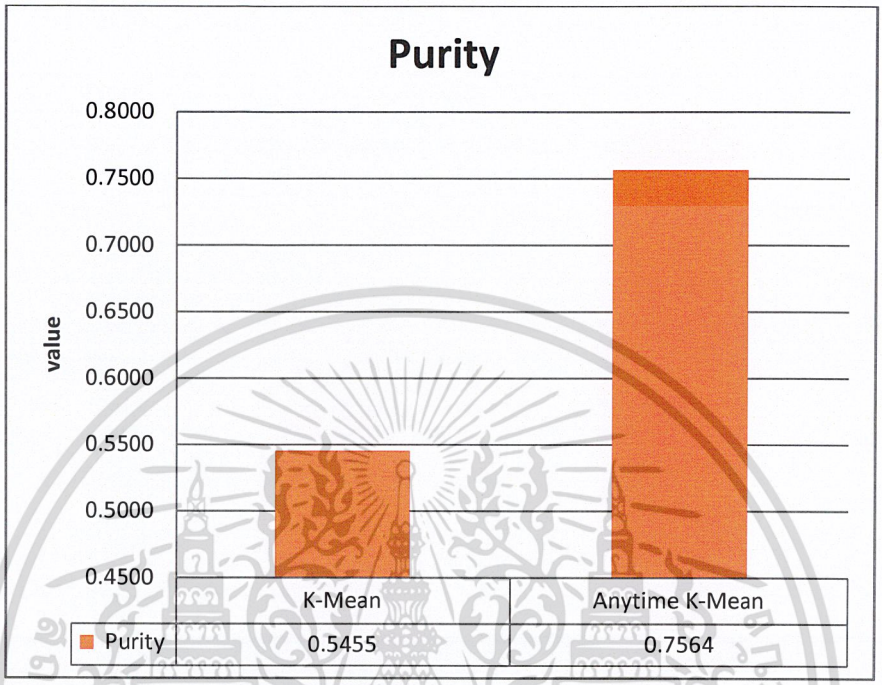
	β=1	β=2	β=3	β=4	β=5
F Measure(K-Mean)	0.6098	0.6105	0.6107	0.6108	0.6109
F Measure(Anytime K-Mean)	0.7206	0.7673	0.7843	0.7915	0.7952

รูปที่ 4.12 กราฟแสดงผลการวัดประสิทธิภาพการจัดกลุ่มด้วย F โดยใช้ขั้นตอนวิธีการเคมีนกับขั้นตอนวิธีการเอนนีไทม์เคมีน

จากรูปที่ 4.12 การใช้ตัววัดประสิทธิภาพการจัดกลุ่ม โดยวัดจาก F Measure ของกลุ่มคลัสเตอร์สำหรับข้อมูล Haberman's Survival นั้น ผลการจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีการเคมีน เมื่อค่า β เท่ากันผลการจัดกลุ่มของการจัดกลุ่มด้วยขั้นตอนวิธีการเคมีนจะเห็นได้ว่า ผลการจัดกลุ่มข้อมูลของขั้นตอนวิธีการเอนนีไทม์เคมีน มีประสิทธิภาพมากกว่าการจัดกลุ่มด้วยขั้นตอนวิธีการเคมีน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

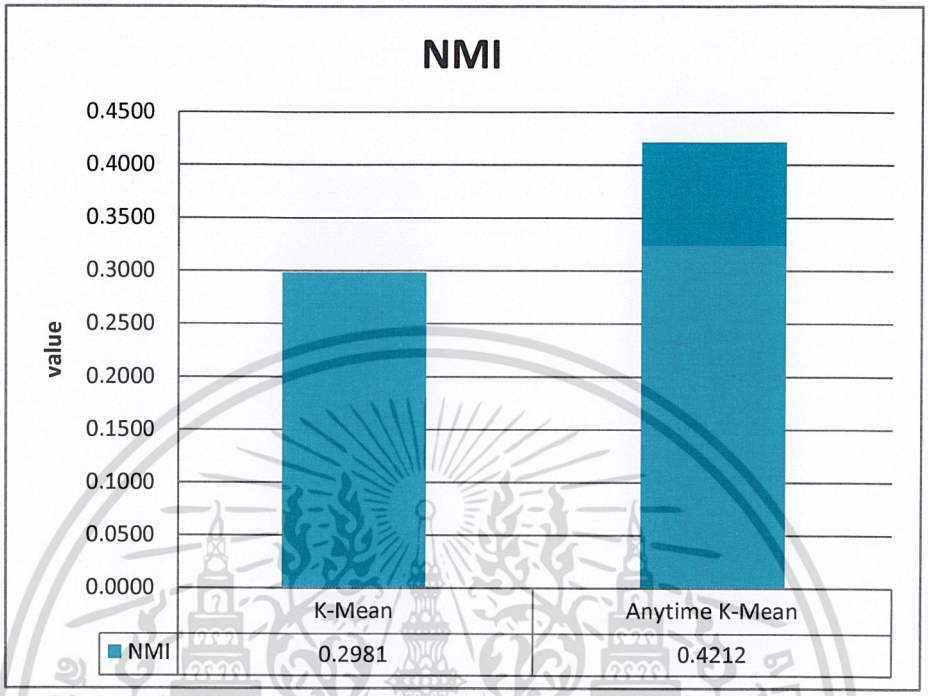
4.5.4 ผลการทดลองกับชุดข้อมูล Hayes-Roth



รูปที่ 4.13 กราฟแสดงผลการวัดประสิทธิภาพการจัดกลุ่มด้วย Purity โดยใช้ขั้นตอนวิธีการเคมีนกับขั้นตอนวิธีการเอนนีไทม์เคมีน

จากรูปที่ 4.13 การใช้ตัววัดประสิทธิภาพการจัดกลุ่ม โดยวัดจากค่าเพียวริตีของกลุ่มคลัสเตอร์สำหรับข้อมูล Hayes-Roth นั้น ผลการจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีการเคมีนได้ค่าเพียวริตีเท่ากับ 0.5455 ซึ่งผลการจัดกลุ่มข้อมูลของขั้นตอนวิธีการเอนนีไทม์เคมีน ผลปรากฏว่า เมื่อแบ่งชุดข้อมูลออกเป็นหลายส่วน จะได้ว่าประสิทธิภาพในการจัดกลุ่มดีกว่า ซึ่งจะได้ค่าเพียวริตีโดยเฉลี่ยเท่ากับ 0.7564

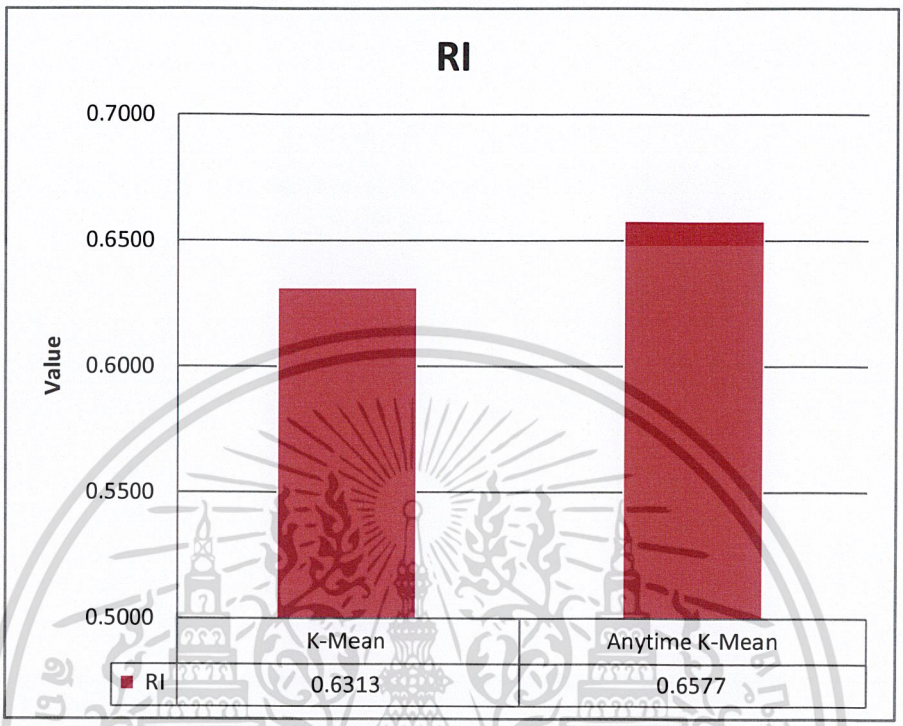
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.14 กราฟแสดงผลการวัดประสิทธิภาพการจัดกลุ่มด้วย NMI โดยใช้ขั้นตอนวิธีการเคมีนกับขั้นตอนวิธีการเอนนีไทม์เคมีน

จากรูปที่ 4.14 การใช้ตัววัดประสิทธิภาพการจัดกลุ่ม โดยวัดจากNMI ของกลุ่มคลัสเตอร์สำหรับข้อมูล Hayes-Roth นั้น ผลการจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีการเคมีน NMI เท่ากับ 0.2981 ซึ่งผลการจัดกลุ่มข้อมูลของขั้นตอนวิธีการเอนนีไทม์เคมีน ผลปรากฏว่า เมื่อแบ่งชุดข้อมูลออกเป็นหลายส่วน จะได้ว่าประสิทธิภาพในการจัดกลุ่มดีกว่า ซึ่งจะได้ NMIโดยเฉลี่ยเท่ากับ 0.4212

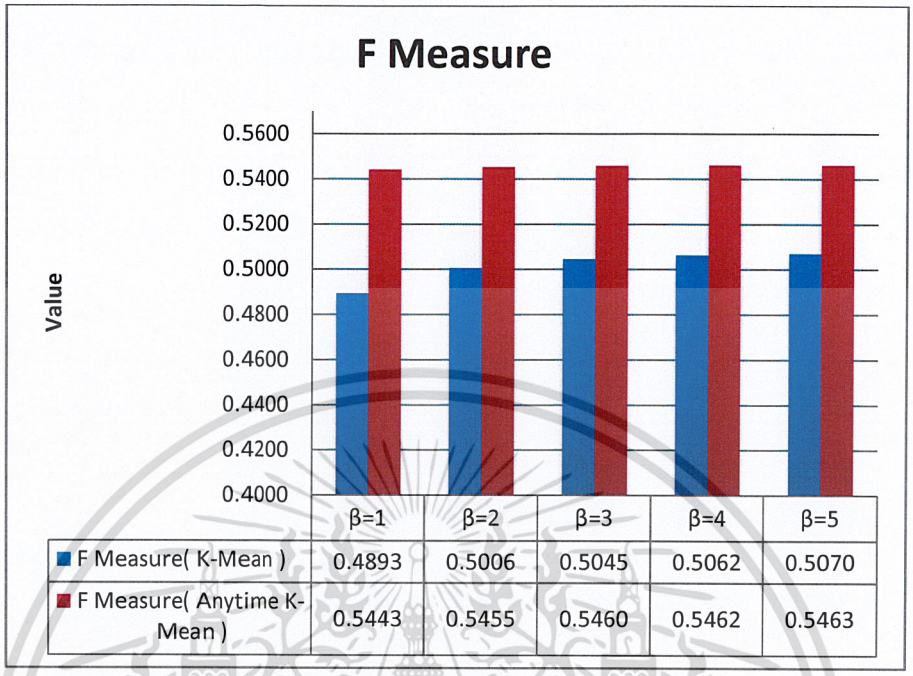
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.15 กราฟแสดงผลการวัดประสิทธิภาพการจัดกลุ่มด้วย RI โดยใช้ขั้นตอนวิธีการเคมีนกับขั้นตอนวิธีการเอนนีไทม์เคมีน

จากรูปที่ 4.15 การใช้ตัววัดประสิทธิภาพการจัดกลุ่ม โดยวัดจาก RI ของกลุ่มคลัสเตอร์ สำหรับข้อมูล Hayes-Roth นั้น ผลการจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีการเคมีน RI เท่ากับ 0.6313 ซึ่งผลการจัดกลุ่มข้อมูลของขั้นตอนวิธีการเอนนีไทม์เคมีน ผลปรากฏว่า เมื่อแบ่งชุดข้อมูลออกเป็นหลายส่วน จะได้ว่าประสิทธิภาพในการจัดกลุ่มดีกว่า ซึ่งจะได้ RIโดยเฉลี่ยเท่ากับ 0.6577

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



	β=1	β=2	β=3	β=4	β=5
F Measure(K-Mean)	0.4893	0.5006	0.5045	0.5062	0.5070
F Measure(Anytime K-Mean)	0.5443	0.5455	0.5460	0.5462	0.5463

รูปที่ 4.16 กราฟแสดงผลการวัดประสิทธิภาพการจัดกลุ่มด้วย F โดยใช้ขั้นตอนวิธีการเคมีนกับขั้นตอนวิธีการเอนนีไทม์เคมีน

จากรูปที่ 4.16 การใช้ตัววัดประสิทธิภาพการจัดกลุ่ม โดยวัดจาก F Measure ของกลุ่มคลัสเตอร์สำหรับข้อมูล Hayes-Roth นั้น ผลการจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีการเคมีน เมื่อค่า β เท่ากันผลการจัดกลุ่มของการจัดกลุ่มด้วยขั้นตอนวิธีการเคมีนจะเห็นได้ว่า ผลการจัดกลุ่มข้อมูลของขั้นตอนวิธีการเอนนีไทม์เคมีน มีประสิทธิภาพมากกว่าการจัดกลุ่มด้วยขั้นตอนวิธีการเคมีน

การจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีการเอนนีไทม์เคมีน ที่ได้นำการแบ่งกลุ่มข้อมูลในการประมวลเข้ามาช่วยเพื่อต้องการให้ผลประสิทธิภาพการจัดกลุ่มที่ดีขึ้นมากกว่าการใช้ขั้นตอนวิธีการเคมีนแบบเดิมที่มีอยู่ และมีความยืดหยุ่นในเรื่องของจำนวนข้อมูลที่ใช้ในการประมวลผลในแต่ละรอบด้วย ซึ่งในการแบ่งจำนวนข้อมูลมาประมวลผลในแต่ละรอบนั้น มีผลกระทบต่อผลการจัดกลุ่มเป็นอย่างมาก

สรุปได้ว่า การทดลองโดยใช้ชุดข้อมูลทดลองแสดงให้เห็นว่าผลการใช้ขั้นตอนวิธีการเอนนีไทม์เคมีนที่มีการแบ่งชุดข้อมูลออกเป็นส่วนๆ ตามความเหมาะสมในขั้นตอนวิธีการจัดกลุ่มนั้น ทำให้ประสิทธิภาพการจัดกลุ่มเพิ่มขึ้นได้จริง โดยวัดจากการวัดประสิทธิภาพการจัดกลุ่มข้อมูลจากการใช้ค่าเพียวริตี ผลที่ได้ทำให้เห็นว่าการจัดกลุ่มข้อมูลนั้นอยู่ถูกกลุ่มมากขึ้น ซึ่งบอกได้ว่าประสิทธิภาพการจัดกลุ่มนั้นดีขึ้นหลังจากการใช้การจัดกลุ่มด้วยขั้นตอนวิธีการเอนนีไทม์เคมีนที่มีการแบ่งชุดข้อมูลออกเป็นกลุ่มย่อย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 5
สรุปและข้อเสนอแนะ

5.1 สรุปผลการทดลอง

การจัดกลุ่มข้อมูลโดยใช้ขั้นตอนวิธีการเคมีน ซึ่งเป็นขั้นตอนวิธีหนึ่งที่ใช้ในการจัดกลุ่มข้อมูลและเพื่อให้ผลลัพธ์ที่ได้จากการใช้ขั้นตอนวิธีดังกล่าวมีข้อบกพร่องลดลง และเพิ่มความยืดหยุ่นด้านเวลาในการจัดกลุ่มข้อมูลให้มากขึ้น จึงได้มีการประยุกต์ใช้ขั้นตอนวิธีเอนนี้ไทม์กับขั้นตอนวิธีการเคมีนโดยเรียกว่า ขั้นตอนวิธีการเอนนีไทม์เคมีน ซึ่งจะเห็นได้ว่าการนำขั้นตอนวิธีเอนนี้ไทม์เข้ามา จะช่วยทำให้ประสิทธิภาพในการจัดกลุ่มดีขึ้น และจากหลายงานวิจัยพบว่าการนำขั้นตอนวิธีเอนนี้ไทม์มาช่วยให้เกิดความยืดหยุ่นทางด้านเวลา และแบ่งชุดข้อมูลออกเป็นกลุ่มๆนั้นทำให้ประสิทธิภาพในการจัดกลุ่มดีขึ้น ดังนั้นจึงทำการทดลองโดยใช้ขั้นตอนวิธีเอนนี้ไทม์เคมีนโดยทำการแบ่งชุดข้อมูลออกเป็นกลุ่มย่อย แล้วทำการทดลองเปลี่ยนจำนวนการแบ่งกลุ่มย่อย นั่นคือ เปลี่ยนข้อมูลในแต่ละรอบการประมวลผล เพื่อดูว่าการแบ่งชุดข้อมูลออกเป็นอย่างไรถึงจะให้ผลการจัดกลุ่มที่ดีที่สุด ในการทดลองนี้ได้ทำการปรับจำนวนการแบ่งกลุ่มย่อยที่หลากหลาย และทำการทดลองการจัดกลุ่มกับชุดข้อมูลที่ได้เตรียมไว้ โดยจะใช้ขั้นตอนวิธีการเคมีน และขั้นตอนวิธีการเอนนีไทม์เคมีน มาทำการเปรียบเทียบผลการทดลองของขั้นตอนวิธี ซึ่งจากการทดลองกับขั้นตอนวิธีแล้ว ผลที่ได้คือ ผลการจัดกลุ่มโดยขั้นตอนวิธีการเอนนีไทม์เคมีนให้ผลการจัดกลุ่มดีกว่าผลการจัดกลุ่มโดยขั้นตอนวิธีการเคมีน โดยผลการจัดกลุ่มที่ดีที่สุดจากขั้นตอนวิธีการเอนนีไทม์เคมีนนั้น จำนวนการแบ่งกลุ่มชุดข้อมูลออกเป็นสองกลุ่มย่อยไม่มีผลต่อความถูกต้องในการจัดกลุ่ม แต่เป็นจำนวนข้อมูลในกลุ่มย่อยกลุ่มที่หนึ่งที่มีผลต่อความถูกต้องในการจัดกลุ่ม หากจำนวนข้อมูลในกลุ่มย่อยกลุ่มแรกมีค่ามากขึ้น ความถูกต้องในการจัดกลุ่มจะดีขึ้นด้วย ซึ่งเราสามารถวัดประสิทธิภาพการจัดกลุ่มข้อมูลได้จากตัววัดประสิทธิภาพการจัดกลุ่มข้อมูลที่มีทั้งหมด 4 ตัว

5.2 ข้อเสนอแนะ

ในการประยุกต์ใช้ขั้นตอนวิธีการเคมีนกับขั้นตอนวิธีการเอนนีไทม์ ได้เป็นขั้นตอนวิธีการเอนนีไทม์เคมีน ในขั้นตอนการนำจุดศูนย์กลางมาคำนวณควรมีการถ่วงน้ำหนักให้กับข้อมูลด้วย และนอกจากนี้ในขั้นตอนวิธีการเอนนีไทม์เคมีน ควรมีขั้นตอนการประมวลผลที่มีการเพิ่มหรือลดจำนวนกลุ่มที่ใช้ในการจัดกลุ่มได้ด้วย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เอกสารอ้างอิง

Shlomo Zilberstein. 1996. Using Anytime Algorithms in Intelligent Systems. AI Magazine Volume 17 Number 3 (1996): 73-83.

Michael C., Horsch David Poole. An Anytime Algorithm for DecisionMaking under Uncertainty, Department of Computer Science ,University of British Columbia.

Kardi Teknomo. 2006. A Study on K-Mean Clustering. [Online] Available: http:// people. revoledu.com/kardi/tutorial/kMean/

Andrew Moore., Jia Li. 2002. Data Mining - Clustering by Mixture Models. [Online] Available: http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html

Maria Irene Miranda. 1999. Clustering methods and algorithms. [Online] Available: http:// www.cse.iitb.ac.in/dbms/Data/Courses/CS632/1999/clustering/dbms.html

T. Kanungo, D. M. Mount, N. Netanyahu, C. Piatko, R. Silverman. 2004. A Local Search Approximation Algorithm for k-Means Clustering. 28 (2004), 89-112. [Online] Available: http://www.cs.umd.edu/~mount/Projects/KMeans/

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ภาคผนวก

รายละเอียดตัวอย่างโปรแกรมที่ใช้

```
package iris;

import java.io.File;
import java.util.Random;
import java.util.Vector;
import org.w3c.dom.Document;
import org.w3c.dom.*;
import java.io.*;

import javax.xml.parsers.DocumentBuilderFactory;
import javax.xml.parsers.DocumentBuilder;
import org.xml.sax.SAXException;
import org.xml.sax.SAXParseException;

public class Iris {

    static Random ran = new Random();
    static double[][] dataR = null;
    static int numRecord;
    static int numGroup;
    static int numData;
    static int[] num_subRecord;

public static void main(String argv[]) throws IOException {
        CreateTable();
        Writer output = null;
        File file = new File("result.txt");
```

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

```
output = new BufferedWriter(new FileWriter(file));
double[][] G = new double[3][4];
double[][][] center = new double[numGroup][3][4];
int[] count = new int[3];
Vector dataV[] = new Vector[numGroup];
int num = 0;
int all = 0;

for (int i = 0; i < numGroup; i++) {
    dataV[i] = new Vector();
    System.out.println("ข้อมูลดิบชุดที่ " + (i + 1));
    output.write("ข้อมูลดิบชุดที่ " + (i + 1)+"\r\n");
    for (int j = 0; j < num_subRecord[i]; j++) {
        double tmp[] = new double[9];
        for (int q = 0; q < 9; q++) {
            tmp[q] = dataR[num][q];
            System.out.print(tmp[q] + "\t");
            output.write(tmp[q] + "\t");
        }
        System.out.println();
        output.write("\r\n");
        dataV[i].add(tmp);
        num++;
    }
}
System.out.println();
output.write("\r\n");
while (all < numGroup) {
    System.out.println("_____");
    output.write("_____\r\n");
    System.out.println("รอบที่ " + (all + 1) + "");
```

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งไม่อนุญาตให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

```
output.write("          รอบที่ " + (all + 1) + "          \r\n");
System.out.println("_____");
output.write("_____\r\n");
System.out.println();
output.write("\r\n");
double[][] data = new double[dataV[all].size()][9];
for (int j = 0; j < dataV[all].size(); j++) {
   data[j] = (double[]) dataV[all].elementAt(j);
}
for (int i = 0; i < data.length; i++) {
   if (data[i][4] == 1.0) {
      G[0][0] = G[0][0] + data[i][0];
      G[0][1] = G[0][1] + data[i][1];
      G[0][2] = G[0][2] + data[i][2];
      G[0][3] = G[0][3] + data[i][3];
      count[0]++;
   } else if (data[i][4] == 2.0) {
      G[1][0] = G[1][0] + data[i][0];
      G[1][1] = G[1][1] + data[i][1];
      G[1][2] = G[1][2] + data[i][2];
      G[1][3] = G[1][3] + data[i][3];
      count[1]++;
   } else if (data[i][4] == 3.0) {
      G[2][0] = G[2][0] + data[i][0];
      G[2][1] = G[2][1] + data[i][1];
      G[2][2] = G[2][2] + data[i][2];
      G[2][3] = G[2][3] + data[i][3];
      count[2]++;
   }
}
```

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

```
for (int i = 0; i < 3; i++) {
    for (int j = 0; j < 4; j++) {
        if (count[i] == 0) {
            count[i] = 1;
        }
        center[all][i][j] = G[i][j] / count[i];
    }
}
int pos = 0;
double min = 2000;
double tmp = 0;
for (int i = 0; i < data.length; i++) {
    for (int j = 0; j < 3; j++) {
        tmp = Math.sqrt(
            Math.pow(data[i][0] - center[all][j][0], 2) +
            Math.pow(data[i][1] - center[all][j][1], 2) +
            Math.pow(data[i][2] - center[all][j][2], 2) +
            Math.pow(data[i][3] - center[all][j][3], 2));
        data[i][j + 5] = tmp;
        if (tmp < min) {
            pos = j + 1;
            min = tmp;
        }
    }
    data[i][8] = pos;
    min = 2000;
    pos = 0;
}
```

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

```
int countTu = 0;
for (int j = 0; j < data.length; j++) {
   if (data[j][4] != data[j][8]) {
      data[j][4] = data[j][8];
      countTu++;
   }
}
G = new double[3][4];
count = new int[3];
System.out.println("ตัวที่แตกต่าง : " + countTu);
output.write("ตัวที่แตกต่าง : " + countTu+"\r\n");
System.out.println("<< ข้อมูล Center >>");
output.write("<< ข้อมูล Center >>\r\n");
System.out.println();
output.write("\r\n");
for (int j = 0; j < 3; j++) {
   for (int q = 0; q < 4; q++) {
      System.out.print(center[all][j][q] + "\t");
      output.write(center[all][j][q] + "\t");
   }
   System.out.println();
   output.write("\r\n");
}
System.out.println();
output.write("\r\n");
System.out.println("<< ข้อมูลจัดกลุ่ม >>");
output.write("<< ข้อมูลจัดกลุ่ม >>\r\n");
System.out.println();
output.write("\r\n");
for (int j = 0; j < data.length; j++) {
   for (int q = 0; q < 9; q++) {
```

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งไม่อนุญาตให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

```
            System.out.print(data[j][q] + "\t");
            output.write(data[j][q] + "\t");
        }
        System.out.println();
        output.write("\r\n");
    }
    if (countTu == 0) {
        System.out.println();
        output.write("\r\n");
        System.out.println("<< ข้อมูล Center รอบสุดท้าย>>");
        output.write("<< ข้อมูล Center รอบสุดท้าย>>\r\n");
        all++;
        for (int i = 0; i < 3; i++) {
            double t[] = new double[9];
            for (int j = 0; j < 4; j++) {
                t[j] = center[all - 1][i][j];
            }
            t[4] = i + 1;
            for (int k = 0; k < 9; k++) {
                System.out.print(t[k] + "\t");
                output.write(t[k] + "\t");
            }
            System.out.println();
            output.write("\r\n");
            if (all < numGroup) {
                dataV[all].add(t);
            }
        }
    }
}
output.close(); }
```

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

```
static void CreateTable() {
    try {
        DocumentBuilderFactory docBuilderFactory = DocumentBuilderFactory.newInstance();
        DocumentBuilder docBuilder = docBuilderFactory.newDocumentBuilder();
        Document doc = docBuilder.parse(new File("iris.xml")); // normalize text representation
        doc.getDocumentElement().normalize();
        NodeList listOfRows = doc.getElementsByTagName("Row");
        for (int s = 0; s < listOfRows.getLength(); s++) {
            Node listOfCells = listOfRows.item(s);
            Element Cell = (Element) listOfCells;
            NodeList Data = Cell.getElementsByTagName("Data");

            if (s == 0) {
                Element elementData = (Element) Data.item(0);
                NodeList textFNList = elementData.getChildNodes();
                numRecord = Integer.parseInt(((Node) textFNList.item(0)).getNodeValue().trim());

                elementData = (Element) Data.item(1);
                textFNList = elementData.getChildNodes();
                numGroup = Integer.parseInt(((Node) textFNList.item(0)).getNodeValue().trim());

                dataR = new double[numRecord][9];
                num_subRecord = new int[numGroup];
            } else if (s > 0 && s < numGroup + 1) {
                Element elementData = (Element) Data.item(0);
                NodeList textFNList = elementData.getChildNodes();
                num_subRecord[s - 1] = Integer.parseInt(((Node)
textFNList.item(0)).getNodeValue().trim());
            }
```

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

```
            else {
                for (int i = 0; i < 5; i++) {
                    Element elementData = (Element) Data.item(i);
                    NodeList textFNList = elementData.getChildNodes();
                    dataR[s - (1 + numGroup)][i] = Double.parseDouble(((Node)
textFNList.item(0)).getNodeValue().trim());
                }
            }
        }
    } catch (SAXParseException err) {
        System.out.println("** Parsing error" + ", line " + err.getLineNumber() + ",uri " +
err.getSystemId());
        System.out.println(" " + err.getMessage());
    } catch (SAXException e) {
        Exception x = e.getException();
        ((x == null) ? e : x).printStackTrace();
    } catch (Throwable t) {
        t.printStackTrace();
    }
  }
}
```

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้