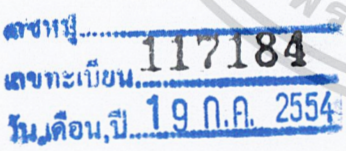
สำนักหอสมุดกลาง พระจอมเกล้าลาดกระบัง

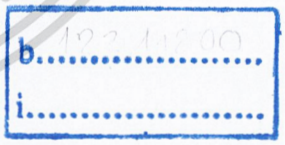
การศึกษาวิธีการจัดกลุ่มข้อมูลโดยทฤษฎีฝูงมด

STUDIES ON ANT-BASED CLUSTERING AND ITS APPLICATIONS

นายปกรณ์ จงธนเศรษฐกุล

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง





โครงงานพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต

สาขาวิชาวิทยาการคอมพิวเตอร์

คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ปีการศึกษา 2553

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

STUDIES ON ANT-BASED CLUSTERING AND ITS APPLICATIONS

MR.PAKORN JONGTANASAETAKUL

A SPECIAL PROJECT SUBMITTED IN PARTIAL FULFILLMENT
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 อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

หัวข้อโครงงานพิเศษ	การศึกษาวิธีการจัดกลุ่มข้อมูลโดยทฤษฎีฝูงมด
	Studies on Ant-Based Clustering and its Applications
ชื่อนักศึกษา	นายปกรณ์ จงธนเศรษฐกุล 47050780
ปริญญา	วิทยาศาสตรบัณฑิต
สาขาวิชา	วิทยาการคอมพิวเตอร์
อาจารย์ที่ปรึกษา	ดร.วรางคณา กิ้มปาน

คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง อนุมัติให้โครงงานพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร วิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ ประจำปีการศึกษา 2553

คณะกรรมการสอบ	ลายมือชื่อ
ผศ.ดร.จีรพร วีระพันธุ์ ประธานกรรมการ	
อ.วิสันต์ ตั้งวงษ์เจริญ กรรมการ	
ดร.วรางคณา กิ้มปาน กรรมการและอาจารย์ที่ปรึกษา	

ลิขสิทธิ์ของคณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

หัวข้อโครงงานพิเศษ	การศึกษาวิธีการจัดกลุ่มข้อมูลโดยทฤษฎีฝูงมด
	Studies on Ant-Based Clustering and its Applications
ชื่อนักศึกษา	นายปกรณ์ จงธนเศรษฐกุล 47050780
ปริญญา	วิทยาศาสตรบัณฑิต
สาขาวิชา	วิทยาการคอมพิวเตอร์
ปีการศึกษา	2553
อาจารย์ที่ปรึกษา	ดร.วรางคณา กิ้มปาน

บทคัดย่อ

ในการจัดกลุ่มตามประเภทของข้อมูล มีอัลกอริทึมต่างๆที่สามารถนำมาจัดกลุ่มข้อมูลได้ ซึ่งในแต่ละอัลกอริทึมก็จะมีวิธีการ หรือวิธีคิดที่แตกต่างกัน และจะมีข้อดีข้อเสียแตกต่างกันด้วย ทำให้มีแนวคิดที่จะนำอัลกอริทึมใหม่ คือ Ant-Based Clustering ซึ่งได้มาจากการศึกษาพฤติกรรมของมดจริงๆในการเก็บซากศพของมดด้วยกันเอง

ในโครงงานชิ้นนี้ ได้ทำการศึกษาขั้นตอนการทำงานของ Ant-Based Clustering เพื่อนำมาใช้ในการจำลองการจัดกลุ่มตามประเภทของข้อมูล มีการปรับค่าพารามิเตอร์ และชุดข้อมูลที่นำมาทดสอบ เพื่อให้ได้ผลการทดลองที่ดี จากการทดลองสรุปได้ว่า การจัดกลุ่มโดยใช้อัลกอริทึมฝูงมด สามารถเพิ่มประสิทธิภาพในการจัดกลุ่ม และยังลดเวลาในการจัดกลุ่มอีกด้วย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Title	Studies on Ant-Based Clustering and its Applications
Student	Mr.Pakorn Jongtanasaetakul 47050780
Degree	Bachelor of Science
Major Program	Computer Science
Academic Year	2010
Advisor	Dr.Warangkhana Kimpan

ABSTRACT

There are many types of clustering algorithms. Each has a different algorithm for clustering data with both advantages and disadvantages. So, the idea of using the new algorithm which is derived from real ants' behavior of gathering the dead body of ants, called Ant-Based Clustering, was proposed.

This project is about the studying of the Ant-Based Clustering algorithm from the simulation of the clustering. The parameters and data sets were changed to see how well the algorithm processed. From the experiments, it can be concluded that Ant-Based Clustering can enhance the performance in clustering effectively and within a shorter time.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

กิตติกรรมประกาศ

ปริญญานิพนธ์ฉบับนี้สำเร็จได้อย่างดีได้ด้วยคำแนะนำคำปรึกษาจาก ดร. วรางคณา กิ้มปาน ซึ่งเป็นอาจารย์ผู้ควบคุมปริญญานิพนธ์ ที่ให้คำปรึกษาและการช่วยเหลือ ข้าพเจ้าขอกราบขอบพระคุณเป็นอย่างสูง

ขอกราบขอบพระคุณคณาจารย์สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ทุกๆท่านที่ได้ประสิทธิ์ประสาทวิชาให้แก่ข้าพเจ้า

ขอขอบคุณเพื่อนๆ และรุ่นพี่ในภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบังทุกคน สำหรับคำแนะนำ และกำลังใจแก่ข้าพเจ้า

สุดท้ายนี้ข้าพเจ้าขอกราบขอบพระคุณ บุคคลสำคัญที่สุดที่ทำให้ข้าพเจ้าได้มีวันนี้คือ บิดามารดา อันเป็นที่เคารพรักยิ่ง ซึ่งได้เลี้ยงดูข้าพเจ้ามาพร้อมทั้งให้โอกาสในการศึกษาอย่างเต็มที่และให้กำลังใจเอาใจใส่เสมอมา รวมทั้งทุกท่านที่มีส่วนเกี่ยวข้องในการทำโครงงานนี้ที่ไม่ได้กล่าวชื่อมา ณ ที่นี้ด้วย ซึ่งทำให้ปริญญานิพนธ์ฉบับนี้สำเร็จลุล่วงด้วยดี และคุณค่าที่ได้จากปริญญานิพนธ์นี้ ข้าพเจ้าขอมอบแด่ผู้มีอุปการคุณทุกท่าน

ปกรณ์ จงธนเศรษฐกุล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ

หน้า

บทคัดย่อภาษาไทย I

บทคัดย่อภาษาอังกฤษ II

กิตติกรรมประกาศ III

สารบัญ IV

สารบัญตาราง VII

สารบัญรูป VIII

บทที่ 1 บทนำ 1

1.1 ที่มาและความสำคัญของปัญหาพิเศษ 1

1.2 วัตถุประสงค์ของปัญหาพิเศษ 1

1.3 ขอบเขตของปัญหาพิเศษ 1

1.4 ประโยชน์ที่คาดว่าจะได้รับ 2

1.5 ขั้นตอนในการดำเนินการ 2

1.6 อุปกรณ์ที่ใช้ในการทำปัญหาพิเศษ 2

บทที่ 2 ทฤษฎีพื้นฐาน 3

2.1 การจัดกลุ่มข้อมูล (Clustering) 3

2.1.1 การวัดค่าความแตกต่างของข้อมูล 3

2.1.1.1 Euclidean distance 4

2.1.2 k - Means Clustering 4

2.1.3 Hierarchical Clustering 7

2.1.4 Ant – based Clustering 9

2.1.4.1 อัลกอริทึมพื้นฐานของ Ant – based Clustering 11

2.1.4.2 Short – term memory 12

2.1.4.3 Increasing radius of perception 13

2.1.4.4 Spatial separation 13

2.1.4.5 Parameter settings 14

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ (ต่อ)

หน้า

2.1.4.5.1 Parameter to be set as a function of the size of the data set 14

2.1.4.5.2 Activity – based α - adaptation 14

2.1.5 การประเมินผล (Evaluation) 15

2.1.5.1 Inner Cluster Variance 15

2.1.5.2 Dunn Index 15

2.1.5.3 Pearson correlation 16

บทที่ 3 โครงสร้างและหลักการทำงานของโปรแกรมที่ศึกษา 18

3.1 รูปแบบอินเตอร์เฟสที่ใช้ในการติดต่อกับผู้ใช้งาน 18

3.1.1 ค่าพารามิเตอร์ต่างๆ 19

3.2 Flowchart 21

3.2.1 หลักการทำงานของอัลกอริทึม 21

บทที่ 4 การทดลองและผลการทดลอง 24

4.1 การทดลอง 24

4.2 ผลการทดลอง 24

4.2.1 Test Set A 24

4.2.1.1 จำลองการจัดกลุ่มข้อมูลครั้งที่ 1 25

4.2.1.2 จำลองการจัดกลุ่มข้อมูลครั้งที่ 2 27

4.2.1.3 จำลองการจัดกลุ่มข้อมูลครั้งที่ 3 29

4.2.1.4 จำลองการจัดกลุ่มข้อมูลครั้งที่ 4 31

4.2.1.5 จำลองการจัดกลุ่มข้อมูลครั้งที่ 5 33

4.2.1.6 จำลองการจัดกลุ่มข้อมูลครั้งที่ 6 35

4.2.2 Test Set B 37

4.2.2.1 จำลองการจัดกลุ่มข้อมูลครั้งที่ 7 38

4.2.2.2 จำลองการจัดกลุ่มข้อมูลครั้งที่ 8 40

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ (ต่อ)

หน้า

4.2.2.3 จำลองการจัดกลุ่มข้อมูลครั้งที่ 9 42

4.2.2.4 จำลองการจัดกลุ่มข้อมูลครั้งที่ 10 44

4.2.2.5 จำลองการจัดกลุ่มข้อมูลครั้งที่ 11 46

4.2.3 Test Set C 48

4.2.3.1 จำลองการจัดกลุ่มข้อมูลครั้งที่ 12 49

4.2.3.2 จำลองการจัดกลุ่มข้อมูลครั้งที่ 13 51

4.2.3.3 จำลองการจัดกลุ่มข้อมูลครั้งที่ 14 53

4.2.3.4 จำลองการจัดกลุ่มข้อมูลครั้งที่ 15 55

บทที่ 5 ปัญหาที่พบและบทสรุป 57

5.1 สรุปผลการทดลอง 57

5.2 ข้อจำกัดของโครงงาน 58

5.3 ข้อเสนอแนะและแนวทางในการพัฒนาต่อไปในอนาคต 58

บรรณานุกรม 59

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญตาราง

ตารางที่ **หน้า**

4.1 ผลค่า Pearson correlation ที่ได้จากการทดลอง 24

4.2 ผลค่า Pearson correlation ที่ได้จากการทดลอง 37

4.3 ผลค่า Pearson correlation ที่ได้จากการทดลอง 48

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป

รูปที่		หน้า
2.1	ขั้นตอนการทำงานของอัลกอริทึม k-Means Clustering	4
2.2	การจัดกลุ่มข้อมูลโดยวิธี k-Means Clustering ขั้นตอนที่ 1	5
2.3	การจัดกลุ่มข้อมูลโดยวิธี k-Means Clustering ขั้นตอนที่ 2	5
2.4	การจัดกลุ่มข้อมูลโดยวิธี k-Means Clustering ขั้นตอนที่ 3	6
2.5	การจัดกลุ่มข้อมูลโดยวิธี k-Means Clustering ขั้นตอนที่ 4	6
2.6	การจัดกลุ่มข้อมูลโดยวิธี Hierarchical Clustering	7
2.7	Dendrogram ใช้สำหรับดูค่าความต่าง (ค่าdissimilarity)	8
2.8	Venn Diagram ใช้สำหรับดูการจัดกลุ่ม	9
2.9	แสดงขนาดของ local neighbourhood	10
2.10	ขั้นตอนการทำงานของอัลกอริทึม Ant-based Clustering	11
2.11	แผนภาพกระจัดกระจายแสดงความสัมพันธ์ระหว่างตัวแปร x และ y	17
3.1	รูปแบบอินเตอร์เฟสที่ใช้ในการติดต่อกับผู้ใช้งานในหน้าหลัก	18
3.2	รูปแบบอินเตอร์เฟสที่ใช้ในการติดต่อกับผู้ใช้งานในการกำหนดพารามิเตอร์	19
3.3	รูปแบบอินเตอร์เฟสที่ใช้ในการติดต่อกับผู้ใช้งานในการเลือกชุดข้อมูล	20
3.4	ระบบการทำงานของอัลกอริทึม	21
3.5	ตัวอย่างการกระจายข้อมูลลงในตาราง	22
3.6	ตัวอย่างแสดงการตัดสินใจของมดว่าจะเก็บข้อมูลหรือไม่	22
3.7	ตัวอย่างแสดงการเคลื่อนย้ายข้อมูลของมด	23
4.1	เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 1	25
4.2	ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 1	26
4.3	เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 2	27
4.4	ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 2	28
4.5	เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 3	29
4.6	ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 3	30
4.7	เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 4	31
4.8	ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 4	32
4.9	เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 5	33

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป (ต่อ)

รูปที่		หน้า
4.10	ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 5	34
4.11	เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 6	35
4.12	ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 6	36
4.13	เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 7	38
4.14	ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 7	39
4.15	เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 8	40
4.16	ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 8	41
4.17	เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 9	42
4.18	ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 9	43
4.19	เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 10	44
4.20	ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 10	45
4.21	เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 11	46
4.22	ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 11	47
4.23	เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 12	49
4.24	ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 12	50
4.25	เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 13	51
4.26	ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 13	52
4.27	เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 14	53
4.28	ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 14	54
4.29	เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 15	55
4.30	ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 15	56

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหาพิเศษ

ในปัจจุบันเป็นยุคของข้อมูลข่าวสาร มีการนำข้อมูลมาวิเคราะห์ในแง่มุมต่างๆ แต่ในธุรกิจบางประเภทก็จำเป็นที่จะต้องมีการแบ่งแยกข้อมูล หรือจำแนกข้อมูลออกเป็นประเภทๆไป เพื่อนำไปใช้วิเคราะห์ประกอบการทำธุรกิจนั้นๆ

ตัวอย่างเช่น ถ้าเรามีข้อมูลอยู่จำนวนมากๆ กระจัดกระจายอยู่ แล้วเราต้องการจัดกลุ่มของข้อมูลออกเป็นกลุ่มๆ โดยที่ข้อมูลที่มีลักษณะเหมือนกันหรือคล้ายกัน จะถูกจัดไว้ในกลุ่มเดียวกัน เราจะใช้วิธีการจัดกลุ่ม (Clustering) ข้อมูล ซึ่งจะมีอยู่หลายวิธีด้วยกัน แต่ในที่นี้จะขอยกตัวอย่างประกอบ 2 วิธี คือ k-Means Clustering กับ Hierarchical Clustering วิธี k-Means Clustering จะมีข้อเสียอยู่ที่เราต้องกำหนดค่า k ด้วยตนเอง โดยที่ k คือ จำนวนกลุ่มที่จะทำการClustering ซึ่งจริงๆ เราไม่รู้ว่าค่า k ควรจะเป็นค่าใด จึงจะเหมาะสม และถ้าข้อมูลไม่เป็นตัวเลข ก็จะคำนวณได้ยาก

โครงงานนี้จะเป็นการศึกษาวิธีการ Ant-Based Clustering หรือชื่อไทยว่า การจัดกลุ่มข้อมูลโดยทฤษฎีฝูงมด ซึ่งจะสามารถแก้ไขจุดด้อยบางประการของวิธีที่กล่าวมาก่อนหน้านี้ และเพิ่มประสิทธิภาพของการ Clustering ข้อมูลให้มากยิ่งขึ้น

1.2 วัตถุประสงค์ของปัญหาพิเศษ

1) เพื่อที่จะศึกษา Ant-Based Clustering ซึ่งเป็น Algorithm ใหม่

2) เพื่อที่จะสามารถเปรียบเทียบความแม่นยำและความรวดเร็วในการแบ่งกลุ่มของข้อมูลแต่ละประเภทได้

3) เพื่อศึกษาข้อดีและข้อจำกัดในการเพิ่มประสิทธิภาพในการแบ่งกลุ่มประเภทของข้อมูล

1.3 ขอบเขตของปัญหาพิเศษ

1) ทดลองนำข้อมูลซึ่งเป็นข้อมูลต่างชนิดกัน มาทำการจัดกลุ่มด้วยวิธีการ Ant-based Clustering

2) นำผลที่ได้มาเปรียบเทียบว่าการจัดกลุ่มข้อมูลประเภทใดดีที่สุด โดยเกณฑ์ในการวัด คือ ความแม่นยำและความรวดเร็วในการแบ่งกลุ่ม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1) ทำให้เข้าใจหลักการทำงานของ Ant-Based Clustering

2) สามารถนำ Applications ที่ได้จากการจัดกลุ่ม โดยวิธี Ant-Based Clustering ไปประยุกต์ใช้กับงานที่จะต้องมีการแบ่งกลุ่มตามชนิดของข้อมูลได้อย่างมีประสิทธิภาพมากยิ่งขึ้น

1.5 ขั้นตอนในการดำเนินการ

1) ศึกษาขอบเขตของปัญหาพิเศษ

2) ศึกษาทฤษฎีของ Ant-Based Algorithm

3) ศึกษาทฤษฎีการจัดกลุ่ม ด้วยวิธี k-Means Clustering และ Hierarchical Clustering

4) ทดลองจัดกลุ่มข้อมูลแต่ละชนิด แล้วนำผลการทดลองที่ได้มาเปรียบเทียบและสรุปผล

5) จัดทำเอกสารคู่มือการใช้งานโปรแกรม

6) นำเสนอผลงาน

1.6 อุปกรณ์ที่ใช้ในการทำปัญหาพิเศษ

1) ฮาร์ดแวร์

- เครื่องคอมพิวเตอร์สำหรับพัฒนาโปรแกรมAnt-Based Clustering จำนวน 1 เครื่อง

2) ซอฟต์แวร์

- ระบบปฏิบัติการ Window XP
- โปรแกรมภาษา Java

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 2
ทฤษฎีพื้นฐาน

2.1 การจัดกลุ่มข้อมูล (Clustering)

การจัดกลุ่มข้อมูล เป็นวิธีการวิเคราะห์ข้อมูล ซึ่งใช้ในการเรียนรู้เกี่ยวกับการทำเหมืองข้อมูล โดยจะแบ่งชุดข้อมูล (มักจะเป็นเวคเตอร์) ออกเป็นกลุ่ม (cluster) นำข้อมูลที่มีคุณลักษณะเหมือนกัน หรือคล้ายกันจัดไว้ในกลุ่มเดียวกัน ขั้นตอนวิธีที่ใช้ในการแบ่งกลุ่มจะอาศัยความเหมือน (similarity) หรือความใกล้ชิด (proximity) โดยคำนวณจากการวัดระยะระหว่างเวคเตอร์ของข้อมูลเข้า โดยใช้การวัดระยะแบบต่างๆ เช่น การวัดระยะแบบยูคลิด (Euclidean distance) การวัดระยะแบบแมนฮัตตัน (Manhattan distance) การวัดระยะแบบเชบิเชฟ (Chebychev distance)

การจัดกลุ่มข้อมูล หรือ การแบ่งกลุ่มข้อมูล จะแตกต่างจากการแบ่งประเภทข้อมูล (classification) โดยจะแบ่งกลุ่มข้อมูลจากความคล้าย โดยไม่มีการกำหนดประเภทของข้อมูลไว้ก่อนจึงกล่าวได้ว่าการแบ่งกลุ่มข้อมูล เป็นการเรียนรู้แบบไม่มีผู้สอน

ตัวอย่างวิธีการจัดกลุ่มข้อมูล ได้แก่ k-Means clustering และ Hierarchical clustering

การจัดกลุ่มข้อมูลอาจใช้เป็นขั้นตอนเบื้องต้นของการวิเคราะห์ข้อมูล เพื่อช่วยในการลดขนาดข้อมูล(แยกเป็นหลายๆกลุ่ม และคัดเฉพาะบางกลุ่มเพื่อทำการวิเคราะห์ต่อไป หรือแยกการวิเคราะห์ออกเป็นสำหรับแต่ละกลุ่ม) ก่อนที่จะนำไปวิเคราะห์ด้วยวิธีการอื่นต่อไป

อัลกอริทึมในการจัดกลุ่มข้อมูล โดยทั่วไปแบ่งได้เป็น 2 ประเภทใหญ่ๆ คือ การแบ่งแบบเป็นลำดับขั้น (hierarchical) และ การแบ่งแบบตัดเป็นส่วน (partitional) การแบ่งแบบเป็นลำดับขั้นนั้นจะทำการแบ่งกลุ่มจากกลุ่มย่อยที่ถูกแบ่งไว้ก่อนหน้านั้นซ้ำหลายครั้ง ส่วนการแบ่งแบบตัดเป็น – ส่วนนั้น การแบ่งจะทำเพียงครั้งเดียว การแบ่งแบบเป็นลำดับขั้น จะมี 2ลักษณะคือ แบบล่างขึ้นบน (bottom-up) หรือ เป็นการแบ่งแบบรวมกลุ่มจากกลุ่มย่อยให้ใหญ่ขึ้นไปเรื่อยๆ โดยเริ่มจากกลุ่มเล็กสุด คือ ในแต่ละกลุ่มมีข้อมูลเพียงค่าเดียว และ แบบบนลงล่าง (top-down) หรือ เป็นการแบ่งแบบกลุ่ม จากกลุ่มใหญ่ให้ย่อยไปเรื่อยๆ โดยเริ่มจากกลุ่มใหญ่ที่สุด คือ กลุ่มเดียวมีข้อมูลทุกค่าอยู่ในกลุ่ม

2.1.1 การวัดค่าความแตกต่างของข้อมูล

ในการจัดกลุ่มข้อมูล ข้อมูลที่นำมาจัดกลุ่มจะมีความแตกต่างกัน ความแตกต่างของข้อมูลจะต้องถูกตีค่าเป็นตัวเลข เพราะในอัลกอริทึมของการจัดกลุ่มจะต้องมีการคำนวณค่าความแตกต่างของข้อมูล ดังนั้นจะต้องมีฟังก์ชันสำหรับใช้หาค่าความแตกต่างของข้อมูล ซึ่งจะขอยกตัวอย่าง 1 ฟังก์ชัน ได้แก่

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.1.1.1 Euclidean distance

การคำนวณค่าจะใช้ฟังก์ชัน(2.1) [5]

$$d(i,j)=\sqrt{(|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+...+|x_{ip}-x_{jp}|^2)} \quad \text{..... (2.1)}$$

จากที่ได้กล่าวมาก่อนหน้านี้ว่าชุดข้อมูลมักจะถูกแปลงให้อยู่ในรูปของเวคเตอร์ ดังนั้น ฟังก์ชัน (2.1) จะเป็นการนำตำแหน่งของชุดข้อมูลมาหาค่าความแตกต่าง โดยที่ i แทนข้อมูลชุดที่ 1 และ j แทนข้อมูลชุดที่ 2 p คือจำนวนมิติของชุดข้อมูล ถ้าค่า d(i,j) ที่คำนวณได้ มีค่าน้อย แสดงว่าข้อมูลมีความคล้ายกัน

2.1.2 k-Means Clustering

อัลกอริทึมแสดงการทำงานของ k-Means Clustering แสดงดังรูปที่ 2.1 [5]

function Direct-k-means()

Initialize k prototypes $\{w_1,\ldots,w_k\}$ such that $w_j = i_l, j \in \{1,\ldots,k\}, l \in \{1,\ldots,n\}$

Each cluster C_j is associated with prototype w_j

Repeat

for each input vector i_l, where $l \in \{1,\ldots,n\}$, *do*

Assign i_l to the cluster C_j with nearest prototype w_j (i.e., $|i_l - w_j| \leq |i_l - w_j|, j' \in \{1,\ldots,k\}$)

for each cluster C_j, where $j \in \{1,\ldots,k\}$, *do*

Update the prototype w_j to be the centroid of all samples currently in C_j, so that $w_j = \sum_{i_l \in C_j} i_l / |C_j|$

Compute the error function:

$$E = \sum_{j=1}^{k} \sum_{i_l \in C_j} |i_l - w_j|^2$$

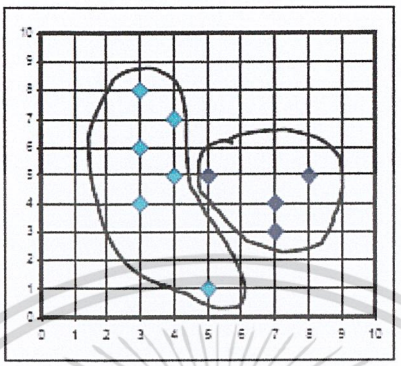
Until E does not change significantly or cluster membership no longer changes

รูปที่ 2.1 ขั้นตอนการทำงานของอัลกอริทึม k-Means Clustering

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

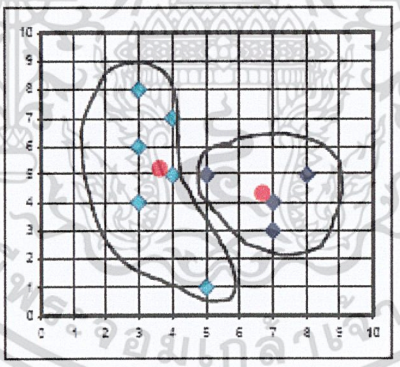
ขั้นตอนการทำงานของอัลกอริทึมจะแสดงเป็นขั้นตอนดังนี้

1) กำหนดค่า k ซึ่ง k คือ จำนวนของกลุ่มที่ต้องการแบ่ง จากรูปตัวอย่างด้านล่างจะกำหนดค่า k = 2 ดังแสดงในรูปที่ 2.2



รูปที่ 2.2 การจัดกลุ่มข้อมูลโดยวิธี k-Means Clustering ขั้นตอนที่ 1

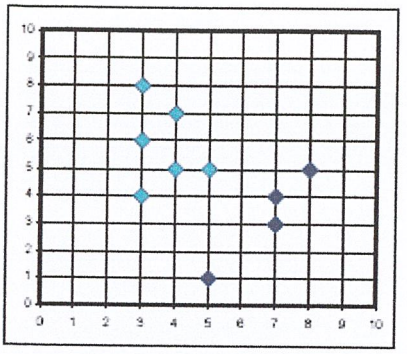
2) หาค่าตำแหน่งจุดศูนย์กลางของแต่ละกลุ่ม ดังแสดงในรูปที่ 2.3



รูปที่ 2.3 การจัดกลุ่มข้อมูลโดยวิธี k-Means Clustering ขั้นตอนที่ 2

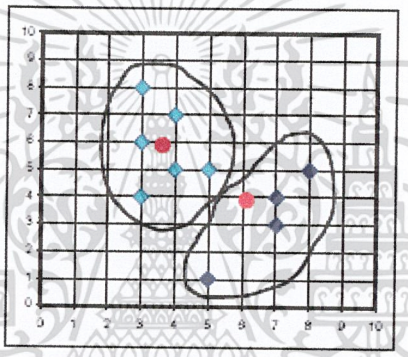
3) ดูว่ามีข้อมูลใดอยู่ใกล้จุดศูนย์กลางของกลุ่มอื่นมากกว่า กลุ่มของตนเอง ก็ให้ย้ายข้อมูลนั้นจากกลุ่มเดิมไปยังกลุ่มใหม่ที่มีระยะจากจุดศูนย์กลางมายังข้อมูลนั้นใกล้กว่า ดังแสดงในรูปที่ 2.4

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 2.4 การจัดกลุ่มข้อมูลโดยวิธี k-Means Clustering ขั้นตอนที่ 3

4) หาค่าตำแหน่งจุดศูนย์กลางของกลุ่มใหม่ ดังแสดงในรูปที่ 2.5



รูปที่ 2.5 การจัดกลุ่มข้อมูลโดยวิธี k-Means Clustering ขั้นตอนที่ 4

5) ทำวนซ้ำขั้นตอนที่ 3 กับขั้นตอนที่ 4 ไปเรื่อยๆ จนกระทั่งไม่มีการย้ายกลุ่มเกิดขึ้น ก็จะถือว่าการ Clustering ครั้งนี้เสร็จสมบูรณ์

วิธีการ k-Means Clustering จะมีข้อดี คือ สามารถใช้กับจำนวนข้อมูลที่มีปริมาณมากๆได้ โดยใช้รอบในการทำซ้ำไม่มาก แต่มีข้อเสียหลายประการ เช่น ถ้าข้อมูลไม่เป็นประเภทตัวเลขก็จะทำการแบ่งกลุ่มได้ยาก ต้องมีการกำหนดค่า k ก่อน ซึ่งจริงๆเราไม่รู้ว่าค่า k ควรจะเป็นเท่าใดจึงจะเหมาะสม ไม่ทนต่อข้อมูลรบกวน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.1.3 Hierarchical Clustering

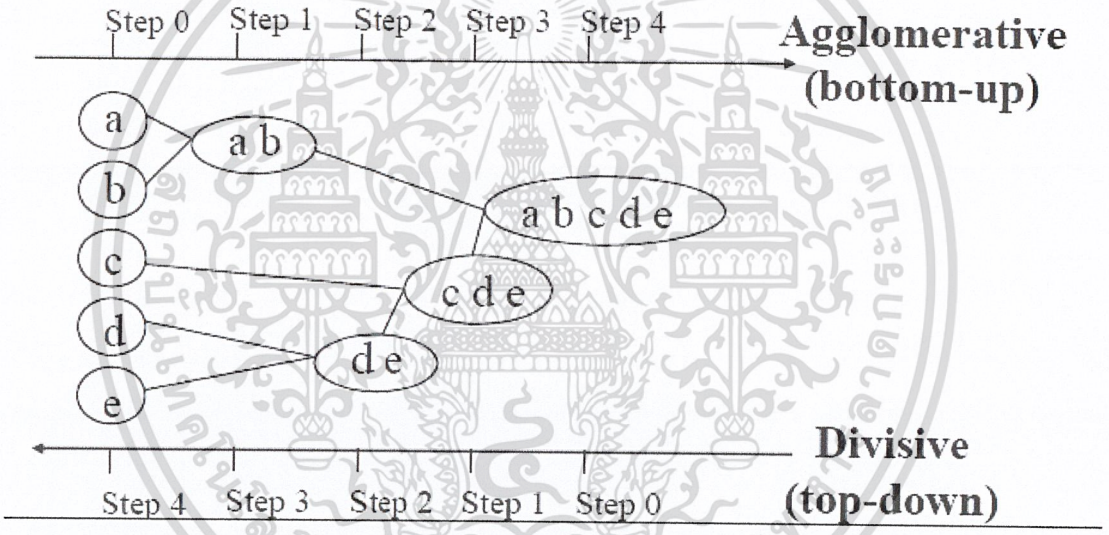
ขั้นตอนการทำงานของ Hierarchical Clustering [5] จะเป็นดังนี้

1) เริ่มต้นแต่ละกลุ่มจะมีข้อมูล 1 ตัว

2) รวมกลุ่มที่มีค่าความต่างน้อยสุดเข้าด้วยกัน ค่าความต่าง (ค่าdissimilarity) แสดงดังรูปที่ 2.7

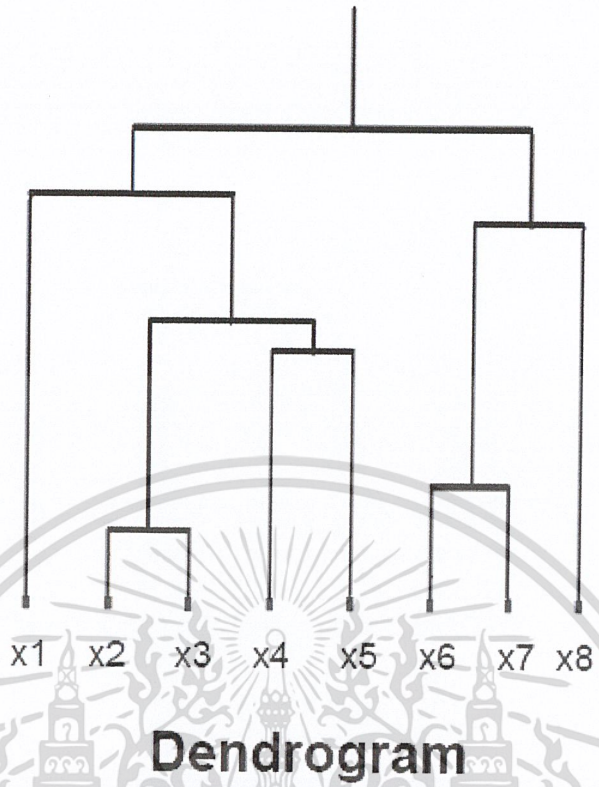
3) รวมไปเรื่อยๆจนกระทั่งกลายเป็นกลุ่มเดียว

ขั้นตอนการทำงานแสดงดังรูปที่ 2.6



รูปที่ 2.6 การจัดกลุ่มข้อมูลโดยวิธี Hierarchical Clustering

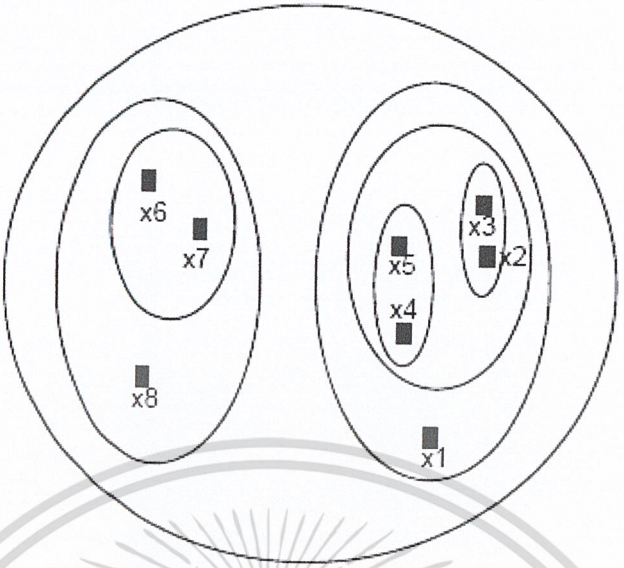
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 2.7 Dendrogram ใช้สำหรับดูค่าความต่าง (ค่าdissimilarity)

จากรูปที่ 2.7 X2 กับ X3 จะมีค่า dissimilarity น้อยที่สุด รองลงมาคือ X6 กับ X7 และ X4 กับ X5 ตามลำดับ ดังนั้นเมื่อมีการจัดกลุ่ม X2 กับ X3 จะรวมกลุ่มกันก่อน ตามด้วย X6 กับ X7 และ X4 กับ X5 ตามลำดับ หลังจากนั้นกลุ่มที่ได้จากการรวม X2 กับ X3 และกลุ่มที่ได้จากการรวม X4 กับ X5 ก็จะรวมกันต่อ และจะมีการรวมกลุ่มต่อไปเรื่อยๆ จนกลาย เป็นกลุ่มเดียว ดังแสดงในรูปที่ 2.8

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 2.8 Venn Diagram ใช้สำหรับดูการจัดกลุ่ม

วิธีการ Hierarchical Clustering จะมีข้อดี คือ ไม่ต้องกำหนดค่าของจำนวนกลุ่มที่ต้องการแบ่ง เหมือนอย่างในกรณีของ k-Means แต่จะมีข้อเสีย คือ ใช้เวลาในการแบ่งกลุ่มค่อนข้างมาก

2.1.4 Ant-based Clustering

แรกเริ่ม Ant-based Clustering and sorting algorithm ถูกคิดค้นขึ้นโดย Deneubourg [3] ในปี ค.ศ. 1990 เดิมทีถูกนำมาใช้กับหุ่นยนต์ เพื่อแบ่งกลุ่มชนิดของวัตถุที่พบเห็น แต่ได้มีการเสนอแบบจำลองใหม่ขึ้นมา โดยมดจะถูกจำลองเป็นตัวแทนอย่างง่าย ให้เคลื่อนที่แบบสุ่มในสภาพแวดล้อมที่เป็นตารางสี่เหลี่ยม เริ่มแรก Data items ที่จะนำมาแบ่งกลุ่มจะถูกกระจายอย่างสุ่มลงใน ตารางสี่เหลี่ยม โดยในตารางแต่ละช่องจะใส่ได้เพียง 1 Data item เท่านั้น และ มดสามารถที่จะเก็บ เคลื่อนย้าย และ วาง Data items ซึ่งมดจะเก็บ Data item ที่ไม่เหมือนกัน หรือ ถูกล้อมรอบด้วยข้อมูลที่ไม่เหมือนกัน และ จะวางในบริเวณที่มีข้อมูลที่เหมือนกันหรือคล้ายกัน

มดจะตัดสินใจเก็บข้อมูล หรือวางข้อมูล จะต้องดูที่ค่าความน่าจะเป็น ซึ่งหาได้จากฟังก์ชัน (2.2) และ (2.3) โดยที่ Ppick(i) แทนความน่าจะเป็นในการเก็บข้อมูล และ Pdrop(i) แทนความน่าจะเป็นในการวางข้อมูล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ ดัดแปลง เผยแพร่ หรือแสวงหาประโยชน์ทางการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น โดยไม่ได้รับอนุญาตจากเจ้าของเอกสาร และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

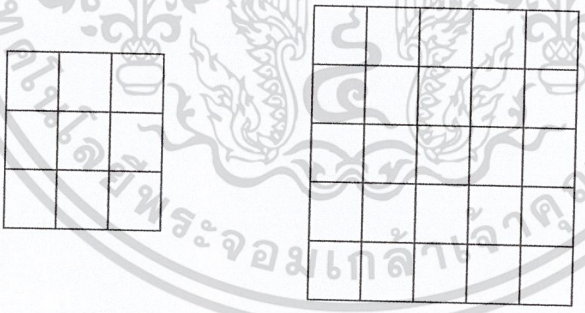
$$p_{pick}(i) = (\frac{k^+}{k^+ + f(i)})^2 \quad \text{..... (2.2)}$$

$$p_{drop}(i) = (\frac{f(i)}{k^- + f(i)})^2 \quad \text{..... (2.3)}$$

k+ และ k- เป็นค่าพารามิเตอร์ ซึ่งเป็นค่าที่บ่งบอกอิทธิพลของ neighbourhood function f(i) โดย Deneubourg ได้กำหนดค่า k+ และ k- ไว้ที่ 0.1 และ 0.3 ตามลำดับ ค่า f(i) จะเป็นค่าที่บอกความแตกต่างของข้อมูลที่วางอยู่ ณ ตำแหน่งนั้นๆ กับข้อมูลที่อยู่รอบข้าง ถ้าค่า f(i) มีค่าน้อย แสดงว่า ข้อมูลมีความแตกต่างกันมาก ในทางตรงกันข้าม ถ้าค่า f(i) มีค่ามาก แสดงว่า ข้อมูลมีความแตกต่างกันน้อย ซึ่งค่า f(i) สามารถหาได้จากฟังก์ชัน (2.4)

$$f(i) = \max\left(0, \frac{1}{\sigma^2}\sum_j (1 - \frac{d(i,j)}{\alpha})\right) \quad \text{..... (2.4)}$$

จากฟังก์ชันที่ (2.4) d(i , j) คือ dissimilarity function จะมีค่าอยู่ระหว่าง 0 ถึง 1 , α คือ ค่าพารามิเตอร์ที่เราต้องกำหนดขึ้น จะมีค่าอยู่ระหว่าง 0 ถึง 1 และ σ^2 คือ ขนาดของ local - neighbourhood ดังแสดงในรูปที่ 2.9



$\sigma^2 = 9$

$\sigma^2 = 25$

รูปที่ 2.9 แสดงขนาดของ local neighbourhood

โดยมดจะอยู่ที่ตำแหน่งศูนย์กลางของตาราง และค่า Radius of perception จะมีค่าเท่ากับ $(\sigma - 1)/2$ และต่อไปจะกล่าวถึงส่วนของ Ant-based Algorithm

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Algorithm 1 basic_ant

```
1: begin
2: INITIALISATION PHASE
3: Randomly scatter data items on the toroidal grid
4: for each j in 1 to #agents do
5:    i := random_select(remaining_items)
6:    pick_up(agent(j), i)
7:    g := random_select(remaining_empty_grid_locations)
8:    place_agent(agent(j), g)
9: end for
10: MAIN LOOP
11: for each it_ctr in 1 to #iterations do
12:    j := random_select(all_agents)
13:    step(agent(j), stepsize)
14:    i := carried_item(agent(j))
15:    drop := drop_item?(f*(i)) // see equations 2 and 3
16:    if drop = TRUE then
17:       while pick = FALSE do
18:          i := random_select(free_data_items)
19:          pick := pick_item?(f*(i)) // see equations 1 and 3
20:       end while
21:    end if
22: end for
23: end
```

รูปที่ 2.10 ขั้นตอนการทำงานของอัลกอริทึม Ant-based Clustering

2.1.4.1 อัลกอริทึมพื้นฐานของ Ant-based Clustering

The basic ant algorithm (รูปที่ 2.10) เริ่มต้นด้วย Initialisation phase ซึ่งข้อมูลทั้งหมดจะถูกกระจายลงในตารางสี่เหลี่ยม มดแต่ละตัวจะสุ่มเก็บข้อมูลขึ้นมา 1 ชิ้น มดแต่ละตัวจะสุ่มหาตำแหน่งที่จะไปตั้งอยู่ในตารางสี่เหลี่ยม ต่อมาจะเข้าสู่ sorting phase มดจะถูกสุ่มเลือกขึ้นมา 1 ตัว มดจะเดินไปตามทิศทางที่ถูกกำหนดอย่างสุ่ม เมื่อไปถึงอีกตำแหน่ง มดจะตัดสินใจวางข้อมูลหรือไม่ จะดูที่ค่าความน่าจะเป็น ในกรณีที่ตัดสินใจวางข้อมูล มดจะวางข้อมูลในตำแหน่งปัจจุบันที่อยู่ (grid cell ที่อยู่ปัจจุบันจะต้องไม่มีข้อมูลอื่นอยู่ก่อนหน้า) หรือในตำแหน่งรอบๆที่ใกล้สุด โดยการสุ่ม เมื่อมดวางข้อมูลแล้วจะต้องค้นหาข้อมูลใหม่ที่จะเก็บทันที ซึ่งจะต้องใช้ index โดยจะมีการเก็บตำแหน่งของข้อมูลทั้งหมด ที่ไม่มีมดถืออยู่ มดจะต้องสุ่มเลือกข้อมูล 1 ชิ้น ที่ไม่มี index แล้วเคลื่อนที่ไปที่ตำแหน่งนั้นๆ และหาค่า neighbourhood function f*(i) เพื่อนำไปหาค่าความน่าจะเป็น ในการตัดสินใจว่าจะเก็บข้อมูลนั้นหรือไม่ มันจะค้นหาไปเรื่อยๆ จนกระทั่งสามารถเก็บข้อมูลได้สำเร็จ loop นี้จะถูกทำซ้ำด้วยมดตัวอื่นๆ

มดจะตัดสินใจเก็บข้อมูล หรือวางข้อมูล จะต้องดูที่ค่าความน่าจะเป็น ซึ่งหาได้จากฟังก์ชัน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

(2.5) และ (2.6) โดยที่ Ppick(i) แทนความน่าจะเป็นในการเก็บข้อมูล และ Pdrop(i) แทนความน่าจะเป็นในการวางข้อมูล

$$p^*_{pick}(i) = \begin{cases} 1.0 & \text{if} f^*(i) \leq 1.0 \\ \frac{1}{f^*(i)^2} & \text{else} \end{cases} \quad \text{..... (2.5)}$$

$$p^*_{drop}(i) = \begin{cases} 1.0 & \text{if} f^*(i) \geq 1.0 \\ f^*(i)^4 & \text{else,} \end{cases} \quad \text{..... (2.6)}$$

ฟังก์ชัน (2.7) เป็น neighbourhood function ที่ถูกแก้ไขโดย Lumer and Faieta

$$f^*(i) = \begin{cases} \frac{1}{\sigma^2} \sum_j (1 - \frac{d(i,j)}{\alpha}) & \text{if } (f^*(i) > 0 \wedge \forall j \ (1 - \frac{d(i,j)}{\alpha}) > 0) \\ 0 & \text{otherwise} \end{cases} \quad \text{..... (2.7)}$$

neighbourhood function ด้านบนจะแตกต่างจาก original neighbourhood function ตรงที่ มีการเพิ่มเงื่อนไข $\forall_j \left(1 - \frac{d(i,j)}{\alpha}\right) > 0$ ซึ่งจะเป็นการปรับปรุง Spatial separation ระหว่างกลุ่มให้ดีขึ้น ซึ่งจะได้กล่าวถึงต่อไป

2.1.4.2 Short-term memory

เวอร์ชันที่ถูกแก้ไขของ Short-term memory ถูกคิดขึ้นโดย Lumer และ Faieta ได้ถูกนำมาใช้ โดยที่มดแต่ละตัวสามารถจำข้อมูลที่ตนเองเคยขน และตำแหน่งที่เคยวางข้อมูลนั้นๆตามลำดับ เมื่อข้อมูลใหม่ถูกเก็บขึ้นมา มดจะนำข้อมูลใน memory มาช่วยในการตัดสินใจเลือกเส้นทางเดิน โดยจะเลือกเดินไปทางที่มีข้อมูลลักษณะเหมือนกันหรือคล้ายกัน ซึ่งจะดูจากค่า d(i , j)

ในระบบที่มีมดเยอะๆ ข้อมูลที่ถูกเก็บในหน่วยความจำ อาจจะมีการถูกลบไปบ้าง เพื่อให้การเลือกเส้นทางมีประสิทธิภาพมากขึ้นเราจะอนุญาตให้มดใช้ประโยชน์จากหน่วยความจำของมัน ดังนี้ ถ้ามดอยู่ที่ตำแหน่ง grid cell p และกำลังขนข้อมูล i หน่วยความจำของมันจะถูกใช้ดำเนินการ โดยตำแหน่งที่จำได้ทั้งหมดจะถูกนำมาหาค่า f*(i) เพื่อที่จะได้รู้ว่ามีตำแหน่งที่เหมาะสมที่จะนำข้อมูลไปวางหรือไม่ ต่อมามดก็จะกลับไปที่จุด p

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

นอกเหนือจากตำแหน่งที่ถูกหาค่า ก็อาจจะมีบาง grid cell ที่ให้ค่า neighbourhood function สูงสุด มดสามารถที่จะกระโดดไปที่ grid cell นั้นได้ โดยที่ไม่ต้องใช้การเดินอย่างสุ่ม การกระ-โดดนี้จะทำได้ด้วยค่าความน่าจะเป็นบางค่าเท่านั้น ซึ่งจะใช้สูตร dropping operation p*drop (i) ในการหาค่าความน่าจะเป็นนั้นๆ แต่ถ้าไม่มีการกระโดด มดก็จะสุ่มหาตำแหน่งในการวางข้อมูลไปเรื่อยๆ จนกระทั่งวางได้สำเร็จ

2.1.4.3 Increasing radius of perception

มดสามารถรับรู้ขนาดของ local neighbourhood ซึ่งจะจำกัดการใช้ข้อมูลระหว่าง sorting - process ดังนั้นจึงน่าสนใจที่จะใช้ Neighbourhood ขนาดใหญ่ เพื่อปรับปรุงคุณภาพของการ - Clustering และ Sorting บน grid อย่างไรก็ตาม การใช้ Neighbourhood ที่กว้าง ไม่เพียงแต่จะแพงมากเท่านั้น แต่ยังทำให้การจัดกลุ่มทำได้ช้า ระหว่างการเรียงลำดับเฟสแรก

เราจึงใช้ Radius of perception ที่ค่อยๆ เพิ่มขึ้นในช่วงเวลา สิ่งนี้จะช่วยควบคุมการคำนวณในขั้นตอนแรกของการจัดกลุ่ม และป้องกันปัญหาการสร้างกลุ่มแรก ในขณะเดียวกันจะช่วยเร่งให้กลุ่มเล็กๆ ถูกสร้างขึ้นก่อน ในการดำเนินงานปัจจุบัน เราจะเริ่มต้นค่า Perceptive radius ด้วย 1 และเพิ่มขึ้นทีละ 1 จนถึง 5 ในขณะที่กำลังทำเช่นนั้น ซึ่งจะละไว้ว่า $1/\sigma^2$ ในฟังก์ชัน(2.7) ไม่ถูกเปลี่ยน การเพิ่มขึ้นของ Perceptive radius จะส่งผลเสียต่อ Spatial separation

วิธีนี้จะนำไปสู่การค่อยๆเปลี่ยนช่วงของค่า f*(i) ในเฟสเริ่มต้นของ algorithm ค่า f*(i) จะถูกจำกัดอยู่ในช่วง 0 ถึง 1 แต่การเพิ่มค่าของ neighbourhood radius ส่งผลให้ขอบเขตบนของ f*(i) มีค่าเพิ่มขึ้น การเลื่อนของค่า f*(i) รวมกับการใช้สูตรฟังก์ชัน (2.5) และ (2.6) จะส่งผลให้ค่า Impact of density ของสูตรการวางลดลง แต่สูตรการเก็บจะเพิ่มขึ้น สิ่งนี้จะส่งผลให้ Spatial - Separation ระหว่างกลุ่มถูกปรับปรุง

2.1.4.4 Spatial separation

ตามที่ระบุไว้ข้างต้น การแยกพื้นที่ของกลุ่มในตารางเป็นสิ่งสำคัญ เพื่อให้กลุ่มแต่ละกลุ่มถูกกำหนดอย่างดี ความใกล้ชิดของพื้นที่เมื่อเกิดในระดับที่มาก จะทำให้เกิดการรวมกลุ่มของกลุ่มแรกเร็วขึ้น เพราะว่ากลุ่มต่างๆ มีแนวโน้มที่จะตั้งที่ใดก็ตาม ที่มีพื้นที่แน่นหนาของข้อมูลที่คล้ายกันและหลังจากนั้นกลุ่มต่างๆ เหล่านี้ มีแนวโน้มที่จะเคลื่อนย้ายอย่างช้ามากบนตาราง หลังจากเฟสเริ่มต้นของการแบ่งกลุ่ม ค่าของ $1/\sigma^2$ ในฟังก์ชัน (2.7) จะถูกแทนด้วย $\frac{1}{N_{occ}}$ โดยที่ N_{occ} คือค่า Actual observed number ของ Cell ที่มีข้อมูลวางอยู่ ดังนั้นความเหมือนกันเท่านั้นที่จะนำมาพิจารณา จะไม่นำความหนาแน่นมาพิจารณา ซึ่งจะส่งผลต่อการกระจายของข้อมูลบนตาราง ข้อมูลที่เป็นของกลุ่มที่แตกต่างกัน จะครอบครองพื้นที่แต่ละส่วนบนตาราง ผลที่ตามมาคือ จะเกิดการจัดตั้งกลุ่มอย่างชัดเจนขึ้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.1.4.5 Parameter settings

ใน Ant-based Clustering จะต้องมีการตั้งค่าของพารามิเตอร์หลายตัว เช่น จำนวนของมด จำนวนของ Cell ในตาราง ขนาดของ Short-term memory ของมด เวลาในการเริ่มและจบ

2.1.4.5.1 Parameters to be set as a function of the size of the data set

มีพารามิเตอร์หลายตัวที่ควรจะถูกกำหนดตามขนาดของข้อมูล ไม่เช่นนั้น จะทำให้ความเร็วเสียไป เช่น ถ้ามีข้อมูล N items (ในตารางมีทั้งหมด M cells) ในตารางควรจะต้องมีพื้นที่ว่างมากเท่าไหร่ จึงจะทำให้การวางข้อมูลเกิดขึ้นอย่างรวดเร็ว (ในแต่ละCellจะใส่ได้เพียง 1 data item เท่านั้น) ค่าที่ดีที่ได้จากการทดลอง อัตราส่วนระหว่าง N ต่อ M ควรจะมีค่าไม่เกิน 1 ต่อ 10 ดังนั้น ถ้าเราใช้ตารางแบบสี่เหลี่ยมจัตุรัส ก็ควรจะมีขนาดเท่ากับ $\sqrt{10N} \times \sqrt{10N}$ Cells และค่า Stepsize ควรจะมีค่าเท่ากับ $\sqrt{20N}$ และจำนวนการทำซ้ำจะเพิ่มขึ้นตามขนาดของชุดข้อมูล

2.1.4.5.2 Activity-based α -adaptation

ปัญหาหนึ่งที่ถูกยกมาคือ การกำหนดอย่างอัตโนมัติของค่า α (α คือ พารามิเตอร์ที่เป็นตัวบอกขนาดความแตกต่าง ภายใน neighbourhood function f*(i)) ในช่วงกระบวนการ Sorting α จะกำหนดเปอร์เซ็นต์ของ Data items ในตารางที่ถูกแบ่งว่าคล้ายกัน ตัวอย่างเช่น ถ้าเลือกค่า α น้อยเกิน จะทำให้การจัดกลุ่มเกิดขึ้นยาก ในทางตรงกันข้าม ถ้าเลือกค่า α มากเกิน ก็จะทำให้รวมกลุ่มได้ง่าย

การเลือกค่า α ที่ไม่ดี จะส่งผลกระทบต่อกิจกรรมบนตาราง แต่ค่า α จะสามารถปรับตัวอย่างอัตโนมัติเมื่อกิจกรรมดำเนินไปมากๆ โดยจะดูที่ความถี่ในการเก็บหรือวางข้อมูลสำเร็จ ซึ่งจะกล่าวถึงต่อไป

มดแต่ละตัวจะมีค่า α ของตัวเองที่ได้จากการสุ่ม โดยจะมีค่าอยู่ในช่วง 0 ถึง 1 มดแต่ละตัวจะพิจารณาปรับค่า α ของตัวเอง หลังจากเคลื่อนที่ไปแล้ว N_{active} moves ในระหว่างนี้ มันก็จะเก็บค่าจำนวนครั้งที่วางล้มเหลว N_{fail} ซึ่งอัตราส่วนความล้มเหลว จะถูกแทนด้วย $r_{fail} = \frac{N_{fail}}{N_{active}}$ ในที่นี้ N_{active} จะถูกกำหนดไว้ที่ 100 การปรับค่า α จะเป็นไปตามกฎ ใน (2.8)

$$\alpha \leftarrow \begin{cases} \alpha + 0.01 & \text{if } r_{fail} > 0.99 \\ \alpha - 0.01 & \text{if } r_{fail} \leq 0.99 \end{cases} \quad \text{..... (2.8)}$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากการทดลอง α จะมีการปรับค่าในช่วงกระบวนการ Sorting ทั้งหมด สรุปก็คือ การปรับค่าของ α จะช่วยเพิ่มประสิทธิภาพในการแบ่งกลุ่มของข้อมูล

2.1.5 การประเมินผล (Evaluation)

เมื่อได้ทำการ Clustering ข้อมูล ด้วยวิธีการทั้ง 3 แบบ ที่ได้กล่าวมาข้างต้น ก็จะต้องมีตัววัดคุณภาพ และ ประสิทธิภาพ ของการแบ่งกลุ่ม ในที่นี้ จะขอยกตัวอย่าง ตัววัดประสิทธิภาพ 2 ตัว ได้แก่

2.1.5.1 Inner Cluster Variance

จะเป็นการหาค่าผลรวมของค่าเบี่ยงเบนยกกำลังสอง ระหว่างรายการข้อมูลทั้งหมด กับค่าจุดศูนย์กลางของกลุ่มที่รายการข้อมูลนั้นๆ อยู่ ซึ่งจะบ่งบอกว่า การแบ่งกลุ่มที่ดี รายการข้อมูลภายในกลุ่มเดียวกัน ควรจะมีความแตกต่างกันน้อย หมายความว่า ถ้าค่า I ที่คำนวณได้จากฟังก์ชัน(2.9) ยิ่งมีค่าน้อย แสดงว่าการจัดกลุ่มจะยิ่งมีประสิทธิภาพมากขึ้น

$$I = \sum_{c \in C} \sum_{i \in c} \delta(i, \mu_c)^2 \qquad \text{..... (2.9)}$$

จากฟังก์ชัน (2.9) i คือ รายการข้อมูลใดๆ μ_c คือ จุดศูนย์กลางของกลุ่มที่รายการข้อมูลนั้นๆอยู่

2.1.5.2 Dunn Index

จะเป็นการหาค่าเส้นผ่านศูนย์กลางของแต่ละกลุ่ม และระยะระหว่างจุดศูนย์กลางของแต่ละกลุ่ม ตัววัดตัวนี้ จะชี้ว่า การแบ่งกลุ่มที่ชัดเจนจะเกิดขึ้น ถ้าระยะ ระหว่างจุดศูนย์กลางของแต่ละกลุ่ม มีค่ามากกว่า ระยะเส้นผ่านศูนย์กลางของแต่ละกลุ่ม มากๆ ซึ่งคำนวณได้จากฟังก์ชัน (2.10)

$$D = \min_{c,d \in C} \left[\frac{\delta(\mu_c, \mu_d)}{\max_{e \in C} [diam(e)]} \right] \qquad \text{..... (2.10)}$$

จากฟังก์ชัน (2.10) diam(e) คือ ระยะเส้นผ่านศูนย์กลางของแต่ละกลุ่ม μ_c μ_d คือ จุดศูนย์กลางของกลุ่ม ที่จะนำมาหาค่าระยะห่าง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.1.5.3 Pearson correlation

แผนภาพกระจัดกระจายแสดงความสัมพันธ์ระหว่างข้อมูล 2 ชุดคือ X และ Y ดังแสดงในรูปที่ 2.11 จะบ่งบอกถึงความสัมพันธ์ในรูปแบบต่างๆ อย่างในรูปที่ 2.11(a) จะเป็นความสัมพันธ์ทางบวกระหว่าง X และ Y นั่นคือในทุกๆคู่ของค่า X และ Y ที่พล็อตลงไปนั้น เมื่อค่า X มีค่าเพิ่มขึ้นแล้วค่า Y ก็จะมีค่าเพิ่มขึ้นด้วย ส่วนในรูปที่ 2.11(b) จะแสดงการเพิ่มของค่า X และ Y ในทิศทางตรงกันข้าม ก็คือเมื่อ X มีค่าเพิ่มขึ้นแล้ว Y จะมีค่าลดลง จะเรียกว่ามีความสัมพันธ์ทางลบ ส่วนในรูปที่ 2.11(c) แสดงความไม่สัมพันธ์กันระหว่าง X และ Y ก็คือเมื่อX เพิ่มขึ้น Y อาจจะลดหรือเพิ่มขึ้นก็ได้

ให้ $(X_1, Y_1), (X_2, Y_2), \ldots,(X_n, Y_n)$ แทนคู่ของตัวแปร (X, Y) ในรูปที่ 2.11(a) จะมีเส้นตรงตั้งฉากกับแกนนอนที่ $\bar{X}$ และเส้นตรงตั้งฉากกับแกนตั้งที่ $\bar{Y}$ แบ่งแผนภาพออกเป็น 4 ส่วน ในส่วนที่ 1 ทั้งค่า X และ Y จะมีค่าสูงกว่าค่าเฉลี่ย ดังนั้นส่วนเบี่ยงเบน $X - \bar{X}$ และ $Y - \bar{Y}$ จะมีค่าเป็นบวก ดังนั้น $(X - \bar{X})(Y - \bar{Y})$ จึงมีค่าเป็นบวก เหมือนกับในส่วนที่ 3 ที่มีค่าเป็นบวกเช่นเดียวกัน เพราะส่วนเบี่ยงเบน $X - \bar{X}$ และ $Y - \bar{Y}$ มีค่าติดลบ เมื่อนำมาคูณกันจึงมีค่าเป็นบวก ในส่วนที่ 2 และ 4 จะมีค่า $(X - \bar{X})(Y - \bar{Y})$ เป็นลบ ในรูปที่ 2.11(a) นี้ เกือบทุกจุดจะอยู่ในส่วนที่ 1 และ 3 ซึ่งจะให้ผลเป็นบวก ดังนั้น $\Sigma(X - \bar{X})(Y - \bar{Y})$ จึงมีค่าเป็นบวก

เหตุผลเดียวกัน ในรูปที่ 2.11(b) ซึ่งจะแสดงความสัมพันธ์ทางลบ เกือบทุกจุดจะอยู่ในส่วนที่ 2 และ 4 ซึ่งจะมีค่า $(X - \bar{X})(Y - \bar{Y})$ เป็นลบ ดังนั้น $\Sigma(X - \bar{X})(Y - \bar{Y})$ จึงมีค่าเป็นลบ ส่วนในรูปที่ 2.11(c) ทุกๆจุดจะอยู่ในส่วนที่ 1 ถึง 4 ในปริมาณที่พอๆ กัน ดังนั้นผลของ $\Sigma(X - \bar{X})(Y - \bar{Y})$ จึงมีค่าเข้าใกล้ศูนย์ โดยสรุปแล้ว $\Sigma(X - \bar{X})(Y - \bar{Y})$ สามารถบอกระดับของความสัมพันธ์ระหว่างตัวแปร X และ Y ได้ว่ามีความสัมพันธ์ทางบวก มีความสัมพันธ์ทางลบ หรือไม่มีความสัมพันธ์

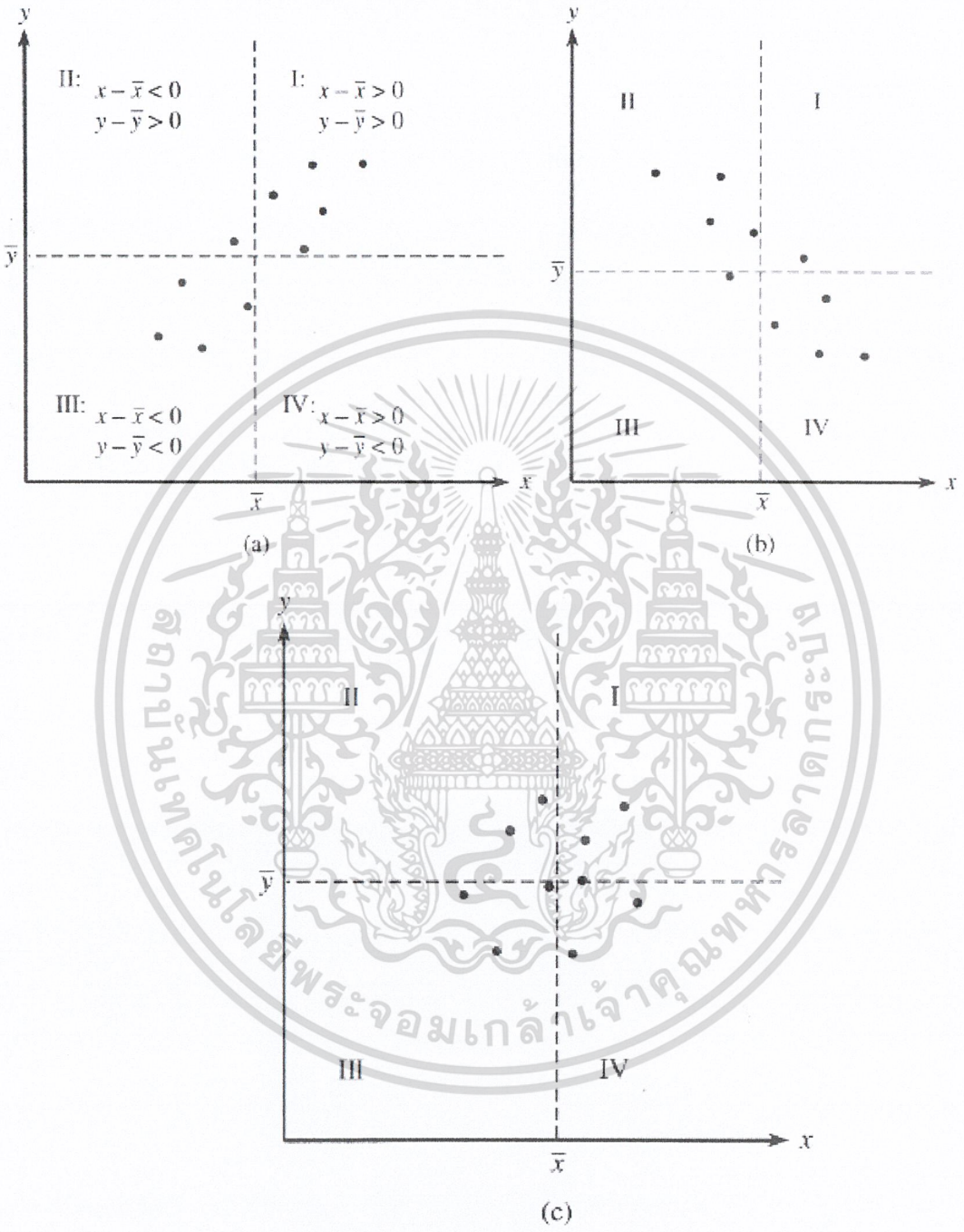
ต่อมาได้มีการนำ $\Sigma(X - \bar{X})(Y - \bar{Y})$ มาพัฒนาเป็นสมการในการคำนวณหาความสัมพันธ์ที่เป็นที่นิยมกันอย่างแพร่หลาย เรียกว่า "สัมประสิทธิ์สหสัมพันธ์เพียร์สัน"

เมื่อ $\bar{X}$ และ S_X แทนค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของค่า X ในคู่ของ (X_1, Y_1), $(X_2, Y_2), \ldots,(X_n, Y_n)$ และให้ $\bar{Y}$ และ S_Y แทนค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของค่า Y แล้ว สัมประสิทธิ์สหสัมพันธ์เพียร์สัน (r) เขียนเป็นสมการได้ดังฟังก์ชัน (2.11)

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{(N - 1)S_X S_Y} \qquad \ldots\ldots\ldots\ldots\ldots (2.11)$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สำนักหอสมุดกลาง พระจอมเกล้าลาดกระบัง



รูปที่ 2.11 แผนภาพกระจัดกระจายแสดงความสัมพันธ์ระหว่างตัวแปร x และ y

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

117184

บทที่ 3
โครงสร้างและหลักการทำงานของโปรแกรมที่ศึกษา

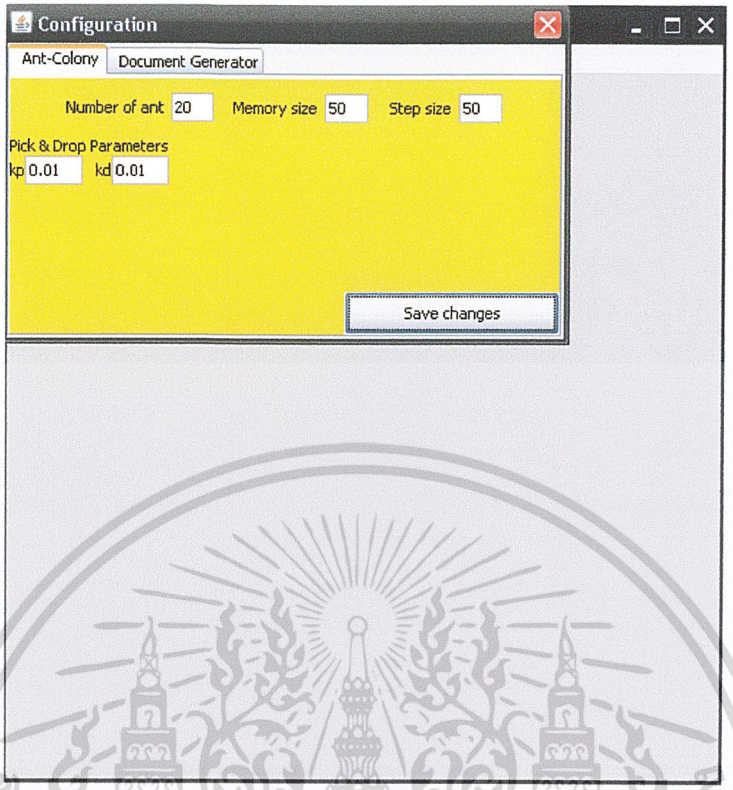
เนื่องจากโครงงานนี้ได้ทำการศึกษางานวิจัยของ Julia Handl และ Bernd Meyer [2] ซึ่งมีหลักการทำงานของโปรแกรมแบ่งเป็น 2 ส่วน คือ ส่วนของขั้นตอนวิธี Ant-based Clustering และส่วนของการติดต่อกับผู้ใช้งาน โดยมีโครงสร้างสามารถอธิบายได้ดังนี้

3.1 รูปแบบอินเตอร์เฟสที่ใช้ในการติดต่อกับผู้ใช้งาน

รูปที่ 3.1 รูปแบบอินเตอร์เฟสที่ใช้ในการติดต่อกับผู้ใช้งานในหน้าหลัก

สามารถปรับค่าพารามิเตอร์ต่างๆ ได้ตามต้องการ เช่น ค่าความน่าจะเป็นในการที่มดจะเก็บหรือวางข้อมูล ค่าจำนวนของมดที่ใช้ในการจัดกลุ่ม ชุดข้อมูลที่จะใช้ในการแบ่งกลุ่ม โดยคลิกเลือกที่ Configuration > Modify และทำการปรับค่าต่างๆ ดังแสดงในรูปที่ 3.2 เมื่อปรับค่าเสร็จสิ้นแล้ว สามารถคลิกเลือกที่ Data > Generate Test Data เพื่อทำการจำลองการจัดกลุ่มข้อมูลต่อไปได้ ขณะที่กำลัง Generate ถ้าต้องการหยุดการ Generate สามารถคลิกที่ Ants > Stop Sorting เพื่อหยุดการ Generate ได้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 3.2 รูปแบบอินเตอร์เฟสที่ใช้ในการติดต่อกับผู้ใช้งานในการกำหนดพารามิเตอร์

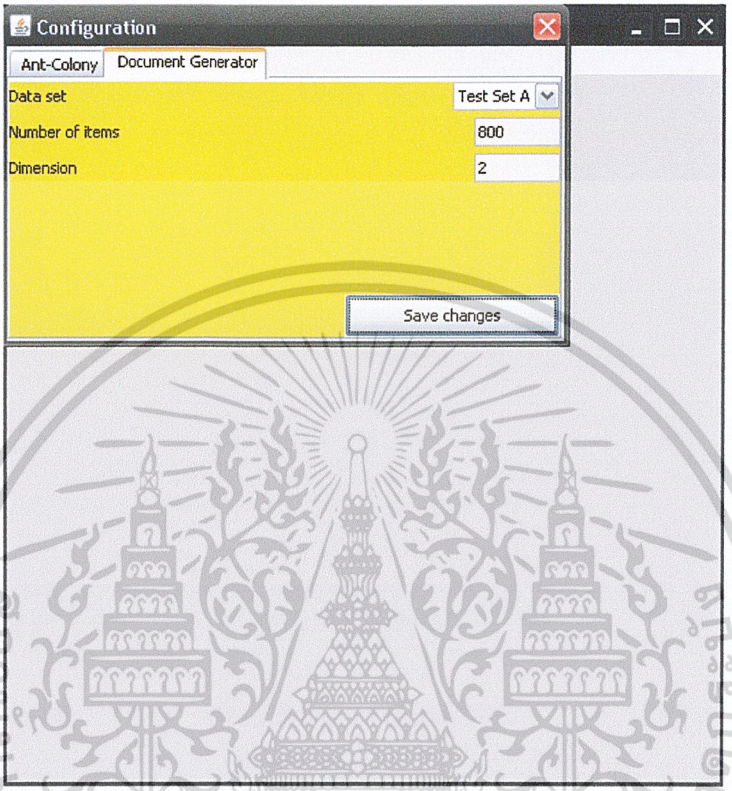
3.1.1 ค่าพารามิเตอร์ต่างๆ

เมื่อผู้ใช้ต้องการปรับค่าพารามิเตอร์ในส่วนของมด สามารถคลิก Configuration > Modify > Ant-Colony จะแสดงหน้าจอดังรูปที่ 3.2 ซึ่งจะมีพารามิเตอร์ให้ใส่ดังนี้

- Number of ant ใช้ในการกำหนดจำนวนมด
- Memory size ใช้ในการกำหนดขนาดหน่วยความจำให้มด
- Step size ใช้ในการกำหนดความเร็วของมด
- kp ใช้ในการกำหนดค่าความน่าจะเป็นในการที่มดจะเก็บข้อมูล
- kd ใช้ในการกำหนดค่าความน่าจะเป็นในการที่มดจะวางข้อมูล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เมื่อผู้ใช้ต้องการเลือกชุดข้อมูลที่จะนำมาจัดกลุ่ม สามารถคลิก Configuration > Modify > Document Generator จะแสดงหน้าจอดังรูปที่ 3.3 ซึ่งจะสามารถคลิกเลือก Data set ที่จะนำมาใช้ทดลองจัดกลุ่มได้

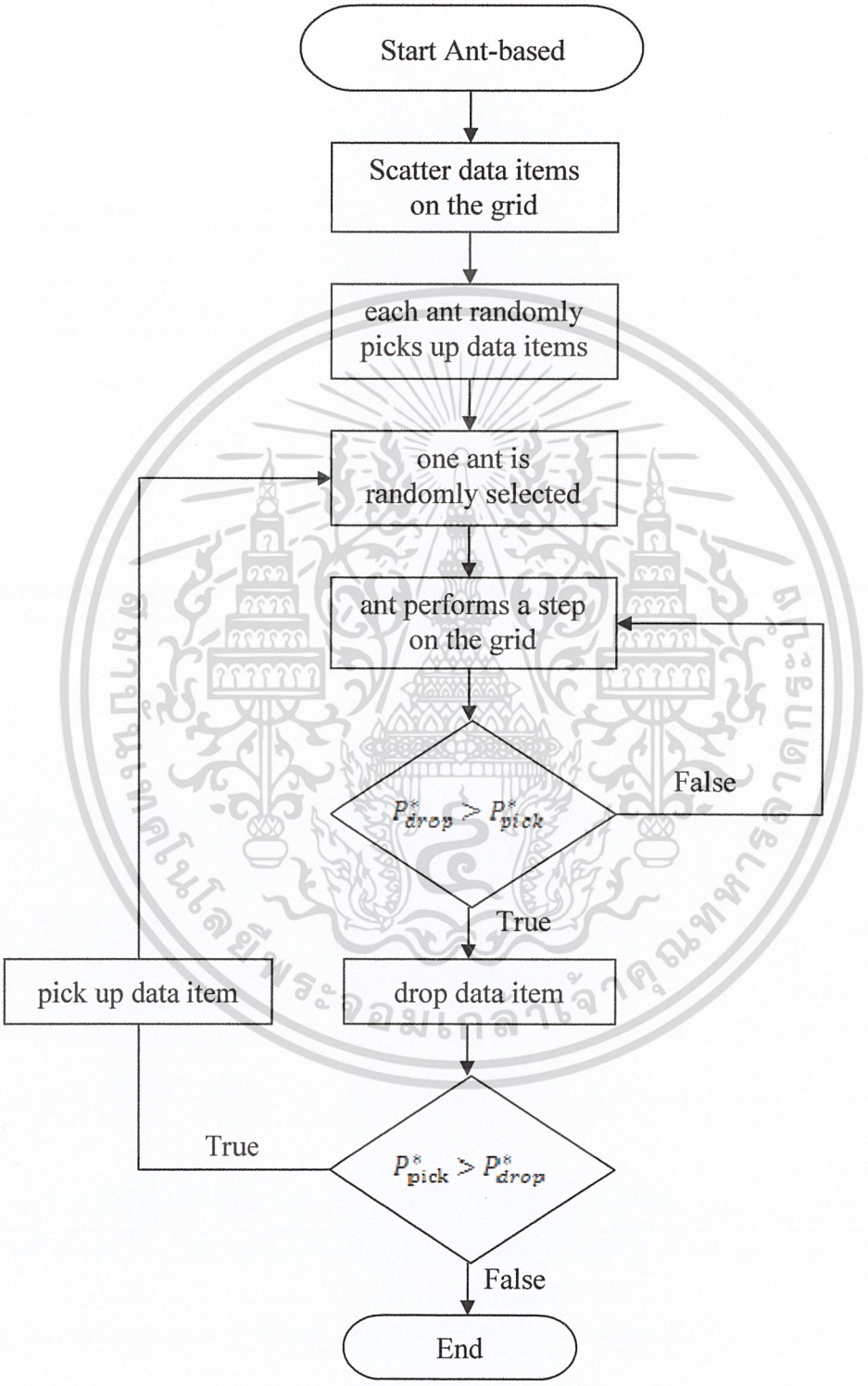


รูปที่ 3.3 รูปแบบอินเตอร์เฟสที่ใช้ในการติดต่อกับผู้ใช้งานในการเลือกชุดข้อมูล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.2 Flowchart

3.2.1 หลักการทำงานของอัลกอริทึม



รูปที่ 3.4 ระบบการทำงานของอัลกอริทึม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Ant – based Algorithm เริ่มต้นจะทำการกระจาย data items ลงในตารางสี่เหลี่ยม แล้วปล่อยมดทั้งหมดลงมาเก็บ data items ในตาราง จากนั้นจะสุ่มมดขึ้นมา 1 ตัว ให้เดินไป 1 ก้าว ในตาราง โดยการสุ่มทิศทางที่จะเดิน จากนั้นจะดูค่าความน่าจะเป็น ถ้าความน่าจะเป็นในการวางมากกว่าการเก็บ ก็จะทำการวางข้อมูล และจะต้องหาข้อมูลใหม่เพื่อจะเก็บในทันที แต่ถ้าความน่าจะเป็นในการเก็บมากกว่าการวาง ก็จะวางข้อมูลไม่ได้ และจะต้องเดินต่อไปเพื่อหาตำแหน่งที่จะวางข้อมูลให้ได้ ในกรณีที่มดทุกตัววางข้อมูลได้แล้วแต่ไม่สามารถที่จะเก็บข้อมูลใหม่ขึ้นมาได้ แสดงว่ากระบวนการจัดกลุ่มข้อมูลได้เสร็จสิ้นลงแล้ว ดังแสดงในรูปที่ 3.5 ถึง 3.7

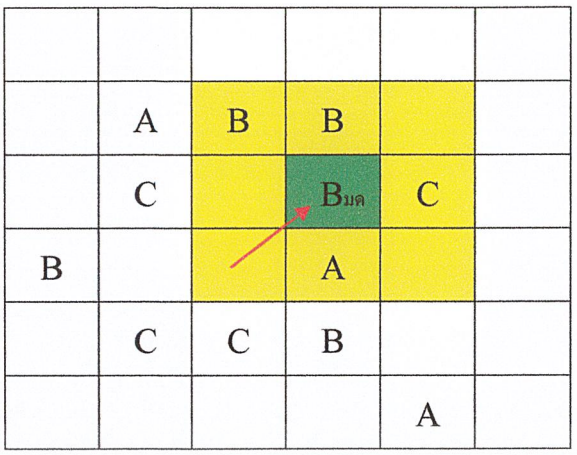
	A	B	B		
	C			C	
B		B	A		
	C	C	B		
				A	

รูปที่ 3.5 ตัวอย่างการกระจายข้อมูลลงในตาราง

	A	B	B		
	C			C	
B		$B_{มด}$	A		
	C	C	B		
				A	

รูปที่ 3.6 ตัวอย่างการกระจายข้อมูลลงในตาราง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 3.7 ตัวอย่างการกระจายข้อมูลลงในตาราง

จากรูปที่ 3.5 เป็นการแสดงการกระจายข้อมูลโดยใช้การสุ่ม ในขั้นตอนเริ่มต้นของกระ--บวนการจัดกลุ่มข้อมูล

จากรูปที่ 3.6 เมื่อมดเดินมาถึงตำแหน่งดังรูป มดจะดูเซลล์รอบข้าง (8 เซลล์) ในกรณีนี้ตำแหน่งที่มดอยู่ปัจจุบันมีข้อมูล B อยู่ แต่รอบข้างมีจำนวนข้อมูลของ C ถึง 3 ค่า ขณะที่มีข้อมูล B เพียงแค่ 1 ค่า มดจึงตัดสินใจเก็บข้อมูล B ขึ้นมา แล้วทำการเคลื่อนย้ายข้อมูลไปยังตำแหน่งอื่น โดยการเคลื่อนย้ายตำแหน่งของมดทุกครั้งจะใช้การสุ่ม

จากรูปที่ 3.7 เมื่อมดทำการเคลื่อนย้ายข้อมูล B มายังตำแหน่งดังรูป มดจะดูเซลล์รอบข้าง (8 เซลล์) ในกรณีนี้ข้อมูลรอบข้างที่มีจำนวนมากที่สุด คือ B ดังนั้นมดจึงตัดสินใจวางข้อมูล B ลงในตำแหน่งปัจจุบันที่อยู่ หลังจากนั้นมดก็จะเคลื่อนที่โดยการสุ่มไปยังตำแหน่งอื่นต่อไป

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 4

การทดลองและผลการทดลอง

4.1 การทดลอง

ทำการทดลองโปรแกรม Ant-based clustering ที่ศึกษา [2] โดยใช้ข้อมูล Test Set 3 ชุด ในการทดลอง ได้แก่ Test Set A Test Set B Test Set C

4.2 ผลการทดลอง

4.2.1 Test Set A

เป็นข้อมูลชนิดตัวเลข ได้แก่ 1 2 3 4 โดยหลังการทดลองข้อมูลจะถูกจัดเป็น 4 กลุ่ม ในการทดลองได้ทำการปรับค่า kp และ kd ให้มีค่าระหว่าง 0.01 ถึง 0.04 รวมทดลองทั้งหมด 16 ครั้ง ได้ผลการทดลองดังตารางที่ 4.1 โดยผลตัวเลขที่นิยม 4 ตำแหน่งที่ได้ คือค่า Pearson correlation ซึ่งผลการจำลองการจัดกลุ่มจะแบ่งออกเป็น 3 ลักษณะหลักๆ ได้แก่ ค่า kp เท่ากับ kd ค่า kp น้อยกว่า kd และ ค่า kp มากกว่า kd โดยจะแสดงตัวอย่างการจำลองการจัดกลุ่มข้อมูลกรณีละ 2 ตัวอย่าง รวมทั้งหมด 6 ตัวอย่าง ซึ่งจะแสดงในหัวข้อที่ 4.2.1.1 ถึง 4.2.1.6 ต่อไป

ตารางที่ 4.1 ผลค่า Pearson correlation ที่ได้จากการทดลอง

	kd=0.01	kd=0.02	kd=0.03	kd=0.04
kp=0.01	0.4506	0.3670	0.4097	0.5347
kp=0.02	0.5686	0.4281	0.4891	0.5096
kp=0.03	0.3774	0.4147	0.4332	0.4688
kp=0.04	0.3699	0.4700	0.5056	0.4512

หมายเหตุ ตัวเลขสีแดงแสดงถึงค่า Pearson correlation ของ 6 ตัวอย่าง ที่จะแสดงการจำลองจัดกลุ่มในหัวข้อที่ 4.2.1.1 ถึง 4.2.1.6

ในตารางที่ 4.1 ผลค่า Pearson correlation ที่ได้จากการทดลอง โดยรวมแล้ว ถือว่าอยู่ในเกณฑ์ไม่สูง เนื่องจากการจำลองการจัดกลุ่มข้อมูลประเภทตัวเลข 4 กลุ่ม ผลการจำลองที่ได้นั้น ข้อ-มูลทำการจัดกลุ่มกันได้ไม่สมบูรณ์ คือ มีทั้งข้อมูลที่ไม่สามารถหากลุ่มได้ และข้อมูลที่อยู่ผิดกลุ่ม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งไม่อนุญาตให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.2.1.1 จำลองการจัดกลุ่มข้อมูลครั้งที่ 1

โดยใช้ค่าพารามิเตอร์ต่างๆดังนี้

Number of ant : 20

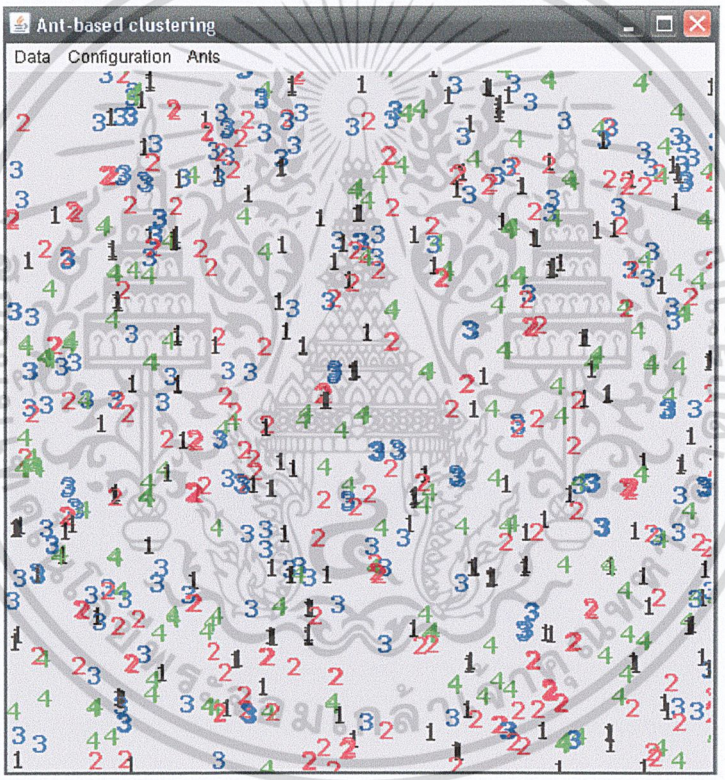
Memory size : 50

Step size : 50

kp : 0.01

kd : 0.01

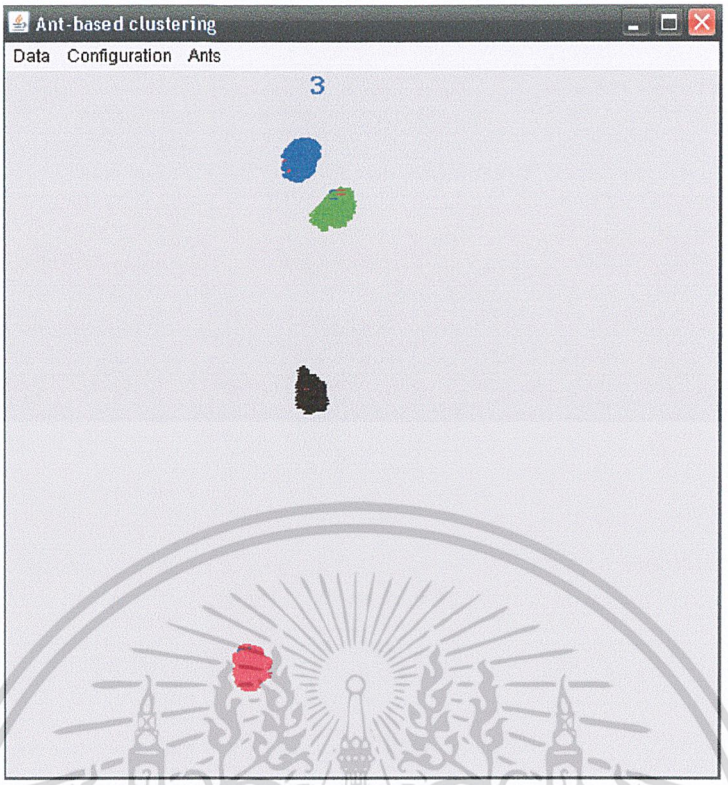
จะได้ผลการทดลองดังรูปที่ 4.1 และ 4.2



รูปที่ 4.1 เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 1

จากรูปที่ 4.1 แสดงการกระจายข้อมูลอย่างสุ่มในขั้นตอนเริ่มต้นของการจำลองการจัดกลุ่มข้อมูล ซึ่งในการกระจายข้อมูลอย่างสุ่มแต่ละครั้งจะได้ผลออกมาไม่เหมือนกัน ทำให้มีผลต่อการจัดกลุ่มของข้อมูลในขั้นตอนต่อไป

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.2 ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 1

หลังจากทำการจำลองการจัดกลุ่มข้อมูลโดยใช้ค่าความน่าจะเป็นในการเก็บข้อมูลเท่ากับ 0.01 และค่าความน่าจะเป็นในการวางข้อมูล เท่ากับ 0.01 พบว่าได้ค่า Pearson correlation เท่ากับ 0.4506 และใช้เวลาในการจัดกลุ่ม 21.80 วินาที

จากรูปที่ 4.2 การจัดกลุ่มค่อนข้างมีความชัดเจน แต่การที่ได้ค่า Pearson correlation ไม่มาก อาจเกิดจากการที่มีข้อมูลต่างประเภทกันแทรกตัวอยู่ภายในแต่ละกลุ่ม ซึ่งจากรูปจะมองเห็นไม่ชัดเจน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.2.1.2 จำลองการจัดกลุ่มข้อมูลครั้งที่ 2

โดยใช้ค่าพารามิเตอร์ต่างๆดังนี้

Number of ant : 20

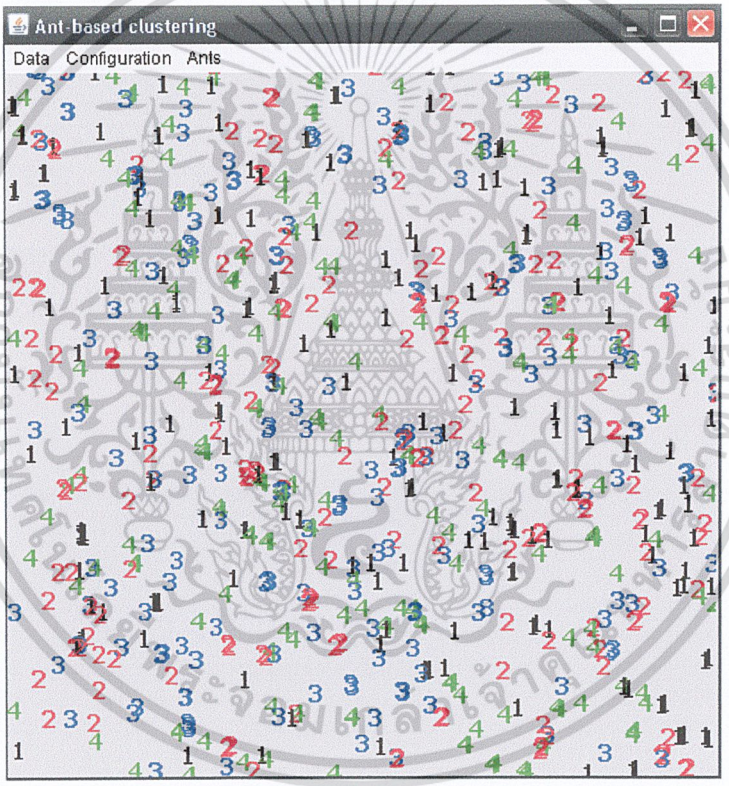
Memory size : 50

Step size : 50

kp : 0.01

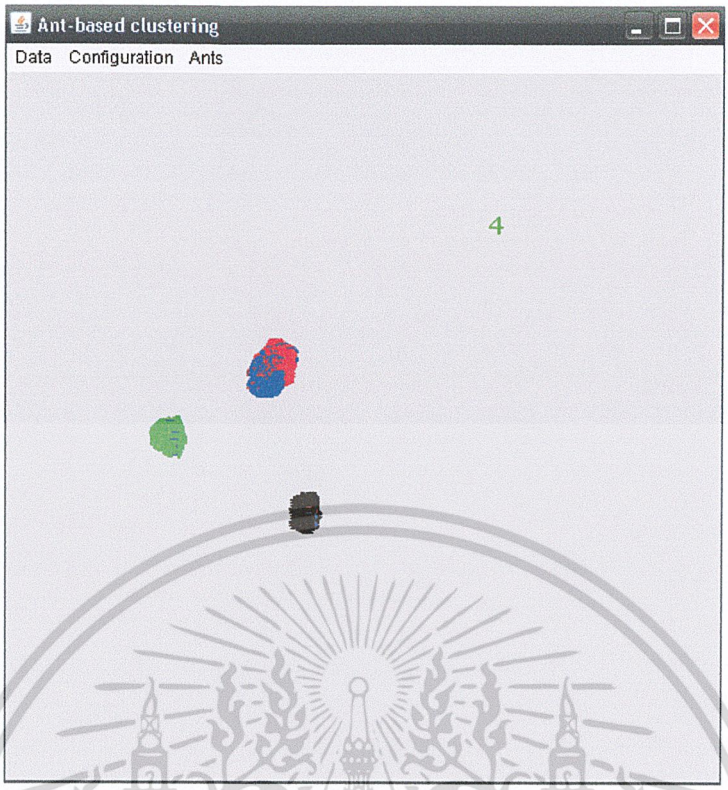
kd : 0.04

จะได้ผลการทดลองดังรูปที่ 4.3 และ 4.4



รูปที่ 4.3 เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 2

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.4 ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 2

หลังจากทำการจำลองการจัดกลุ่มข้อมูลโดยใช้ค่าความน่าจะเป็นในการเก็บข้อมูลเท่ากับ 0.01 และค่าความน่าจะเป็นในการวางข้อมูล เท่ากับ 0.04 พบว่าได้ค่า Pearson correlation เท่ากับ 0.5347 และใช้เวลาในการจัดกลุ่ม 61.53 วินาที

จากรูปที่ 4.4 การจัดกลุ่มมีความไม่ชัดเจน เนื่องจากมีการซ้อนทับกันระหว่างกลุ่ม 2 กลุ่ม ซึ่งเป็นผลมาจากการปรับค่า kd มากกว่า kp ในอัตราส่วนที่มาก ทำให้แนวโน้มที่ข้อมูลจะถูกวางมีมากกว่าถูกเก็บ เป็นผลให้เกิดการวางข้อมูลผิดพลาด แต่การที่ได้ค่า Pearson correlation มาก เนื่องจากค่า Pearson correlation เป็นการวัดการกระจายของข้อมูล ซึ่งในกรณีนี้ข้อมูลมีการเกาะกลุ่มกันค่อนข้างดี ทำให้ได้ค่า Pearson correlation มาก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.2.1.3 จำลองการจัดกลุ่มข้อมูลครั้งที่ 3

โดยใช้ค่าพารามิเตอร์ต่างๆดังนี้

Number of ant : 20

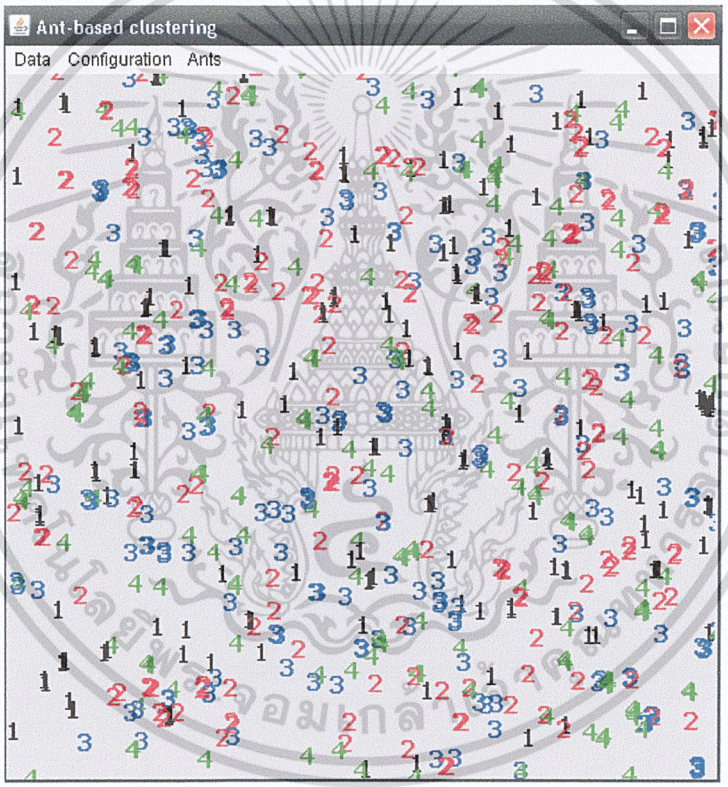
Memory size : 50

Step size : 50

kp : 0.02

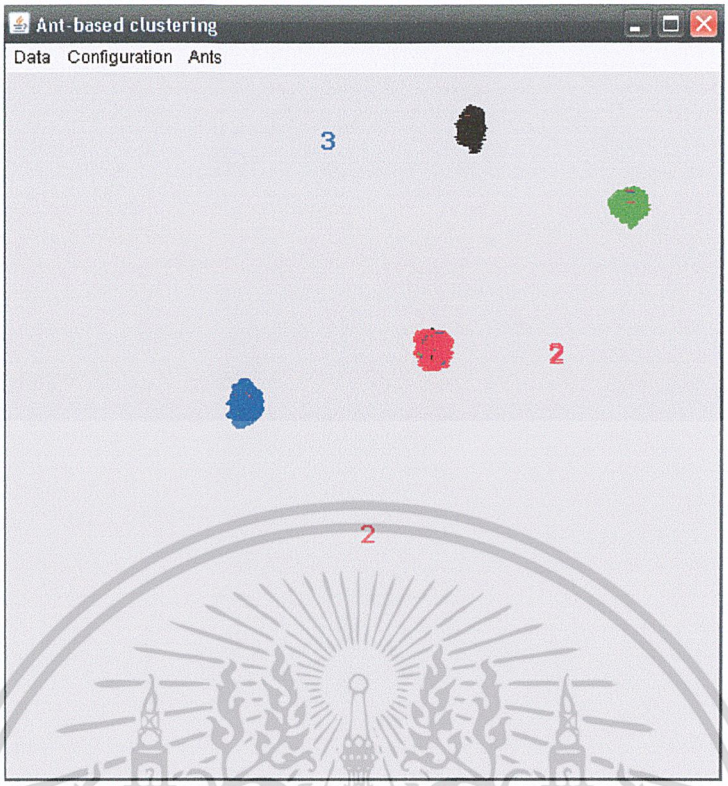
kd : 0.02

จะได้ผลการทดลองดังรูปที่ 4.5 และ 4.6



รูปที่ 4.5 เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 3

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.6 ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 3

หลังจากทำการจำลองการจัดกลุ่มข้อมูลโดยใช้ค่าความน่าจะเป็นในการเก็บข้อมูลเท่ากับ 0.02 และค่าความน่าจะเป็นในการวางข้อมูล เท่ากับ 0.02 พบว่าได้ค่า Pearson correlation เท่ากับ 0.4281 และใช้เวลาในการจัดกลุ่ม 17.65 วินาที

จากรูปที่ 4.6 การจัดกลุ่มค่อนข้างมีความชัดเจน แต่การที่ได้ค่า Pearson correlation ไม่มาก อาจเกิดจากการที่มีข้อมูลต่างประเภทกันแทรกตัวอยู่ภายในแต่ละกลุ่ม และมีข้อมูลบางค่าที่ไม่สามารถหากลุ่มได้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.2.1.4 จำลองการจัดกลุ่มข้อมูลครั้งที่ 4

โดยใช้ค่าพารามิเตอร์ต่างๆดังนี้

Number of ant : 20

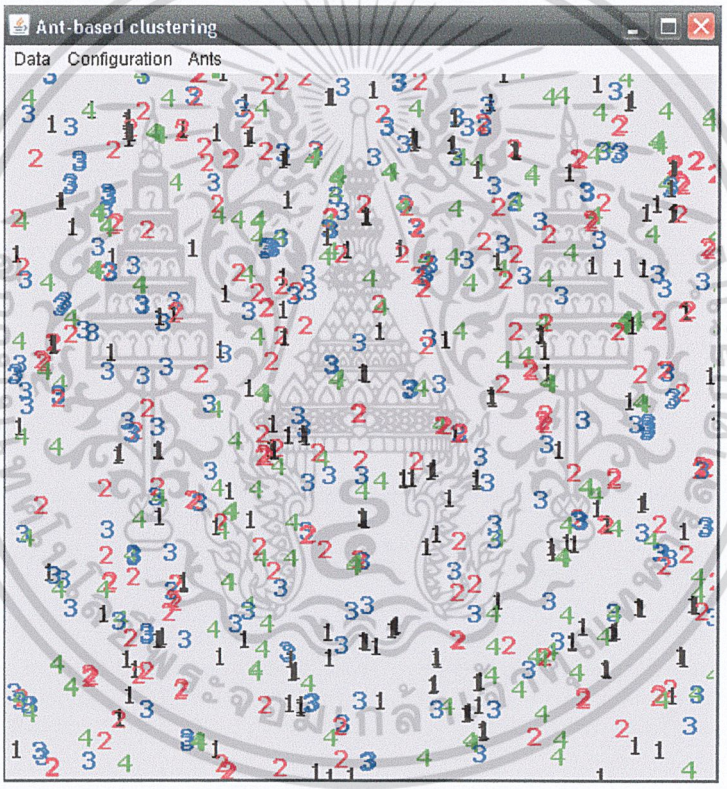
Memory size : 50

Step size : 50

kp : 0.02

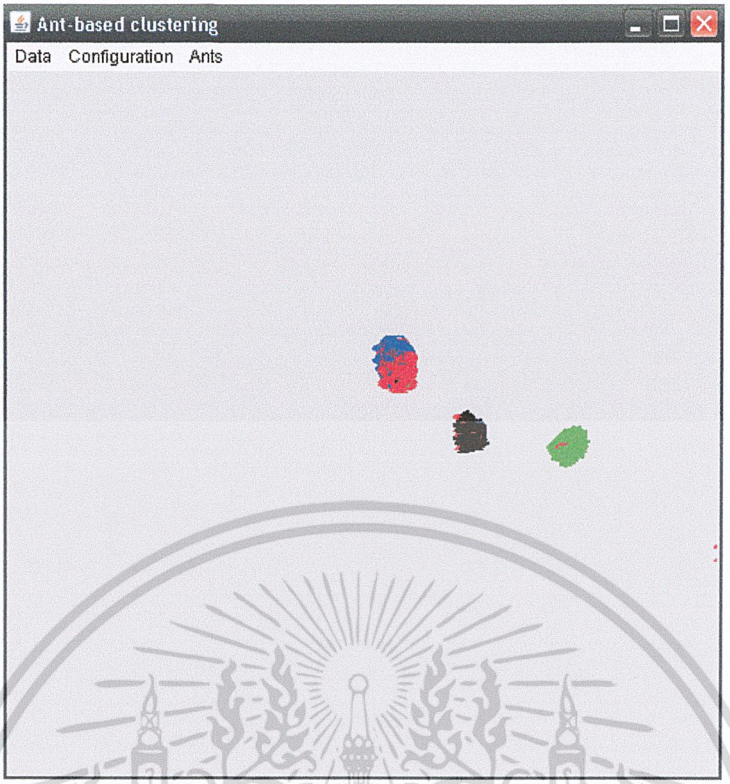
kd : 0.04

จะได้ผลการทดลองดังรูปที่ 4.7 และ 4.8



รูปที่ 4.7 เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 4

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.8 ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 4

หลังจากทำการจำลองการจัดกลุ่มข้อมูลโดยใช้ค่าความน่าจะเป็นในการเก็บข้อมูลเท่ากับ 0.02 และค่าความน่าจะเป็นในการวางข้อมูล เท่ากับ 0.04 พบว่าได้ค่า Pearson correlation เท่ากับ 0.5096 และใช้เวลาในการจัดกลุ่ม 30.48 วินาที

จากรูปที่ 4.8 การจัดกลุ่มมีความไม่ชัดเจน เนื่องจากมีการซ้อนทับกันระหว่างกลุ่ม 2 กลุ่ม ซึ่งเป็นผลมาจากการปรับค่า kd มากกว่า kp ในอัตราส่วนที่มาก ทำให้แนวโน้มที่ข้อมูลจะถูกวางมีมากกว่าถูกเก็บ เป็นผลให้เกิดการวางข้อมูลผิดพลาด แต่การที่ได้ค่า Pearson correlation มาก เนื่องจากค่า Pearson correlation เป็นการวัดการกระจายของข้อมูล ซึ่งในกรณีนี้ข้อมูลมีการเกาะกลุ่มกันค่อนข้างดี ทำให้ได้ค่า Pearson correlation มาก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.2.1.5 จำลองการจัดกลุ่มข้อมูลครั้งที่ 5

โดยใช้ค่าพารามิเตอร์ต่างๆดังนี้

Number of ant : 20

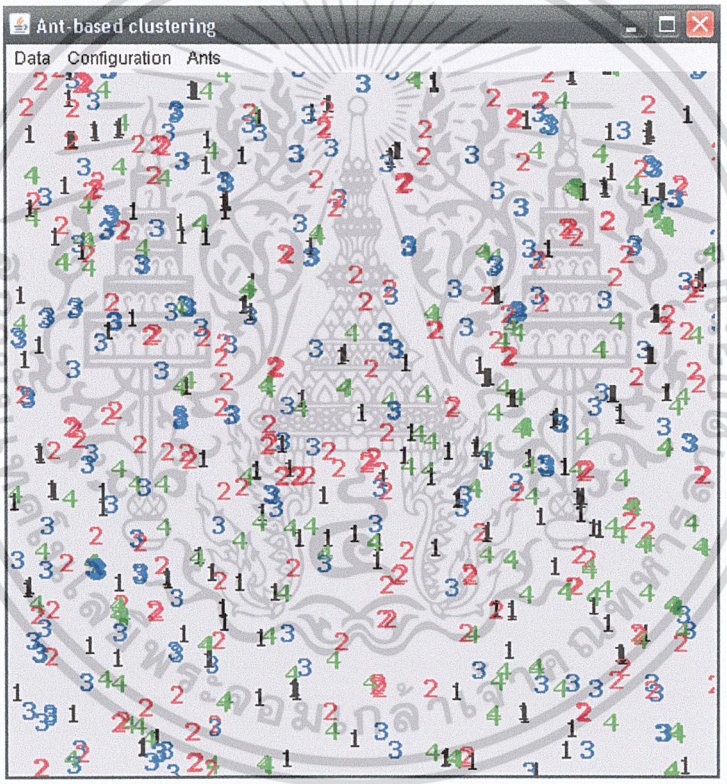
Memory size : 50

Step size : 50

kp : 0.03

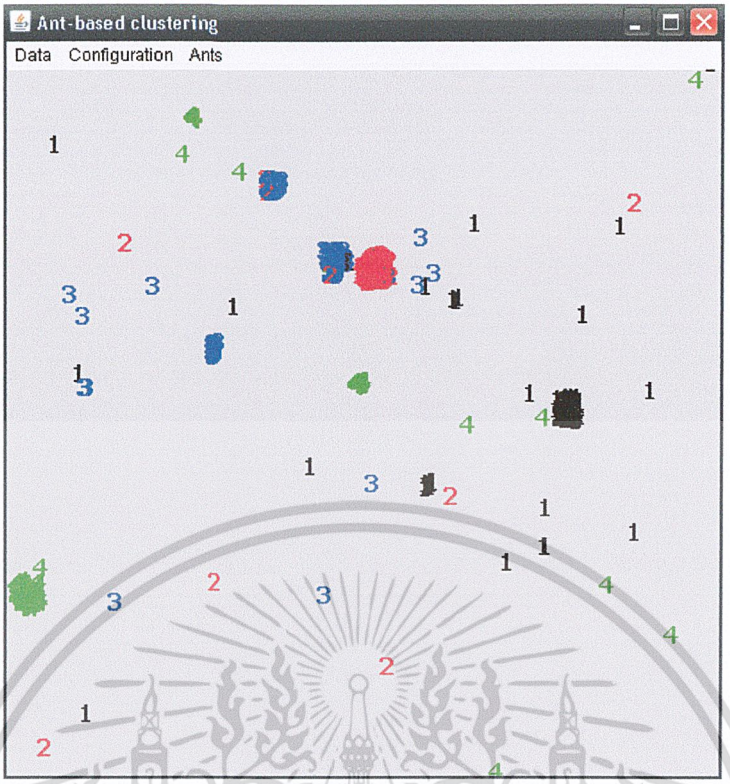
kd : 0.01

จะได้ผลการทดลองดังรูปที่ 4.9 และ 4.10



รูปที่ 4.9 เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 5

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.10 ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 5

หลังจากทำการจำลองการจัดกลุ่มข้อมูลโดยใช้ค่าความน่าจะเป็นในการเก็บข้อมูลเท่ากับ 0.03 และค่าความน่าจะเป็นในการวางข้อมูล เท่ากับ 0.01 พบว่าได้ค่า Pearson correlation เท่ากับ 0.3774 และใช้เวลาในการจัดกลุ่ม 15.64 วินาที

จากรูปที่ 4.10 จะเห็นว่าข้อมูลไม่สามารถที่จะจัดกลุ่มได้อย่างสมบูรณ์ เนื่องจากการที่ค่า kp มากกว่า kd ส่งผลให้แนวโน้มที่ข้อมูลจะถูกเก็บมีมากกว่าถูกวาง ทำให้มีข้อมูลจำนวนมากที่ไม่สามารถจัดกลุ่มได้ เป็นผลให้ได้ค่า Pearson correlation ค่อนข้างน้อย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.2.1.6 จำลองการจัดกลุ่มข้อมูลครั้งที่ 6

โดยใช้ค่าพารามิเตอร์ต่างๆดังนี้

Number of ant : 20

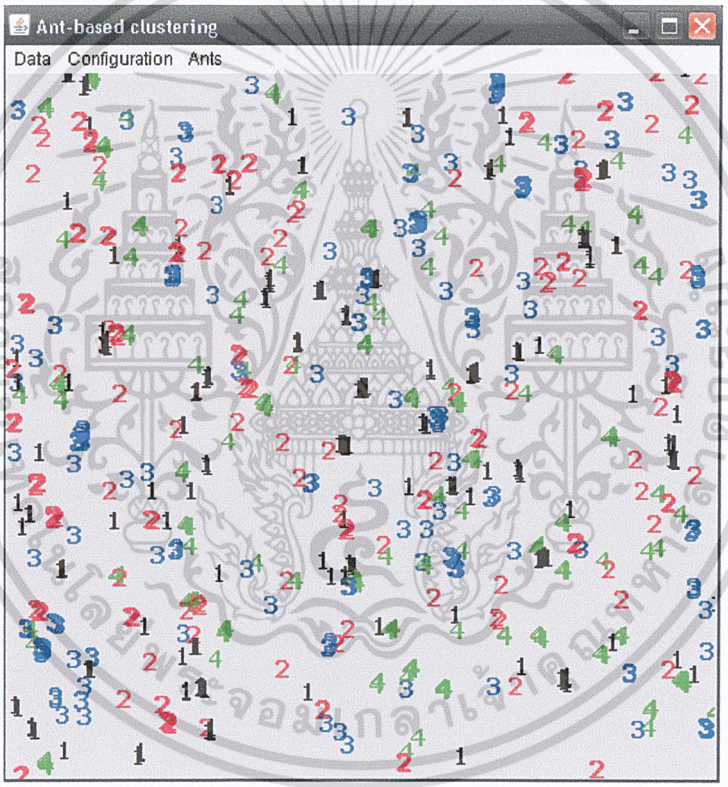
Memory size : 50

Step size : 50

kp : 0.04

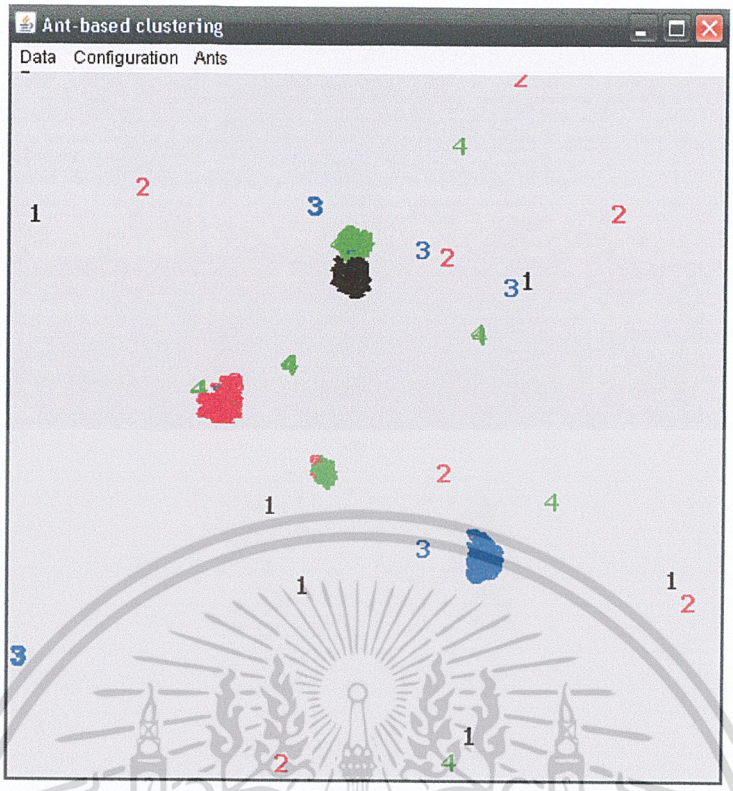
kd : 0.01

จะได้ผลการทดลองดังรูปที่ 4.11 และ 4.12



รูปที่ 4.11 เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 6

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.12 ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 6

หลังจากทำการจำลองการจัดกลุ่มข้อมูลโดยใช้ค่าความน่าจะเป็นในการเก็บข้อมูลเท่ากับ 0.04 และค่าความน่าจะเป็นในการวางข้อมูล เท่ากับ 0.01 พบว่าได้ค่า Pearson correlation เท่ากับ 0.3699 และใช้เวลาในการจัดกลุ่ม 13.51 วินาที

จากรูปที่ 4.12 จะเห็นว่าข้อมูลไม่สามารถที่จะจัดกลุ่มได้อย่างสมบูรณ์ เนื่องจากการที่ค่า kp มากกว่า kd ในอัตราส่วนที่มาก ส่งผลให้แนวโน้มที่ข้อมูลจะถูกเก็บมีมากกว่าถูกวาง ทำให้มีข้อมูลจำนวนมากที่ไม่สามารถจัดกลุ่มได้ เป็นผลให้ได้ค่า Pearson correlation น้อย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.2.2 Test Set B

เป็นข้อมูลชนิดตัวเลข ได้แก่ 1 2 3 4 หลังการทดลองข้อมูลจะถูกจัดเป็น 2 กลุ่ม โดยที่เลข 1 จะรวมกับเลข 3 และเลข 2 รวมกับเลข 4 ในการทดลองได้ทำการปรับค่า kp และ kd ให้มีค่าระหว่าง 0.01 ถึง 0.09 รวมทดลองทั้งหมด 81 ครั้ง ได้ผลการทดลองดังตารางที่ 4.2 โดยผลตัวเลขทศนิยม 4 ตำแหน่งที่ได้คือค่า Pearson correlation เนื่องจากรูปผลการจำลองการจัดกลุ่มข้อมูลที่ได้แต่ละครั้ง มีลักษณะค่อนข้างคล้ายกัน จึงจะนำค่า kp และ kd ที่ทำการทดลองแล้วให้ค่า Pearson correlation มากที่สุด 5 ค่า มาจำลองการจัดกลุ่มในหัวข้อ 4.2.2.1 ถึง 4.2.2.5 ต่อไป

ตารางที่ 4.2 ผลค่า Pearson correlation ที่ได้จากการทดลอง

	kd=0.01	kd=0.02	kd=0.03	kd=0.04	kd=0.05	kd=0.06	kd=0.07	kd=0.08	kd=0.09
kp=0.01	0.7466	0.1379	0.7240	0.7700	0.7823	0.8302	0.8406	0.8409	0.7865
kp=0.02	0.7677	0.4412	0.7960	0.4682	0.8467	0.8234	0.8322	0.5731	0.8290
kp=0.03	0.3192	0.4951	0.7959	0.8054	0.8135	0.8467	0.8271	0.8376	0.8502
kp=0.04	0.7242	0.8341	0.7417	0.8267	0.8236	0.8654	0.8300	0.8303	0.8404
kp=0.05	0.1449	0.7859	0.7133	0.8487	0.6282	0.7712	0.8244	0.2298	0.7502
kp=0.06	0.2652	0.7073	0.5033	0.8248	0.7535	0.7970	0.5339	0.8430	0.7856
kp=0.07	0.3154	0.4761	0.7450	0.4390	0.7981	0.7031	0.5334	0.1282	0.8289
kp=0.08	0.2338	0.7437	0.3780	0.8433	0.8542	0.7728	0.2786	0.7882	0.8079
kp=0.09	0.3398	0.8177	0.6020	0.8038	0.7790	0.7353	0.7922	0.3649	0.8264

หมายเหตุ ตัวเลขสีแดงแสดงถึงค่า Pearson correlation มากที่สุด 5 ค่า ที่ได้จากการทดลอง ซึ่งจะแสดงการจำลองจัดกลุ่มในหัวข้อที่ 4.2.2.1 ถึง 4.2.2.5

ผลค่า Pearson correlation ที่ได้จากการทดลองในตารางที่ 4.2 โดยรวมแล้ว ถือว่าได้ค่าค่อนข้างมาก เป็นผลมาจากการจัดกลุ่มของข้อมูลที่นำเลข 1 มารวมกับเลข 3 และเลข 2 รวมกับเลข 4 ทำให้ข้อมูลมีการเกาะกลุ่มกันมาก แต่ในกรณีที่ปรับค่า kd = 0.01 แล้วปรับค่า kp สูงขึ้น จะเห็นว่าแนวโน้มค่า Pearson correlation มีค่าน้อย เนื่องจากการที่ค่า kp มากกว่า kd ในอัตราส่วนที่มาก ส่งผลให้แนวโน้มที่ข้อมูลจะถูกเก็บมีมากกว่าถูกวาง ทำให้มีข้อมูลจำนวนมากที่ไม่สามารถจัดกลุ่มได้ เป็นผลให้ได้ค่า Pearson correlation ค่อนข้างน้อย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.2.2.1 จำลองการจัดกลุ่มข้อมูลครั้งที่ 7

โดยใช้ค่าพารามิเตอร์ต่างๆดังนี้

Number of ant : 20

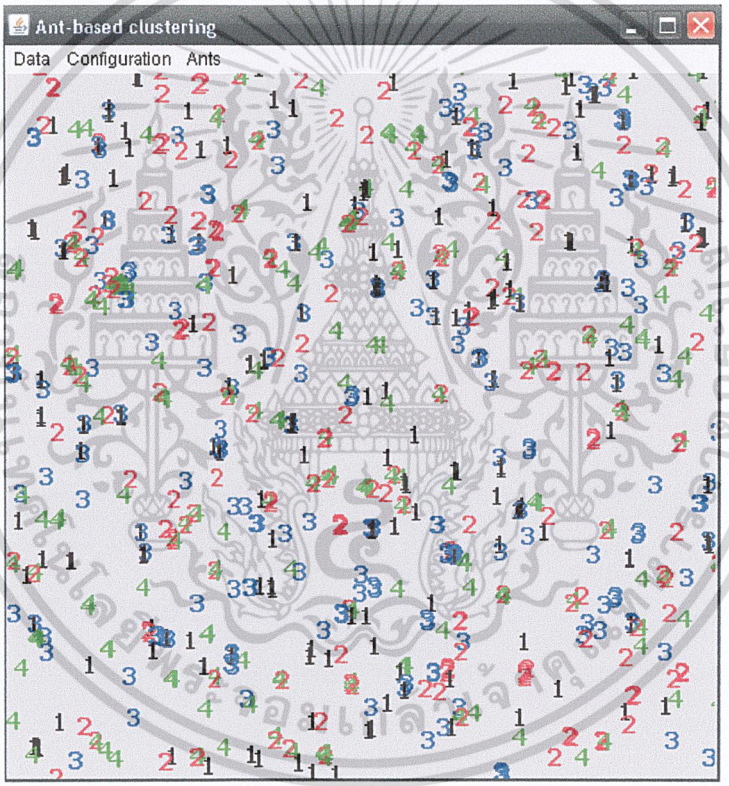
Memory size : 50

Step size : 50

kp : 0.02

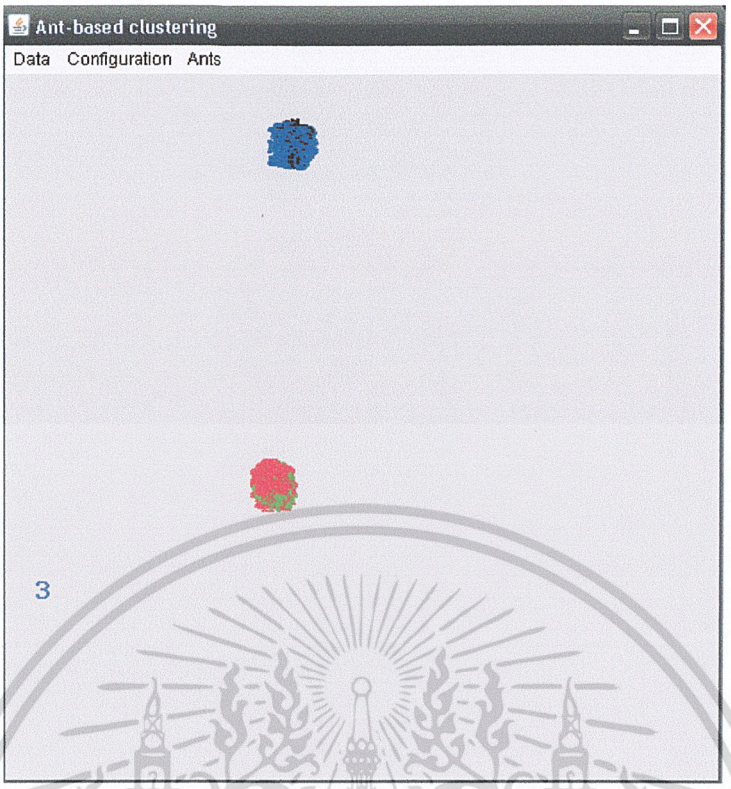
kd : 0.05

จะได้ผลการทดลองดังรูปที่ 4.13 และ 4.14



รูปที่ 4.13 เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 7

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.14 ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 7

หลังจากทำการจำลองการจัดกลุ่มข้อมูลโดยใช้ค่าความน่าจะเป็นในการเก็บข้อมูลเท่ากับ 0.02 และค่าความน่าจะเป็นในการวางข้อมูล เท่ากับ 0.05 พบว่าได้ค่า Pearson correlation เท่ากับ 0.8467 และใช้เวลาในการจัดกลุ่ม 62.70 วินาที

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.2.2.2 จำลองการจัดกลุ่มข้อมูลครั้งที่ 8

โดยใช้ค่าพารามิเตอร์ต่างๆดังนี้

Number of ant : 20

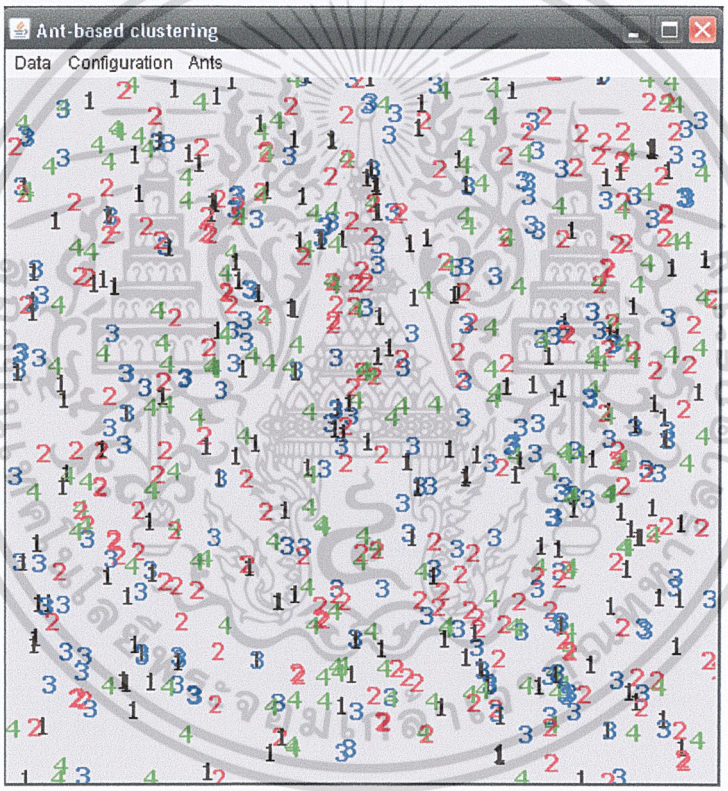
Memory size : 50

Step size : 50

kp : 0.03

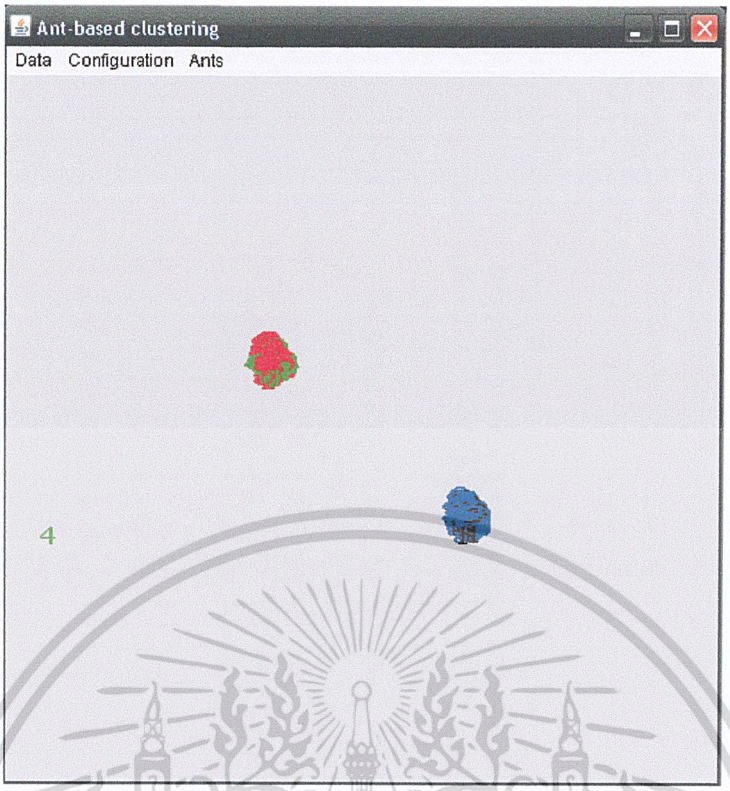
kd : 0.09

จะได้ผลการทดลองดังรูปที่ 4.15 และ 4.16



รูปที่ 4.15 เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 8

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.16 ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 8

หลังจากทำการจำลองการจัดกลุ่มข้อมูลโดยใช้ค่าความน่าจะเป็นในการเก็บข้อมูลเท่ากับ 0.03 และค่าความน่าจะเป็นในการวางข้อมูล เท่ากับ 0.09 พบว่าได้ค่า Pearson correlation เท่ากับ 0.8502 และใช้เวลาในการจัดกลุ่ม 30.47 วินาที

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.2.2.3 จำลองการจัดกลุ่มข้อมูลครั้งที่ 9

โดยใช้ค่าพารามิเตอร์ต่างๆดังนี้

Number of ant : 20

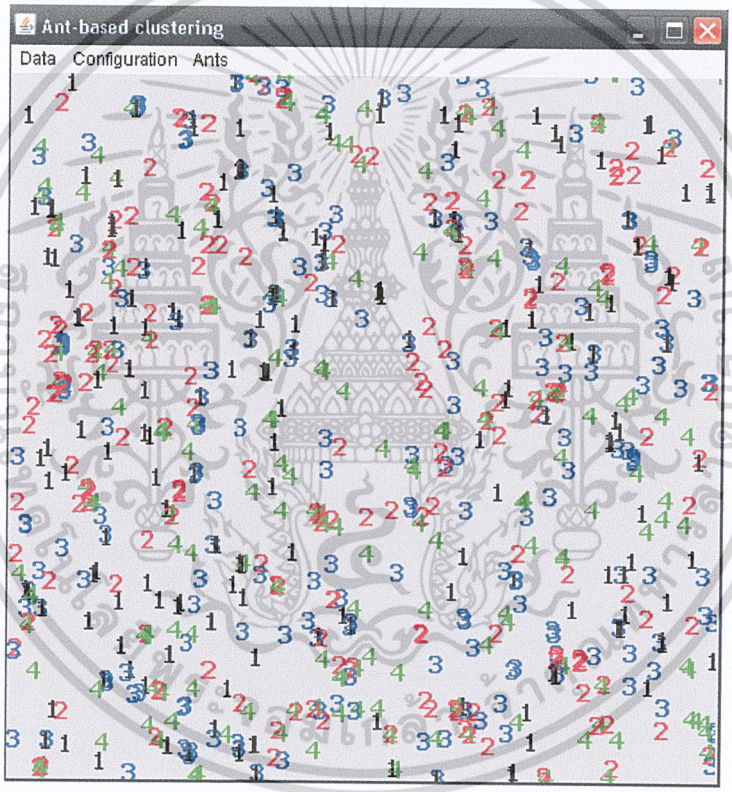
Memory size : 50

Step size : 50

kp : 0.04

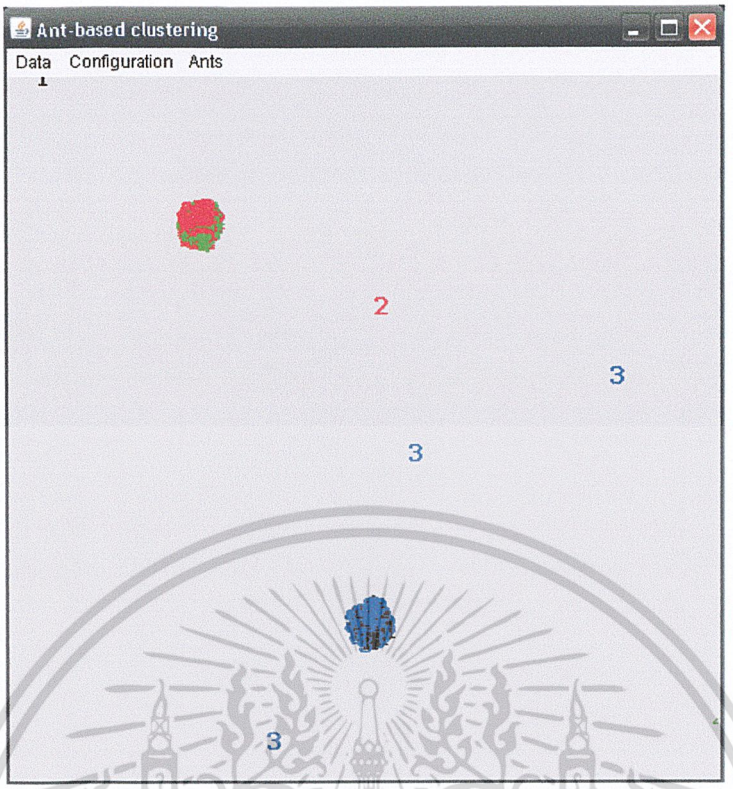
kd : 0.06

จะได้ผลการทดลองดังรูปที่ 4.17 และ 4.18



รูปที่ 4.17 เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 9

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.18 ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 9

หลังจากทำการจำลองการจัดกลุ่มข้อมูลโดยใช้ค่าความน่าจะเป็นในการเก็บข้อมูลเท่ากับ 0.04 และค่าความน่าจะเป็นในการวางข้อมูล เท่ากับ 0.06 พบว่าได้ค่า Pearson correlation เท่ากับ 0.8654 และใช้เวลาในการจัดกลุ่ม 28.27 วินาที

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.2.2.4 จำลองการจัดกลุ่มข้อมูลครั้งที่ 10

โดยใช้ค่าพารามิเตอร์ต่างๆดังนี้

Number of ant : 20

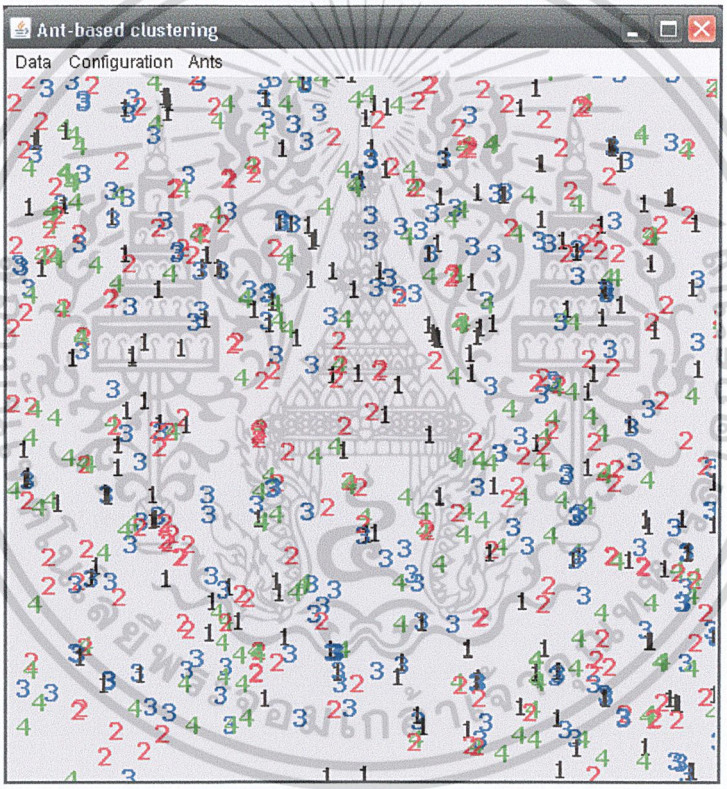
Memory size : 50

Step size : 50

kp : 0.05

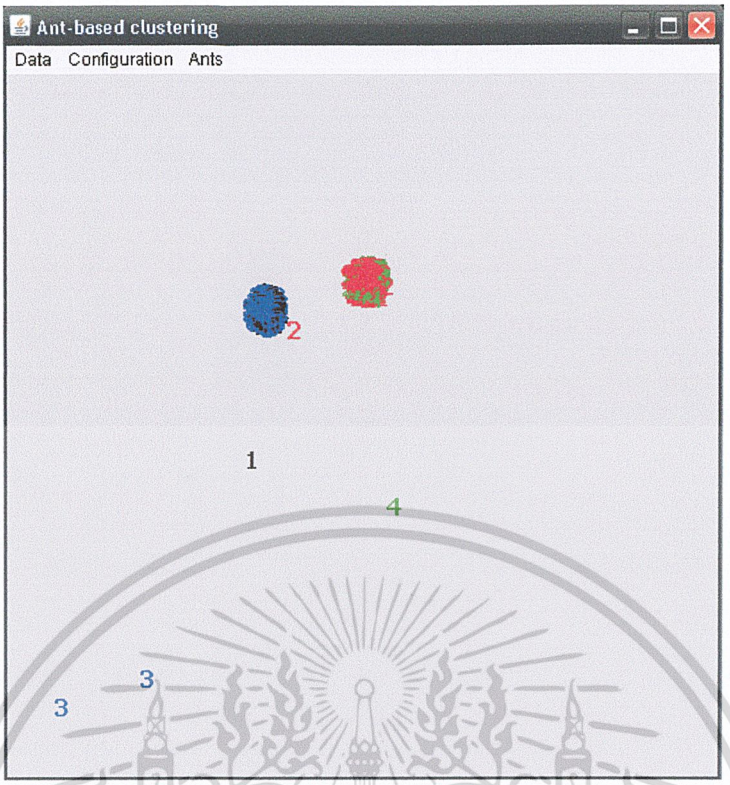
kd : 0.04

จะได้ผลการทดลองดังรูปที่ 4.19 และ 4.20



รูปที่ 4.19 เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 10

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.20 ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 10

หลังจากทำการจำลองการจัดกลุ่มข้อมูลโดยใช้ค่าความน่าจะเป็นในการเก็บข้อมูลเท่ากับ 0.05 และค่าความน่าจะเป็นในการวางข้อมูล เท่ากับ 0.04 พบว่าได้ค่า Pearson correlation เท่ากับ 0.8487 และใช้เวลาในการจัดกลุ่ม 16.86 วินาที

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.2.2.5 จำลองการจัดกลุ่มข้อมูลครั้งที่ 11

โดยใช้ค่าพารามิเตอร์ต่างๆดังนี้

Number of ant : 20

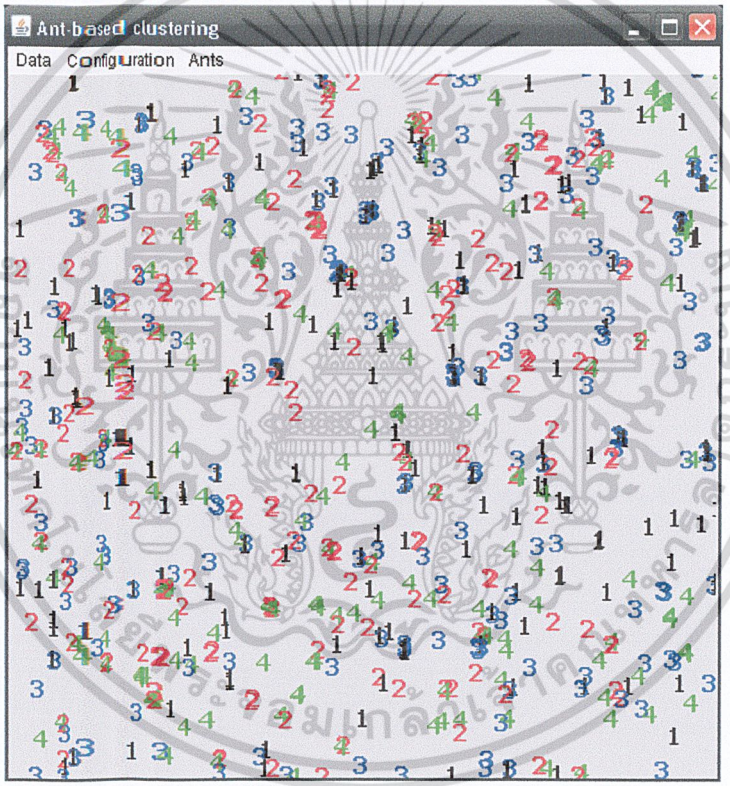
Memory size : 50

Step size : 50

kp : 0.08

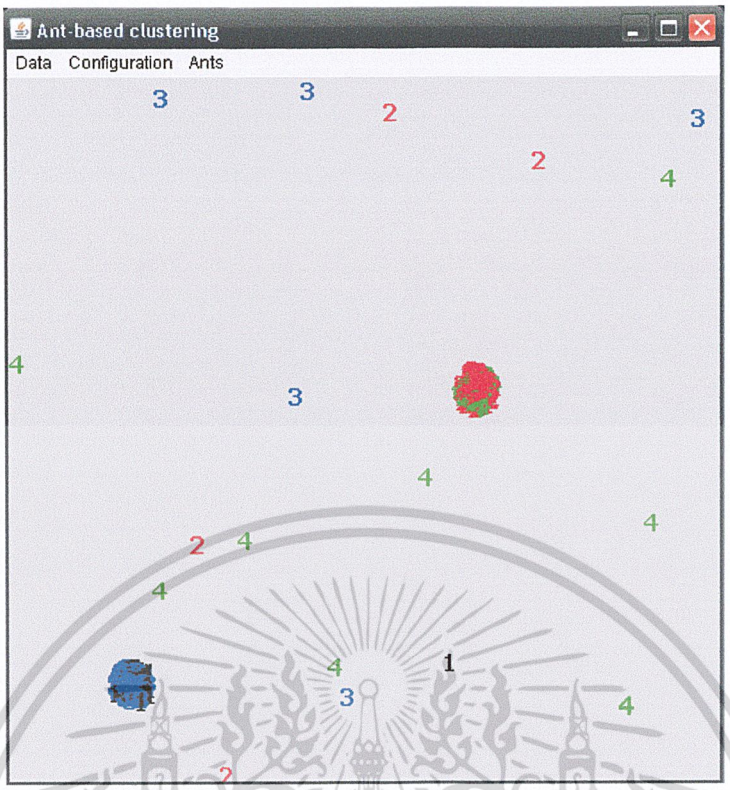
kd : 0.05

จะได้ผลการทดลองดังรูปที่ 4.21 และ 4.22



รูปที่ 4.21 เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 11

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.22 ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 11

หลังจากทำการจำลองการจัดกลุ่มข้อมูลโดยใช้ค่าความน่าจะเป็นในการเก็บข้อมูลเท่ากับ 0.08 และค่าความน่าจะเป็นในการวางข้อมูล เท่ากับ 0.05 พบว่าได้ค่า Pearson correlation เท่ากับ 0.8542 และใช้เวลาในการจัดกลุ่ม 17.59 วินาที

จากรูปที่ 4.22 จะเห็นว่ามีข้อมูลจำนวนหนึ่งที่ไม่สามารถหากลุ่มได้ เป็นผลมาจากการปรับค่า kp มากกว่า kd เหมือนในบางการทดลองที่กล่าวถึงก่อนหน้านี้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.2.3 Test Set C

เป็นข้อมูลชนิด Image ได้แก่ ภาพผลไม้ 3 ชนิด หลังการทดลองข้อมูลจะถูกจัดเป็น 3 กลุ่ม ในการทดลองได้ทำการปรับค่า kp และ kd ให้มีค่าระหว่าง 0.01 ถึง 0.04 รวมทดลองทั้งหมด 16 ครั้ง ได้ผลการทดลองดังตารางที่ 4.3 โดยผลตัวเลขทศนิยม 4 ตำแหน่งที่ได้ คือค่า Pearson - correlation และจะนำค่า kp และ kd ที่ทำการทดลองแล้วให้ค่า Pearson correlation สูงที่สุด 4 ค่า มาจำลองการจัดกลุ่มในหัวข้อ 4.2.3.1 ถึง 4.2.3.4 ต่อไป

ตารางที่ 4.3 ผลค่า Pearson correlation ที่ได้จากการทดลอง

	kd=0.01	**kd=0.02**	**kd=0.03**	**kd=0.04**
kp=0.01	0.3280	0.5024	0.3367	0.5742
kp=0.02	0.3236	0.5698	0.7254	0.0821
kp=0.03	0.4211	0.4353	0.4779	0.3030
kp=0.04	0.3609	0.4422	0.3250	0.4646

หมายเหตุ ตัวเลขสีแดงแสดงถึงค่า Pearson correlation มากที่สุด 4 ค่า ที่ได้จากการทดลอง ซึ่งจะแสดงการจำลองจัดกลุ่มในหัวข้อที่ 4.2.3.1 ถึง 4.2.3.4

จากตารางที่ 4.3 ผลค่า Pearson correlation ที่ได้จากการทดลองมีค่าค่อนข้างน้อย เนื่องจากการรวมข้อมูลชนิด Image ทำได้ยากกว่าข้อมูลชนิดตัวเลข แต่การที่ค่า Pearson correlation ยังไม่ต่ำมาก เนื่องจากในการจัดกลุ่มมีข้อมูลเพียงแค่ 3 ประเภทเท่านั้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.2.3.1 จำลองการจัดกลุ่มข้อมูลครั้งที่ 12

โดยใช้ค่าพารามิเตอร์ต่างๆดังนี้

Number of ant : 20

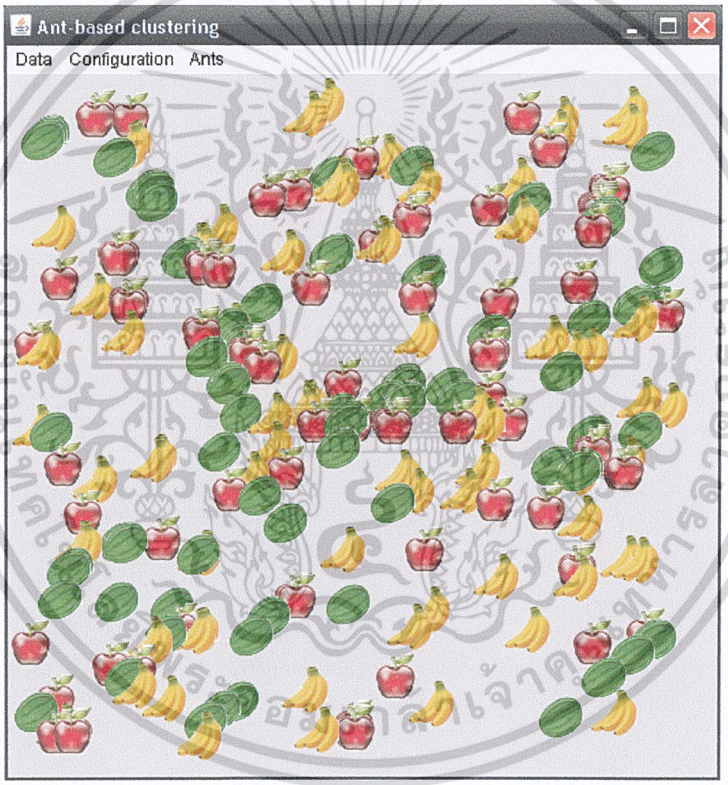
Memory size : 50

Step size : 50

kp : 0.01

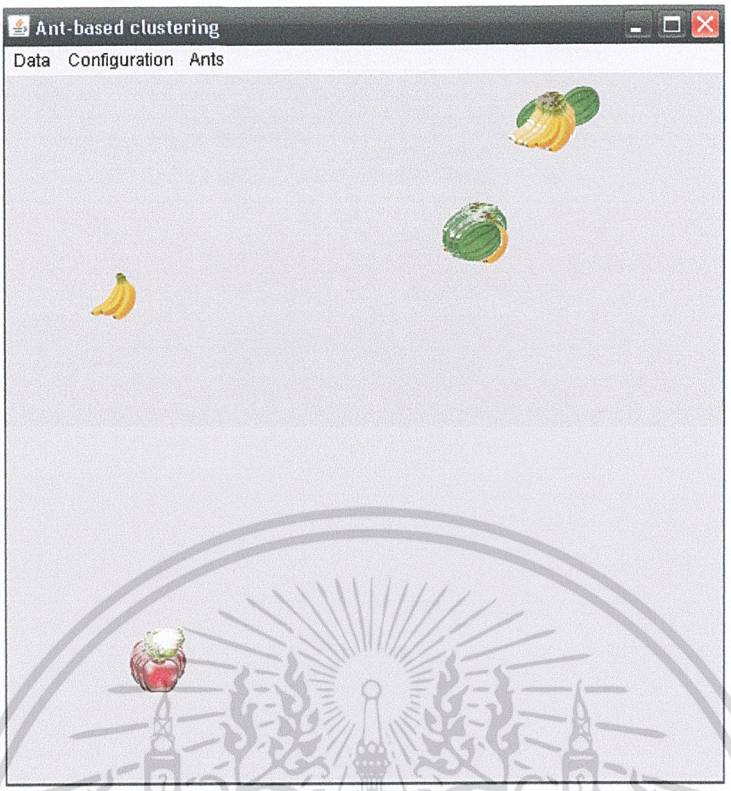
kd : 0.02

จะได้ผลการทดลองดังรูปที่ 4.23 และ 4.24



รูปที่ 4.23 เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 12

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.24 ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 12

หลังจากทำการจำลองการจัดกลุ่มข้อมูล โดยใช้ค่าความน่าจะเป็นในการเก็บข้อมูลเท่ากับ 0.01 และค่าความน่าจะเป็นในการวางข้อมูล เท่ากับ 0.02 พบว่าได้ค่า Pearson correlation เท่ากับ 0.5024 และใช้เวลาในการจัดกลุ่ม 28.91 วินาที

จากรูปที่ 4.24 จะเห็นว่าในกลุ่มของรูปแตงโม มีรูปกล้วยแทรกตัวอยู่ และในกลุ่มของรูปกล้วย ก็มีรูปแตงโมแทรกตัวอยู่ เนื่องจากการจัดกลุ่มข้อมูลแต่ละครั้งจะไม่สมบูรณ์ 100%

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.2.3.2 จำลองการจัดกลุ่มข้อมูลครั้งที่ 13

โดยใช้ค่าพารามิเตอร์ต่างๆดังนี้

Number of ant : 20

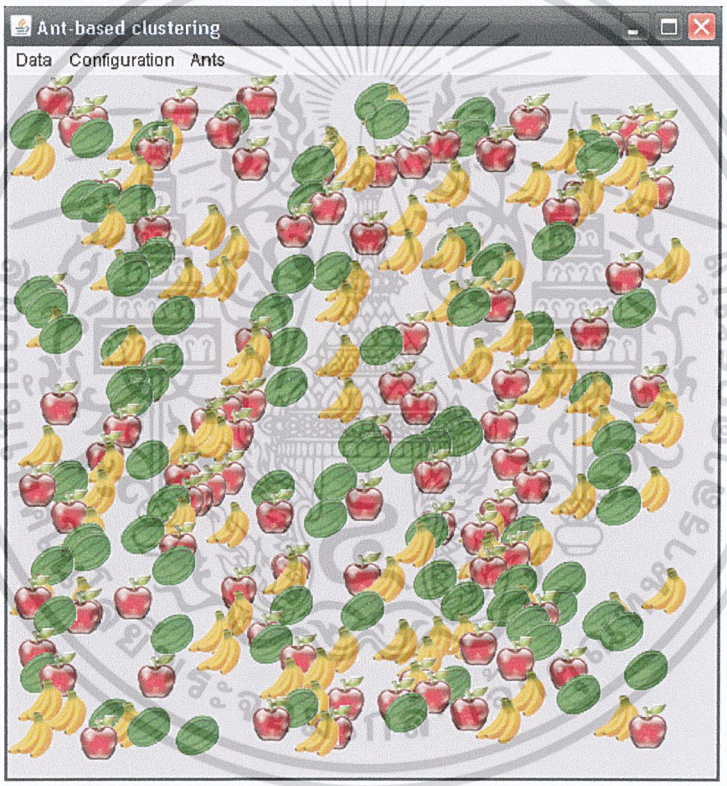
Memory size : 50

Step size : 50

kp : 0.01

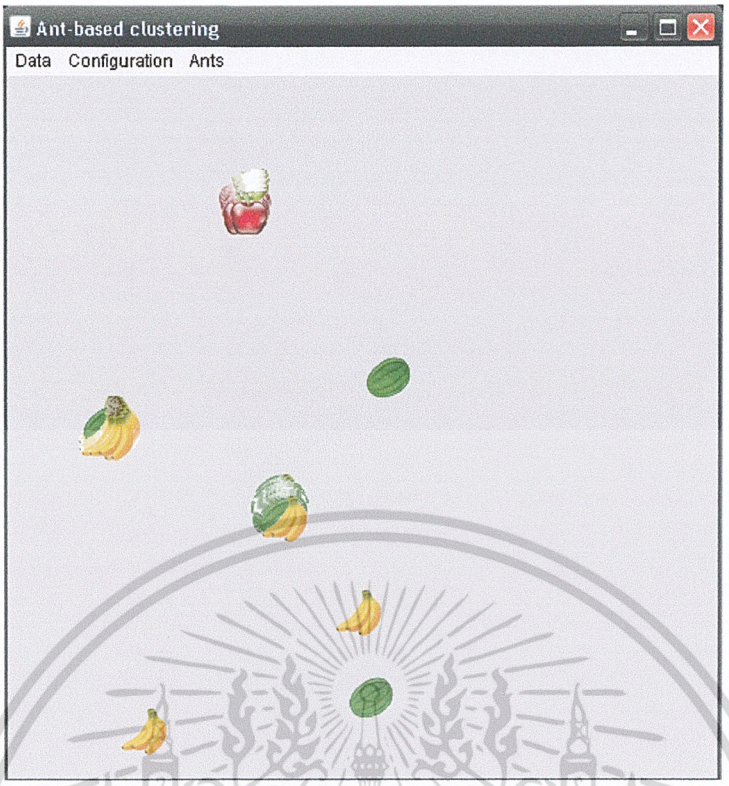
kd : 0.04

จะได้ผลการทดลองดังรูปที่ 4.25 และ 4.26



รูปที่ 4.25 เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 13

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.26 ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 13

หลังจากทำการจำลองการจัดกลุ่มข้อมูลโดยใช้ค่าความน่าจะเป็นในการเก็บข้อมูลเท่ากับ 0.01 และค่าความน่าจะเป็นในการวางข้อมูล เท่ากับ 0.04 พบว่าได้ค่า Pearson correlation เท่ากับ 0.5742 และใช้เวลาในการจัดกลุ่ม 26.39 วินาที

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.2.3.3 จำลองการจัดกลุ่มข้อมูลครั้งที่ 14

โดยใช้ค่าพารามิเตอร์ต่างๆดังนี้

Number of ant : 20

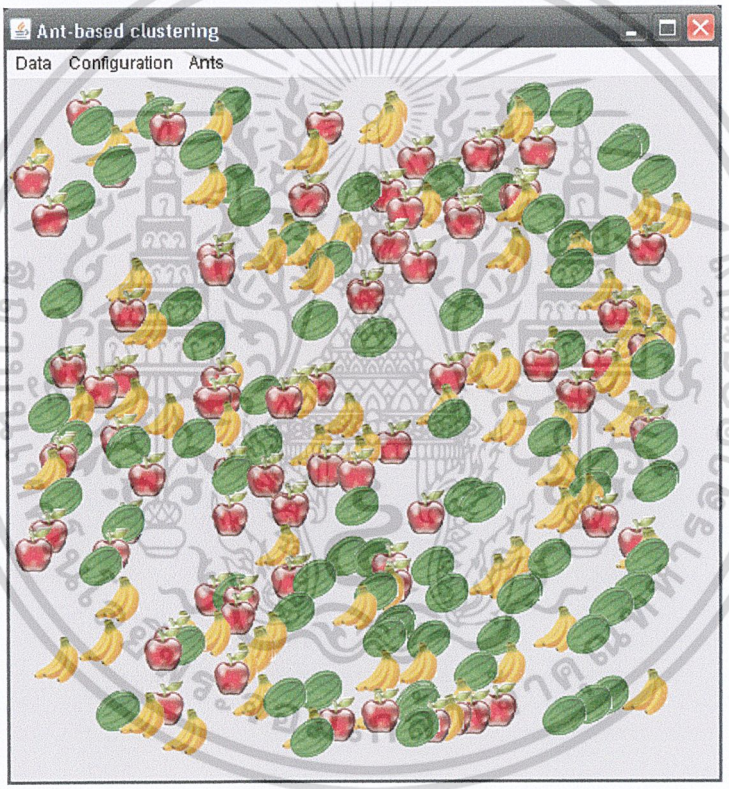
Memory size : 50

Step size : 50

kp : 0.02

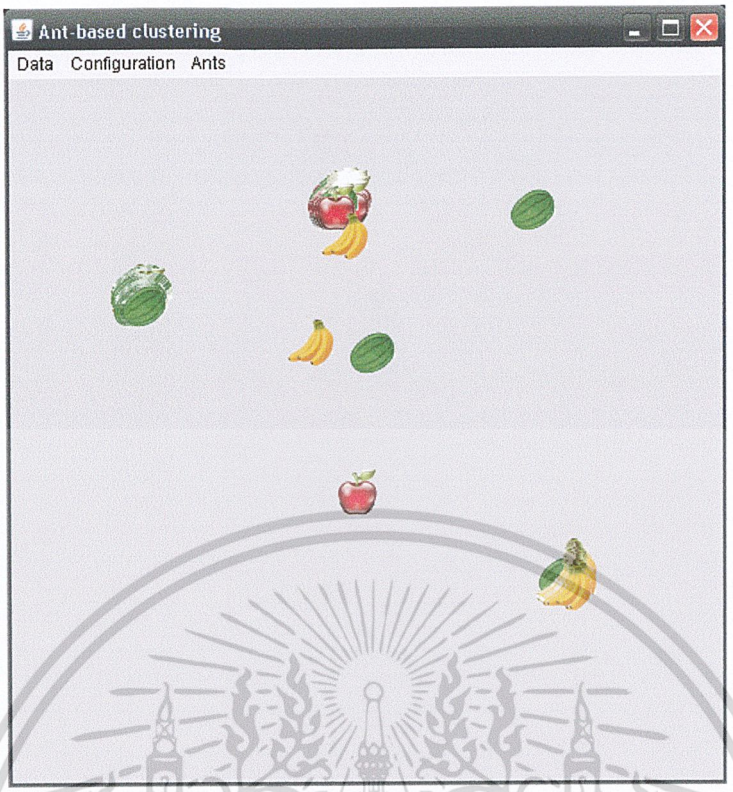
kd : 0.02

จะได้ผลการทดลองดังรูปที่ 4.27 และ 4.28



รูปที่ 4.27 เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 14

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.28 ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 14

หลังจากทำการจำลองการจัดกลุ่มข้อมูล โดยใช้ค่าความน่าจะเป็นในการเก็บข้อมูลเท่ากับ 0.02 และค่าความน่าจะเป็นในการวางข้อมูล เท่ากับ 0.02 พบว่าได้ค่า Pearson correlation เท่ากับ 0.5698 และใช้เวลาในการจัดกลุ่ม 22.27 วินาที

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.2.3.4 จำลองการจัดกลุ่มข้อมูลครั้งที่ 15

โดยใช้ค่าพารามิเตอร์ต่างๆดังนี้

Number of ant : 20

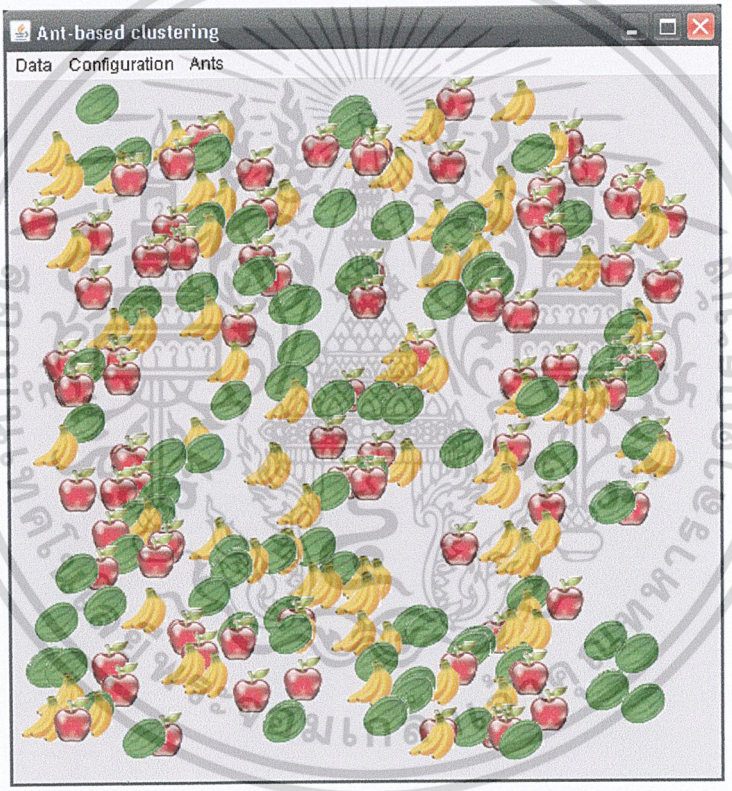
Memory size : 50

Step size : 50

kp : 0.02

kd : 0.03

จะได้ผลการทดลองดังรูปที่ 4.29 และ 4.30



รูปที่ 4.29 เริ่มต้นการจำลองการจัดกลุ่มข้อมูลครั้งที่ 15

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

รูปที่ 4.30 ผลการจำลองการจัดกลุ่มข้อมูลครั้งที่ 15

หลังจากทำการจำลองการจัดกลุ่มข้อมูล โดยใช้ค่าความน่าจะเป็นในการเก็บข้อมูลเท่ากับ 0.02 และค่าความน่าจะเป็นในการวางข้อมูล เท่ากับ 0.03 พบว่าได้ค่า Pearson correlation เท่ากับ 0.7254 และใช้เวลาในการจัดกลุ่ม 23.08 วินาที

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 5
ปัญหาที่พบและบทสรุป

5.1 สรุปผลการทดลอง

การทำโครงงานครั้งนี้ ได้ศึกษาทฤษฎีฝูงมด เพื่อนำมาประยุกต์ใช้กับการจัดกลุ่มตามประเภทของข้อมูล โดยได้นำโปรแกรมสำหรับการจำลองการแบ่งกลุ่มโดยใช้ทฤษฎีฝูงมด มาทดลองจัดกลุ่มกับข้อมูลที่มีลักษณะต่างกัน

การทดสอบโปรแกรม พบว่าการนำอัลกอริทึมฝูงมดมาใช้นั้น จำเป็นต้องมีการกำหนดค่าพารามิเตอร์ต่างๆ ซึ่งจะมีผลต่อการจัดกลุ่มที่มีประสิทธิภาพมากที่สุด และจะแสดงผลการจัดกลุ่มออกมา พร้อมกับแสดงค่า Pearson correlation ซึ่งเป็นค่าที่ใช้วัดประสิทธิภาพในการจัดกลุ่ม

จากการทดลองสามารถสรุปได้ดังนี้

1) Test Set A และ Test Set C สามารถปรับค่า kp และ kd ให้อยู่ในช่วง 0.01 ถึง 0.04 ได้เท่านั้น ถ้ามากกว่า 0.04 จะทำการจัดกลุ่มได้ยาก สำหรับ Test Set B สามารถปรับค่า kp และ kd ได้ในช่วง 0.01 ถึง 0.09

2) ผลค่า Pearson correlation ของ Test Set A ที่มากที่สุด คือ 0.5686 เกิดจากการปรับค่า kp และ kd เท่ากับ 0.02 และ 0.01 ตามลำดับ

3) ผลค่า Pearson correlation ของ Test Set B ที่มากที่สุด คือ 0.8654 เกิดจากการปรับค่า kp และ kd เท่ากับ 0.04 และ 0.06 ตามลำดับ

4) ผลค่า Pearson correlation ของ Test Set C ที่มากที่สุด คือ 0.7254 เกิดจากการปรับค่า kp และ kd เท่ากับ 0.02 และ 0.03 ตามลำดับ

5) ผลค่า Pearson correlationที่ได้จากการปรับค่า kp และ kdไม่มีความสัมพันธ์เชิงเส้นต่อกัน

6) ถ้ากำหนดค่า Step size น้อยกว่า 30 จะทำให้ใช้เวลาในการจัดกลุ่มนาน และอาจเกิดการรวมกันของข้อมูลต่างประเภทได้

7) ถ้ากำหนดค่า kp มากกว่า kd ในอัตราส่วนที่มากกว่า 3 ต่อ 1 เวลาทำการทดลองจัดกลุ่มข้อมูลเสร็จ อาจมีข้อมูลจำนวนมาก ที่ไม่สามารถหากลุ่มได้

8) ถ้ากำหนดค่า kp น้อยกว่า kd ในอัตราส่วนน้อยกว่า 1 ต่อ 2 เวลาทำการทดลองจัดกลุ่มข้อมูลเสร็จ อาจเกิดการรวมกันของข้อมูลต่างประเภทได้ และใช้เวลาในการจัดกลุ่มนาน

9) แม้ว่าจะกำหนดค่าพารามิเตอร์ต่างๆเท่ากันทั้งหมด แล้ว Generate ข้อมูลออกมา ภาพของการจำลองจัดกลุ่มข้อมูล และค่า Pearson correlation ก็จะออกมาไม่เหมือนกัน เนื่องจากก่อนที่จะทำการจัดกลุ่ม ข้อมูลทั้งหมดได้ถูกกระจายอย่างสุ่ม ซึ่งการสุ่มแต่ละครั้งตำแหน่งของข้อมูลจะไม่เหมือนกัน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

5.2 ข้อจำกัดของโครงงาน

1) โปรแกรมสามารถทดลองจัดกลุ่มข้อมูลได้เฉพาะ Data set 3 ชุด ที่สร้างขึ้นมาเท่านั้น

2) ถ้าในชุดข้อมูลมีมากกว่า 4 กลุ่ม จะทำการจัดกลุ่มได้ยากมาก

3) ไม่สามารถลดความเร็วในการเคลื่อนที่ของข้อมูลขณะทำการ generate

4) การจัดกลุ่มของข้อมูล ไม่สามารถจัดกลุ่มได้ 100% หมายความว่าในการทดลองแต่ละครั้งจะมีข้อมูลที่ไม่สามารถหากลุ่มได้

5) ไม่สามารถเพิ่มขนาดของหน้าจอที่แสดงผลการจำลองจัดกลุ่มได้

5.3 ข้อเสนอแนะและแนวทางในการพัฒนาต่อไปในอนาคต

1) ควรระวังการนำชุดข้อมูลที่มีจำนวนกลุ่มมากกว่า 4 กลุ่ม มาทำการทดลอง เพราะการจัดกลุ่มจะทำได้ยากมาก

2) ในการทดลองควรระวังการปรับค่า kp และ kd ถ้าปรับค่าต่างกันมาก จะทำให้การจัดกลุ่มมีประสิทธิภาพลดลง

3) สามารถเปลี่ยนแปลงข้อมูล Input ที่สร้างขึ้นให้เป็นแบบการแจกแจงชนิดอื่น เพื่อให้การจัดกลุ่มมีประสิทธิภาพเพิ่มขึ้นได้

4) ปรับปรุงอัลกอริทึมโดยใช้หลักการวางฟีโรโมนของมด มาช่วยเพิ่มประสิทธิภาพในการจัดกลุ่มข้อมูล

5) ปรับปรุงแก้ไขโปรแกรมให้สามารถนำ Data set จริงๆ ซึ่งเป็น Data set ที่ไม่ได้เกิดจากการสร้างขึ้นมา เพื่อมาทำการทดลองจัดกลุ่มข้อมูลได้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บรรณานุกรม

[1] J. Handl. Ant-based methods for tasks of clustering and topographic mapping: extensions, analysis and comparison with alternative methods. Master thesis. Chair of Artificial Intelligence, University of Erlangen-Nuremberg, Germany. November 2003.

[2] J. Handl. and B. Meyer. Improved ant-based clustering and sorting in a document retrieval Interface. In Proceedings of the Seventh International Conference on Parallel Problem Solving from Nature, volume 2439 of LNCS, pages 913-923. Springer-Verlag, Berlin, Germany, 2002.

[3] J.-L. Deneubourg, S. Goss, N. Franks, A. Sendova-Franks, C. Detrain, and L. Chretien. The dynamics of collective sorting: Robot-like ants and ant-like robots. In J.-A. Meyer and S. Wilson, editors, Proceedings of the First International Conference on Simulation of Adaptive Behaviour: From Animals to Animats 1, pages 356-365. MIT Press, Cambridge, MA, 1991.

[4] M. Dorigo and G. Di Caro. Ant Colony Optimization: A new meta-heuristic. In D. Corne, M. Dorigo, and F. Glover, editors, New Ideas in Optimization, pages 11-32. McGraw - Hill, London, UK, 1999.

[5] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", the second Edition, Morgan Kaufmann Publishers, March 2006.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้