ห้องสมุดคณะเทคโนโลยีสารสนเทศ พระจอมเกล้าลาดกระบัง

การเปรียบเทียบการจัดกลุ่มข้อมูลแบบแบ่งส่วนข้อมูลในดาต้าไมน์นิ่ง

A COMPARISON OF CLUSTERING WITH THE PARTITIONING METHOD IN DATA MINING

โดย

สิรารมย์ สมบัติ

SIRAROM SOMBAT

อาจารย์ที่ปรึกษา

รศ.ดร.อาริต ธรรมโน

H005957

วท.
ส๗๒๗๑
2551

เลขหมู่.....
เลขทะเบียน 05957
วัน.เดือน.ปี 3 ก.พ. 2553

b.12174841
i.....

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ภาคฤดูร้อน ปีการศึกษา 2551

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

A COMPARISON OF CLUSTERING WITH THE PARTITIONING METHOD IN DATA MINING

SIRAROM SOMBAT

A SYSTEM DEVELOPMENT PROJECT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF SCIENCE PROGRAM IN INFORMATION TECHNOLOGY

FACULTY OF INFORMATION TECNOLOGY

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

SUMMER/ 2008

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

COPYRIGHT 2009

FACULTY OF INFORMATION TECHNOLOGY

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

หัวข้อ	การเปรียบเทียบการจัดกลุ่มข้อมูลแบบแบ่งส่วนข้อมูลในดาต้าไมน์นิ่ง
นักศึกษา	นางสาวสิรารมย์ สมบัติ
รหัสนักศึกษา	49066825
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2551
อาจารย์ที่ปรึกษา	รศ.ดร. อาริต ธรรมโน

บทคัดย่อ

ดาต้าไมน์นิ่งเป็นวิธีอย่างหนึ่งที่ช่วยในการค้นหาความรู้ ความสัมพันธ์และคุณลักษณะของข้อมูลที่ซ่อนอยู่ในฐานข้อมูล เพื่อนำความรู้ที่ได้มาช่วยประกอบการตัดสินใจในองค์กร ซึ่งรูปแบบที่ใช้ในการทำดาต้าไมน์นิ่งนั้นมีมากมาย วิธีการจัดกลุ่มข้อมูลเป็นเทคนิคหนึ่งในดาต้าไมน์นิ่ง ซึ่งนำมาใช้สังเกตคุณลักษณะของแต่ละกลุ่มข้อมูลที่มีลักษณะใกล้เคียงกัน ให้อยู่ในกลุ่มข้อมูลเดียวกัน แต่จะมีความแตกต่างเมื่อเปรียบเทียบกับกลุ่มอื่น ซึ่งการจัดกลุ่มที่ดีนั้นจำเป็นที่จะต้องได้ศูนย์กลางของกลุ่มที่ดีที่มีระยะห่างใกล้กับข้อมูลสมาชิกมากที่สุดเพื่อให้การจัดกลุ่มข้อมูลนั้นมีประสิทธิภาพ

ในโครงการศึกษานี้ได้นำหลักการของวิธีการจัดกลุ่มข้อมูล พร้อมทั้งนำเสนอวิธีการและอัลกอริทึมต่างๆในการจัดกลุ่มข้อมูล ได้แก่ การจัดกลุ่มข้อมูลแบบ K-Means, การจัดกลุ่มข้อมูลแบบโครงข่ายประสาทเทียมโดยวิธีเซลฟ์ออร์แกไนซิ่งแม็บ และการจัดกลุ่มข้อมูลแบบฟัซซี่ซีมีน เพื่อใช้ในการวิเคราะห์จัดกลุ่มข้อมูลขนาดใหญ่และข้อมูลที่ซ้ำซ้อน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Title	A Comparison Of Clustering with The Partitioning Method in Data Mining
Student	Miss Sirarom Sombat
Student ID	49066825
Degree	Master of Science
Programme	Information Science
Academic Year	2009
Advisor	Assoc. Prof. Dr.Arit Thammano

ABSTRACT

Currently, there are a lot of data mining techniques to help in analyzing data for making the decision in the organization. However, there are many methods to build data mining. Clustering is one of data mining which observe a division of data into groups of similar objects while it's different from another group. Thus, the way of data arrangement needs to be in the center of a good data that keep the distance with data member most in order to effectively arrange data.

In this project, focus on studying clustering technique with K-Means, Self-organizing Map and Fuzzy C-Means algorithms which is one of the operations to execute of the complex datasets.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

กิตติกรรมประกาศ

โครงการพัฒนาระบบงานนี้สำเร็จลุล่วงได้อย่างดีด้วยคำปรึกษา คำแนะนำ และคำตักเตือนในสิ่งที่บกพร่องจาก ผศ.ดร.อาริต ธรรมโน อาจารย์ที่ปรึกษา และขอกราบพระคุณคณาจารย์ทุกๆท่านที่ได้ประสิทธิ์ประสาทวิชาให้ความรู้ต่างๆ เพื่อนำมาใช้ในการพัฒนาโครงการ ตลอดจนเจ้าหน้าที่ของคณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบังทุกคนที่ให้ความช่วยเหลือในเรื่องต่างๆ

สุดท้ายนี้ขอกราบขอบพระคุณ บิดา มารดา และครอบครัวที่เป็นกำลังใจที่ดีที่สุดในโลก ซึ่งให้การสนับสนุนในทุกๆเรื่อง เพื่อนๆพี่ๆน้องๆทุกคนในสาขาวิทยาการสารสนเทศ รวมถึง บูโร่, นุช, หญิงวิ, เพลี้ย ที่คอยให้กำลังใจและช่วยเหลือจนทำให้การพัฒนาโครงการสำเร็จลุล่วงด้วยดี

คุณค่าและประโยชน์อันพึงมาจากโครงการฉบับนี้ ขอมอบแด่ผู้มีพระคุณทุกท่าน

สิรารมย์ สมบัติ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ

หน้า

บทคัดย่อภาษาไทย.....I

บทคัดย่อภาษาอังกฤษ.....II

กิตติกรรมประกาศ.....III

สารบัญ.....IV

สารบัญตาราง.....VI

สารบัญรูป.....VII

บทที่ 1 บทนำ.....1

1.1 ความเป็นมาของโครงการ.....1

1.2 วัตถุประสงค์ของโครงการ.....1

1.3 ขอบเขตของโครงการ.....2

1.4 ขั้นตอนการดำเนินงาน.....2

1.5 ประโยชน์ที่คาดว่าจะได้รับ.....2

บทที่ 2 ทฤษฎีที่เกี่ยวข้อง.....3

2.1 คาต้าไมน์นิ่ง (Data Mining).....3

2.2 ประเภทข้อมูลที่สามารถทำคาต้าไมน์นิ่ง.....4

2.3 ลักษณะเฉพาะของข้อมูลที่สามารถทำคาต้าไมน์นิ่ง.....4

2.4 กระบวนการของคาต้าไมน์นิ่ง.....4

2.5 เทคนิคต่าง ๆ ของคาต้าไมน์นิ่ง.....8

2.6 ประโยชน์ของคาต้าไมน์นิ่ง.....9

2.7 นิวรอลเน็ตเวิร์ก (Neural Network).....9

2.8 ฟัซซี่ลอจิก.....12

บทที่ 3 ทฤษฎีที่นำมาใช้.....14

3.1 การจัดกลุ่มข้อมูล (Clustering).....14

3.1.1 K-Means Algorithm.....14

3.1.2 Fuzzy C-Means Algorithm.....15

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ (ต่อ)

หน้า

3.1.3 Self Organizing Map Algorithm.....21

บทที่ 4 การพัฒนาระบบและผลดำเนินการศึกษา.....33

4.1 กำหนดวัตถุประสงค์.....33

4.2 ขั้นตอนและรายละเอียดการใช้งาน.....33

4.3 ข้อมูลที่นำมาทดลอง.....41

4.4 ผลการจัดกลุ่มข้อมูล.....43

4.4.1 ผลการจัดกลุ่มข้อมูล Iris.....43

4.4.2 ผลการจัดกลุ่มข้อมูล Wine.....50

4.4.3 ผลการจัดกลุ่มข้อมูลSegmentation.....53

4.4.4 ผลการจัดกลุ่มข้อมูล Diabetes.....60

บทที่ 5 สรุปผลการศึกษาและข้อเสนอแนะ.....63

5.1 สรุปผลการดำเนินงาน.....63

5.2 สรุปผลการทดลอง.....63

5.3 ข้อเสนอแนะ.....64

บรรณานุกรม.....65

ประวัติผู้เขียน.....66

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญตาราง

ตารางที่ หน้า

3.1 คำนวณหาค่าจุดศูนย์กลางของกลุ่ม.....18

3.2 คำนวณหาระยะห่างระหว่างข้อมูลชุดที่ 1.....19

3.3 คำนวณหาระยะห่างระหว่างข้อมูลชุดที่ 2.....19

3.4 คำนวณหาระยะห่างระหว่างข้อมูลชุดที่ 3.....19

3.5 คำนวณหาระยะห่างระหว่างข้อมูลชุดที่ 4.....19

3.6 ตารางแสดงค่า Matrix U.....20

3.7 ตารางข้อมูล Input.....23

3.8 ตารางแสดงนิวรอนที่ชนะ.....23

3.9 ตารางแสดงนิวรอนที่ชนะ.....25

3.10 ตารางแสดงนิวรอนที่ชนะ.....27

3.11 ตารางแสดงนิวรอนที่ชนะ.....29

3.12 ตารางแสดงนิวรอนที่ชนะ.....31

4.1 แสดงผลการจัดกลุ่มข้อมูล Iris แบบ K-Means Algorithm.....44

4.2 แสดงผลการจัดกลุ่มข้อมูล Iris แบบ Fuzzy C-Means Algorithm.....46

4.3 แสดงผลการจัดกลุ่มข้อมูล Iris แบบ Self Organizing Map Algorithm.....48

4.4 แสดงผลการจัดกลุ่มข้อมูล Iris ที่ดีที่สุด.....49

4.5 แสดงผลการจัดกลุ่มข้อมูล Wine แบบ K-Means Algorithm.....50

4.6 แสดงผลการจัดกลุ่มข้อมูล Wine แบบ Fuzzy C-Means Algorithm.....51

4.7 แสดงผลการจัดกลุ่มข้อมูล Wine แบบ Self Organizing Map.....52

4.8 แสดงผลการจัดกลุ่มข้อมูล Wine ที่ดีที่สุด.....53

4.9 แสดงผลการจัดกลุ่มข้อมูล Segmentation แบบ K-Means Algorithm.....54

4.10 แสดงผลการจัดกลุ่มข้อมูล Segmentation แบบ Fuzzy C-Means Algorithm.....55

4.11 แสดงผลการจัดกลุ่มข้อมูล Segmentation แบบ Self Organizing Map.....57

4.12 แสดงผลการจัดกลุ่มข้อมูล Segmentation ที่ดีที่สุด.....59

4.13 แสดงผลการจัดกลุ่มข้อมูล Diabetes แบบ K-Means Algorithm.....60

4.14 แสดงผลการจัดกลุ่มข้อมูล Diabetes แบบ Fuzzy C-Means Algorithm.....61

4.15 แสดงผลการจัดกลุ่มข้อมูล Diabetes แบบ Self Organizing Map.....61

4.16 แสดงผลการจัดกลุ่มข้อมูล Diabetes ที่ดีที่สุด.....62

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป

รูปที่ หน้า

2.1 แสดงขั้นตอนของกระบวนการทำดาต้าไมน์นิ่ง.....5

2.2 แสดงขั้นตอนการทำงานของดาต้าไมน์นิ่ง.....5

2.3 แสดง Model ของ Neuron ในสมองมนุษย์.....10

2.4 แสดง Model ของ Neuron ในคอมพิวเตอร์.....10

2.5 แสดงการเรียนรู้แบบมีการสอน (Supervised Learning).....11

2.6 แสดงการเรียนรู้แบบไม่มีการสอน Unsupervised Learning.....12

2.7 แสดงตรรกะแบบจริงเท็จ (บูลีนลอจิก) และตรรกะแบบฟัซซี่(ฟัซซี่ลอจิก).....13

3.1 ลักษณะของโครงข่ายของ Self Organizing Map.....21

3.2 ลักษณะของ neighborhood $t_1 \langle t_2 \langle t_3 \ldots$ 22

3.3 แสดงการกำหนดค่าให้โครงข่าย.....23

3.4 แสดงการปรับค่า weight โครงข่ายเมื่อ input คือ $[0,6]$.....25

3.5 แสดงการปรับค่า weight โครงข่ายเมื่อ input คือ $[0,0]$.....27

3.6 แสดงการปรับค่า weight โครงข่ายเมื่อ input คือ $[4,1]$.....29

3.7 แสดงการปรับค่า weight โครงข่ายเมื่อ input คือ $[5,0]$.....31

3.8 แสดงการปรับค่า weight โครงข่ายเมื่อ input คือ $[5,5]$.....32

4.1 หน้าจอเริ่มต้นของระบบ.....33

4.2 หน้าจอหลักของโปรแกรม.....34

4.3 หน้าจอการทำงานของ K-Means อัลกอริทึม.....35

4.4 การเปิดไฟล์ข้อมูลที่ต้องการนำมาประมวลผล.....35

4.5 แสดงหน้าจอหลังการประมวลผลจัดกลุ่ม.....36

4.6 แสดงข้อมูลที่อยู่ในรูปแบบ XML.....36

4.7 หน้าจอการทำงานของ Fuzzy C-Means อัลกอริทึม.....37

4.8 การเปิดไฟล์ข้อมูลที่ต้องการนำมาประมวลผล.....37

4.9 แสดงหน้าจอหลังการประมวลผลจัดกลุ่ม.....38

4.10 แสดงข้อมูลที่อยู่ในรูปแบบ XML.....38

4.11 หน้าจอการทำงานของ SOM อัลกอริทึม.....39

4.12 การเปิดไฟล์ข้อมูลที่ต้องการนำมาประมวลผล.....39

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป(ต่อ)

รูปที่ หน้า

4.13 แสดงหน้าจอหลังการประมวลผลจัดกลุ่ม.....40

4.14 หน้าจอการแสดงวิธีการใช้งานโปรแกรม Data Mining Clustering Application.....40

4.15 หน้าจอรายละเอียดโปรแกรมที่พัฒนา, เวอร์ชั่น.....41

4.16 ข้อมูล Iris ที่แบ่งตามคลาส.....41

4.17 ข้อมูล Wine ที่แบ่งตามคลาส.....42

4.18 ข้อมูล Segmentation ที่แบ่งตามคลาส.....42

4.19 ข้อมูล Diabetes ที่แบ่งตามคลาส.....43

4.20 การใส่ค่าพารามิเตอร์ของ K-Means Algorithm.....43

4.21 แสดงหน้าจอสรุปผลการจัดกลุ่มของ K-Means Algorithm.....44

4.22 การใส่ค่าพารามิเตอร์ของ Fuzzy C-Means Algorithm.....45

4.23 แสดงหน้าจอสรุปผลการจัดกลุ่มของ Fuzzy C-Means Algorithm.....46

4.24 การใส่ค่าพารามิเตอร์ของ Self Organizing Map Algorithm.....47

4.25 แสดงหน้าจอสรุปผลการจัดกลุ่มของ Self Organizing Map Algorithm.....48

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 1
บทนำ

1.1 ความเป็นมาของโครงการ

ในปัจจุบันองค์กรต่างๆ ได้มีการนำเอาระบบสารสนเทศเข้ามาใช้งานเพื่อเพิ่มประสิทธิภาพในการดำเนินงานขององค์กรเพิ่มมากขึ้น จึงทำให้มีข้อมูลจำนวนมากเกิดขึ้นภายในระบบ โดยข้อมูลที่เกิดขึ้นนั้นได้ถูกจัดเก็บลงฐานข้อมูลที่มีความแตกต่างกันไปแล้วแต่ความต้องการของแต่ละองค์กรในปัจจุบัน ซึ่งข้อมูลต่างๆ ที่เกิดขึ้นในระบบนั้นต่างก็เป็นข้อมูลที่มีประโยชน์อย่างมาก หากแต่ละองค์กรรู้จักวิธีการที่จะจัดการกับข้อมูลที่ถูกต้อง แต่ในความเป็นจริงแล้วในหลายองค์กรยังไม่สามารถที่จะนำข้อมูลที่มีประโยชน์เหล่านี้มาใช้งานได้อย่างคุ้มค่าเท่าที่ควร ดังนั้นจึงได้มีแนวคิดเกี่ยวกับการนำกระบวนการทางด้าน Data Mining ซึ่งเป็นกระบวนการในการเพิ่มประสิทธิภาพของระบบข้อมูลข่าวสารที่มีอยู่ให้มีประโยชน์สูงสุด มาใช้กับข้อมูลต่างๆ ที่เกิดของภายในองค์กร โดยได้นำเอาเทคนิคที่ใช้ในกระบวนการจัดกลุ่มข้อมูล (Clustering) คือ K-Means Algorithm, Fuzzy C-Means Algorithm และ Self Organizing Map Algorithm มาประยุกต์ใช้ในการวิเคราะห์ข้อมูลที่เกิดขึ้นของแต่ละองค์กร รวมถึงช่วยในการสนับสนุนการตัดสินใจของผู้บริหารอีกด้วย โดยในปัจจุบันได้มีการพัฒนา เครื่องมือที่ช่วยสนับสนุนกระบวนการทางด้านดาต้าไมน์นิ่งมากมาย ไม่ว่าจะเป็น Microsoft SQL Server และอื่นๆ เพื่อที่จะช่วยอำนวยความสะดวกแก่องค์กรที่ต้องการใช้ความสามารถทางด้านดาต้าไมน์นิ่งในการวิเคราะห์ข้อมูลที่ตนมีอยู่ให้เกิดประโยชน์ และใช้ในการวางแผนทางการตลาดขององค์กรต่อไป

1.2 วัตถุประสงค์ของโครงการ

1. เพื่อศึกษาวิธีการและขั้นตอนการทำงานของดาต้าไมน์นิ่ง
2. เพื่อศึกษากระบวนการในการจัดกลุ่มข้อมูลของ K-Means Algorithm
3. เพื่อศึกษากระบวนการในการจัดกลุ่มข้อมูลของ Fuzzy C-Means Algorithm
4. เพื่อศึกษากระบวนการในการจัดกลุ่มข้อมูลของ Self Organizing Map Algorithm
5. เพื่อนำข้อมูลที่มีอยู่มาใช้ให้เกิดประโยชน์มากที่สุด
6. เพื่อนำเอาความรู้ในกระบวนการจัดกลุ่มไปประยุกต์ใช้ในการพัฒนาแอพพลิเคชั่น
7. เพื่อนำข้อมูลที่ได้จากการจัดกลุ่มไปใช้ให้เกิดประโยชน์ต่อองค์กร

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

1.3 ขอบเขตของโครงการ

โครงการนี้จะเป็นการศึกษาและพัฒนาระบบโดยใช้เทคนิคทางด้านดาต้าไมน์นิ่งเพื่อใช้ในการจัดกลุ่มข้อมูล (Clustering) โดยจะทำการพัฒนาแอพพลิเคชั่นจัดกลุ่มข้อมูล ที่ข้อมูลถูกจัดเก็บอยู่ในรูปแบบไฟล์ CSV คือเป็นเท็กซ์ไฟล์ประเภทหนึ่ง ที่มีโครงสร้างแน่นอน โดยมีคอมม่า ',' เป็นตัวคั่นระหว่างข้อมูล ซึ่งจะศึกษาเกี่ยวกับวิธีใช้ไฟล์ CSV และการนำข้อมูลมาใช้เพื่อให้เกิดประสิทธิภาพสูงสุด

โดยขอบเขตของระบบจะเป็นการดำเนินการศึกษาเปรียบเทียบการทำงานเทคนิคที่ใช้ในกระบวนการจัดกลุ่มข้อมูล (Clustering) คือ K-Means Algorithm, Fuzzy C-Means Algorithm และ Self Organizing Map Algorithm

1.4 ขั้นตอนการดำเนินงาน

1. กำหนดหัวข้อ เป้าหมาย และวัตถุประสงค์ ตลอดจนขอบเขตของโครงการ
2. ศึกษาทฤษฎี บทความ งานวิจัย และหนังสือที่เกี่ยวข้องกับกระบวนการและเทคนิคทางด้านดาต้าไมน์นิ่ง รวมถึงเทคนิคของการจัดกลุ่มข้อมูล K-Means, Fuzzy C-Means Algorithm, Self Organizing Map Algorithm
3. ศึกษาเทคนิคต่างๆ ในการพัฒนาโปรแกรม
4. ทำการออกแบบและพัฒนาโปรแกรม
5. ทำการทดสอบโปรแกรม เพื่อหาข้อผิดพลาดที่เกิดขึ้นและทำการปรับปรุงแก้ไขข้อผิดพลาดที่เกิดขึ้นนั้น
6. จัดทำเอกสาร

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. เพื่อได้เรียนรู้ทฤษฎี เทคนิคและวิธีการของการจัดกลุ่มข้อมูล
2. เพื่อนำเอาความรู้ที่มีไปประยุกต์ใช้ในการพัฒนาโปรแกรมเพื่อทำการจัดกลุ่มข้อมูลต่างๆ
3. เพื่อนำเอาข้อมูลที่มีอยู่มาวิเคราะห์และนำไปใช้ให้เกิดประโยชน์มากขึ้น
4. เพื่อให้การทำงานในองค์กรเร็วขึ้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 2
ทฤษฎีที่เกี่ยวข้อง

ความสำคัญในการรวบรวมข้อมูลในการดำเนินธุรกิจเป็นสิ่งที่ควรให้ความสำคัญอย่างยิ่ง เนื่องจากเป็นองค์ประกอบสำคัญที่จะทำให้ธุรกิจบรรลุความสำเร็จ ซึ่งในการที่จะเปลี่ยนข้อมูลจำนวนมากมาเป็นกุญแจแห่งความสำเร็จ จำเป็นต้องหาวิธีที่จะดึงความรู้เหล่านั้นออกมาจากข้อมูลที่เก็บรวบรวมไว้ซึ่งมีอยู่จำนวนมาก ซึ่งวิธีการที่กล่าวถึงคือหลักการของดาต้าไมน์นิ่ง

2.1 ดาต้าไมน์นิ่ง (Data Mining)

ดาต้าไมน์นิ่งเป็นกระบวนการในการค้นหาสารสนเทศที่มีประโยชน์จากฐานข้อมูลที่มีอยู่ ซึ่งการนำเอาดาต้าไมน์นิ่งมาใช้ในการค้นหาสารสนเทศจากฐานข้อมูลนั้นเนื่องมาจากว่า

1. ปัจจุบันในการดำเนินการทางธุรกิจทั้งองค์กรขนาดเล็ก ขนาดกลาง และขนาดใหญ่ ซึ่งจะต้องใช้ฐานข้อมูลมาเก็บรวบรวมข้อมูลเกี่ยวกับกิจกรรมต่างๆ ที่เกิดขึ้น ซึ่งในแต่ละวันข้อมูลที่เก็บก็จะเพิ่มขึ้นเรื่อยๆ และจากข้อมูลทั้งหมดที่เก็บอยู่ในฐานข้อมูลอาจจะมีลักษณะบางอย่างแอบแฝงอยู่ เช่น ความสัมพันธ์ แนวโน้ม (Trend) หรือรูปแบบเฉพาะ (Pattern) ซึ่งสารสนเทศเหล่านี้มีประโยชน์อย่างมากในการตัดสินใจเชิงธุรกิจ เพื่อให้องค์กรมีโอกาสในการแข่งขันมากยิ่งขึ้น แต่เนื่องจากการที่จะวิเคราะห์หาประโยชน์จากฐานข้อมูลขนาดใหญ่นี้จำเป็นต้องใช้หลักการดาต้าไมน์นิ่งช่วยในการวิเคราะห์ให้ได้สารสนเทศที่ถูกต้อง และสามารถนำไปใช้ประโยชน์ได้จริงในการดำเนินธุรกิจ

2. จากการที่มีข้อมูลจำนวนมากในฐานข้อมูลอีกทั้งข้อมูลเหล่านั้นเป็นข้อมูลที่ซับซ้อน ทำให้การวิเคราะห์ข้อมูลให้ถูกต้องนั้นยากเกินกว่าที่จะวิเคราะห์โดยมนุษย์ได้ ซึ่งมนุษย์อาจจะวิเคราะห์ข้อมูลเหล่านั้นได้ไม่ครอบคลุม หรือมองข้ามตัวแปรบางตัวที่มีผลกับผลลัพธ์ที่จะเกิดขึ้น แต่ถ้าใช้หลักการดาต้าไมน์นิ่งในการวิเคราะห์สามารถมั่นใจได้ว่าการวิเคราะห์นั้นถูกต้องครอบคลุมทุกตัวแปร

3. เมื่อเปรียบเทียบกันระหว่างองค์กรที่จ้างผู้เชี่ยวชาญหลายๆ คนกับองค์กรที่พัฒนาเครื่องมือโดยใช้หลักการดาต้าไมน์นิ่งมาใช้วิเคราะห์ข้อมูลจำนวนมากและซับซ้อนนั้นจะเห็นว่าค่าใช้จ่ายในการจ้างผู้เชี่ยวชาญหลายๆ คนอาจจะสูงกว่าการพัฒนาเครื่องมือดาต้าไมน์นิ่งก็ได้ รวมทั้งการวิเคราะห์โดยมนุษย์จะใช้เวลานานกว่าการวิเคราะห์ของเครื่องมือดาต้าไมน์นิ่งมาก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.2 ประเภทข้อมูลที่สามารถทำดาต้าไมน์นิ่ง

1. Relational Database ซึ่งเป็นฐานข้อมูลที่จัดเก็บอยู่ในรูปแบบของตาราง โดยในแต่ละตารางจะประกอบไปด้วยแถวและคอลัมน์ ความสัมพันธ์ของข้อมูลทั้งหมดสามารถแสดงได้โดย Entity-relationship (ER) model

2. Data Warehouses เป็นการเก็บรวบรวมข้อมูลจากหลายแหล่งมาเก็บไว้ในรูปแบบเดียวกันและรวบรวมไว้ในที่ๆ เดียวกัน

3. Transactional Database ประกอบด้วยข้อมูลที่แต่ละทรานเซกชันแทนด้วยเหตุการณ์ในขณะใดขณะหนึ่ง เช่น ใบเสร็จรับเงิน จะเก็บข้อมูลในรูป ชื่อลูกค้าและรายการสินค้าที่ลูกค้ารายนั้นซื้อ เป็นต้น

4. Advanced Database เป็นฐานข้อมูลที่จัดเก็บในรูปแบบอื่นๆ เช่น ข้อมูลแบบ Object-oriented, ข้อมูลที่เป็น Text file, ข้อมูลมัลติมีเดีย, ข้อมูลในรูปของ Web

2.3 ลักษณะเฉพาะของข้อมูลที่สามารถทำดาต้าไมน์นิ่ง

1. ข้อมูลขนาดใหญ่เกินกว่าจะพิจารณาความสัมพันธ์ที่ซ่อนอยู่ภายในข้อมูลได้ด้วยตาเปล่า หรือโดยการใช้ Database Management System (DBMS) ในการจัดการฐานข้อมูล

2. ข้อมูลที่มาจากหลายแหล่งโดยอาจจะรวบรวมมาจากหลายระบบปฏิบัติการหรือหลาย DBMS เช่น Oracle, DB2, MS SQL, MS Access เป็นต้น

3. ข้อมูลที่ไม่มีการเปลี่ยนแปลงตลอดช่วงเวลาที่ทำการไมน์นิ่ง (Mining) หากข้อมูลที่มีอยู่นั้นเป็นข้อมูลที่เปลี่ยนแปลงตลอดเวลาจะต้องแก้ปัญหานี้ก่อน โดยบันทึกฐานข้อมูลนั้นไว้และนำฐานข้อมูลที่บันทึกไว้มาทำไมน์นิ่ง แต่เนื่องจากข้อมูลนั้นมีการเปลี่ยนแปลงอยู่ตลอดเวลา จึงทำให้ผลลัพธ์ที่ได้จากการทำไมน์นิ่งสมเหตุสมผลในช่วงเวลาหนึ่งเท่านั้น ดังนั้นเพื่อให้ได้ผลลัพธ์ที่มีความถูกต้องเหมาะสมอยู่ตลอดเวลาจึงต้องทำไมน์นิ่งใหม่ทุกครั้งในช่วงเวลาที่เหมาะสม

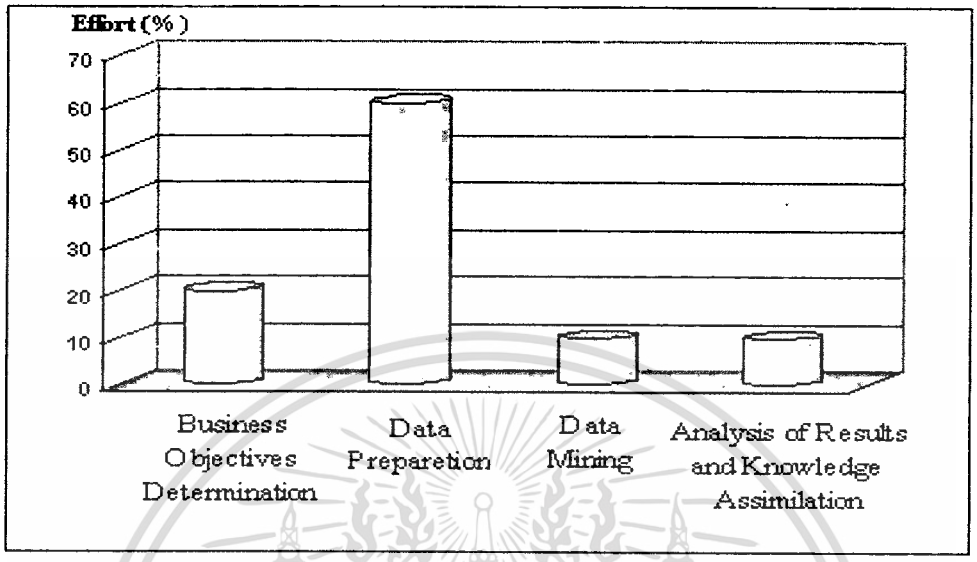
4. ข้อมูลที่มีโครงสร้างซับซ้อน เช่น ข้อมูลรูปภาพ ข้อมูลมัลติมีเดีย ข้อมูลเหล่านี้สามารถนำมาทำไมน์นิ่งได้เช่นกันแต่ต้องใช้เทคนิคการทำดาต้าไมน์นิ่งขั้นสูง

2.4 กระบวนการของดาต้าไมน์นิ่ง

ถ้ากล่าวถึงดาต้าไมน์นิ่งส่วนใหญ่จะให้ความสำคัญกับการไมน์นิ่งข้อมูล แต่ที่จริงแล้วการไมน์นิ่งข้อมูลเป็นเพียงขั้นตอนหนึ่งของกระบวนการดาต้าไมน์นิ่งเท่านั้น

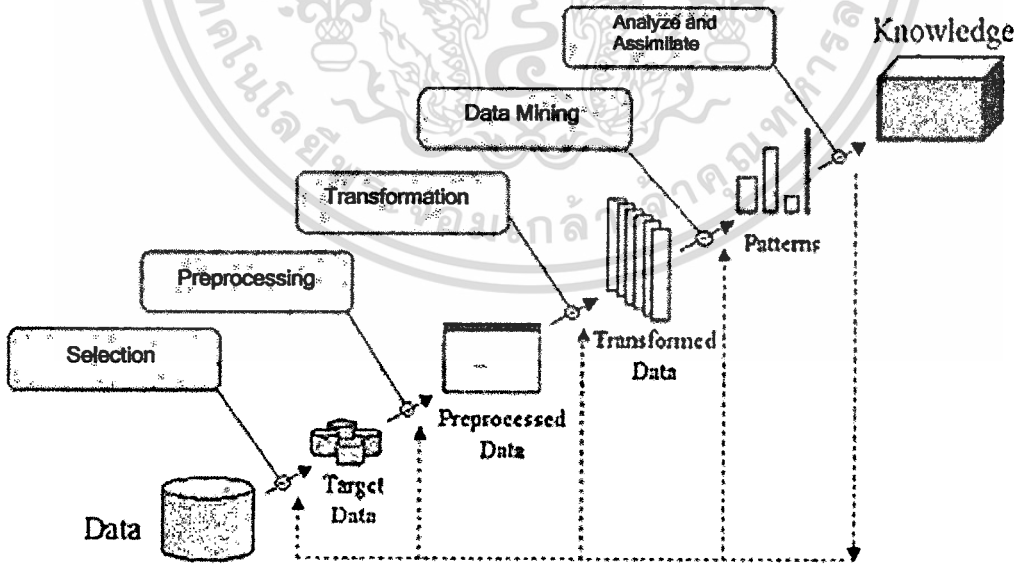
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

โดยขั้นตอนหลักในการทำดาต้าไมน์นิ่งมี 5 ขั้นตอน ดังรูปที่ 2.1 ซึ่งจะแสดงเปอร์เซ็นต์ในการทำงานในแต่ละขั้นตอนด้วยดังนี้



รูปที่ 2.1 แสดงขั้นตอนของกระบวนการทำดาต้าไมน์นิ่ง

จากรูปที่ 2.1 มี 5 ขั้นตอนหลัก และสามารถแบ่งการทำงานออกเป็นขั้นตอนย่อยได้ดังรูปที่ 2.2 โดยสามารถอธิบายในแต่ละขั้นตอนได้ดังนี้



รูปที่ 2.2 แสดงขั้นตอนการทำงานของดาต้าไมน์นิ่ง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ขั้นตอนที่ 1: กำหนดวัตถุประสงค์ทางธุรกิจ (Business Objectives Determination)

การกำหนดวัตถุประสงค์ทางธุรกิจจะต้องเข้าใจถึงปัญหาและความต้องการทางธุรกิจ เพราะจะเป็นตัวกำหนดทิศทางในการทำดาต้าไมน์นิ่งและสามารถกำหนดได้ว่าเมื่อไหร่จะใช้ดาต้าไมน์นิ่งในการแก้ปัญหา เนื่องจากในทุกปัญหาไม่สามารถแก้ไขได้ด้วยหลักการดาต้าไมน์นิ่งทั้งหมด ซึ่งในส่วนนี้จะประกอบด้วยการวิเคราะห์ทางธุรกิจและวิเคราะห์ข้อมูลเบื้องต้นว่าเรามีข้อมูลอะไรอยู่และต้องการอะไรจากข้อมูล ปัญหามีขอบเขตแค่ไหน ซึ่งในขั้นตอนนี้จะสามารถมองถึงอัลกอริทึมและฐานข้อมูลที่จะใช้เบื้องต้นที่สัมพันธ์กับวัตถุประสงค์ทางธุรกิจได้ การกำหนดวัตถุประสงค์ทางธุรกิจนี้จะใช้เวลาประมาณ 20 % ของการทำดาต้าไมน์นิ่ง

ขั้นตอนที่ 2: การเตรียมข้อมูล (Data preparation)

การเตรียมข้อมูลเป็นขั้นตอนที่ใช้เวลานานที่สุดประมาณ 60 % ของการทำดาต้าไมน์นิ่ง เพราะเป็นส่วนสำคัญที่สุดในการทำดาต้าไมน์นิ่ง เนื่องจากบางครั้งอาจมีการนำข้อมูลจากหลายแหล่งมารวมกันเพื่อดูความสัมพันธ์ของข้อมูล ข้อมูลที่ได้จากขั้นตอนนี้จะต้องมีความชัดเจนถูกต้อง เหมาะสม โดยขั้นตอนการเตรียมข้อมูลจะแบ่งการทำงานเป็น 3 ขั้นตอนย่อย ดังนี้

1. การเลือกข้อมูล (Data Selection)

ข้อมูลก่อนทำดาต้าไมน์นิ่งอาจจะมาจากข้อมูลหลายๆ แหล่งรวมกันจึงต้องมีการเลือกข้อมูลที่สำคัญออกมา วัตถุประสงค์ของการเลือกข้อมูลคือ การระบุลักษณะข้อมูล, เลือกข้อมูลที่ต้องการ และนำข้อมูลที่ไม่ต้องการออกไป ซึ่งการเลือกข้อมูลจะขึ้นอยู่กับวัตถุประสงค์ของแต่ละธุรกิจ การเลือกข้อมูลจำเป็นต้องมีความเข้าใจกับชนิดของข้อมูล ค่าที่เป็นไปได้ แหล่งกำเนิดของข้อมูล รูปแบบของข้อมูลและลักษณะอื่นๆ ซึ่งเราสามารถแบ่งลักษณะและรูปแบบของข้อมูลเป็น 2 ลักษณะคือ

1) **ตัวแปรแบบ Categorical**
   - Nominal: เป็นตัวแปรที่ลำดับของข้อมูลไม่มีผลกับค่า เช่น สถานะการแต่งงาน (single, married, divorced)
   - Ordinal: เป็นตัวแปรที่ลำดับของข้อมูลมีผลกับค่า เช่น เกรดของนักศึกษา (A, B, C, D, F) หรือลำดับคุณภาพของสินค้า (ดี, ปานกลาง, เลว)

2) **ตัวแปรแบบ Quantitative**
   - Continuous: ค่าที่เก็บจะเป็นจำนวนจริง (real number) หรือเป็นค่าต่อเนื่อง เช่น น้ำหนักของนักศึกษา
   - Discrete: ค่าที่เก็บเป็นเลขจำนวนเต็ม (Integer) เช่น จำนวนนักศึกษา

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2. การกลั่นกรองข้อมูล (Data preprocessing)

เมื่อเลือกข้อมูลที่ต้องการแล้วจะต้องกลั่นกรองข้อมูลเพื่อให้ข้อมูลที่เลือกมามีคุณภาพเหมาะกับจะนำไปทำดาต้าไมน์นิ่งเนื่องจากข้อมูลที่เลือกมาอาจมีค่าที่ไม่ถูกต้อง โดยประเด็นที่ต้องพิจารณาเพิ่มเติม 2 ประเด็น คือ

1) Noisy Data เป็นข้อมูลที่มีลักษณะต่างจากข้อมูลที่คาดการณ์ไว้ ซึ่งอาจจะเป็นข้อมูลในแง่ดี คือ สิ่งที่เรามองหาอยู่ หรือเป็นข้อมูลในแง่ร้าย คือเป็นข้อมูลที่ไม่สมบูรณ์ อาจเกิดจากการบันทึกข้อมูลผิดพลาด หรือความผิดพลาดในการรับส่งข้อมูลผ่าน network ดังนั้นจะต้องจัดการกับข้อมูลเหล่านี้ก่อน โดยตัดข้อมูลเหล่านี้ทิ้งถ้ามีจำนวนน้อย หรือใช้เทคนิคทางสถิติแก้ไขให้เหมาะสมก่อนนำมาวิเคาะห์

2) Missing Data เป็นข้อมูลที่มีลักษณะบางส่วนของข้อมูลขาดหายไป สาเหตุอาจเกิดคล้ายกับ Noisy Data แก้ไขโดยการตัดข้อมูลเหล่านี้ทิ้งถ้ามีจำนวนน้อย หรือ ถ้าข้อมูลที่ขาดหายมีมากก็ให้แก้ไขโดยถ้าเป็นข้อมูล Quantitative ก็ให้แทนข้อมูลเหล่านี้ด้วยค่าเฉลี่ยหรือค่าที่ปรากฏบ่อย ถ้าเป็นข้อมูลแบบ Categorical อาจแทนด้วยค่าที่ปรากฏบ่อยหรือแทนด้วย Unknown

3. การแปลงข้อมูล (Data Transformation)

เป็นการแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมกับการทำดาต้าไมน์นิ่งเช่น การแปลง Quantitative เป็น Categorical เช่นแปลงเกรดของนักศึกษาจาก A, B เป็น high แปลง C เป็น medium แปลง D, F เป็น low หรือการแปลงจาก Categorical เป็น Quantitative เช่น รถยนต์ แปลงเป็น 01 รถจักรยานยนต์ แปลงเป็น 02

ขั้นตอนที่ 3: การทำดาต้าไมน์นิ่ง

เป็นขั้นตอนการประมวลผลข้อมูลตามอัลกอริทึมที่กำหนดไว้ โดยถ้าผลลัพธ์ที่ได้จากขั้นตอนนี้ไม่เป็นไปอย่างที่คาดหวังก็อาจย้อนกลับไปทำขั้นตอนกลั่นกรองข้อมูลใหม่ได้ ในขั้นตอนนี้จะเกี่ยวข้องกับการเลือกอัลกอริทึมในการทำดาต้าไมน์นิ่งซึ่งจะต้องพิจารณาลักษณะของปัญหาเป็นหลัก เพราะในแต่ละปัญหาต้องเลือกใช้อัลกอริทึมที่เหมาะสมจึงจะได้ผลการวิเคราะห์ที่ถูกต้อง ซึ่งอาจจะใช้หลายอัลกอริทึมเพื่อเปรียบเทียบผลลัพธ์ได้

ขั้นตอนที่ 4: การวิเคราะห์ผลลัพธ์ที่ได้จากการทำดาต้าไมน์นิ่ง (Analysis of Results)

เป็นการวิเคราะห์ผลของการประมวลผล ซึ่งจะทำการแปลความหมายและประเมินผลที่ได้จากการทำไมน์นิ่ง โดยที่เมื่อเราได้รูปแบบของความสัมพันธ์หรือโมเดลแล้ว ขั้นตอนนี้จะเป็นการทดสอบโมเดลที่ได้กับชุดข้อมูลอีกชุดหนึ่งที่เราทราบผลลัพธ์อยู่แล้ว ว่าเมื่อทดสอบกับโมเดลแล้วได้ผลลัพธ์ถูกต้องหรือยอมรับได้หรือไม่ หากยอมรับไม่ได้ก็อาจจะแก้ไขโดยเพิ่มจำนวนข้อมูลให้มากขึ้น หรือเปลี่ยนอัลกอริทึมใหม่

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ขั้นตอนที่ 5: การปรับความรู้ที่ได้เข้ากับธุรกิจ (Assimilation of Knowledge)

เป็นการรวบรวมความเข้าใจทางธุรกิจที่มีผลจากขั้นตอนการวิเคราะห์ผลลัพธ์เพื่อนำไปใช้ในโอกาสต่อไป โดยขั้นตอนนี้จะมีหลักอยู่ 2 ประการ คือ

1) การนำเสนอแนวคิดที่ค้นพบใหม่

2) การหาแนวทางที่จะนำความรู้ที่ค้นพบใหม่ไปใช้ให้เกิดประโยชน์สูงสุดต่อไป

2.5 เทคนิคต่างๆ ของดาต้าไมน์นิ่ง

ปัญหาต่างๆ ที่นำดาต้าไมน์นิ่งมาช่วยแก้ไขมีแตกต่างกันออกไป ขึ้นกับลักษณะของปัญหา และวัตถุประสงค์ของการแก้ปัญหา

การแก้ปัญหาของงานชนิดต่างๆ โดยใช้วิธีดาต้าไมน์นิ่งนั้น ในแต่ละงานก็มีเทคนิคที่จะทำไปใช้ได้อย่างเหมาะสมแตกต่างกันไป

1) Association Rule Discovery

เป็นเทคนิคการค้นหาความสัมพันธ์ของข้อมูลจากข้อมูลขนาดใหญ่ที่มีอยู่ เพื่อนำไปใช้ในการวิเคราะห์ทำนายปรากฏการณ์ต่างๆ หรือมาจากการวิเคราะห์ การซื้อสินค้า ของลูกค้าเรียกว่า Market Basket Analysis ซึ่งประเมินจากข้อมูลในตารางที่รวบรวมไว้ ผลการวิเคราะห์ที่ได้จะเป็นคำตอบของปัญหา ซึ่งการวิเคราะห์แบบนี้เป็นการใช้กฎความสัมพันธ์ (Association Rule) เพื่อหาความสัมพันธ์ของข้อมูล

2) Classification / Predictive

เป็นกระบวนการสร้าง Model จัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ ตัวอย่างเช่น การแบ่งประเภทลูกค้าว่า เชื่อถือได้หรือไม่ โดยใช้ Model Construction ซึ่งเป็นการสร้าง Model โดยการเรียนรู้จากข้อมูล ที่ได้กำหนดกลุ่มไว้เรียบร้อยแล้ว ซึ่ง Model ที่ได้อาจแสดงในรูปของแบบต้นไม้ (Decision Tree) หรือ แบบนิวรอลเน็ต (Neural Net)

3) Database Clustering / Segmentation

เป็นเทคนิคการลดขนาดของข้อมูลด้วยการรวมกลุ่มตัวแปรที่มีลักษณะเดียวกันไว้ด้วยกัน ทำให้สามารถค้นหาข้อมูลที่ถูกละเลยไปได้ เทคนิคนี้มักถูกใช้เป็นขั้นตอนเบื้องต้นในการทำดาต้าไมน์นิ่ง

4) Deviation Detection

เป็นกรรมวิธีในการหาค่าที่แตกต่างไปจากค่ามาตรฐานหรือค่าที่คาดคิดไว้ว่าต่างไปมากน้อยเพียงใด โดยทั่วไปมักใช้วิธีการทางสถิติ หรือการแสดงให้เห็นภาพ สำหรับเทคนิคนี้ใช้ในการตรวจสอบ ลายเซ็นปลอม หรือบัตรเครดิตปลอม เป็นต้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

5) Sequential Analysis

ในการวิเคราะห์ลำดับ เพื่อค้นพบรูปแบบของการปรากฏของข้อมูล ซึ่งปรากฏในรายการที่แยกออกมา เช่นถ้าผู้ซื้อของซื้อสินค้า A แล้วเขาจะซื้อสินค้า B ในภายหลัง

2.6 ประโยชน์ของดาต้าไมน์นิ่ง

1) ผลลัพธ์ที่ได้อยู่ในรูปของตัวแบบที่ง่ายต่อการเข้าใจ ผู้ที่ไม่มีความรู้พื้นฐานทางสถิติก็สามารถแปลความจากตัวแบบได้ สามารถนำสารสนเทศที่ได้ไปใช้ในกระบวนการทางธุรกิจได้

2) สามารถวิเคราะห์ข้อมูลจำนวนมากได้

3) ดาต้าไมนิ่ง จะค้นพบในสิ่งที่เราคาดไม่ถึง เนื่องจากการที่มีตัวแบบและการตรวจสอบที่หลากหลาย จะพบว่าการค้นหาจากตัวแปรที่นำมารวมกัน ทำให้ได้สารสนเทศที่เป็นประโยชน์ต่อธุรกิจ

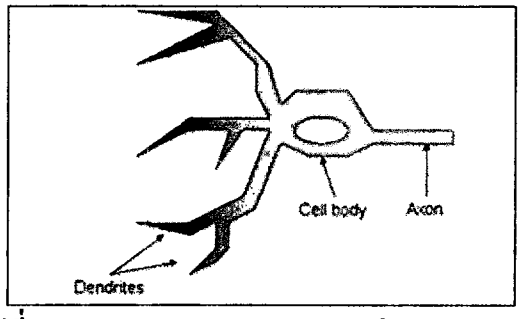
4) ตัวแบบถูกสร้างได้อย่างรวดเร็ว ตัวแบบต่างๆ จะถูกทดสอบและเลือกตัวแบบที่ดีที่สุดสำหรับผู้ใช้

2.7 นิวรอลเน็ตเวิร์ก (Neural Network)

โครงข่ายประสาทเทียม (Artificial neural network) คือโมเดลทางคณิตศาสตร์ สำหรับประมวลผลสารสนเทศด้วยการคำนวณแบบคอนเนคชันนิสต์ (connectionist) เพื่อจำลองการทำงานของเครือข่ายประสาทในสมองมนุษย์ ด้วยวัตถุประสงค์ที่จะสร้างเครื่องมือซึ่งมีความสามารถในการเรียนรู้การจดจำแบบรูป (Pattern Recognition) และการอุปมานความรู้ (Knowledge deduction) เช่นเดียวกับความสามารถที่มีในสมองมนุษย์ แนวคิดเริ่มต้นของเทคนิคนี้ได้มาจากการศึกษาข่ายงานไฟฟ้าชีวภาพ (bioelectric network) ในสมอง ซึ่งประกอบด้วย เซลล์ประสาท หรือ "นิวรอน" (neurons) และจุดประสานประสาท (synapses) แต่ละเซลล์ประสาทประกอบด้วยปลายในการรับกระแสประสาท เรียกว่า "เดนไดรท์" (Dendrite) ซึ่งเป็น input และปลายในการส่งกระแสประสาทเรียกว่า "แอคซอน" (Axon) ซึ่งเป็นเหมือน output ของเซลล์ เซลล์เหล่านี้ทำงานด้วยปฏิกิริยาไฟฟ้าเคมี เมื่อมีการกระตุ้นด้วยสิ่งเร้าภายนอกหรือกระตุ้นด้วยเซลล์ด้วยกัน กระแสประสาทจะวิ่งผ่านเดนไดรท์เข้าสู่นิวเคลียสซึ่งจะเป็นตัวตัดสินว่าต้องกระตุ้นเซลล์อื่นๆ ต่อหรือไม่ ถ้ากระแสประสาทแรงพอ นิวเคลียสก็จะกระตุ้นเซลล์อื่นๆ ต่อไปผ่านทางแอคซอนของมัน

ตามโมเดลนี้ข่ายงานประสาทเกิดจากการเชื่อมต่อระหว่างเซลล์ประสาท จนเป็นเครือข่ายที่ทำงานร่วมกัน

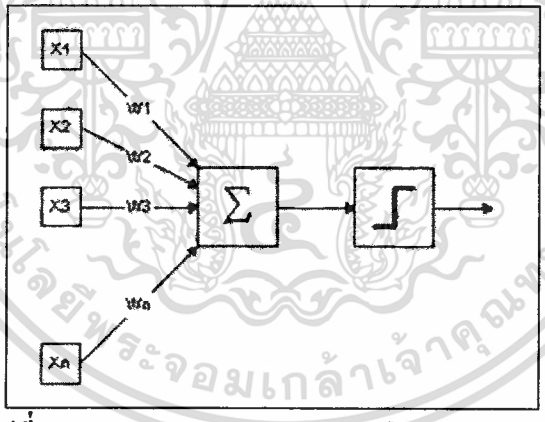
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 2.3 แสดง Model ของ Neuron ในสมองมนุษย์

โครงสร้าง

นักวิจัยส่วนใหญ่ในปัจจุบันเห็นตรงกันว่าโครงข่ายประสาทเทียมมีโครงสร้างแตกต่างจากโครงข่ายในสมอง แต่ก็ยังเหมือนสมองในแง่ที่ว่าโครงข่ายประสาทเทียม คือการรวมกลุ่มแบบขนานของหน่วยประมวลผลย่อยๆ และการเชื่อมต่อนี้เป็นส่วนสำคัญที่ทำให้เกิดสติปัญญาของโครงข่าย เมื่อพิจารณาขนาดแล้วสมองมีขนาดใหญ่กว่าโครงข่ายประสาทเทียมอย่างมาก รวมทั้งเซลล์ประสาทยังมีความซับซ้อนกว่าหน่วยย่อยของโครงข่าย อย่างไรก็ดีหน้าที่สำคัญของสมอง เช่น การเรียนรู้ยังคงสามารถถูกจำลองขึ้นอย่างง่ายด้วยโครงข่ายประสาทนี้



รูปที่ 2.4 แสดง Model ของ Neuron ในคอมพิวเตอร์

หลักการ

สำหรับในคอมพิวเตอร์ Neuron ประกอบด้วย input และ output โดยจำลองให้ input แต่ละอันมี weight เป็นตัวกำหนดน้ำหนักของ input โดย neuron แต่ละหน่วยจะมีค่า threshold เป็นตัวกำหนดว่าน้ำหนักรวมของ input ต้องมากขนาดไหนจึงจะสามารถส่ง output ไปยัง neuron ตัวอื่นได้ เมื่อนำ neuron แต่ละหน่วยมาต่อกันให้ทำงานร่วมกันการทำงานนี้ในทางตรรกะแล้วก็จะเหมือนกับปฏิกิริยาเคมีที่เกิดในสมอง เพียงแต่ในคอมพิวเตอร์ทุกอย่างเป็นตัวเลขเท่านั้นเอง โครงสร้างของนิวรอลเน็ตจะประกอบด้วยโหนด (node) สำหรับ Input – Output และการ

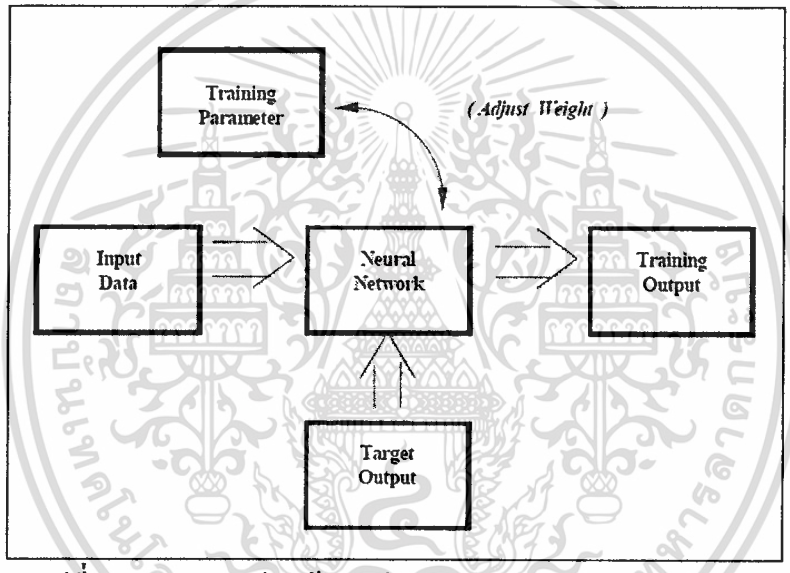
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ประมวลผล กระจายอยู่ในโครงสร้างเป็นชั้นๆ ได้แก่ Input layer, Output layer และ Hidden layer การประมวลผลของนิวรอลเน็ตจะอาศัยการส่งการทำงานผ่านโหนดต่างๆ ใน layer

การเรียนรู้สำหรับ Neural Network

1. การเรียนแบบมีการสอน (Supervised Learning)

เป็นการเรียนแบบที่มีการตรวจคำตอบเพื่อให้วงจรข่ายปรับตัว ชุดข้อมูลที่ใช้สอนวงจรข่ายจะมีคำตอบไว้คอยตรวจดูว่าวงจรข่ายให้คำตอบที่ถูกหรือไม่ ถ้าตอบไม่ถูก วงจรข่ายก็จะปรับตัวเองเพื่อให้ได้คำตอบที่ดีขึ้น (เปรียบเทียบกับคน เหมือนกับการสอนนักเรียนโดยมีครูผู้สอนคอยแนะนำ)

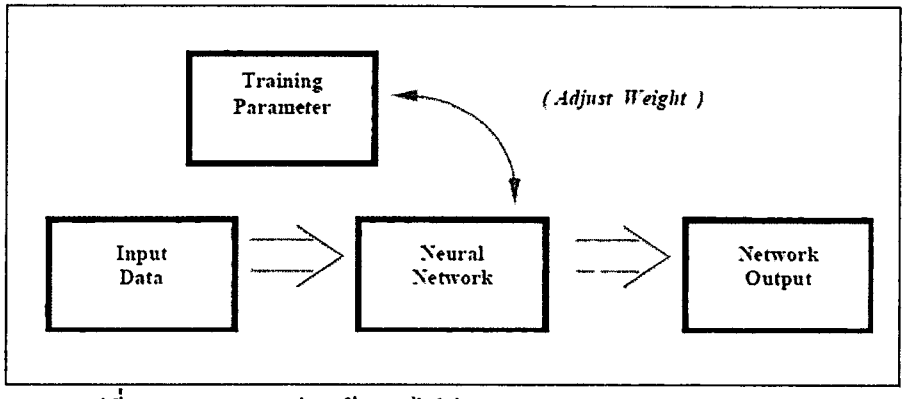


รูปที่ 2.5 แสดงการเรียนรู้แบบมีการสอน (Supervised Learning)

2. การเรียนแบบไม่มีการสอน (Unsupervised Learning)

เป็นการเรียนแบบไม่มีผู้แนะนำ ไม่มีการตรวจคำตอบว่าถูกหรือผิด วงจรข่ายจะจัดเรียงโครงสร้างด้วยตัวเองตามลักษณะของข้อมูล ผลลัพธ์ที่ได้ วงจรข่ายจะสามารถจัดหมวดหมู่ของข้อมูลได้ (เปรียบเทียบกับคน เช่น การที่เราสามารถแยกแยะพันธุ์พืช พันธุ์สัตว์ตามลักษณะรูปร่างของมันได้เองโดยไม่มีใครสอน)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 2.6 แสดงการเรียนรู้แบบไม่มีการสอน Unsupervised Learning

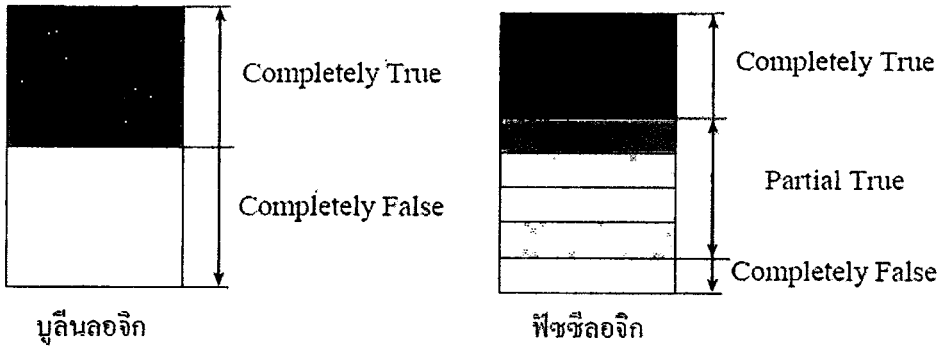
2.8 ฟัซซี่ลอจิก

ฟัซซีลอจิกหรือตรรกะแบบคลุมเครือ เป็นคณิตศาสตร์แขนงใหม่ที่มีความสำคัญต่อเทคโนโลยีสารสนเทศเป็นอย่างมาก ทั้งนี้เพื่อนำไปประยุกต์ใช้งานด้านต่างๆ ซึ่งนับวันจะยิ่งมีความต้องการระบบคอมพิวเตอร์ที่มีความสามารถในการปรับเปลี่ยนระบบได้โดยอัตโนมัติตามสภาพแวดล้อมที่เปลี่ยนไป มีการตัดสินใจแบบชาญฉลาดเยี่ยงมนุษย์ได้มากขึ้น ซึ่งมนุษย์สามารถแก้ปัญหาต่างๆ ที่ไม่เคยพบได้โดยอาศัยความรู้เก่าที่ได้เรียนรู้มาประยุกต์ในการแก้ปัญหาได้อย่างมีประสิทธิภาพ คิดค้นโดย L. A. Zadeh ในปี ค.ศ. 1965

ฟัซซีลอจิกเป็นตรรกะที่อยู่บนพื้นฐานความเป็นจริงที่ว่า "ทุกสิ่งบนโลกแห่งความเป็นจริงไม่ใช่มีเฉพาะสิ่งมีความแน่นอนเท่านั้น แต่มีหลายสิ่งหลายเหตุการณ์ที่เกิดขึ้นอย่างไม่เที่ยงและไม่แน่นอน (uncertain) อาจเป็นสิ่งที่คลุมเครือ (fuzzy) และไม่ชัดเจน (exact)" เช่น เซตของอายุคนอาจแบ่งเป็นวัยทารก วัยเด็ก วัยรุ่น วัยกลางคน และวัยชรา จะเห็นได้ว่าในแต่ละช่วงอายุคนไม่สามารถระบุได้แน่ชัดว่าวัยทารกกับวัย เด็กแยกจากกันแน่ชัดช่วงใด วัยทารกอาจถูกตีความว่าเป็นอายุระหว่าง 0 ถึง 1 ปี บางคนอาจตีความว่าวัยทารกอยู่ในช่วงอายุ 0 ถึง 2 ปี ในทำนองเดียวกัน วัยเด็กและวัยรุ่น ก็ไม่สามารถระบุได้ชัดเจนว่าช่วงต่อของอายุควรจะอยู่ในช่วงใด อาจตีความว่าวัยเด็กมีอายุอยู่ในช่วง 1 ถึง 12 ปี หรืออาจเป็น 2 ถึง 10 ปี เป็นต้น เซตของเหตุการณ์ที่ไม่แน่นอนเช่นนี้เรียกว่าฟัซซีเซต (fuzzy set)

ตรรกะแบบฟัซซี (fuzzy logic) เป็นเครื่องมือที่ช่วยในการตัดสินใจภายในใต้ความไม่แน่นอนของข้อมูล โดยยอมให้มีความยืดหยุ่นได้ ใช้หลักเหตุผลที่คล้ายการเลียนแบบวิธีความคิดที่ซับซ้อนของมนุษย์ ฟัซซีลอจิกมีลักษณะที่พิเศษกว่าตรรกะแบบจริงเท็จ (Boolean logic) เป็นแนวคิดที่มีการต่อขยายในส่วนของความจริง (partial true) โดยค่าความจริงจะอยู่ในช่วงระหว่างจริง (completely true) กับเท็จ (completely false) ส่วนตรรกศาสตร์เดิมจะมีค่าเป็นจริงกับเท็จเท่านั้น แสดงดังภาพที่ 2.7

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 2.7 แสดงตรรกะแบบจริงเท็จ (บูลีนลอจิก) และตรรกะแบบฟัซซี่ (ฟัซซี่ลอจิก)

ความเป็นฟัซซี (Fuzziness) มีชื่อเรียกว่า มัลติวาลานซ์ (multivalance) ซึ่งมีค่าที่ความเป็นสมาชิกมากกว่า 2 ค่า และแตกต่างกับไบวาลานซ์ (bivalance) ที่มีความเป็นสมาชิกเพียง 2 ค่า ฟัซซีเซต (Fuzzy set) เป็นเครื่องมือทางคณิตศาสตร์ที่สื่อถึง "ความไม่แน่นอน (uncertainty)" สามารถที่จะสร้างและกำหนดรูปแบบ (modeling) ของลักษณะความไม่แน่นอนที่เป็นความคลุมเครือ ความไม่ตายตัว รวมถึงความขาดข้อมูลบางส่วน โดยทฤษฎีของฟัซซีเซตจะใช้ลักษณะความหมายตัวแปร (linguistic) มากกว่าปริมาณ (quantitative) ของตัวแปร

ฟังก์ชันความเป็นสมาชิก (Membership function) เป็นฟังก์ชันที่มีการกำหนดระดับความเป็นสมาชิกของตัวแปรที่ต้องการใช้งาน โดยเริ่มจากการแทนที่กับตัวแทนที่มีความไม่ชัดเจน ไม่แน่นอน และคลุมเครือ ดังนั้นส่วนที่สำคัญต่อคุณสมบัติหรือการดำเนินการของฟัซซี เพราะรูปร่างของฟังก์ชันความเป็นสมาชิกมีความสำคัญต่อกระบวนการคิดและแก้ไขปัญหา โดยฟังก์ชันความเป็นสมาชิกจะไม่สมมาตรกันหรือสมมาตรกันทุกประการก็ได้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 3
ทฤษฎีที่นำมาใช้

3.1 การจัดกลุ่มข้อมูล (Clustering)

ในโครงการศึกษานี้จะเน้นถึงวิธีการจัดกลุ่มข้อมูล (Clustering) และนำข้อมูลที่ได้จากการจัดกลุ่มมาวิเคราะห์ ซึ่งวิธีการจัดกลุ่มของข้อมูลที่ได้ศึกษามีดังต่อไปนี้

3.1.1 K-Means Algorithm

K-Means อัลกอริทึมที่แบ่ง n ออบเจ็กต์ ให้อยู่ใน k กลุ่มข้อมูล ดังนั้นผลภายในกลุ่มข้อมูล จึงมีความคล้ายคลึงกันสูง แต่ว่าระหว่างกลุ่มข้อมูลมีความแตกต่างกัน โดยที่ค่าเฉลี่ยของข้อมูลในแต่ละกลุ่มข้อมูล จะถูกมองว่าเป็นศูนย์กลางของกลุ่ม คำนวณค่าเฉลี่ยได้จากสูตรดังนี้

$$\mathbf{m} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (3.1)$$

ที่ซึ่ง m คือ ค่าเฉลี่ยของกลุ่มข้อมูล

x_i คือ ค่าข้อมูล เมื่อ i = 1,...,n

n คือ จำนวนข้อมูลในกลุ่มข้อมูล

K-Means เริ่มต้นจากการกำหนดจำนวนกลุ่มที่ต้องการแบ่งกลุ่ม จากนั้นให้ทำการสุ่มข้อมูลที่เป็นตัวแทนค่าเฉลี่ยของกลุ่มข้อมูล หรือเป็นจุดศูนย์กลางของกลุ่มข้อมูล ซึ่งจะรวมข้อมูลที่มีระยะห่างใกล้เคียงกับศูนย์กลางของกลุ่มข้อมูลมากที่สุด ซึ่งจะคำนวณระยะห่างได้จากสูตร Euclidean distance ดังนี้

$$\mathbf{d} = \sqrt{\sum_{i=1}^{n} (x_i - m)^2} \qquad (3.2)$$

ที่ซึ่ง d คือ ระยะทาง

x_i คือ ตำแหน่งของข้อมูลสมาชิก

n คือ จำนวนข้อมูลสมาชิก

m คือ ตำแหน่งค่าเฉลี่ยของกลุ่มข้อมูล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากนั้นอัลกอริทึมจะคำนวณศูนย์กลางใหม่ของแต่ละกลุ่มข้อมูล และพยายามระบุข้อมูลที่ใกล้เคียงมากที่สุดให้กับกลุ่มของข้อมูล คำนวณซ้ำไปซ้ำมาจนกระทั่งศูนย์กลางของกลุ่มข้อมูลไม่เปลี่ยนแปลง ขั้นตอนของ K-Means อัลกอริทึมอธิบายได้ดังนี้

1. กำหนดจำนวน k กลุ่ม

2. เลือกข้อมูล k ตัวแรก เพื่อเริ่มจุดศูนย์กลางของกลุ่มข้อมูล

3. กำหนดสมาชิกของข้อมูลให้อยู่ในกลุ่มข้อมูล โดยวัดระยะห่างจากสมาชิกข้อมูลกับศูนย์กลางของกลุ่ม หากสมาชิกอยู่ใกล้ศูนย์กลางของกลุ่มใดมากที่สุดก็นำไปรวมกับกลุ่มนั้น

4. ในแต่ละกลุ่มคำนวณค่าเฉลี่ยใหม่เพื่อสร้างศูนย์กลางของกลุ่มใหม่

5. ทำขั้นตอน 3-5 ซ้ำจนกระทั่ง ศูนย์กลางของกลุ่มข้อมูลไม่เปลี่ยนแปลง

3.1.2 Fuzzy C-Means Algorithm

Fuzzy C-Means Algorithm หรือที่รู้จักกันในอีกชื่อหนึ่งว่า Fuzzy ISODATA เป็นเทคนิคหนึ่งซึ่งนิยมใช้กันอย่างแพร่หลายมาก สำหรับการจำแนกข้อมูลที่มีความคลุมเครือออกเป็นกลุ่มๆ ลักษณะพิเศษของ Fuzzy C-Means Algorithm ที่แตกต่างจากวิธีการอื่นๆ นั้นก็คือ สามารถที่จะบ่งบอกถึงระดับความเป็นสมาชิก (Membership Grade) ของแต่ละกลุ่ม (Clusters) ได้นั่นเอง ซึ่งทำให้วิธีนี้แตกต่างจากวิธีการอื่นๆ ซึ่งไม่มีความสามารถในลักษณะนี้ สำหรับวิธีการอื่นๆ โดยทั่วไปนั้นจะสามารถบอกค่าระดับของความเป็นสมาชิกได้เพียงแค่ค่าเดียวเท่านั้น ซึ่งค่าที่บ่งบอกนี้ก็จะเป็นเฉพาะของกลุ่มเดียวเท่านั้นอีกด้วย นอกจากนี้วิธีการแบบ Fuzzy C-Means Algorithm ยังมีการใช้ค่าของ Membership Grade ในการถ่วงน้ำหนัก เพื่อให้ค่าของความเป็นสมาชิกมีความเป็น Fuzzy น้อยลงอีกด้วย

กำหนดให้ X เป็นกลุ่มของข้อมูลเข้า ซึ่งมีลักษณะเป็น Input Vectors และมีขนาด N ตัว เขียนในรูปแบบของเซต จะได้เป็น

$$X = \{x_1, x_2, ..., x_N\} \tag{3.3}$$

และถูกจัดแบ่งเป็นกลุ่มของข้อมูลได้ C กลุ่ม (C Clusters) แต่ละ Cluster จะมีศูนย์กลาง หรือ Cluster Center เป็นของตัวเอง ซึ่งสามารถเขียนแสดงในรูปแบบของเซตได้เป็น

$$V = \{v_1, v_2, ..., v_C\} \tag{3.4}$$

ดังนั้นสูตรการคำนวณค่า Cost Function ของวิธีการแบบ Fuzzy C-Means Clustering Algorithm จะสามารถแสดงเป็นสัญลักษณ์ทางคณิตศาสตร์ได้ดังนี้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$$J(U,V)=\sum_{j=1}^{C}\sum_{i=1}^{N}\left(u_{ij}\right)^{m}\left\|x_i-v_j\right\|^2 \quad (3.5)$$

โดยที่ U เป็นเมตริกซ์ ที่เรียกว่า Fuzzy Partition Matrix ประกอบไปด้วยสมาชิก u_{ij} ซึ่งเป็นค่าบอกระดับความเป็นสมาชิก (Membership Grade) ของข้อมูลตัวที่ i กับ Cluster ตัวที่ j สำหรับเมตริกซ์ U ที่ว่านี้จะมีมิติเป็น NxC มิติ เราสามารถแสดงเป็นสัญลักษณ์ทางคณิตศาสตร์ได้เป็น

$$U=\left(u_{ij}\right)_{NxC} \quad (3.6)$$

นอกจากนี้เมตริกซ์ U ยังต้องมีคุณสมบัติดังต่อไปนี้ จึงจะเป็นเมตริกซ์ที่สมบูรณ์ และถูกต้องตามสูตรการคำนวณ

1) ค่าของแต่ละสมาชิกที่อยู่ภายในเมตริกซ์ U นั้นจะต้องอยู่ในช่วงปิด $[0,1]$ เสมอทุกๆ ข้อมูลที่ i กับ Cluster ที่ j

$$u_{ij}\in[0,1],\forall_i=1,...,N,\forall_j=1,...,C \quad (3.7)$$

2) ผลรวมของ Membership Grade สำหรับข้อมูลตัวที่ i ของทุกๆ Cluster จะต้องมีค่าเท่ากับ 1 เสมอ

$$\sum_{j=1}^{C}u_{ij}=1,\forall i=1,...,N \quad (3.8)$$

ตัวแปร m (Fuzziness Degree) เป็นตัวแปรที่ใช้กำหนดระดับของความเป็นฟัซซี่ ซึ่งจะมีค่าอยู่ในช่วงตั้งแต่ $[1,\infty)$ การกำหนดค่าของตัวแปร m จะขึ้นอยู่กับลักษณะของแต่ละแอปพลิเคชั่น ค่าของตัวแปร m ที่สามารถทำงานได้ดีจะอยู่ในช่วงระหว่าง 1.5-2.5 แต่โดยปกติทั่วไปแล้วจะกำหนดค่าไว้ให้เท่ากับ 2 เพราะถ้าหากกำหนดค่ามากเกินไปจะทำให้แต่ละกลุ่มไม่เกิดความแตกต่างกัน ส่งผลให้ไม่สามารถจัดแบ่งกลุ่มได้นั่นเอง

ส่วนการคำนวณ $\left\|x_i-v_j\right\|$ นั้นเป็นการคำนวณเพื่อหาระยะห่าง หรือระยะทางระหว่างจุดสองจุดที่เรียกว่า การหา Distance Norm สำหรับวิธีการหา Distance ในรูปแบบของ Fuzzy C-Means Algorithm นั้น จะใช้วิธีการหาแบบ Euclidean Distance เพราะเป็นมาตรฐานในการหาระยะห่างระหว่างจุดสองจุดหรือขนาดของเวกเตอร์สองเวกเตอร์ที่เอามา ลบกันนั่นคือ ข้อมูลตัวที่ i กับ Cluster Center ตัวที่ j

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ในการคำนวณโดยใช้สูตรดังกล่าวมานี้ จะเป็นการแก้ปัญหาแบบ nonlinear ซึ่งจะใช้กระบวนการในการทำงานซ้ำๆ หรือวนลูปในการทำงานนั่นเอง โดยสามารถที่จะอธิบายเป็นขั้นตอนต่างๆ ได้ดังต่อไปนี้

ขั้นที่ 1 กำหนดค่าแสดงความเป็นสมาชิกแต่ละตัว ในเมตริกซ์ U โดยอาจจะทำการสุ่มเอาค่าต่างๆ ขึ้นมาใช้ได้ แต่ต้องตรงกับคุณสมบัติที่ได้กล่าวไว้แล้วข้างต้น และกำหนดค่าระดับความเป็นฟัซซี่ หรือค่าตัวแปร m (Fuzziness Degree) ที่เหมาะสมเพื่อใช้ในการคำนวณด้วย

ขั้นที่ 2 คำนวณหาค่า Cluster Center (V_j) ของทุกๆ Cluster ตามสมการข้างล่างนี้

$$v_j = \frac{\sum_{i=1}^{N} (u_{ij})^m x_i}{\sum_{i=1}^{N} (u_{ij})^m}, \forall j = 1,...,C \qquad (3.9)$$

ขั้นที่ 3 คำนวณหาค่า Distance Norm ตามที่ได้อธิบายไว้แล้วข้างต้น โดยใช้สมการข้างล่างดังต่อไปนี้

$$d_{ij} = \|x_i - v_j\| = \sqrt{(x - v1)^2 + (y - v2)^2} \qquad (3.10)$$

ขั้นที่ 4 อัพเดตค่า U_{ij} ในเมตริกซ์ U โดยการตรวจสอบค่าของ Distance Norm ที่คำนวณได้จากขั้นตอนที่ 3 ว่ามีค่ามากกว่า 0 หรือไม่ ถ้าค่า Distance Norm มีค่ามากกว่า 0 จริง นั่นหมายความว่า $X_i \neq V_j$ นั่นเอง ก็ให้อัพเดตค่าของ U_{ij} ตามสมการข้างล่าง

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{d_{ij}}{d_{ik}} \right)^{\frac{2}{m-1}}} \qquad (3.11)$$

แต่ถ้าค่าของ Distance Norm มีค่าเท่ากับ 0 ก็ให้กำหนดค่า U_{ij} ของข้อมูลตัวที่ i กับ Cluster ตัวที่ j เป็น 1

ขั้นที่ 5 ตรวจสอบค่าของ J(U,V) ว่ามีค่าเปลี่ยนแปลงจากเดิมหรือไม่ ถ้ามีการเปลี่ยนแปลงก็ให้กลับไปเริ่มทำในขั้นตอนที่ 2 อีกครั้งหนึ่งจนกว่าของ J(U,V) จะไม่มีการเปลี่ยนแปลงอีกจึงหยุดทำงาน ค่าของ J(U,V) ในที่นี้นั้นจริงๆแล้วก็คือ ค่าของเมตริกซ์ U นั่นเอง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

05957

หลังจากที่เราได้ทำการคำนวณตามขั้นตอนทั้ง 5 เสร็จเรียบร้อยแล้ว ให้ทำการวัด และประเมินผลค่าต่างๆ ในเมตริกซ์ U โดยให้ตรวจสอบข้อมูลแต่ละตัวว่ามีค่าของความเป็นสมาชิกใน Cluster กลุ่มใดสูงที่สุด แสดงว่าข้อมูลตัวนั้นน่าจะปรากฏอยู่ใน Cluster หรือกลุ่มนั้นนั่นเอง

ตัวอย่างการทำงานของ Fuzzy C-Means Algorithm

กำหนดข้อมูลเข้า (input vectors) ต่างๆ ดังนี้

DATA INPUT: (1,1), (1,2), (3,4), (4,5)

ต้องการแบ่งข้อมูลที่กำหนดมาให้ออกเป็น 2 กลุ่มหรือ 2 Cluster เมื่อทำตามขั้นตอนการทำงานตามที่ได้กล่าวมาสามารถแสดงได้ดังนี้

1) กำหนดค่าเริ่มต้นของเมตริกซ์ U

$$U=\begin{bmatrix}0 & 1\\ 1 & 0\\ 1 & 0\\ 0 & 1\end{bmatrix} \quad \text{และ} \quad m=2 \tag{3.12}$$

2) คำนวณหาค่าจุดศูนย์กลางของกลุ่ม (Cluster Center) ของทุกๆ Cluster (V_j) ตามสูตร

$$v_j=\frac{\sum_{i=1}^{N}(u_{ij})^m x_i}{\sum_{i=1}^{N}(u_{ij})^m}, \forall j=1,\ldots,C \tag{3.13}$$

ตารางที่ 3.1 คำนวณหาค่าจุดศูนย์กลางของกลุ่ม

โหนด	Cluster Center	Center
v_1	[0(1,1)+1(1,2)+1(3,4)+0(4,5)]/[0+1+1+0]	(2,3)
v_2	[1(1,1)+0(1,2)+0(3,4)+1(4,5)]/[1+0+0+1]	(2.5,3)

3) คำนวณหาค่า Distance Norm ซึ่งเป็นระยะห่างระหว่างจุดสองจุด นั่นคือ ข้อมูลตัวที่ i กับ Cluster Center ที่ j

- ระยะห่างระหว่างข้อมูลชุดที่ 1 $(x1)$ กับ Cluster

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.2 คำนวณหาระยะห่างระหว่างข้อมูลชุดที่ 1

โหนด	Euclidean distance	Distance
d11 = ||x1-v1||	$\left[(1-2)^2+(1-3)^2\right]^{1/2}$	2.24
d12 = ||x1-v2||	$\left[(1-2.5)^2+(1-3)^2\right]^{1/2}$	2.5

➢ ระยะห่างระหว่างข้อมูลชุดที่ 2 $(x2)$ กับ Cluster

ตารางที่ 3.3 คำนวณหาระยะห่างระหว่างข้อมูลชุดที่ 2

โหนด	Euclidean distance	Distance
d21 = ||x2-v1||	$\left[(1-2)^2+(2-3)^2\right]^{1/2}$	1.41
d22 = ||x2-v2||	$\left[(1-2.5)^2+(2-3)^2\right]^{1/2}$	1.8

➢ ระยะห่างระหว่างข้อมูลชุดที่ 3 $(x3)$ กับ Cluster

ตารางที่ 3.4 คำนวณหาระยะห่างระหว่างข้อมูลชุดที่ 3

โหนด	Euclidean distance	Distance
d31 = ||x3-v1||	$\left[(3-2)^2+(4-3)^2\right]^{1/2}$	1.41
d32 = ||x3-v2||	$\left[(3-2.5)^2+(4-3)^2\right]^{1/2}$	1.12

➢ ระยะห่างระหว่างข้อมูลชุดที่ 4 $(x4)$ กับ Cluster

ตารางที่ 3.5 คำนวณหาระยะห่างระหว่างข้อมูลชุดที่ 4

โหนด	Euclidean distance	Distance
d41 = ||x4-v1||	$\left[(4-2)^2+(5-3)^2\right]^{1/2}$	2.83
d42 = ||x4-v2||	$\left[(4-2.5)^2+(5-3)^2\right]^{1/2}$	2.5

4) เมื่อตรวจสอบค่าของ Distance Norm ที่คำนวณได้ พร้อมทั้งการอัพเดตค่าของ U_{ij} ก็จะได้ค่า Matrix U ใหม่แสดงได้ดังนี้

จากสูตร $u_{ij} = \dfrac{1}{\sum_{k=1}^{C}\left(\dfrac{d_{ij}}{d_{ik}}\right)^{\frac{2}{m-1}}}$ **(3.14)**

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.6 ตารางแสดงค่า Matrix U

U_{ij}	อัพเดทค่าของ u_{ij}	
u_{11}	$\frac{1}{\left(\frac{2.24}{2.24}\right)^{\left(\frac{2}{2-1}\right)}+\left(\frac{2.24}{2.5}\right)^{\left(\frac{2}{2-1}\right)}}$	0.56
u_{12}	$\frac{1}{\left(\frac{2.5}{2.24}\right)^{\left(\frac{2}{2-1}\right)}+\left(\frac{2.5}{2.5}\right)^{\left(\frac{2}{2-1}\right)}}$	0.44
u_{21}	$\frac{1}{\left(\frac{1.41}{1.41}\right)^{\left(\frac{2}{2-1}\right)}+\left(\frac{1.41}{1.8}\right)^{\left(\frac{2}{2-1}\right)}}$	0.62
u_{22}	$\frac{1}{\left(\frac{1.8}{1.41}\right)^{\left(\frac{2}{2-1}\right)}+\left(\frac{1.8}{1.8}\right)^{\left(\frac{2}{2-1}\right)}}$	0.38
u_{31}	$\frac{1}{\left(\frac{1.41}{1.41}\right)^{\left(\frac{2}{2-1}\right)}+\left(\frac{1.41}{1.12}\right)^{\left(\frac{2}{2-1}\right)}}$	0.39
u_{32}	$\frac{1}{\left(\frac{1.12}{1.41}\right)^{\left(\frac{2}{2-1}\right)}+\left(\frac{1.12}{1.12}\right)^{\left(\frac{2}{2-1}\right)}}$	0.61
u_{41}	$\frac{1}{\left(\frac{2.83}{2.83}\right)^{\left(\frac{2}{2-1}\right)}+\left(\frac{2.83}{2.5}\right)^{\left(\frac{2}{2-1}\right)}}$	0.44
u_{42}	$\frac{1}{\left(\frac{2.5}{2.83}\right)^{\left(\frac{2}{2-1}\right)}+\left(\frac{2.5}{2.5}\right)^{\left(\frac{2}{2-1}\right)}}$	0.56

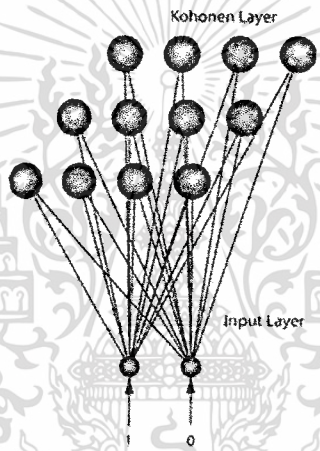
5) กลับไปทำในขั้นตอนที่สองอีกครั้งหนึ่งเพราะภายใน Matrix U ยังมีการเปลี่ยนแปลงค่าของสมาชิกอยู่ ดังนั้นจึงต้องกลับไปทำในขั้นตอนที่สองจนกว่า Matrix U จะไม่มีการเปลี่ยนแปลงอีก

หลังจากที่ได้ทำการปรับค่าใน Matrix U จนค่าภายในไม่มีการเปลี่ยนแปลงแล้ว ก็พอที่จะสรุปได้ว่าข้อมูลตัวที่ 1 และตัวที่ 2 นั้นน่าจะมีโอกาสที่จะอยู่ในกลุ่มเดียวกันคืออยู่ใน Cluster กลุ่มที่ 1 ส่วนข้อมูลตัวที่ 3 และตัวที่ 4 นั้น ก็มีโอกาสที่น่าจะอยู่ใน Cluster กลุ่มที่ 2

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.1.3 Self Organizing Map Algorithm

Self Organizing Map (SOM) ถูกนำเสนอโดย Kohonen เป็น model ที่ได้รับความนิยมตัวหนึ่งลักษณะของ model เป็นลักษณะของโครงข่ายมีจำนวน 2 ชั้น คือ ชั้นอินพุต (Input layer) และชั้นเอาท์พุต (Output layer) หรือ Kohonen layer เท่านั้น กระบวนการเรียนรู้เมื่อมีการกระตุ้นโครงข่ายนิวรอนแต่ละโหนดในชั้นเอาท์พุตจะแข่งขันกันขึ้น นิวรอนที่ตอบสนองต่อการกระตุ้นที่ดีที่สุดจะเป็นผู้ที่ชนะ โดยโหนดที่ชนะจะมีการปรับค่า weight เพื่อให้เข้าใกล้อินพุตมากขึ้น โหนดที่ชนะจะมีการปรับค่า weight มากกว่าโหนดที่อยู่รอบๆ โหนดที่ชนะ ซึ่งโหนดที่อยู่ไกลออกไปจากโหนดที่ชนะการปรับค่า weight ก็จะน้อยลง จนกระทั่งไม่ต้องมีการปรับค่า weight ลักษณะของโครงข่ายของ Self Organizing Map ดังแสดงในรูปที่ 3.1



รูปที่ 3.1 ลักษณะของโครงข่ายของ Self Organizing Map

ขั้นตอนการทำงานของ Self Organizing Map

ขั้นที่ 1 ขั้นเตรียมการ จะกำหนดค่าต่างๆ ให้กับโครงข่ายเพื่อการเรียนรู้ เช่น กำหนดค่า weight ให้กับโครงข่ายโดยการสุ่มค่าขึ้นมา กำหนดค่า α คือ อัตราการเรียนรู้ ณ เวลาใดๆ ซึ่งกำหนดให้มีค่าบวกในช่วงที่แคบๆ ปกติอยู่ในช่วง 0 ถึง 1 กำหนดจำนวนนิวรอนในชั้นเอาท์พุตและขนาดความกว้างของ Gaussian function ณ เวลาใดๆ

ขั้นที่ 2 ขั้นกระตุ้นโครงข่าย จะทำการกระตุ้นโครงข่ายแล้วดูว่านิวรอนตัวไหนตอบสนองการกระตุ้นได้ดีที่สุดนิวรอนตัวนั้นคือผู้ชนะ โดยใช้ Euclidean distance

$$\|x - w_m\| = \min_i \{\|x - w_i\|\} \tag{3.15}$$

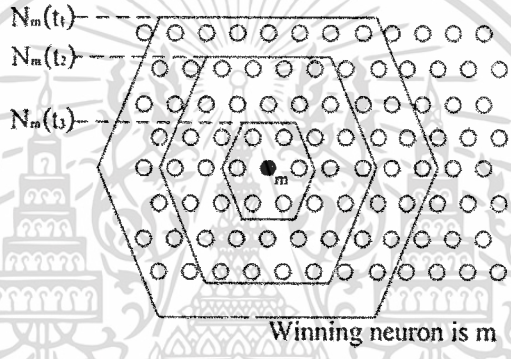
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

โดยที่ i คือ จำนวนนิวรอนในชั้นเอาท์พุต

เมื่อผ่านการกระตุ้นโครงข่ายจะได้นิวรอนที่มี distance ต่ำที่สุดเป็นนิวรอนที่ชนะแล้วก็จะทำการปรับค่า weight ต่อไป

ขั้นที่ 3 ขั้นการเรียนรู้ในการปรับปรุงค่า weight

Kohonen ได้เสนอวิธีในการปรับค่า weight ไว้ คือถ้า m คือ นิวรอนตัวที่ชนะแล้ว m จะได้รับการปรับค่า weight มากที่สุดและนิวรอนที่อยู่บริเวณรอบๆ นิวรอนตัวที่ชนะ (neighborhood) ก็ควรปรับปรุงค่า weight ด้วย โดยให้นิวรอนที่อยู่บริเวณรอบๆ (เพื่อนบ้าน) ก็ปรับค่า weight น้อยกว่าตัวที่ชนะ นิวรอนตัวถัดออกไปอีกก็ปรับค่า weight น้อยลงไปอีกจนถึง ณ จุดหนึ่งจะไม่มีการปรับค่า weight แล้วจะถือว่าเป็นเพื่อนบ้านที่อยู่ไกล โดยลักษณะของ neighborhood ได้แสดงไว้ในรูปที่ 3.2



รูปที่ 3.2 ลักษณะของ neighborhood $t_1 \langle t_2 \langle t_3 \ldots$

การปรับค่า weight ณ เวลา t ใดๆ โดยใช้สมการ

$$w_i(t+1) = w_i(t) + \Delta w_i(t) \tag{3.16}$$

โดยที่ $\Delta w_i(t)$ คือ การปรับปรุงค่า weight

$$\Delta w_i(t) = \alpha(N_i, t)[x(t) - w_i(t)] \quad for \quad i \in N_m(t) \tag{3.17}$$
$$\alpha(N_i, t) = \alpha(t)\exp\left[-\|r_i - r_m\| / \sigma^2(t)\right]$$

โดย r_m คือ ตำแหน่งของนิวรอนที่ชนะ

r_i คือ ตำแหน่งของนิวรอนรอบนิวรอนที่ชนะ

$\sigma(t)$ คือ ขนาดความกว้างของ gaussian function

$\alpha(t)$ คือ อัตราการเรียนรู้ขณะใดๆ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ขั้นที่ 4 ขั้นทำซ้ำ ทำการฝึกสอนไปเรื่อยๆ จนกระทั่งค่า weight มีการเปลี่ยนแปลงน้อยมากหรือได้ค่าที่สามารถยอมรับได้

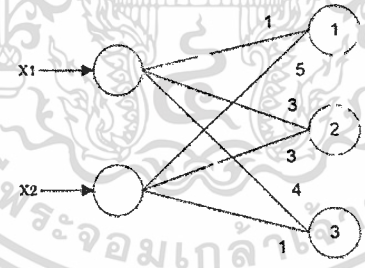
ตัวอย่างการคำนวณ Self Organizing Map (SOM)

กำหนดให้รูปแบบของ input ดังตาราง

ตารางที่ 3.7 ตารางข้อมูล Input

x_1	x_2
0	6
0	0
4	1
5	0
5	5

1. กำหนดค่าให้นิวรอนในชั้นเอาท์พุต $(k)=3$ และอัตราการเรียนรู้ $\alpha(t)=0.5$ และ $\sigma(t)=0.7$ ตั้งแต่ $t=1$ ถึง 4 และสุ่มค่า weight ให้โครงข่าย



รูปที่ 3.3 แสดงการกำหนดค่าให้โครงข่าย

2. หานิวรอนที่ชนะโดยใช้ Euclidean distance เมื่อ input คือ $[0,6]$ จะได้โหนดที่ชนะคือโหนดที่ 1

ตารางที่ 3.8 ตารางแสดงนิวรอนที่ชนะ

โหนด	Euclidean distance	Distance
1	$\left[(0-1)^2+(6-5)^2\right]^{1/2}$	1.41
2	$\left[(0-3)^2+(6-3)^2\right]^{1/2}$	4.24
3	$\left[(0-4)^2+(6-1)^2\right]^{1/2}$	6.4

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3. ปรับค่า weight โดยสูตร $w_i(t+1)=w_i(t)+\Delta w_i(t)$ และ $\alpha(t)=0.5$ และ $\sigma(t)=0.7$ ตั้งแต่ $t=1$ ถึง 4

$$\Delta w_i(t)=\alpha(N_i,t)[x(t)-w_i(t)] \; for \;\; i\in N_m(t)$$
$$\alpha(N_i,t)=\alpha(t)\exp[-\|r_i-r_m\|/\sigma^2(t)]$$

ที่ $m=1, i=1$

$$\begin{aligned}
w_1(t+1) &= w_1(t)+0.5\exp[-\|r_i-r_m\|/\sigma^2(t)][x(t)-w_1(t)] \\
&= w_1(t)+0.5\exp[-\|r_1-r_1\|/(0.7)^2][x(t)-w_1(t)] \\
&= [1,5]+0.5(1)([0,6]-[1,5]) \\
&= [1,5]+0.5([-1,1]) \\
&= [1,5]+[-0.5,0.5]=[0.5,5.5]
\end{aligned}$$

ที่ $m=1, i=2$

$$\begin{aligned}
w_2(t+1) &= w_2(t)+0.5\exp[-\|r_i-r_m\|/\sigma^2(t)][x(t)-w_2(t)] \\
&= w_2(t)+0.5\exp[-\|r_2-r_1\|/(0.7)^2][x(t)-w_2(t)] \\
\|r_2-r_1\| &= [(3-1)^2+(3-5)^2]^{1/2} \\
&= [(2)^2+(-2)^2]^{1/2}=[(4)+(4)]^{1/2}=2.82
\end{aligned}$$

$$\begin{aligned}
w_2(t+1) &= w_2(t)+0.5\exp[(-2.82)/(0.7)^2][x(t)-w_2(t)] \\
&= w_2(t)+0.5\exp(-5.76)[x(t)-w_2(t)] \\
&= [3,3]+0.5(0.003)([0,6]-[3,3]) \\
&= [3,3]+(0.0015)[-3,3] \\
&= [3,3]+[-0.0045,0.0045]=[2.9955,3.0045]
\end{aligned}$$

ที่ $m=1, i=3$

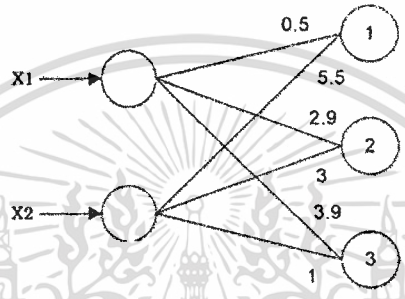
$$\begin{aligned}
w_3(t+1) &= w_3(t)+0.5\exp[-\|r_i-r_m\|/\sigma^2(t)][x(t)-w_3(t)] \\
&= w_3(t)+0.5\exp[-\|r_3-r_1\|/(0.7)^2][x(t)-w_3(t)] \\
\|r_3-r_1\| &= [(4-1)^2+(1-5)^2]^{1/2} \\
&= [(3)^2+(-4)^2]^{1/2}=[(9)+(16)]^{1/2}=5
\end{aligned}$$

$$\begin{aligned}
w_3(t+1) &= w_3(t)+0.5\exp[(-5)/(0.7)^2][x(t)-w_3(t)] \\
&= w_3(t)+0.5\exp[(-10.2)][x(t)-w_3(t)]
\end{aligned}$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$$= [4,1] + 0.5(0.00003)([0,6] - [4,1])$$
$$= [4,1] + (0.000015)[-4,5]$$
$$= [4,1] + [-0.00006, 0.000075]$$
$$= [3.99994, 1.000075]$$

จาก input คือ $[0,6]$ ปรับค่า weight ครบทั้งโหนดที่ชนะและโหนดข้างเคียงจะ weight ได้ใหม่ออกมาดังแสดงรูปที่ 3.4



รูปที่ 3.4 แสดงการปรับค่า weight โครงข่ายเมื่อ input คือ $[0,6]$

เมื่อ input ต่อไปเข้ามาคือ $[0,0]$

1. หานิวรอนที่ชนะโดยใช้ Euclidean distance เมื่อ input คือ $[0,0]$ จะได้โหนดที่ชนะคือโหนดที่ 3

ตารางที่ 3.9 ตารางแสดงนิวรอนที่ชนะ

โหนด	Euclidean distance	Distance
1	$\left[(0-0.5)^2 + (0-5.5)^2\right]^{1/2}$	5.52
2	$\left[(0-2.9)^2 + (0-3.0)^2\right]^{1/2}$	4.17
3	$\left[(0-3.9)^2 + (0-1.0)^2\right]^{1/2}$	4.03

2. ปรับค่า weight โดยสูตร $w_i(t+1) = w_i(t) + \Delta w_i(t)$ และ $\alpha(t) = 0.5$ และ $\sigma(t) = 0.7$

$$\Delta w_i(t) = \alpha(N_i, t)[x(t) - w_i(t)] \quad for \quad i \in N_m(t)$$
$$\alpha(N_i, t) = \alpha(t)\exp\left[-\|r_i - r_m\| / \sigma^2(t)\right]$$

ที่ $m = 3, i = 1$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$$
\begin{aligned}
w_1(t+1) &= w_1(t)+0.5\exp\left[-\|r_i-r_m\|/\sigma^2(t)\right]\left[x(t)-w_1(t)\right] \\
&= w_1(t)+0.5\exp\left[-\|r_1-r_3\|/(0.7)^2\right]\left[x(t)-w_1(t)\right] \\
\|r_1-r_3\| &= \left[(0.5-3.9)^2+(5.5-1.0)^2\right]^{1/2} \\
&= \left[(-3.4)^2+(4.5)^2\right]^{1/2} \\
&= \left[(11.56)+(20.25)\right]^{1/2} = 5.64 \\
w_1(t+1) &= w_1(t)+0.5\exp\left[(-5.64)/(0.7)^2\right]\left[x(t)-w_1(t)\right] \\
&= w_1(t)+0.5\exp(-11.51)\left[x(t)-w_1(t)\right] \\
&= [0.5,5.5]+0.5(0.00001)([0,0]-[0.5,5.5]) \\
&= [0.5,5.5]+(0.000005)[-0.5,-5.5] \\
&= [0.5,5.5]+[-0.0000025,-0.0000275] \\
&= [0.499,5.499]
\end{aligned}
$$

ที่ $m=3, i=2$

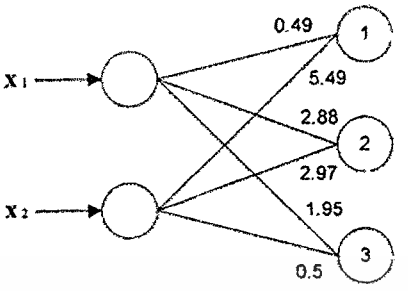
$$
\begin{aligned}
w_2(t+1) &= w_2(t)+0.5\exp\left[-\|r_i-r_m\|/\sigma^2(t)\right]\left[x(t)-w_2(t)\right] \\
&= w_2(t)+0.5\exp\left[-\|r_2-r_3\|/(0.7)^2\right]\left[x(t)-w_2(t)\right] \\
\|r_2-r_3\| &= \left[(2.9-3.9)^2+(3.0-1.0)^2\right]^{1/2} \\
&= \left[(-0.1)^2+(2.0)^2\right]^{1/2} \\
&= \left[(0.01)+(4)\right]^{1/2} = 2.002 \\
w_2(t+1) &= w_2(t)+0.5\exp\left[(-2.002)/(0.7)^2\right]\left[x(t)-w_2(t)\right] \\
&= w_2(t)+0.5\exp(-4.087)\left[x(t)-w_2(t)\right] \\
&= [2.9,3.0]+0.5(0.017)([0,0]-[2.9,3.0]) \\
&= [2.9,3.0]+(0.0085)[-2.9,-3.0] \\
&= [2.9,3.0]+[-0.02465,-0.0255] \\
&= [2.875,2.974]
\end{aligned}
$$

ที่ $m=3, i=3$

$$
\begin{aligned}
w_3(t+1) &= w_3(t)+0.5\exp\left[-\|r_i-r_m\|/\sigma^2(t)\right]\left[x(t)-w_3(t)\right] \\
&= w_3(t)+0.5\exp\left[-\|r_3-r_3\|/(0.7)^2\right]\left[x(t)-w_3(t)\right] \\
&= [3.9,1.0]+0.5(1)([0,0]-[3.9,1.0]) \\
&= [3.9,1.0]+(0.5)[-3.9,-1.0] \\
&= [3.9,1.0]+[-1.95,-0.5] \\
&= [1.95,0.5]
\end{aligned}
$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จาก input คือ $[0,0]$ ปรับค่า weight ครบทั้งโหนดที่ชนะและโหนดข้างเคียงจะ weight ได้ใหม่ออกมาดังแสดงรูปที่ 3.5



รูปที่ 3.5 แสดงการปรับค่า weight โครงข่ายเมื่อ input คือ $[0,0]$

เมื่อ input ต่อไปเข้ามาคือ $[4,1]$

1. หานิวรอนที่ชนะโดยใช้ Euclidean distance เมื่อ input คือ [4,1] จะได้โหนดที่ชนะคือโหนดที่ 3

ตารางที่ 3.10 ตารางแสดงนิวรอนที่ชนะ

โหนด	Euclidean distance	Distance
1	$\left[(4-0.49)^2+(1-5.49)^2\right]^{1/2}$	5.699
2	$\left[(4-2.88)^2+(1-2.97)^2\right]^{1/2}$	2.266
3	$\left[(4-1.95)^2+(1-0.50)^2\right]^{1/2}$	2.110

2.ปรับค่า weight โดยสูตร $w_i(t+1)=w_i(t)+\Delta w_i(t)$ และ $\alpha(t)=0.5$ และ $\sigma(t)=0.7$

$$\Delta w_i(t)=\alpha(N_i,t)[x(t)-w_i(t)] \; for \;\; i\in N_m(t)$$

$$\alpha(N_i,t)=\alpha(t)\exp\left[-\|r_i-r_m\|/\sigma^2(t)\right]$$

ที่ $m=3, i=1$

$$\begin{aligned} w_1(t+1) &= w_1(t)+0.5\exp\left[-\|r_i-r_m\|/\sigma^2(t)\right][x(t)-w_1(t)] \\ &= w_1(t)+0.5\exp\left[-\|r_1-r_3\|/(0.7)^2\right][x(t)-w_1(t)] \\ \|r_1-r_3\| &= \left[(0.49-1.95)^2+(5.49-0.50)^2\right]^{1/2} \\ &= \left[(-1.46)^2+(4.99)^2\right]^{1/2} \\ &= \left[(2.1316)+(24.9001)\right]^{1/2}=5.199 \end{aligned}$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$$
\begin{aligned}
w_1(t+1) &= w_1(t) + 0.5\exp\left[(-5.199)/(0.7)^2\right]\left[x(t) - w_1(t)\right] \\
&= w_1(t) + 0.5\exp\left[(-10.61)\right]\left[x(t) - w_1(t)\right] \\
&= [0.49,5.49] + 0.5(0.000025)([4,1] - [0.49,5.49]) \\
&= [0.49,5.49] + (0.0000125)[3.51,-4.49] \\
&= [0.49,5.49] + [0.000044,-0.000056] \\
&= [0.490044,5.489944]
\end{aligned}
$$

ที่ $m = 3, i = 2$

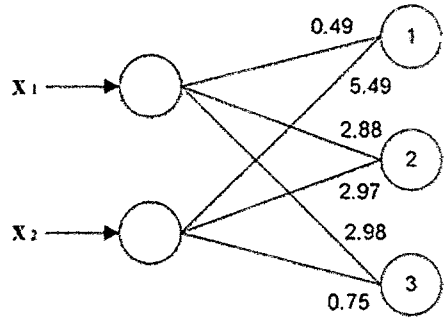
$$
\begin{aligned}
w_2(t+1) &= w_2(t) + 0.5\exp\left[-\|r_i - r_m\|/\sigma^2(t)\right]\left[x(t) - w_2(t)\right] \\
&= w_2(t) + 0.5\exp\left[-\|r_2 - r_3\|/(0.7)^2\right]\left[x(t) - w_2(t)\right] \\
\|r_2 - r_3\| &= \left[(2.88 - 1.95)^2 + (2.97 - 0.50)^2\right]^{1/2} \\
&= \left[(0.93)^2 + (2.47)^2\right]^{1/2} \\
&= \left[(0.8649) + (6.1009)\right]^{1/2} = 2.639 \\
w_2(t+1) &= w_2(t) + 0.5\exp\left[(-2.639)/(0.7)^2\right]\left[x(t) - w_2(t)\right] \\
&= w_2(t) + 0.5\exp\left[(-5.386)\right]\left[x(t) - w_2(t)\right] \\
&= [2.88,2.97] + 0.5(0.0046)([4,1] - [2.88,2.97]) \\
&= [2.88,2.97] + (0.0023)[1.12,-1.97] \\
&= [2.88,2.97] + [-0.0026,-0.0045] \\
&= [2.8826,2.9655]
\end{aligned}
$$

ที่ $m = 3, i = 3$

$$
\begin{aligned}
w_3(t+1) &= w_3(t) + 0.5\exp\left[-\|r_i - r_m\|/\sigma^2(t)\right]\left[x(t) - w_3(t)\right] \\
&= w_3(t) + 0.5\exp\left[-\|r_3 - r_3\|/(0.7)^2\right]\left[x(t) - w_3(t)\right] \\
&= [1.95,0.50] + 0.5(1)([4,1] - [1.95,0.50]) \\
&= [1.95,0.50] + (0.5)[2.05,0.5] \\
&= [1.95,0.50] + [1.025,0.25] \\
&= [2.975,0.75]
\end{aligned}
$$

จาก input คือ [4,1] ปรับค่า weight ครบทั้งโหนดที่ชนะและโหนดข้างเคียงจะ weight ได้ใหม่ออกมาดังแสดงรูปที่ 3.6

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 3.6 แสดงการปรับค่า weight โครงข่ายเมื่อ input คือ [4,1]

เมื่อ input ต่อไปเข้ามาคือ [5,0]

1. หานิวรอนที่ชนะโดยใช้ Euclidean distance เมื่อ input คือ [5,0] จะได้โหนดที่ชนะคือโหนดที่ 3

ตารางที่ 3.11 ตารางแสดงนิวรอนที่ชนะ

โหนด	Euclidean distance	Distance
1	$\left[(5-0.49)^2+(0-5.49)^2\right]^{1/2}$	7.105
2	$\left[(5-2.88)^2+(0-2.97)^2\right]^{1/2}$	3.649
3	$\left[(5-2.98)^2+(0-0.75)^2\right]^{1/2}$	2.155

2. ปรับค่า weight โดยสูตร $w_i(t+1)=w_i(t)+\Delta w_i(t)$ และ $\alpha(t)=0.5$ และ $\sigma(t)=0.7$

$$\Delta w_i(t)=\alpha(N_i,t)[x(t)-w_i(t)] \quad for \quad i\in N_m(t)$$
$$\alpha(N_i,t)=\alpha(t)\exp\left[-\|r_i-r_m\|/\sigma^2(t)\right]$$

ที่ $m=3, i=1$

$$\begin{aligned}
w_1(t+1) &= w_1(t)+0.5\exp\left[-\|r_i-r_m\|/\sigma^2(t)\right][x(t)-w_1(t)] \\
&= w_1(t)+0.5\exp\left[-\|r_1-r_3\|/(0.7)^2\right][x(t)-w_1(t)] \\
\|r_1-r_3\| &= \left[(0.49-2.98)^2+(5.49-0.75)^2\right]^{1/2} \\
&= \left[(-2.49)^2+(4.74)^2\right]^{1/2} \\
&= \left[(6.2001)+(22.4676)\right]^{1/2}=5.354 \\
w_1(t+1) &= w_1(t)+0.5\exp\left[(-5.354)/(0.7)^2\right][x(t)-w_1(t)] \\
&= w_1(t)+0.5\exp\left[(-10.93)\right][x(t)-w_1(t)]
\end{aligned}$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$$= [0.49,5.49] + 0.5(0.000017)([5,0] - [0.49,5.49])$$
$$= [0.49,5.49] + (0.0000085)[4.51,-5.49]$$
$$= [0.49,5.49] + [0.00038,-0.000047]$$
$$= [0.490038,5.489953]$$

ที่ $m = 3, i = 2$

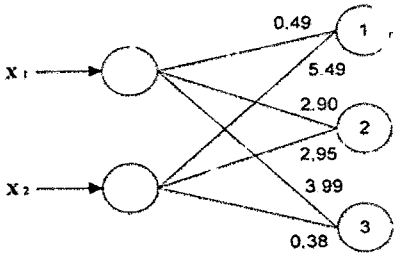
$$w_2(t+1) = w_2(t) + 0.5\exp\left[-\|r_i - r_m\|/\sigma^2(t)\right][x(t) - w_2(t)]$$
$$= w_2(t) + 0.5\exp\left[-\|r_2 - r_3\|/(0.7)^2\right][x(t) - w_2(t)]$$
$$\|r_2 - r_3\| = \left[(2.88 - 2.98)^2 + (2.97 - 0.75)^2\right]^{1/2}$$
$$= \left[(-0.1)^2 + (2.22)^2\right]^{1/2}$$
$$= \left[(0.01) + (4.9284)\right]^{1/2} = 2.222$$
$$w_2(t+1) = w_2(t) + 0.5\exp\left[(-2.222)/(0.7)^2\right][x(t) - w_2(t)]$$
$$= w_2(t) + 0.5\exp[(-4.535)][x(t) - w_2(t)]$$
$$= [2.88,2.97] + 0.5(0.0107)([5,0] - [2.88,2.97])$$
$$= [2.88,2.97] + (0.00535)[2.12,-2.97]$$
$$= [2.88,2.97] + [0.0113,-0.01589]$$
$$= [2.8913,2.95411]$$

ที่ $m = 3, i = 3$

$$w_3(t+1) = w_3(t) + 0.5\exp\left[-\|r_i - r_m\|/\sigma^2(t)\right][x(t) - w_3(t)]$$
$$= w_3(t) + 0.5\exp\left[-\|r_3 - r_3\|/(0.7)^2\right][x(t) - w_3(t)]$$
$$= [2.98,0.75] + 0.5(1)([5,0] - [2.98,075])$$
$$= [2.98,075] + (0.5)[2.02,-0.75]$$
$$= [2.98,0.75] + [1.01,-0.375]$$
$$= [3.99,0.375]$$

จาก input คือ $[5,0]$ ปรับค่า weight ครบทั้งโหนดที่ชนะและโหนดข้างเคียงจะ weight ได้ใหม่ออกมาดังแสดงรูปที่ 3.7

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 3.7 แสดงการปรับค่า weight โครงข่ายเมื่อ input คือ $[5,0]$

เมื่อ input ต่อไปเข้ามาคือ $[5,5]$

1. หานิวรอนที่ชนะโดยใช้ Euclidean distance เมื่อ input คือ $[5,5]$ จะได้โหนดที่ชนะคือโหนดที่ 3

ตารางที่ 3.12 ตารางแสดงนิวรอนที่ชนะ

โหนด	Euclidean distance	Distance
1	$\left[(5-0.49)^2+(5-5.49)^2\right]^{1/2}$	4.536
2	$\left[(5-2.90)^2+(5-2.95)^2\right]^{1/2}$	2.935
3	$\left[(5-3.99)^2+(5-0.38)^2\right]^{1/2}$	4.729

2. ปรับค่า weight โดยสูตร $w_i(t+1)=w_i(t)+\Delta w_i(t)$ และ $\alpha(t)=0.5$ และ $\sigma(t)=0.7$

$$\Delta w_i(t)=\alpha(N_i,t)[x(t)-w_i(t)] \quad for \quad i\in N_m(t)$$

$$\alpha(N_i,t)=\alpha(t)\exp[-\|r_i-r_m\|/\sigma^2(t)]$$

ที่ $m=2, i=1$

$$\begin{aligned}
w_1(t+1) &= w_1(t)+0.5\exp[-\|r_i-r_m\|/\sigma^2(t)][x(t)-w_1(t)] \\
&= w_1(t)+0.5\exp[-\|r_1-r_2\|/(0.7)^2][x(t)-w_1(t)] \\
\|r_1-r_2\| &= \left[(0.49-2.90)^2+(5.49-2.95)^2\right]^{1/2} \\
&= \left[(-2.41)^2+(2.54)^2\right]^{1/2} \\
&= \left[(5.8081)+(6.4516)\right]^{1/2}=3.501 \\
w_1(t+1) &= w_1(t)+0.5\exp[(-3.501)/(0.7)^2][x(t)-w_1(t)] \\
&= w_1(t)+0.5\exp[(-7.145)][x(t)-w_1(t)] \\
&= [0.49,5.49]+0.5(0.00078)([5,5]-[0.49,5.49]) \\
&= [0.49,5.49]+(0.00039)[4.51,-0.49]
\end{aligned}$$

เอกสารนี้เป็นเอกสารที่สงวนลิขสิทธิ์ ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$$= [0.49,5.49] + [0.0017,-0.0002]$$
$$= [0.4917,5.4898]$$

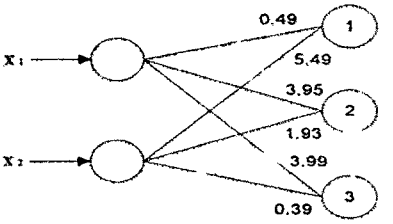
ที่ $m = 2, i = 2$

$$w_2(t+1) = w_2(t) + 0.5\exp\left[-\|r_i - r_m\| / \sigma^2(t)\right]\left[x(t) - w_2(t)\right]$$
$$= w_2(t) + 0.5\exp\left[-\|r_2 - r_2\| / (0.7)^2\right]\left[x(t) - w_2(t)\right]$$
$$= [2.90,2.95] + 0.5(1)([5,5] - [2.90,2.95])$$
$$= [2.90,2.95] + (0.5)[2.1,-2.05]$$
$$= [2.90,2.95] + [1.05,-1.025]$$
$$= [3.95,1.925]$$

ที่ $m = 2, i = 3$

$$w_3(t+1) = w_3(t) + 0.5\exp\left[-\|r_i - r_m\| / \sigma^2(t)\right]\left[x(t) - w_3(t)\right]$$
$$= w_3(t) + 0.5\exp\left[-\|r_3 - r_2\| / (0.7)^2\right]\left[x(t) - w_3(t)\right]$$
$$\|r_3 - r_2\| = \left[(3.99 - 2.9)^2 + (0.38 - 2.95)^2\right]^{1/2}$$
$$= \left[(1.09)^2 + (-2.57)^2\right]^{1/2}$$
$$= \left[(1.1881) + (6.6049)\right]^{1/2} = 2.327$$
$$w_3(t+1) = w_3(t) + 0.5\exp\left[(-2.327)/(0.7)^2\right]\left[x(t) - w_3(t)\right]$$
$$= w_3(t) + 0.5\exp\left[(-4.749)\right]\left[x(t) - w_3(t)\right]$$
$$= [3.99,0.38] + 0.5(0.0086)([5,5] - [3.99,0.38])$$
$$= [3.99,0.38] + (0.0043)[1.01,4.62]$$
$$= [3.99,0.38] + [0.00434,0.0198]$$
$$= [3.9943,0.3998]$$

จาก input คือ $[5,5]$ ปรับค่า weight ครบทั้งโหนดที่ชนะและโหนดข้างเคียงจะ weight ได้ใหม่ออกมาดังแสดงรูปที่ 3.8



รูปที่ 3.8 แสดงการปรับค่า weight โครงข่ายเมื่อ input คือ $[5,5]$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 4
การพัฒนาระบบและผลดำเนินการศึกษา

ในบทนี้จะกล่าวถึงรายละเอียดของการพัฒนาระบบการจัดกลุ่มข้อมูล ทำการเปรียบเทียบประสิทธิภาพของอัลกอริทึมการจัดกลุ่มข้อมูลระหว่าง K-Means Algorithm, Fuzzy C-Means Algorithm และ Self Organizing Map Algorithm โดยจะอธิบายขั้นตอนต่างๆ ในการพัฒนาระบบดังต่อไปนี้

4.1 กำหนดวัตถุประสงค์

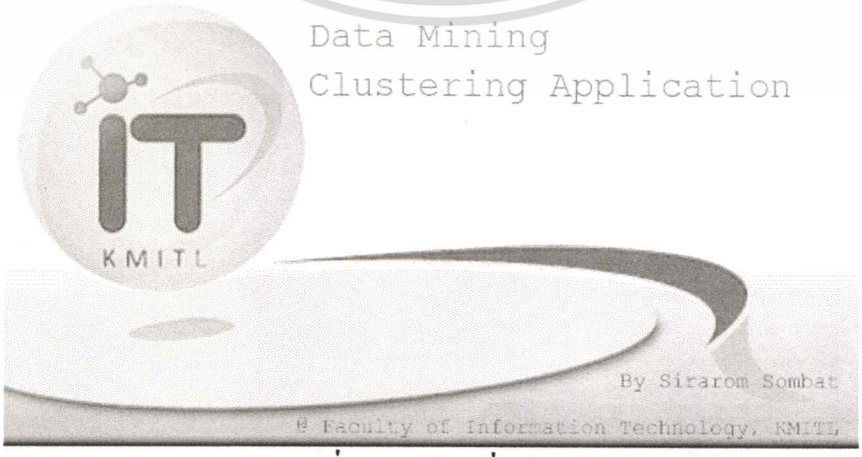
การจัดกลุ่มเป็นวิธีที่นิยมในดาต้าไมน์นิ่ง โดยการพัฒนาระบบงานของโครงการศึกษานี้ มีวัตถุประสงค์หลักเพื่อจะดูประสิทธิภาพของการจำแนกข้อมูลด้วยระหว่าง K-Means Algorithm, Fuzzy C-Means Algorithm และ Self Organizing Map Algorithm

4.2 ขั้นตอนและรายละเอียดการใช้งาน

การพัฒนาระบบงานนี้ แบ่งอัลกอริทึมออกเป็น 3 ส่วนด้วยกันคือ

1. K-Means Algorithm
2. Fuzzy C-Means Algorithm
3. Self Organizing Map Algorithm

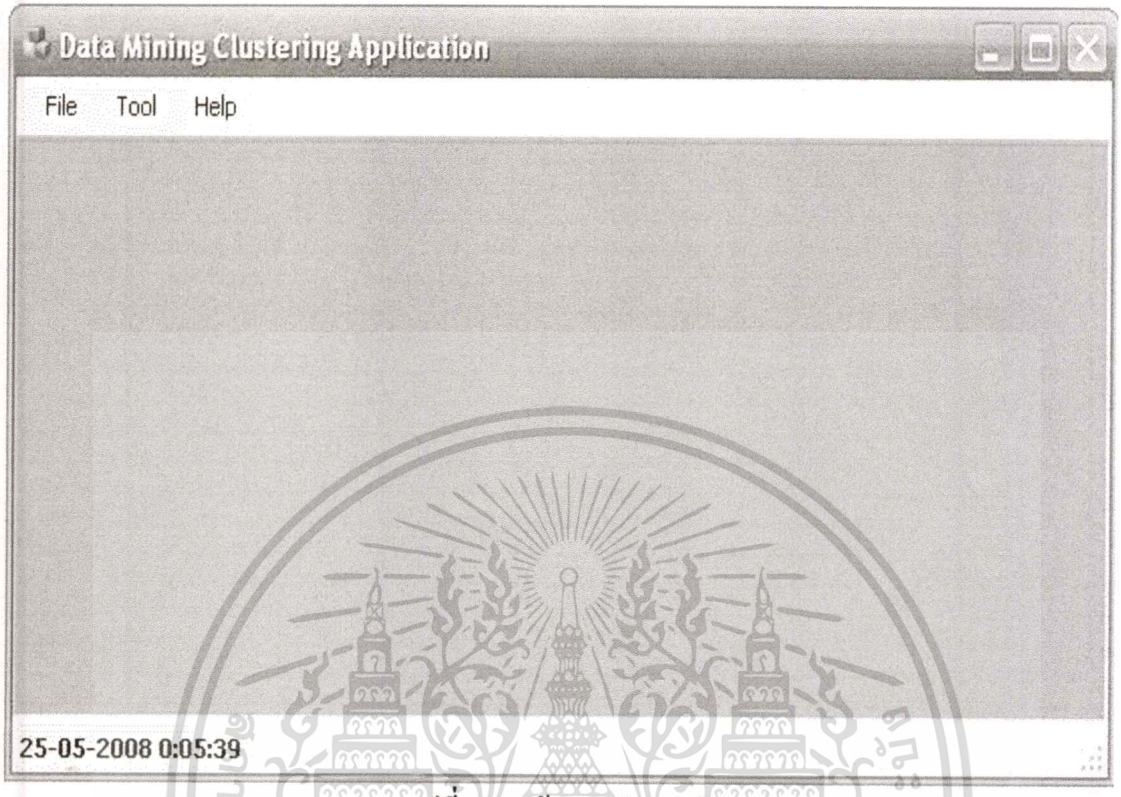
เมื่อเข้าสู่โปรแกรมจะปรากฏหน้าจอดังรูปที่ 4.1



รูปที่ 4.1 หน้าจอเริ่มต้นของระบบ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เมื่อเข้าสู่โปรแกรมจะปรากฏหน้าจอดังรูปที่ 4.2



รูปที่ 4.2 หน้าจอหลักของโปรแกรม

โดยที่ เมนู File > Exit คือ ออกจากโปรแกรม

Tool > K-Means คือ เลือกการทำงาน K-Means อัลกอริทึม

Tool > Fuzzy C-Means คือ เลือกการทำงาน Fuzzy C-Means อัลกอริทึม

Tool > SOM คือ เลือกการทำงาน SOM อัลกอริทึม

Tool > Horizontal คือ การจัดหน้าจอการทำงานแบบ Horizontal

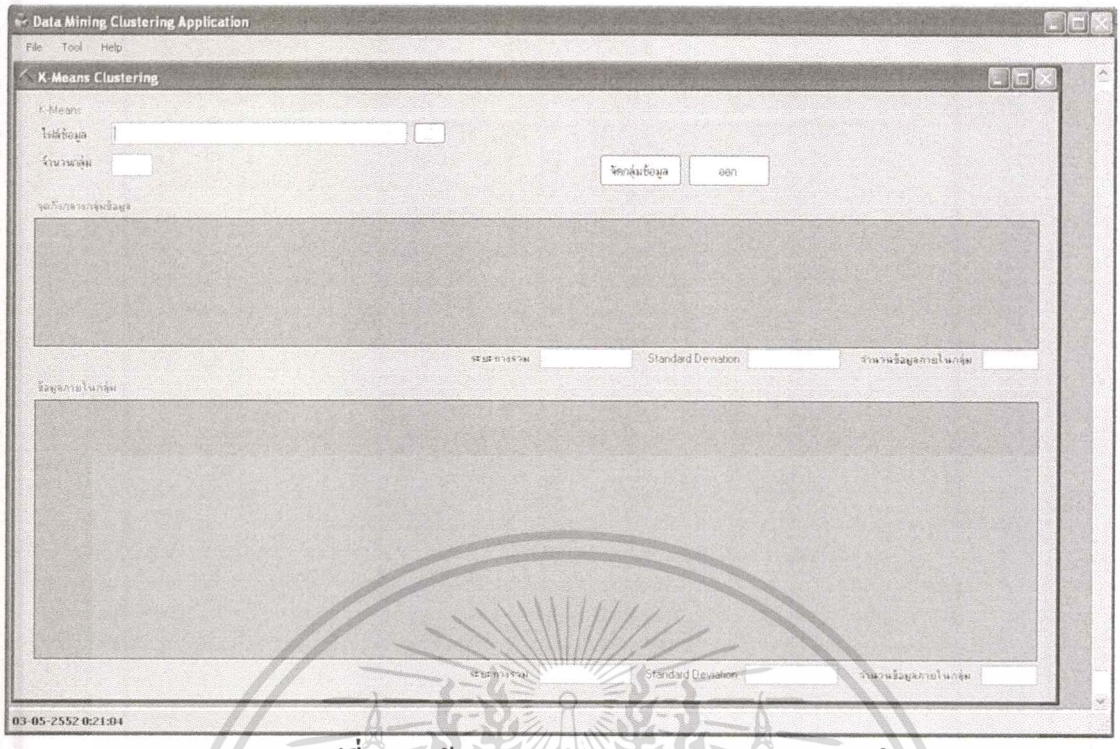
Tool > Vertical คือ การจัดหน้าจอการทำงานแบบ Vertical

Help > How to use คือ แสดงวิธีการใช้งาน

Help > About คือ แสดงรายละเอียดโปรแกรมที่พัฒนา, เวอร์ชั่น

เมื่อเข้าสู่หน้าจอหลัก จะปรากฏหน้าจอที่เริ่มการทำงานของโปรแกรม ก่อนอื่นต้องเลือกอัลกอริทึมที่ต้องการทำงานด้วย โดยไปเลือกที่เมนู Tool > K-Means จะแสดงหน้าจอการทำงานของ K-Means อัลกอริทึม ดังรูปที่ 4.3

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.3 หน้าจอการทำงานของ K-Means อัลกอริทึม

โดยที่ ไฟล์ข้อมูล คือ นำเข้าไฟล์ข้อมูลที่ต้องการนำมาจัดกลุ่ม

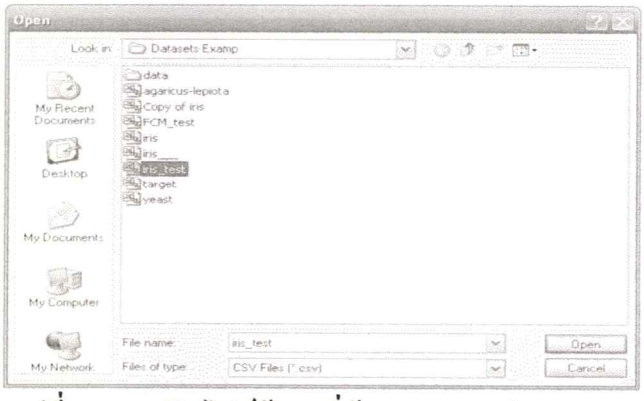
จำนวนกลุ่ม คือ จำนวนกลุ่มข้อมูลที่ต้องาการจัดกลุ่ม

ปุ่มจัดกลุ่มข้อมูล คือ โปรแกรมจะประมวลผลตามพารามิเตอร์ที่รับเข้ามาแสดงผลออกทางหน้าจอ และสร้าง XML ไฟล์ให้ด้วย

ปุ่มออก คือ ออกจากหน้าจอนี้

วิธีการใช้งานในหน้าจอ K-Means นั้น สามารถทำงานได้โดย

1. กด ... โปรแกรมจะเปิดไฟล์ข้อมูลที่ต้องการนำมาประมวลผลดังรูป 4.4



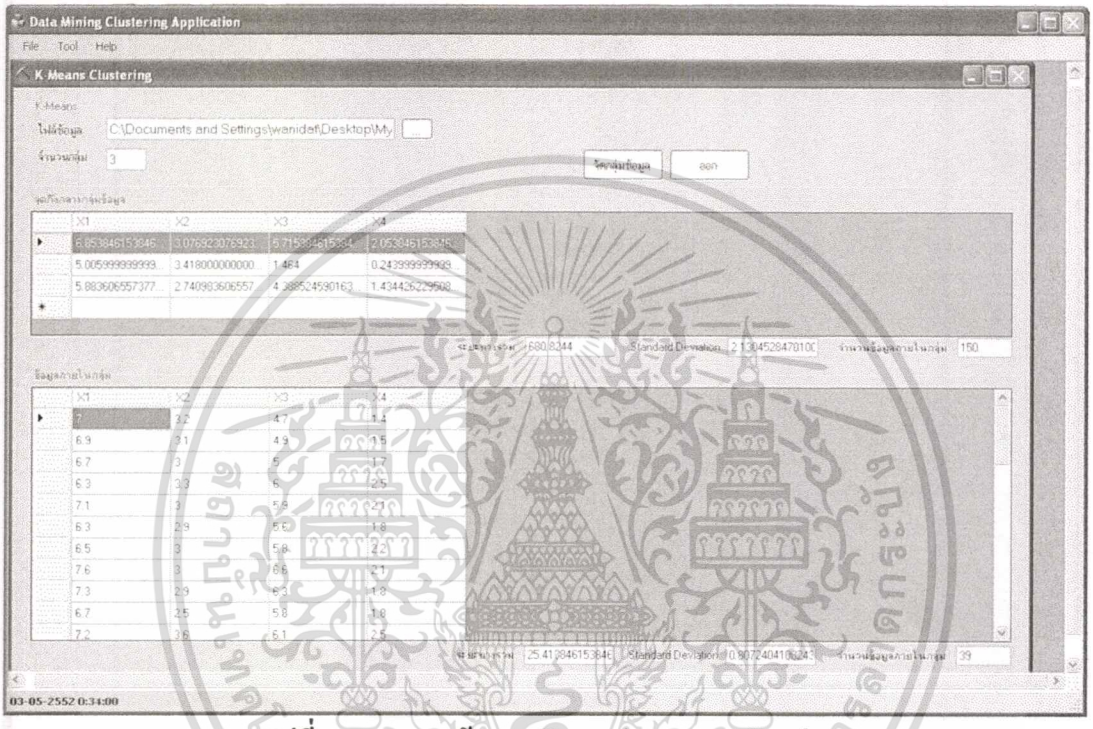
รูปที่ 4.4 การเปิดไฟล์ข้อมูลที่ต้องการนำมาประมวลผล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2. ใส่จำนวนที่ต้องการจัดกลุ่ม

3. กดปุ่ม

โปรแกรมจะประมวลผลแสดงออกหน้าจอ เมื่อผู้ใช้ต้องการดูว่าข้อมูลในกลุ่มนี้มีสมาชิกเป็นตัวใดบ้างสามารถดูได้ด้วยการคลิกที่เรคอร์ดซึ่งแสดงจุดกึ่งกลางของแต่ละกลุ่ม แล้วผลลัพธ์จะปรากฏในส่วนของสมาชิกของกลุ่ม ดังแสดงในรูปที่ 4.5



รูปที่ 4.5 แสดงหน้าจอหลังการประมวลผลจัดกลุ่ม

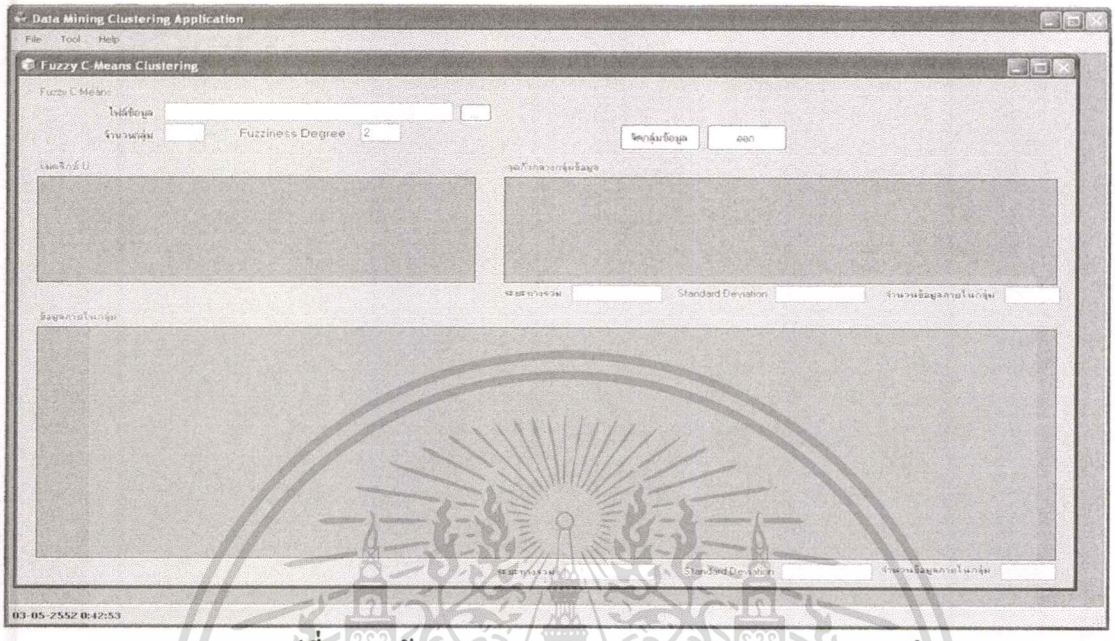
และหากผู้ใช้ต้องการข้อมูลที่เป็น XML ไฟล์ สามารถเข้าไปเปิดได้ที่ \XML\KMeans.xml ดังแสดงในรูป 4.6

```
<?xml version="1.0" encoding="utf-8" ?>
- <ArrayOfArrayOfArrayOfDouble xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema">
  - <ArrayOfArrayOfDouble>
    - <ArrayOfDouble>
        <double>5.1</double>
        <double>3.5</double>
        <double>1.4</double>
        <double>0.2</double>
        <double>1</double>
      </ArrayOfDouble>
    - <ArrayOfDouble>
        <double>4.9</double>
        <double>3</double>
        <double>1.4</double>
        <double>0.2</double>
        <double>1</double>
      </ArrayOfDouble>
    - <ArrayOfDouble>
        <double>4.7</double>
```

รูปที่ 4.6 แสดงข้อมูลที่อยู่ในรูปแบบ XML

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เมื่อเลือกที่เมนู Tool > Fuzzy C-Means จะแสดงหน้าจอการทำงานของ Fuzzy C-Means อัลกอริทึม ดังรูปที่ 4.7



รูปที่ 4.7 หน้าจอการทำงานของ Fuzzy C-Means อัลกอริทึม

โดยที่ ไฟล์ข้อมูล คือ นำเข้าไฟล์ข้อมูลที่ต้องการนำมาจัดกลุ่ม

จำนวนกลุ่ม คือ จำนวนกลุ่มข้อมูลที่ต้องาการจัดกลุ่ม

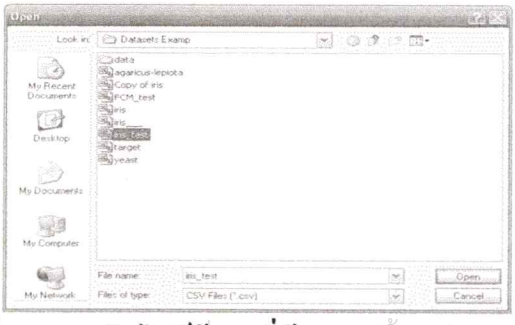
Fuzziness Degree คือ ระดับความเป็นฟัซซี่ ค่าที่สามารถทำงานได้ดีจะอยู่ในช่วงระหว่าง 1.5 – 2.5 โดยปกติกำหนดไว้ให้เท่ากับ 2

ปุ่มจัดกลุ่มข้อมูล คือ โปรแกรมจะประมวลผลตามพารามิเตอร์ที่รับเข้ามาแสดงผลออกทางหน้าจอ และสร้าง XML ไฟล์ให้ด้วย

ปุ่มออก คือ ออกจากหน้าจอนี้

วิธีการใช้งานในหน้าจอ Fuzzy C-Means นั้น สามารถทำงานได้โดย

1. กด ... โปรแกรมจะเปิดไฟล์ข้อมูลที่ต้องการนำมาประมวลผลดังรูป 4.8



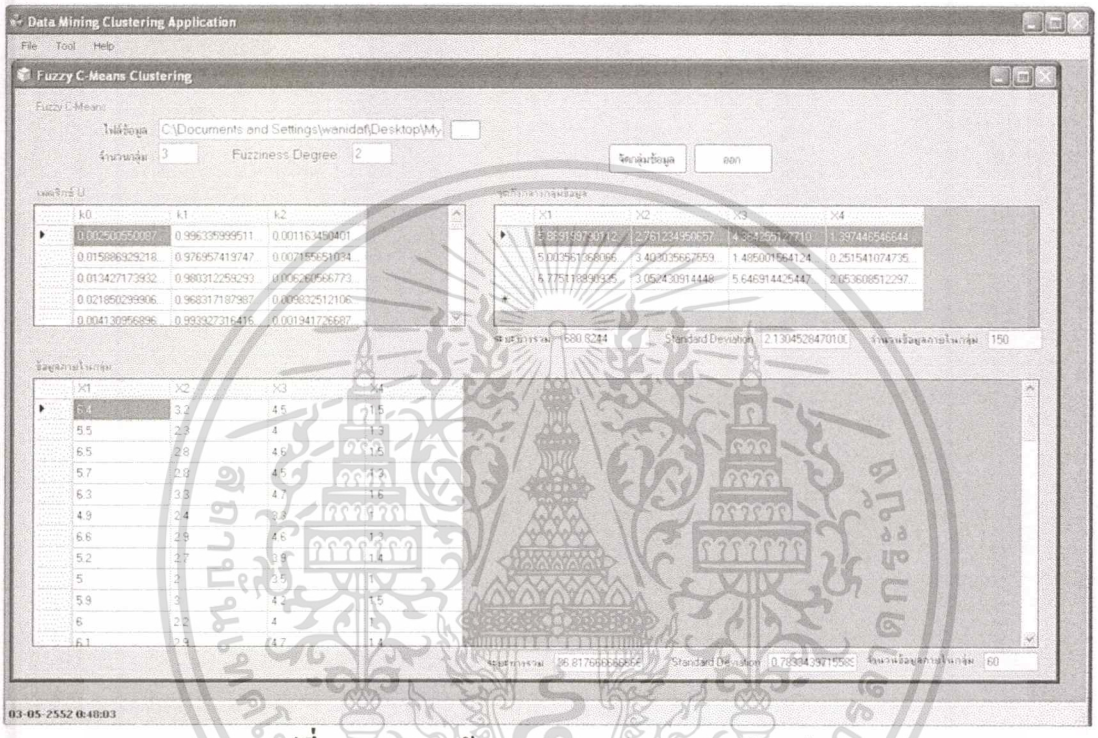
รูปที่ 4.8 การเปิดไฟล์ข้อมูลที่ต้องการนำมาประมวลผล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2. ใส่จำนวนที่ต้องการจัดกลุ่ม และค่า Fuzziness Degree

3. กดปุ่ม

โปรแกรมจะประมวลผลแสดงออกหน้าจอ โดยแสดงค่าเมตริกซ์ U และจุดกึ่งกลางของกลุ่มข้อมูล เมื่อผู้ใช้ต้องการดูว่าข้อมูลในกลุ่มนี้มีสมาชิกเป็นตัวใดบ้าง สามารถดูได้ด้วยการคลิกที่เรคอร์ดซึ่งแสดงจุดกึ่งกลางของแต่ละกลุ่ม แล้วผลลัพธ์จะปรากฏในส่วนของสมาชิกของกลุ่ม ดังแสดงในรูปที่ 4.9



รูปที่ 4.9 แสดงหน้าจอหลังการประมวลผลจัดกลุ่ม

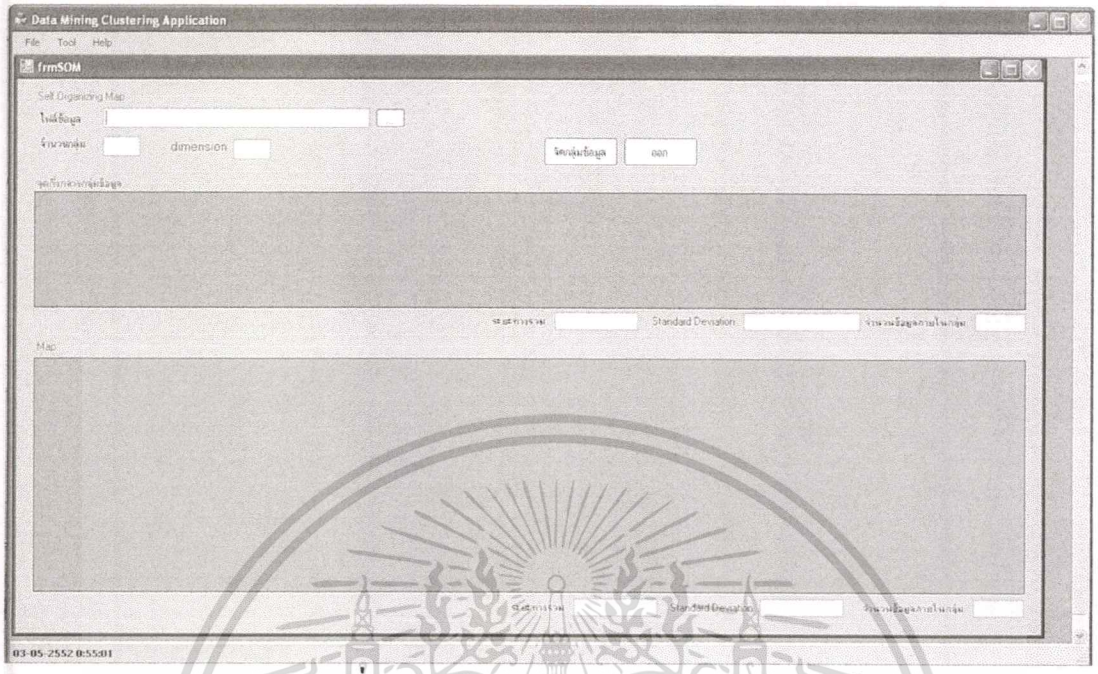
และหากผู้ใช้ต้องการข้อมูลที่เป็น XML ไฟล์ สามารถเข้าไปเปิดได้ที่ \XML\KMeans.xml ดังแสดงในรูป 4.10

```
<?xml version="1.0" encoding="utf-8" ?>
- <ArrayOfArrayOfArrayOfDouble xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema">
  - <ArrayOfArrayOfDouble>
    - <ArrayOfDouble>
        <double>5.1</double>
        <double>3.5</double>
        <double>1.4</double>
        <double>0.2</double>
        <double>1</double>
      </ArrayOfDouble>
    - <ArrayOfDouble>
        <double>4.9</double>
        <double>3</double>
        <double>1.4</double>
        <double>0.2</double>
        <double>1</double>
      </ArrayOfDouble>
    - <ArrayOfDouble>
        <double>4.7</double>
```

รูปที่ 4.10 แสดงข้อมูลที่อยู่ในรูปแบบ XML

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เมื่อเลือกที่เมนู Tool > SOM จะแสดงหน้าจอการทำงานของ SOM อัลกอริทึม ดังรูปที่ 4.11



รูปที่ 4.11 หน้าจอการทำงานของ SOM อัลกอริทึม

โดยที่ ไฟล์ข้อมูล คือ นำเข้าไฟล์ข้อมูลที่ต้องการนำมาจัดกลุ่ม

จำนวนกลุ่ม คือ จำนวนกลุ่มข้อมูลที่ต้องาการจัดกลุ่ม

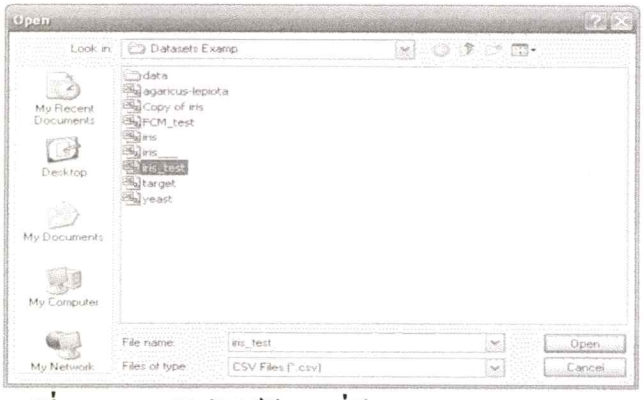
Dimension คือ มิติของข้อมูลที่นำมาทดลอง

ปุ่มจัดกลุ่มข้อมูล คือ โปรแกรมจะประมวลผลตามพารามิเตอร์ที่รับเข้ามาแสดงผลออกทางหน้าจอ

ปุ่มออก คือ ออกจากหน้าจอนี้

วิธีการใช้งานในหน้าจอ SOM นั้น สามารถทำงานได้โดย

1. กด ... โปรแกรมจะเปิดไฟล์ข้อมูลที่ต้องการนำมาประมวลผลดังรูป 4.12



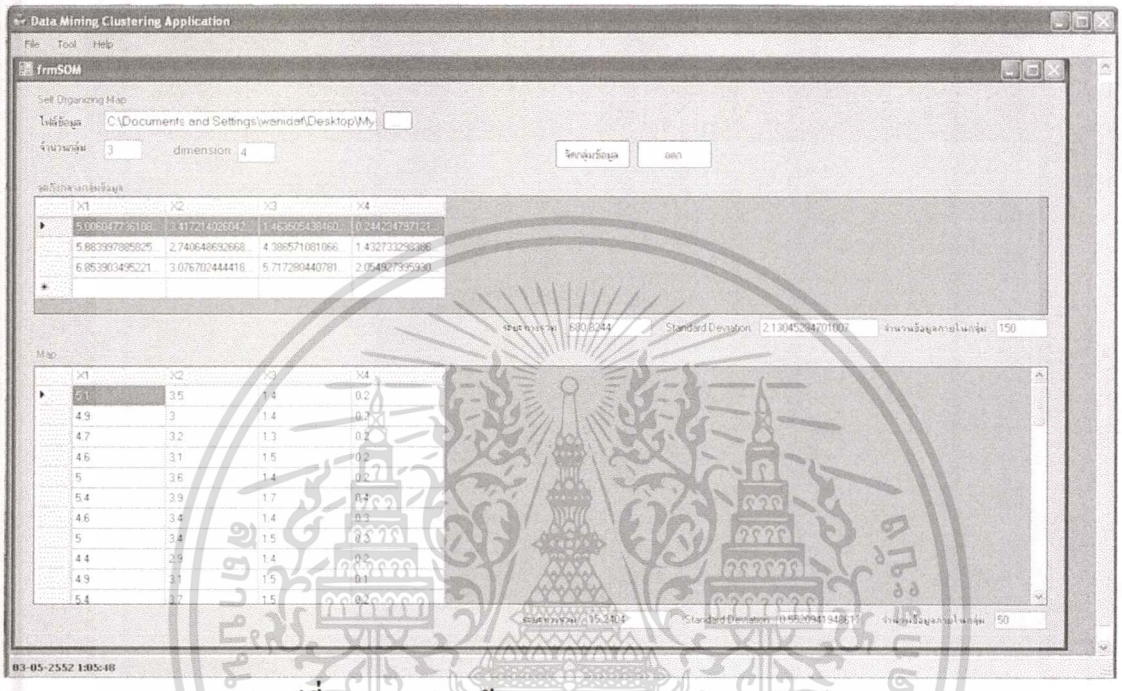
รูปที่ 4.12 การเปิดไฟล์ข้อมูลที่ต้องการนำมาประมวลผล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2. ใส่จำนวนที่ต้องการจัดกลุ่ม และค่า Dimension

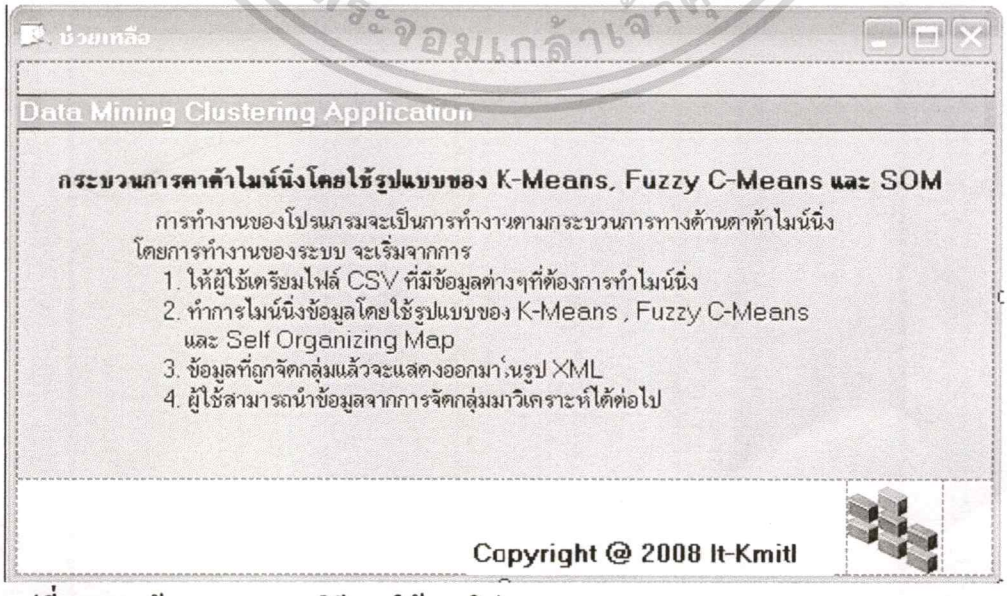
3. กดปุ่ม

โปรแกรมจะประมวลผลแสดงออกหน้าจอ เมื่อผู้ใช้ต้องการดูว่าข้อมูลในกลุ่มนี้มีสมาชิกเป็นตัวใดบ้างสามารถดูได้ด้วยการคลิกที่เรคอร์ดซึ่งแสดงจุดกึ่งกลางของแต่ละกลุ่ม แล้วผลลัพธ์จะปรากฏในส่วนของสมาชิกของกลุ่ม ดังแสดงในรูปที่ 4.13



รูปที่ 4.13 แสดงหน้าจอหลังการประมวลผลจัดกลุ่ม

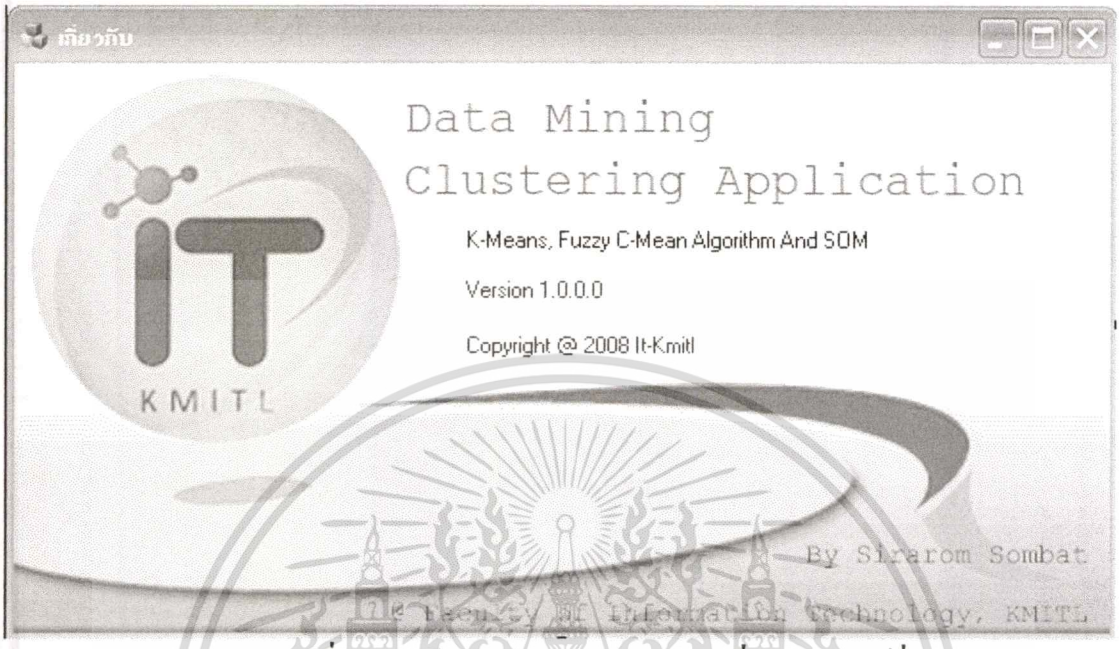
เมื่อเลือกที่เมนู Help > How to use จะแสดงหน้าจอวิธีการใช้งานโปรแกรม ดังรูปที่ 4.14



รูปที่ 4.14 หน้าจอการแสดงวิธีการใช้งานโปรแกรม Data Mining Application

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เมื่อเลือกที่เมนู Help > About จะแสดงหน้าจอรายละเอียดโปรแกรมที่พัฒนา, เวอร์ชั่น ดังรูป 4.15



รูปที่ 4.15 หน้าจอรายละเอียดโปรแกรมที่พัฒนา, เวอร์ชั่น

4.3 ข้อมูลที่นำมาทดลอง

ข้อมูลที่นำมาทดลอง ได้แก่ ข้อมูล Iris, Wine, Segmentation และ Diabetes โดยแต่ละข้อมูลมีลักษณะข้อมูลดังต่อไปนี้

1. ข้อมูล Iris

ข้อมูล Iris มี 4 ค่าข้อมูลที่แตกต่างกัน ได้แก่ ค่าความยาวของกลีบเลี้ยงดอกไม้, ค่าความกว้างของกลีบเลี้ยงดอกไม้, ค่าความยาวของกลีบดอกไม้ และค่าความกว้างของกลีบดอกไม้ ในหน่วยเซนติเมตร มีทั้งหมด 3 คลาส ดังรูปที่ 4.16 (ที่ซึ่งข้อมูลของคลาส 2 และคลาส 3 คาบเกี่ยวกัน) โดยแต่ละคลาสมีข้อมูล 50 ตัวอย่าง ดังนั้นจำนวนกลุ่มข้อมูลของ Iris จึงเท่ากับ 3

1	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	7	3.2	4.7	1.4	versicolor
6	6.4	3.2	4.5	1.5	versicolor
7	6.9	3.1	4.9	1.5	versicolor
8	6.3	3.3	6	2.5	virginica
9	5.8	2.7	5.1	1.9	virginica
10	7.1	3	5.9	2.1	virginica

รูปที่ 4.16 ข้อมูล Iris ที่แบ่งตามคลาส

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2. ข้อมูล Wine

ข้อมูลไวน์รสเลิศปลูกในพื้นที่เดียวกันในอิตาลี มี 13 ค่าข้อมูลที่แตกต่างกัน ได้แก่ Alcohol, Malic acid, Ash, Alcalinity of ash, Magnesium, Total phenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines, Proline มีทั้งหมด 3 คลาส ซึ่งคลาสที่หนึ่งมีข้อมูล 59 ตัวอย่าง คลาสที่ 2 มีข้อมูล 71 ตัวอย่าง และคลาสที่ 3 มีข้อมูล 48 ตัวอย่าง ดังรูปที่ 4.17

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13
14.23	1.71	2.43	15.6	127	2.8	3.06	0.28	2.29	5.64	1.04	3.92	1065
13.2	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.4	1050
13.16	2.36	2.67	18.6	101	2.8	3.24	0.3	2.81	5.68	1.03	3.17	1185
14.37	1.95	2.5	16.8	113	3.85	3.49	0.24	2.18	7.8	0.86	3.45	1480
13.24	2.59	2.87	21	118	2.8	2.69	0.39	1.82	4.32	1.04	2.93	735
14.2	1.76	2.45	15.2	112	3.27	3.39	0.34	1.97	6.75	1.05	2.85	1450
14.39	1.87	2.45	14.6	96	2.5	2.52	0.3	1.98	5.25	1.02	3.58	1290
14.06	2.15	2.61	17.6	121	2.6	2.51	0.31	1.25	5.05	1.06	3.58	1295
14.83	1.64	2.17	14	97	2.8	2.98	0.29	1.98	5.2	1.08	2.85	1045

รูปที่ 4.17 ข้อมูล Wine ที่แบ่งตามคลาส

3. ข้อมูล Segmentation

ข้อมูล Segmentation เป็นข้อมูลที่เกี่ยวข้องกับรูปภาพของสิ่งของกลางแจ้ง มีทั้งหมด 2310 ข้อมูล ประกอบด้วย 19 คุณลักษณะ และมีทั้งหมด 7 คลาส คลาสละ 330 ตัวอย่าง ได้แก่ คลาส หน้าอิฐ, ท้องฟ้า, ใบไม้, ซีเมนต์, หน้าต่าง, ทางเดิน และหญ้า ดังรูปที่ 4.18

```
BRICKFACE,140.0,125.0,9,0.0,0.0,0.0,0.2777779,0.06296301,0.66666675,0.31111118
,6.185185,7.3333335,7.6666665,3.5555556,3.4444444,4.4444447,-
7.888889,7.7777777,0.5456349,-1.1218182
SKY,140.0,25.0,9,0.0,0.0,0.0,0.99999875,1.4666697,1.1111107,0.118517555,128.0,
117.77778,142.33334,123.888885,-30.666666,43.0,-
12.333333,142.33334,0.17250364,-2.3545518
FOLIAGE,101.0,121.0,9,0.11111111,0.0,0.6666667,0.843274,1.5,0.88819396,3.4
074075,1.1111112,6.0,3.1111112,-6.888889,7.7777777,-
0.8888889,6.0,0.8435185,-2.5191648
CEMENT,191.0,119.0,9,0.0,0.0,1.1111107,1.2938615,0.9444459,0.772202,39.851
852,36.22222,48.22222,35.11111,-10.888889,25.11111,-
14.222222,48.22222,0.27171725,-2.0059998
WINDOW,189.0,144.0,9,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0
.0,0.0,0.0,0
PATH,86.0,197.0,9,0.11111111,0.11111111,1.611112,1.4516907,1.2777786,1.103
8647,63.22222,56.22222,77.77778,55.666668,-21.0,43.666668,-
22.666666,77.77778,0.28533262,-2.06802
GRASS,204.0,156.0,9,0.0,0.0,0.5000003,0.27888626,1.9999996,0.55777323,23.7
03703,17.333334,25.444445,28.333334,-
19.11111,5.2222223,13.888889,28.333334,0.38903406,2.8649306
```

รูปที่ 4.18 ข้อมูล Segmentation ที่แบ่งตามคลาส

4. ข้อมูล Diabetes

ข้อมูลนี้เป็นข้อมูลของผู้ป่วยโรคเบาหวาน ที่เป็นเพศหญิง อายุอย่างน้อย 21 ปีขึ้นไป มีทั้งหมด 768 ข้อมูล ประกอบด้วย 8 คุณลักษณะ และมี 2 คลาส ได้แก่ คลาส tested negative และ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

tesed positive โดยข้อมูลคลาสแรกมี 500 ตัวอย่าง และ คลาสที่สองมี 268 ตัวอย่าง ดังรูปที่ 4.19 ดังนั้นจำนวนกลุ่มของ Diabetes จึงมีค่าเท่ากับ 2

```
0.067474048,0.477830328,0.540180006,0.138542286,0.053840449,0.296177786,0.116968905,0.075163399,1
0.125144175,0.561582422,0.636612022,0.504951476,0.172043758,0.585473832,0.050409404,0.058823529,1
0.355824683,0.586215391,0.781260045,0.009803922,0.009803922,0.42913708,0.042874366,0.156862745,1
0.471164937,0.630554734,0.716972035,0.36631016,0.134960367,0.572324012,0.123248104,0.467320261,1
0.009803922,0.72908661,0.572324012,0.009803922,0.009803922,0.563557465,0.116968905,0.124183007,2
0.298154556,0.561582422,0.540180006,0.009803922,0.009803922,0.562096374,0.08641014,0.336601307,2
0.125144175,0.620701547,0.556252009,0.287086552,0.247369397,0.490502908,0.343438656,0.156862745,2
0.067474048,0.857178047,0.556252009,0.495048524,0.680781532,0.629306566,0.27101857,0.124183007,2
0.009803922,0.615774953,0.588396014,0.009803922,0.009803922,0.540180006,0.085154301,0.516339869,2
```

รูปที่ 4.19 ข้อมูล Diabetes ที่แบ่งตามคลาส

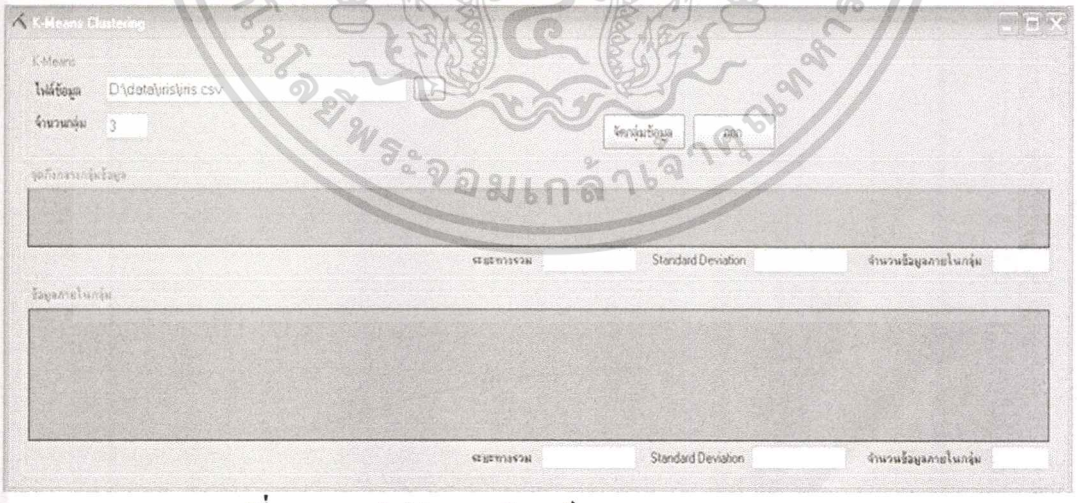
4.4 ผลการจัดกลุ่มข้อมูล

ในการทดลองเปรียบเทียบประสิทธิภาพในการจัดกลุ่มข้อมูลระหว่าง K-Means Algorithm, Fuzzy C-Means Algorithm และ Self Organizing Map Algorithm สำหรับข้อมูล Iris, Wine, Segmentation และ Diabetes ได้ผลการทดลองดังนี้

4.4.1 ผลการจัดกลุ่มข้อมูล Iris

1. การจัดกลุ่มแบบ K-Means Algorithm

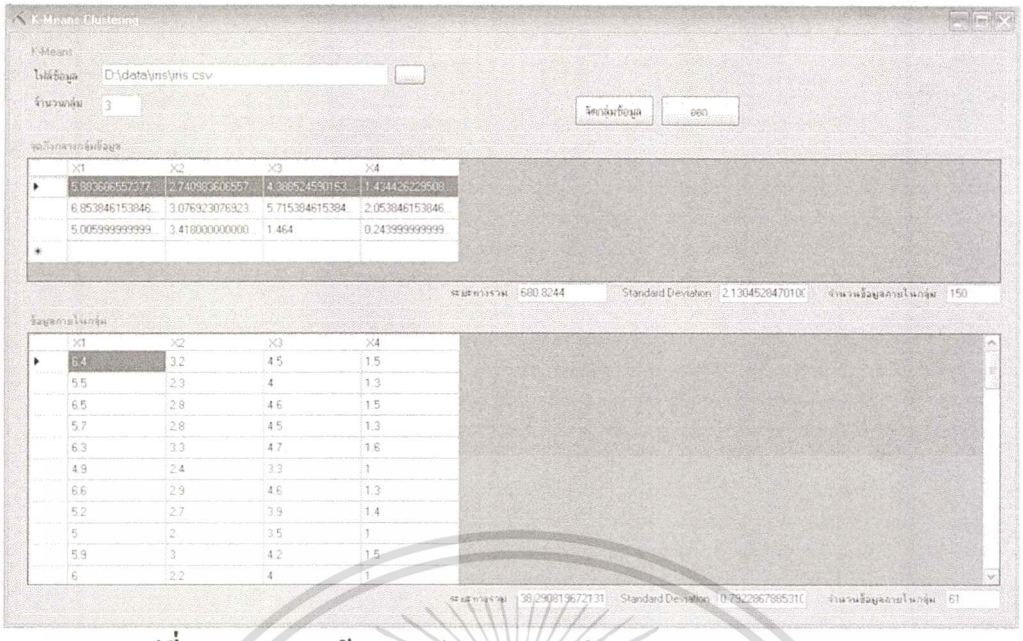
การจัดกลุ่มข้อมูลของข้อมูล Iris ที่มีข้อมูล 150 ตัวอย่าง และกำหนดให้จำนวนกลุ่ม K เท่ากับ 3 แสดงหน้าจอการใส่พารามิเตอร์ดังรูปที่ 4.20



รูปที่ 4.20 การใส่ค่าพารามิเตอร์ของ K-Means Algorithm

เมื่อรันผลการจัดกลุ่มเสร็จ จะแสดงหน้าจอสรุปผลระยะทางรวมของการจัดกลุ่มข้อมูล และส่วนเบี่ยงเบนมาตรฐาน พร้อมทั้ง XML ไฟล์เอาท์พุตของ K-Means Algorithm ได้ดังรูปที่ 4.21

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.21 แสดงหน้าจอสรุปผลการจัดกลุ่มของ K-Means Algorithm

ตารางที่ 4.1 แสดงผลการจัดกลุ่มข้อมูล Iris แบบ K-Means Algorithm

รันครั้งที่	กลุ่มที่	จำนวนสมาชิกภายในกลุ่ม	ระทางรวม	ส่วนเบี่ยงเบนมาตรฐาน
1		150	78.94084142	1.25518479733985
	1	50	15.2404	0.552094195
	2	38	23.87947368	0.79272139
	3	62	39.82096774	0.801419786
2		150	78.94084142	1.25518479733985
	1	50	15.2404	0.552094195
	2	38	23.87947368	0.79272139
	3	62	39.82096774	0.801419786
3		150	78.94084142	1.25518479733985
	1	50	15.2404	0.552094195
	2	38	23.87947368	0.79272139
	3	62	39.82096774	0.801419786
4		150	78.94506582	1.25863554540172
	1	50	15.2404	0.552094195
	2	39	25.41384615	0.807240411
	3	61	38.29081967	0.792286789

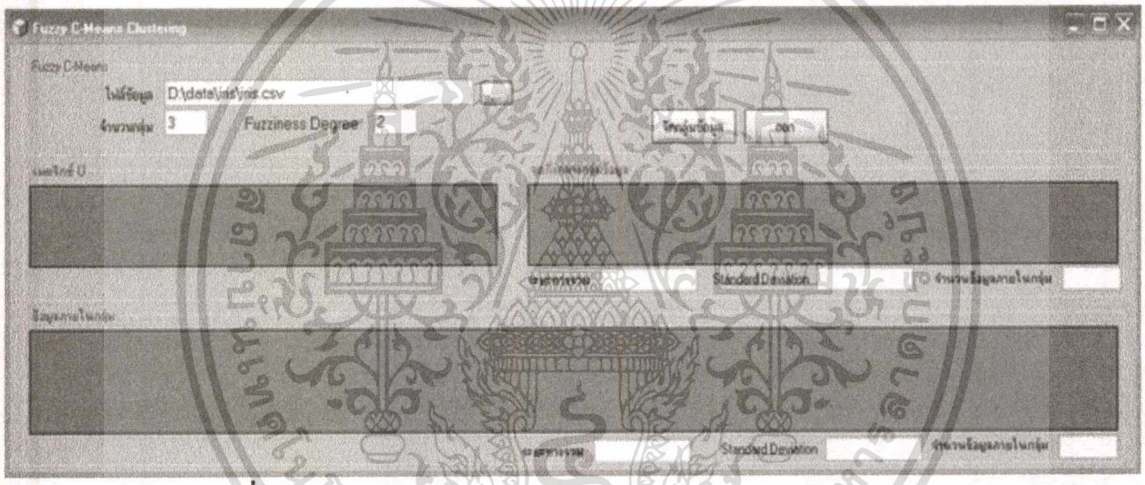
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.1 (ต่อ)

รันครั้งที่	กลุ่มที่	จำนวนสมาชิกภายในกลุ่ม	ระทางรวม	ส่วนเบี่ยงเบนมาตรฐาน
5		150	78.94084142	1.25518479733985
	1	50	15.2404	0.552094195
	2	38	23.87947368	0.79272139
	3	62	39.82096774	0.801419786

2. การจัดกลุ่มข้อมูลแบบ Fuzzy C-Means Algorithm

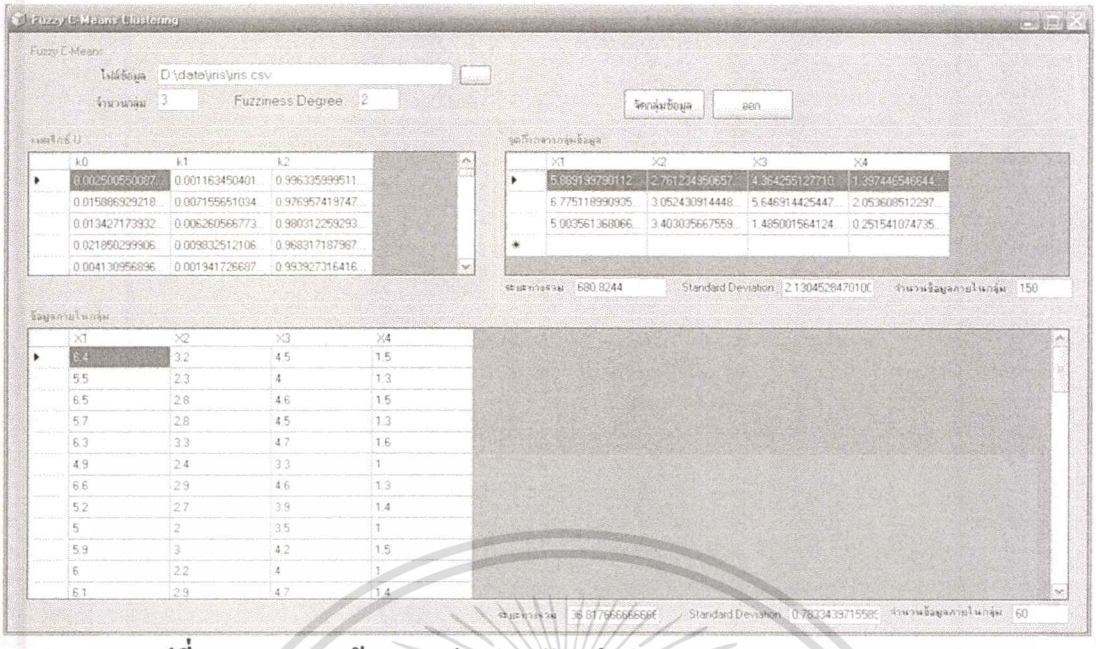
การจัดกลุ่มข้อมูลของข้อมูล Iris ที่มีข้อมูล 150 ตัวอย่าง และกำหนดให้จำนวนกลุ่ม K เท่ากับ 3, Fuzziness Degree เท่ากับ 2 แสดงหน้าจอการใส่พารามิเตอร์ดังรูปที่ 4.22



รูปที่ 4.22 การใส่ค่าพารามิเตอร์ของ Fuzzy C-Means Algorithm

เมื่อรันผลการจัดกลุ่มเสร็จ จะแสดงหน้าจอสรุปผลระยะทางรวมของการจัดกลุ่มข้อมูล และส่วนเบี่ยงเบนมาตรฐาน พร้อมทั้ง XML ไฟล์เอาท์พุตของ Fuzzy C-Means Algorithm ได้ดังรูปที่ 4.23

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.23 แสดงหน้าจอสรุปผลการจัดกลุ่มของ Fuzzy C-Means Algorithm

ตารางที่ 4.2 แสดงผลการจัดกลุ่มข้อมูล Iris แบบ Fuzzy C-Means Algorithm

รันครั้งที่	กลุ่มที่	จำนวนสมาชิกภายในกลุ่ม	ระทางรวม	ส่วนเบี่ยงเบนมาตรฐาน
1		150	79.11556667	1.26288292322676
	1	50	15.2404	0.552094195
	2	40	27.0575	0.822458206
	3	60	36.81766667	0.783343972
2		150	79.11556667	1.26288292322676
	1	50	15.2404	0.552094195
	2	40	27.0575	0.822458206
	3	60	36.81766667	0.783343972
3		150	79.11556667	1.26288292322676
	1	50	15.2404	0.552094195
	2	40	27.0575	0.822458206
	3	60	36.81766667	0.783343972
4		150	79.11556667	1.26288292322676
	1	50	15.2404	0.552094195
	2	40	27.0575	0.822458206
	3	60	36.81766667	0.783343972

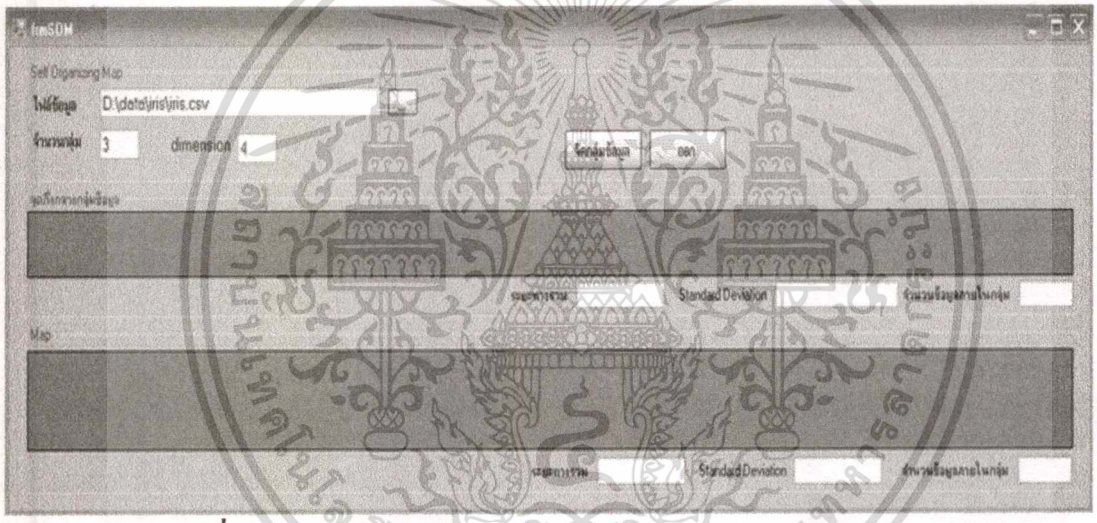
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.2 (ต่อ)

รันครั้งที่	กลุ่มที่	จำนวนสมาชิกภายในกลุ่ม	ระทางรวม	ส่วนเบี่ยงเบนมาตรฐาน
5		150	79.11556667	1.26288292322676
	1	50	15.2404	0.552094195
	2	40	27.0575	0.822458206
	3	60	36.81766667	0.783343972

3. การจัดกลุ่มข้อมูลแบบ Self Organizing Map Algorithm

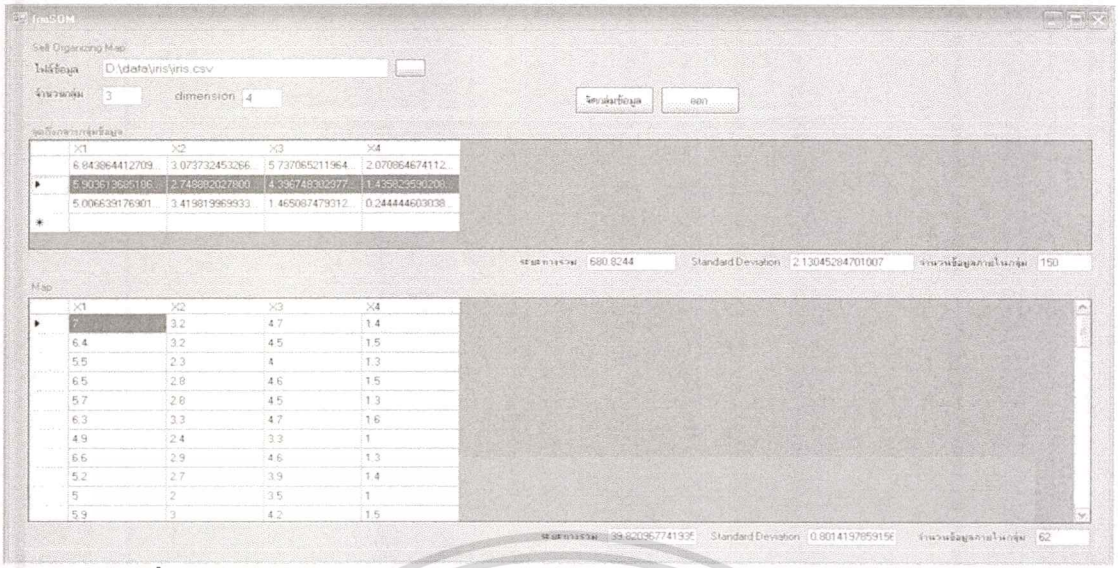
การจัดกลุ่มข้อมูลของข้อมูล Iris ที่มีข้อมูล 150 ตัวอย่าง และกำหนดให้จำนวนกลุ่ม K เท่ากับ 3, Dimension เท่ากับ 4 แสดงหน้าจอการใส่พารามิเตอร์ดังรูปที่ 4.24



รูปที่ 4.24 การใส่ค่าพารามิเตอร์ของ Self Organizing Map Algorithm

เมื่อรันผลการจัดกลุ่มเสร็จ จะแสดงหน้าจอสรุปผลระยะทางรวมของการจัดกลุ่มข้อมูล และส่วนเบี่ยงเบนมาตรฐาน ผลการจัดกลุ่มของ Self Organizing Map Algorithm ได้ดังรูปที่ 4.25

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.25 แสดงหน้าจอสรุปผลการจัดกลุ่มของ Self Organizing Map Algorithm

ตารางที่ 4.3 แสดงผลการจัดกลุ่มข้อมูล Iris แบบ Self Organizing Map Algorithm

รันครั้งที่	กลุ่มที่	จำนวนสมาชิกภายในกลุ่ม	ระทางรวม	ส่วนเบี่ยงเบนมาตรฐาน
1		150	78.94084142	1.25518479733985
	1	50	15.2404	0.552094195
	2	38	23.87947368	0.79272139
	3	62	39.82096774	0.801419786
2		150	78.94084142	1.25518479733985
	1	50	15.2404	0.552094195
	2	38	23.87947368	0.79272139
	3	62	39.82096774	0.801419786
3		150	78.94084142	1.25518479733985
	1	50	15.2404	0.552094195
	2	38	23.87947368	0.79272139
	3	62	39.82096774	0.801419786
4		150	78.94506582	1.25863554540172
	1	50	15.2404	0.552094195
	2	39	25.41384615	0.807240411
	3	61	38.29081967	0.792286789

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.3 แสดงผลการจัดกลุ่มข้อมูล Iris แบบ Self Organizing Map Algorithm (ต่อ)

รันครั้งที่	กลุ่มที่	จำนวนสมาชิกภายในกลุ่ม	ระทางรวม	ส่วนเบี่ยงเบนมาตรฐาน
5		150	78.94506582	1.25863554540172
	1	50	15.2404	0.552094195
	2	39	25.41384615	0.807240411
	3	61	38.29081967	0.792286789

สำหรับข้อมูล Iris ในตารางที่ 4.1, 4.2 และ 4.3 ผลของ K-Means ที่ดีที่ที่สุดคือการรันครั้งที่ 1 , ผลของ Fuzzy C-Means ที่ดีที่สุดคือการรันครั้งที่ 1 และผลของ Self Organizing Map ที่ดีที่สุดคือการรันครั้งที่ 1 ดังแสดงในตารางที่ 4.4

ตารางที่ 4.4 แสดงผลการจัดกลุ่มข้อมูล Iris ที่ดีที่สุด

รันครั้งที่	กลุ่มที่	จำนวนสมาชิกภายในกลุ่ม	ระทางรวม	ส่วนเบี่ยงเบนมาตรฐาน
K-Means				
1		150	78.94084142	1.25518479733985
	1	50	15.2404	0.552094195
	2	38	23.87947368	0.79272139
	3	62	39.82096774	0.801419786
Fuzzy C-Means				
1		150	79.11556667	1.26288292322676
	1	50	15.2404	0.552094195
	2	40	27.0575	0.822458206
	3	60	36.81766667	0.783343972
Self Organizing Map				
1		150	78.94084142	1.25518479733985
	1	50	15.2404	0.552094195
	2	38	23.87947368	0.79272139
	3	62	39.82096774	0.801419786

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.4.2 ผลการจัดกลุ่มข้อมูล Wine

1. การจัดกลุ่มแบบ K-Means Algorithm

การจัดกลุ่มข้อมูลของข้อมูล Wine ที่มีข้อมูล 178 ตัวอย่าง และกำหนดให้จำนวนกลุ่ม K เท่ากับ 3 เมื่อรันผลการจัดกลุ่มเสร็จ จะแสดงหน้าจอสรุปผลระยะทางรวมของการจัดกลุ่มข้อมูล และส่วนเบี่ยงเบนมาตรฐานในการรันแต่ละครั้งได้ดังตารางที่ตารางที่ 4.5

ตารางที่ 4.5 แสดงผลการจัดกลุ่มข้อมูล Wine แบบ K-Means Algorithm

รันครั้งที่	กลุ่มที่	จำนวนสมาชิกภายในกลุ่ม	ระทางรวม	ส่วนเบี่ยงเบนมาตรฐาน
1		178	2370689.687	210.991386328405
	1	62	566572.5032	95.59427901
	2	69	443166.7208	80.14178746
	3	47	1360950.463	170.1657804
2		178	2370689.687	210.991386328405
	1	62	566572.5032	95.59427901
	2	69	443166.7208	80.14178746
	3	47	1360950.463	170.1657804
3		178	2370689.687	210.991386328405
	1	62	566572.5032	95.59427901
	2	69	443166.7208	80.14178746
	3	47	1360950.463	170.1657804
4		178	2370689.687	210.991386328405
	1	62	566572.5032	95.59427901
	2	69	443166.7208	80.14178746
	3	47	1360950.463	170.1657804
5		178	2370689.687	210.991386328405
	1	62	566572.5032	95.59427901
	2	69	443166.7208	80.14178746
	3	47	1360950.463	170.1657804

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2. การจัดกลุ่มแบบ Fuzzy C-Means Algorithm

การจัดกลุ่มข้อมูลของข้อมูล Wine ที่มีข้อมูล 178 ตัวอย่าง และกำหนดให้จำนวนกลุ่ม K เท่ากับ 3, Fuzziness Degree เท่ากับ 2 เมื่อรันผลการจัดกลุ่มเสร็จ จะแสดงหน้าจอสรุปผลระยะทางรวมของการจัดกลุ่มข้อมูล และส่วนเบี่ยงเบนมาตรฐานในการรันแต่ละครั้งได้ดังตารางที่ ตารางที่ 4.6

ตารางที่ 4.6 แสดงผลการจัดกลุ่มข้อมูล Wine แบบ Fuzzy C-Means Algorithm

รันครั้งที่	กลุ่มที่	จำนวนสมาชิกภายในกลุ่ม	ระทางรวม	ส่วนเบี่ยงเบนมาตรฐาน
1		178	2379535.426	211.878149352172
	1	61	588105.6985	98.18898627
	2	46	1308995.625	168.6903276
	3	71	482434.1023	82.43085897
2		178	2379535.426	211.878149352172
	1	61	588105.6985	98.18898627
	2	46	1308995.625	168.6903276
	3	71	482434.1023	82.43085897
3		178	2379535.426	211.878149352172
	1	61	588105.6985	98.18898627
	2	46	1308995.625	168.6903276
	3	71	482434.1023	82.43085897
4		178	2379535.426	211.878149352172
	1	61	588105.6985	98.18898627
	2	46	1308995.625	168.6903276
	3	71	482434.1023	82.43085897
5		178	2379535.426	211.878149352172
	1	61	588105.6985	98.18898627
	2	46	1308995.625	168.6903276
	3	71	482434.1023	82.43085897

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3. การจัดกลุ่มแบบ Self Organizing Map Algorithm

การจัดกลุ่มข้อมูลของข้อมูล Wine ที่มีข้อมูล 178 ตัวอย่าง และกำหนดให้จำนวนกลุ่ม K เท่ากับ 3, Dimension เท่ากับ 13 เมื่อรันผลการจัดกลุ่มเสร็จ จะแสดงหน้าจอสรุปผลระยะทางรวมของการจัดกลุ่มข้อมูล และส่วนเบี่ยงเบนมาตรฐานในการรันแต่ละครั้งได้ดังตารางที่ ตารางที่ 4.7

ตารางที่ 4.7 แสดงผลการจัดกลุ่มข้อมูล Wine แบบ Self Organizing Map

รันครั้งที่	กลุ่มที่	จำนวนสมาชิกภายในกลุ่ม	ระทางรวม	ส่วนเบี่ยงเบนมาตรฐาน
1		178	2370689.687	210.991386328405
	1	69	443166.7208	80.14178746
	2	62	566572.5032	95.59427901
	3	47	1360950.463	170.1657804
2		178	2370689.687	210.991386328405
	1	69	443166.7208	80.14178746
	2	62	566572.5032	95.59427901
	3	47	1360950.463	170.1657804
3		178	2370689.687	210.991386328405
	1	69	443166.7208	80.14178746
	2	62	566572.5032	95.59427901
	3	47	1360950.463	170.1657804
4		178	2370689.687	210.991386328405
	1	69	443166.7208	80.14178746
	2	62	566572.5032	95.59427901
	3	47	1360950.463	170.1657804
5		178	2370689.687	210.991386328405
	1	69	443166.7208	80.14178746
	2	62	566572.5032	95.59427901
	3	47	1360950.463	170.1657804

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สำหรับข้อมูล Wine ในตารางที่ 4.5, 4.6 และ 4.7 ผลของ K-Means ที่ดีที่ที่สุดคือการรันครั้งที่ 1, ผลของ Fuzzy C-Means ที่ดีที่สุดคือการรันครั้งที่ 1 และผลของ Self Organizing Map ที่ดีที่สุดคือการรันครั้งที่ 1 ดังแสดงในตารางที่ 4.8

ตารางที่ 4.8 แสดงผลการจัดกลุ่มข้อมูล Wine ที่ดีที่สุด

รันครั้งที่	กลุ่มที่	จำนวนสมาชิกภายในกลุ่ม	ระทางรวม	ส่วนเบี่ยงเบนมาตรฐาน
K-Means				
1		178	2370689.687	210.991386328405
	1	62	566572.5032	95.59427901
	2	69	443166.7208	80.14178746
	3	47	1360950.463	170.1657804
Fuzzy C-Means				
1		178	2379535.426	211.878149352172
	1	61	588105.6985	98.18898627
	2	46	1308995.625	168.6903276
	3	71	482434.1023	82.43085897
Self Organizing Map				
1		178	2370689.687	210.991386328405
	1	69	443166.7208	80.14178746
	2	62	566572.5032	95.59427901
	3	47	1360950.463	170.1657804

4.4.3 ผลการจัดกลุ่มข้อมูล Segmentation

1. การจัดกลุ่มแบบ K-Means Algorithm

การจัดกลุ่มข้อมูลของข้อมูล Segmentation มีข้อมูล 2,310 ตัวอย่าง และกำหนดให้จำนวนกลุ่ม K เท่ากับ 7 เมื่อรันผลการจัดกลุ่มเสร็จ จะแสดงหน้าจอสรุปผลระยะทางรวมของการจัดกลุ่มข้อมูล และส่วนเบี่ยงเบนมาตรฐานในการรันแต่ละครั้งได้ดังตารางที่ตารางที่ 4.9

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.9 แสดงผลการจัดกลุ่มข้อมูล Segmentation แบบ K-Means Algorithm

รันครั้งที่	กลุ่มที่	จำนวนสมาชิกภายในกลุ่ม	ระทางรวม	ส่วนเบี่ยงเบนมาตรฐาน
1		2310	389.3056043	1.046015142
	1	327	46.18821276	0.375830224
	2	236	31.46226487	0.365122831
	3	383	55.83573291	0.381818286
	4	330	59.03857874	0.422971375
	5	157	24.1345214	0.392075332
	6	580	146.1647814	0.502004227
	7	297	26.48151221	0.298602312
2		2310	385.9393679	1.060711505
	1	403	77.11530925	0.437439281
	2	275	57.81276756	0.458506538
	3	330	59.03857874	0.422971375
	4	548	91.43789143	0.408481904
	5	327	46.18821276	0.375830224
	6	72	7.535041167	0.323501699
	7	355	46.81156694	0.363130237
3		2310	394.7205973	1.074114447
	1	177	16.75536764	0.307673681
	2	153	23.58580856	0.392626555
	3	366	67.47529263	0.429370147
	4	279	58.25627826	0.456950615
	5	558	96.36102953	0.415559884
	6	376	55.23354995	0.383272409
	7	401	77.05327071	0.438352364
4		2310	389.2794685	1.066558776
	1	563	97.34683588	0.415821291
	2	280	58.5770813	0.457388087
	3	192	31.08699268	0.402382182
	4	330	59.03857874	0.422971375

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.9 (ต่อ)

รันครั้งที่	กลุ่มที่	จำนวนสมาชิกภายในกลุ่ม	ระทางรวม	ส่วนเบี่ยงเบนมาตรฐาน
	5	398	76.45762625	0.438297375
	6	370	52.50374503	0.376698837
	7	297	41.67425123	0.374589568
5		2310	393.16458	1.094165552
	1	256	30.94889428	0.347698315
	2	364	66.31240059	0.426821889
	3	252	54.81243169	0.466379298
	4	489	81.47184209	0.40817776
	5	330	59.03857874	0.422971375
	6	335	43.00469359	0.358290692
	7	284	57.57573902	0.450257121

2. การจัดกลุ่มแบบ Fuzzy C-Means Algorithm

การจัดกลุ่มข้อมูลของข้อมูล Segmentation มีข้อมูล 2,310 ตัวอย่าง และกำหนดให้จำนวนกลุ่ม K เท่ากับ 7, Fuzziness Degree เท่ากับ 2 เมื่อรันผลการจัดกลุ่มเสร็จ จะแสดงหน้าจอสรุปผลระยะทางรวมของการจัดกลุ่มข้อมูล และส่วนเบี่ยงเบนมาตรฐานในการรันแต่ละครั้งได้ดังตารางที่ 4.10

ตารางที่ 4.10 แสดงผลการจัดกลุ่มข้อมูล Segmentation แบบ Fuzzy C-Means Algorithm

รันครั้งที่	กลุ่มที่	จำนวนสมาชิกภายในกลุ่ม	ระทางรวม	ส่วนเบี่ยงเบนมาตรฐาน
1		2310	373.6627879	1.064552599
	1	354	0.38277881	51.86794463
	2	294	0.452439981	60.18236926
	3	331	0.381689466	48.22234672
	4	352	0.418639192	61.69108801
	5	330	0.422971375	59.03857874
	6	337	0.429603267	62.19637181
	7	312	0.312476095	30.46408874

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.10 (ต่อ)

รันครั้งที่	กลุ่มที่	จำนวนสมาชิกภายในกลุ่ม	ระทางรวม	ส่วนเบี่ยงเบนมาตรฐาน
2		2310	395.8394721	1.067068857
	1	176	0.374200281	24.64454965
	2	358	0.439102585	69.02636657
	3	166	0.393925249	25.75939887
	4	538	0.484215437	126.1419489
	5	338	0.32710245	36.16465224
	6	403	0.404319755	65.88020914
	7	331	0.381689466	48.22234672
3		2310	373.6627879	1.064552599
	1	352	0.418639192	61.69108801
	2	354	0.38277881	51.86794463
	3	312	0.312476095	30.46408874
	4	337	0.429603267	62.19637181
	5	331	0.381689466	48.22234672
	6	330	0.422971375	59.03857874
	7	294	0.452439981	60.18236926
4		2310	373.6627879	1.064552599
	1	312	0.312476095	30.46408874
	2	337	0.429603267	62.19637181
	3	331	0.381689466	48.22234672
	4	294	0.452439981	60.18236926
	5	330	0.422971375	59.03857874
	6	352	0.418639192	61.69108801
	7	354	0.38277881	51.86794463
5		2310	373.6627879	1.064552599
	1	354	0.38277881	51.86794463
	2	294	0.452439981	60.18236926
	3	312	0.312476095	30.46408874
	4	331	0.381689466	48.22234672

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.10 (ต่อ)

รันครั้งที่	กลุ่มที่	จำนวนสมาชิกภายในกลุ่ม	ระทางรวม	ส่วนเบี่ยงเบนมาตรฐาน
	5	337	0.429603267	62.19637181
	6	352	0.418639192	61.69108801
	7	330	0.422971375	59.03857874

3. การจัดกลุ่มแบบ Self Organizing Map Algorithm

การจัดกลุ่มข้อมูลของข้อมูล Segmentation มีข้อมูล 2,310 ตัวอย่าง และกำหนดให้จำนวนกลุ่ม K เท่ากับ 7, Dimension เท่ากับ 19 เมื่อรันผลการจัดกลุ่มเสร็จ จะแสดงหน้าจอสรุปผลระยะทางรวมของการจัดกลุ่มข้อมูล และส่วนเบี่ยงเบนมาตรฐานในการรันแต่ละครั้งได้ดังตารางที่ 4.11

ตารางที่ 4.11 แสดงผลการจัดกลุ่มข้อมูล Segmentation แบบ Self Organizing Map

รันครั้งที่	กลุ่มที่	จำนวนสมาชิกภายในกลุ่ม	ระทางรวม	ส่วนเบี่ยงเบนมาตรฐาน
1		2310	371.9955616	1.064122542
	1	330	0.380267839	47.7191976
	2	339	0.41640219	58.7794756
	3	319	0.307143421	30.09352887
	4	367	0.389417332	55.65402995
	5	370	0.43132471	68.83517214
	6	255	0.451036187	51.87557871
	7	330	0.422971375	59.03857874
2		2310	373.4666176	1.059667566
	1	327	0.375830224	46.18821276
	2	345	0.418472959	60.41626811
	3	262	0.287867041	21.71126746
	4	413	0.390798825	63.07489699
	5	374	0.432397272	69.92580796
	6	259	0.452839968	53.11158554
	7	330	0.422971375	59.03857874

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.11 (ต่อ)

รันครั้งที่	กลุ่มที่	จำนวนสมาชิกภายในกลุ่ม	ระทางรวม	ส่วนเบี่ยงเบนมาตรฐาน
3		2310	398.2762465	1.113288718
	1	331	0.381689466	48.22234672
	2	547	0.408872314	91.44558311
	3	332	0.415767993	57.39052392
	4	306	0.440885103	59.48018036
	5	332	0.40302679	53.92695709
	6	132	0.46687287	28.77207656
	7	330	0.422971375	59.03857874
4		2310	381.2075534	1.090179922
	1	327	0.375830224	46.18821276
	2	352	0.432048676	65.70645262
	3	319	0.307680975	30.19895877
	4	384	0.394916269	59.88820201
	5	434	0.428154968	79.55943773
	6	164	0.497724767	40.62771078
	7	330	0.422971375	59.03857874
5		2310	371.9843491	1.063465914
	1	327	0.375830224	46.18821276
	2	344	0.419716972	60.59984369
	3	317	0.306407066	29.761637
	4	363	0.387720557	54.56878453
	5	371	0.432075668	69.26176116
	6	258	0.451378299	52.56553124
	7	330	0.422971375	59.03857874

สำหรับข้อมูล Segmentation ในตารางที่ 4.9, 4.10 และ 4.11 ผลของ K-Means ที่ดีที่ที่สุดคือการรันครั้งที่ 5, ผลของ Fuzzy C-Means ที่ดีที่สุดคือการรันครั้งที่ 1 และผลของ Self Organizing Map ที่ดีที่สุดคือการรันครั้งที่ 5 ดังแสดงในตารางที่ 4.12

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.12 แสดงผลการจัดกลุ่มข้อมูล Segmentation ที่ดีที่สุด

รันครั้งที่	กลุ่มที่	จำนวนสมาชิกภายในกลุ่ม	ระทางรวม	ส่วนเบี่ยงเบนมาตรฐาน
K-Means				
5		2310	393.16458	1.094165552
	1	256	30.94889428	0.347698315
	2	364	66.31240059	0.426821889
	3	252	54.81243169	0.466379298
	4	489	81.47184209	0.40817776
	5	330	59.03857874	0.422971375
	6	335	43.00469359	0.358290692
	7	284	57.57573902	0.450257121
Fuzzy C-Means				
1		2310	373.6627879	1.064552599
	1	354	0.38277881	51.86794463
	2	294	0.452439981	60.18236926
	3	331	0.381689466	48.22234672
	4	352	0.418639192	61.69108801
	5	330	0.422971375	59.03857874
	6	337	0.429603267	62.19637181
	7	312	0.312476095	30.46408874
Self Organizing Map				
5		2310	371.9843491	1.063465914
	1	327	0.375830224	46.18821276
	2	344	0.419716972	60.59984369
	3	317	0.306407066	29.761637
	4	363	0.387720557	54.56878453
	5	371	0.432075668	69.26176116
	6	258	0.451378299	52.56553124
	7	330	0.422971375	59.03857874

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.4.4 ผลการจัดกลุ่มข้อมูล Diabetes

1. การจัดกลุ่มแบบ K-Means Algorithm

การจัดกลุ่มข้อมูลของข้อมูล Diabetes ที่มีข้อมูล 768 ตัวอย่าง และกำหนดให้จำนวนกลุ่ม K เท่ากับ 2 เมื่อรันผลการจัดกลุ่มเสร็จ จะแสดงหน้าจอสรุปผลระยะทางรวมของการจัดกลุ่มข้อมูล และส่วนเบี่ยงเบนมาตรฐานในการรันแต่ละครั้งได้ดังตารางที่ 4.13

ตารางที่ 4.13 แสดงผลการจัดกลุ่มข้อมูล Diabetes แบบ K-Means Algorithm

รันครั้งที่	กลุ่มที่	จำนวนสมาชิกภายในกลุ่ม	ระทางรวม	ส่วนเบี่ยงเบนมาตรฐาน
1		768	116.5493096	0.562154968
	1	515	71.93626681	0.373740647
	2	253	44.61304277	0.419923966
2		768	116.5493096	0.562154968
	1	515	71.93626681	0.373740647
	2	253	44.61304277	0.419923966
3		768	116.5493096	0.562154968
	1	515	71.93626681	0.373740647
	2	253	44.61304277	0.419923966
4		768	116.5493096	0.562154968
	1	515	71.93626681	0.373740647
	2	253	44.61304277	0.419923966
5		768	116.5493096	0.562154968
	1	515	71.93626681	0.373740647
	2	253	44.61304277	0.419923966

2. การจัดกลุ่มแบบ Fuzzy C-Means Algorithm

การจัดกลุ่มข้อมูลของข้อมูล Diabetes ที่มีข้อมูล 768 ตัวอย่าง และกำหนดให้จำนวนกลุ่ม K เท่ากับ 2, Fuzziness Degree เท่ากับ 2 เมื่อรันผลการจัดกลุ่มเสร็จ จะแสดงหน้าจอสรุปผลระยะทางรวมของการจัดกลุ่มข้อมูล และส่วนเบี่ยงเบนมาตรฐานในการรันแต่ละครั้งได้ดังตารางที่ 4.14

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.14 แสดงผลการจัดกลุ่มข้อมูล Diabetes แบบ Fuzzy C-Means Algorithm

รันครั้งที่	กลุ่มที่	จำนวนสมาชิกภายในกลุ่ม	ระทางรวม	ส่วนเบี่ยงเบนมาตรฐาน
1		768	117.1632893	0.561230602
	1	460	60.97879309	0.364091463
	2	308	56.18449624	0.427103261
2		768	117.1632893	0.561230602
	1	460	60.97879309	0.364091463
	2	308	56.18449624	0.427103261
3		768	117.1632893	0.561230602
	1	460	60.97879309	0.364091463
	2	308	56.18449624	0.427103261
4		768	117.1632893	0.561230602
	1	460	60.97879309	0.364091463
	2	308	56.18449624	0.427103261
5		768	117.1632893	0.561230602
	1	460	60.97879309	0.364091463
	2	308	56.18449624	0.427103261

3. การจัดกลุ่มแบบ Self Organizing Map Algorithm

การจัดกลุ่มข้อมูลของข้อมูล Diabetes ที่มีข้อมูล 768 ตัวอย่าง และกำหนดให้จำนวนกลุ่ม K เท่ากับ 2, Dimension เท่ากับ 8 เมื่อรันผลการจัดกลุ่มเสร็จ จะแสดงหน้าจอสรุปผลระยะทางรวมของการจัดกลุ่มข้อมูล และส่วนเบี่ยงเบนมาตรฐานในการรันแต่ละครั้งได้ดังตารางที่ ตารางที่ 4.15

ตารางที่ 4.15 แสดงผลการจัดกลุ่มข้อมูล Diabetes แบบ Self Organizing Map

รันครั้งที่	กลุ่มที่	จำนวนสมาชิกภายในกลุ่ม	ระทางรวม	ส่วนเบี่ยงเบนมาตรฐาน
1		768	116.5604959	0.562028748
	1	258	45.59681464	0.420394862
	2	510	70.96368129	0.373020741
2		768	116.5540653	0.562049678
	1	255	44.97345359	0.419960099

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.15 (ต่อ)

รันครั้งที่	กลุ่มที่	จำนวนสมาชิกภายในกลุ่ม	ระทางรวม	ส่วนเบี่ยงเบนมาตรฐาน
	2	513	71.58061169	0.373541639
3		768	116.5493096	0.562154968
	1	253	44.61304277	0.419923966
	2	515	71.93626681	0.373740647
4		768	116.5493096	0.562154968
	1	253	44.61304277	0.419923966
	2	515	71.93626681	0.373740647
5		768	116.5493096	0.562154968
	1	253	44.61304277	0.419923966
	2	515	71.93626681	0.373740647

สำหรับข้อมูล Diabetes ในตารางที่ 4.13, 4.14 และ 4.15 ผลของ K-Means ที่ดีที่ที่สุดคือการรันครั้งที่ 1, ผลของ Fuzzy C-Means ที่ดีที่สุดคือการรันครั้งที่ 1 และผลของ Self Organizing Map ที่ดีที่สุดคือการรันครั้งที่ 1 ดังแสดงในตารางที่ 4.16

ตารางที่ 4.16 แสดงผลการจัดกลุ่มข้อมูล Diabetes ที่ดีที่สุด

รันครั้งที่	กลุ่มที่	จำนวนสมาชิกภายในกลุ่ม	ระทางรวม	ส่วนเบี่ยงเบนมาตรฐาน
K-Means				
1		768	116.5493096	0.562154968
	1	515	71.93626681	0.373740647
	2	253	44.61304277	0.419923966
Fuzzy C-Means				
1		768	117.1632893	0.561230602
	1	460	60.97879309	0.364091463
	2	308	56.18449624	0.427103261
Self Organizing Map				
1		768	116.5604959	0.562028748
	1	258	45.59681464	0.420394862
	2	510	70.96368129	0.373020741

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 5
สรุปผลการศึกษาและข้อเสนอแนะ

การจัดกลุ่มข้อมูล เป็นอัลกอลิทึมที่พยายามจัดข้อมูลที่คล้ายคลึงกันให้อยู่ในกลุ่มเดียวกัน โดยจะพยายามจัดกลุ่มข้อมูลให้ข้อมูลสมาชิกกับศูนย์กลางของกลุ่มนั้นมีระยะห่างกันน้อยที่สุด จากการศึกษา K-Means Algorithm, Fuzzy C-Means Algorithm และ Self Organizing Map Algorithm สามารถสรุปผลการดำเนินงาน และสรุปผลการทดลองรวมถึงข้อเสนอแนะ ดังต่อไปนี้

5.1 สรุปผลการดำเนินงาน

จากการศึกษา K-Means Algorithm, Fuzzy C-Means Algorithm และ Self Organizing Map Algorithm รวมถึงการออกแบบพัฒนาโปรแกรม สามารถสรุปผลการดำเนินงานได้ดังนี้

1. การจัดกลุ่มข้อมูลจะคำนวณระยะห่างจากศูนย์กลางของกลุ่มข้อมูล ที่เรียกว่า Euclidean distance ซึ่งเป็นเครื่องมือที่ใช้ในการจัดกลุ่ม
2. การค้นหาจุดศูนย์กลางของวิธี K-Means จะเป็นการคำนวณค่าเฉลี่ยจากข้อมูลทั้งหมด
3. วิธี Fuzzy C-Means มีความสามารถในการจำแนกข้อมูลที่มีความคลุมเครือออกเป็นกลุ่มๆ สามารถที่จะบ่งบอกถึงระดับความเป็นสมาชิก (Membership Grade) ของแต่ละกลุ่มได้
4. วิธี Self Organizing Map เป็นกระบวนการเรียนรู้แบบโครงข่ายนิวรอน แต่ละโหนดในชั้นเอาท์พุตจะแข่งขันกันเพื่อที่จะเป็นผู้ชนะ

5.2 สรุปผลการทดลอง

ในการทดลองนี้สามารถสรุปได้ว่า การจัดกลุ่มของข้อมูลสมาชิกกับศูนย์กลางของกลุ่มภายในแต่ละกลุ่ม สำหรับข้อมูลที่มีปริมาณข้อมูลขนาดไม่เกิน 1,000 ตัว ได้แก่ Iris, Wine และ Diabetes พบว่าการจัดกลุ่มของ K-Means Algorithm, Fuzzy C-Means Algorithm และ Self Organizing Map Algorithm ให้ผลใกล้เคียงกัน รวมถึงให้ผลระยะทางรวม และส่วนเบี่ยงเบนมาตรฐานที่ใกล้เคียงกันอีกด้วย (ดังสรุปในตารางที่ 4.4, ตารางที่ 4.8 และตารางที่ 4.16) ส่วนข้อมูล Segmentation ซึ่งเป็นข้อมูลที่มีจำนวนข้อมูลทั้งหมด 2,310 ตัวนั้นพบว่าวิธี Fuzzy C-Means Algorithm ให้ผลใกล้เคียงเป้าหมาย และเมื่อทำการทดสอบซ้ำหลายๆครั้งการแบ่งกลุ่มก็ยังคงมีข้อมูลภายในกลุ่มคงที่มากกว่าวิธี K-Means Algorithm และวิธี Self Organizing Map Algorithm (ดังสรุปในตารางที่ 4.12) ทั้งนี้อาจเกิดจากปัจจัยหลายอย่าง เช่น ในการสุ่มจุดศูนย์กลางเริ่มต้นเพื่อ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

มาจัดกลุ่มข้อมูล ถ้าจุดศูนย์กลางที่สุ่มขึ้นมาได้นั้นไม่ใช่ศูนย์กลางที่เป็นตัวแทนที่ดีจะทำให้ผลของการจัดกลุ่มไม่ดีนัก, ปริมาณข้อมูลที่นำมาใช้ทดสอบ หรือเกิดจากการเลือกวิธีที่ไม่เหมาะสมกับปัญหา จึงทำให้ผลการทดสอบทำได้เพียงใกล้เคียงเป้าหมาย

5.3 ข้อเสนอแนะ

ระบบที่พัฒนาขึ้นจะใช้เวลาในการประมวลผลแตกต่างกันในต่ละอัลกอริทึม ขึ้นอยู่กับจำนวนข้อมูลที่นำมาทดสอบ, ความเร็วของ CPU, ขนาดของหน่วยความจำ เป็นต้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บรรณานุกรม

กฤษณะ ไวยมัย และคณะ. 2547. **Data Mining การเตรียมข้อมูลสำหรับดาต้าไมนิง (2).** [Online]. Available: http://micro.se-ed.com/content/mc189/mainframe.asp?tar=mc189_91.asp

Ching-Yi Wu.et al. 2005. **DATA MINING.** [Online]. Available: http://userfs.cec.wustl.edu/~cse530/2004/DataMIning.ppt

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

Jiawei Han and Micheline Kamber. 2001. **Data Mining:Concepts and Techniques.** Morgan Kaufmann. San Mate.

Kardi Teknomo, PhD. 2005. **K-Mean Clustering Tutorial.** [Online]. Available: http://people.revoledu.com/kardi/tutorial/kMean/

Liyan Zhang. 2001. **Comparison of Fuzzy c-means Algorithm and New Fuzzy Clustering and Fuzzy Merging Algorithm.** [Online]. Available: http://www.cse.unr.edu/~lzhang/fuzzyCluster/paper/fuzzypaperMayNoK.doc

Mohanad Alata.et al. 2008. **Optimizing of Fuzzy C-Means Clustering Algorithm Using GA.** [Online]. Available: http://www.waset.org/pwaset/v29/v29-41.pdf

Tom Germano. 1999. **Self Organizing Maps.** [Online]. Available: http://davis.wpi.edu/~matt/courses/soms/

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

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ประวัติผู้เขียน

ชื่อผู้เขียน นางสาวสิรารมย์ สมบัติ

วันเดือนปีเกิด 2 ตุลาคม พ.ศ. 2524

สถานที่เกิด อุดรธานี

ประวัติการศึกษา

- สำเร็จการศึกษาระดับประถม จากโรงเรียนอนุบาลอุดรธานี
- สำเร็จการศึกษาระดับมัธยมต้น จากโรงเรียนสตรีราชินูทิศอุดรธานี
- สำเร็จการศึกษาระดับมัธยมปลาย จากโรงเรียนสตรีราชินูทิศอุดรธานี
- สำเร็จการศึกษาระดับปริญญาตรี คณะวิทยาศาสตร์ สาขาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยขอนแก่น
- กำลังศึกษาระดับปริญญาโท คณะเทคโนโลยีสารสนเทศ สาขาเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีเจ้าคุณทหารลาดกระบัง

ประวัติการทำงาน

- บริษัท ภัทรประกันภัย จำกัด (มหาชน) (ปี พ.ศ. 2547 - 2551)
  ตำแหน่ง เจ้าหน้าที่พัฒนาระบบ
- Hyro Consulting (Thailand) Company Limited. (ปี พ.ศ. 2551 – ปัจจุบัน)
  ตำแหน่ง .Net Developer

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้