การพัฒนาระบบดาต้าไมน์นิ่งโดยใช้ Clustering

Development of Data Mining system using Clustering

โดย

นุชชิดา พิริยะพงศธร

รหัสประจำตัว 47066111

อาจารย์ที่ปรึกษา

รศ.ดร.วรพจน์ กรีสุระเดช

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน

หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

ภาคฤดูร้อน ปีการศึกษา 2548

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ชื่อหัวข้อ	การพัฒนาระบบดาต้าไมน์นิ่งโดยใช้ Clustering
นักศึกษา	นางสาวนุชชิดา พิริยะพงศธร
อาจารย์ที่ปรึกษา	รศ.ดร.วรพจน์ กรีสุระเดช
ระดับการศึกษา	วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2548

บทคัดย่อ

ในการใช้ Data Mining เพื่อช่วยในการวิเคราะห์ข้อมูลที่มีปริมาณมากให้เกิดประโยชน์สูงสุดนั้นเป็นสิ่งที่จำเป็น โดยผลลัพธ์ที่ได้จากการวิเคราะห์นั้นจะถูกนำมาใช้ในการวางแผนกลยุทธ์ทางการตลาดและใช้ในการตัดสินใจเกี่ยวกับการทำธุรกิจต่อไป ซึ่ง Model ที่ใช้ในการทำ Data Mining นั้นมีมากมาย แต่ที่สนใจและนำมาใช้ในการพัฒนาระบบนี้คือ Model ที่ใช้ในการจัดกลุ่มข้อมูล (Clustering) ซึ่งเลือกใช้อัลกอริทึม K-means เพื่อทำการจัดกลุ่มข้อมูล โดยในการโครงงานจะทำการสร้างเครื่องมือที่ใช้ในการจัดกลุ่มข้อมูลและทำงานร่วมกับฐานข้อมูล Microsoft SQL Server ซึ่งใช้สำหรับดึงข้อมูลและจัดเก็บข้อมูล

Title	Development of Data Mining system using Clustering
Student	Miss Nootchida Piriyapongsaton
Advisor	Assoc.Prof.Dr.Worapoj Kreesuradej
Level of Study	Master of Science in Information Technology
Major	Information Science
Academic Year	2005

ABSTRACT

Data Mining necessaries to useful for analyze huge data. We use that result to do marketing strategy and make decision with business plan. There are many Data Mining model to use but we almost interesting and often using is Clustering Model which have K-means algorithm to do data grouping. Job scheming will create tools to do data grouping and run with data base Microsoft SQL server to transfer , collect and save all data.

กิตติกรรมประกาศ

ในโครงงานการพัฒนาเครื่องมือเพื่อช่วยในการจัดกลุ่มข้อมูล (Clustering) โดยใช้อัลกอริทึม K-means นี้ได้รับความช่วยเหลือและแรงสนับสนุนมาจากบุคคลที่สำคัญหลายๆ ท่าน ดังต่อไปนี้

- บิดามารดาผู้อบรมสั่งสอน และบุคคลในครอบครัวที่ให้ความสนับสนุนในด้านต่างๆ รวมถึงเรื่องการเรียนการศึกษา

- รศ.ดร.วรพจน์ กรีสุระเดช อาจารย์ที่ปรึกษาโครงการ ซึ่งให้ความกรุณาให้คำแนะนำและเป็นที่ปรึกษา อันเป็นประโยชน์ต่อการพัฒนาโครงงานนี้ รวมทั้งเป็นผู้ส่งเสริมให้โครงงานนี้สำเร็จลุล่วงไปได้ด้วยดี

- เพื่อนและรุ่นพี่ทุกคนที่คอยให้ความช่วยเหลือในด้านต่างๆ ไม่ว่าจะเป็นกำลังใจและให้คำปรึกษา

และขอขอบพระคุณสถาบันและคณาจารย์ทุกท่านที่คอยช่วยถ่ายทอดวิชาและให้คำปรึกษาที่ดีแก้ข้าพเจ้า จนสามารถพัฒนาโครงงานนี้สำเร็จ

นุชชิดา พิริยะพงศธร

สารบัญ

หน้า

บทคัดย่อภาษาไทย..... I

บทคัดย่อภาษาอังกฤษ..... II

กิตติกรรมประกาศ..... III

สารบัญ..... IV

สารบัญตาราง..... VI

สารบัญภาพ..... VII

บทที่

1. บทนำ

1.1 ความเป็นมาของโครงการ..... 1

1.2 วัตถุประสงค์ของโครงการ..... 1

1.3 ขอบเขตของโครงการ..... 2

1.4 ขั้นตอนการดำเนินงาน..... 2

1.5 ประโยชน์ที่คาดว่าจะได้รับ..... 2

2. ดาต้าไมน์นิ่ง

2.1 ดาต้าไมน์นิ่ง (Data Mining)..... 3

2.2 ประเภทข้อมูลที่สามารถทำดาต้าไมน์นิ่ง..... 3

2.3 ลักษณะเฉพาะของข้อมูลที่สามารถทำดาต้าไมน์นิ่ง..... 4

2.4 การเตรียมข้อมูลให้เหมาะสมสำหรับการทำดาต้าไมน์นิ่ง..... 4

2.5 เทคนิคต่างๆ ของดาต้าไมน์นิ่ง..... 6

2.6 การประยุกต์ใช้งานดาต้าไมน์นิ่ง..... 13

3. Customer Segmentation

3.1 ความหมายของ Customer Segmentation..... 14

3.2 Clustering..... 18

3.3 ความหมายของ K-means algorithm..... 19

3.4 ตัวอย่างการจัดกลุ่มของข้อมูล..... 22

3.5 จุดอ่อนของ K-means algorithm..... 27

สารบัญ (ต่อ)

หน้า

4. การวิเคราะห์และออกแบบระบบ

4.1 ความต้องการของระบบ..... 28

4.1.1 การสร้างโปรเจ็กต์..... 28

4.1.2 การทดสอบโมเดล.....29

4.2 องค์ประกอบของระบบงาน.....30

4.2.1 ส่วนนำข้อมูลเข้า..... 30

4.2.2 ส่วนวิเคราะห์และประมวลผล.....30

4.2.3 ส่วนของการแสดงผล.....32

4.3 ขั้นตอนการทำงานของระบบ..... 32

4.3.1 Use Case Diagram.....32

4.3.2 Sequence Diagram..... 32

4.3.3 Activity Diagram..... 40

4.3.4 Structure Chart.....42

4.4 Data Dcitionary.....46

5. การประยุกต์ใช้ดาต้าไมน์นิ่งเพื่อทำการจัดกลุ่มข้อมูล

5.1 กำหนดวัตถุประสงค์..... 48

5.2 การใช้งานระบบการจัดกลุ่มข้อมูล.....48

5.2.1 ส่วนของการสร้างโปรเจ็กต์.....49

5.2.2 ส่วนของการทดสอบโมเดล.....59

6. สรุปผลการศึกษาและข้อเสนอแนะ

6.1 สรุปผลการศึกษา.....63

6.2 ข้อเสนอแนะ..... 63

บรรณานุกรม.....64

ประวัติผู้เขียน.....65

สารบัญตาราง

หน้า

ตารางที่

2.1 Business_Info 9
2.2 ผลลัพธ์ของ SQL สำหรับ root node.....9
2.3 ผลลัพธ์ของ SQL สำหรับ node ที่เป็น child ทางขวาของ root..... 10
3.1 แสดงตัวอย่างของข้อมูล Medicine ที่จะใช้ในการจัดกลุ่ม 22
3.2 แสดงผลลัพธ์ที่ได้จากการจัดกลุ่ม 27
4.1 RuleName : เป็นตารางซึ่งเก็บชื่อโปรเจ็กต์ 46
4.2 FieldName : เก็บชื่อและชนิดของฟิลด์ที่ถูกใช้ในโปรเจ็กต์นั้นๆ 46
4.3 TbMinMax : เก็บค่าน้อยสุดและมากสุดที่ใช้ทำการแปลงข้อมูลของแต่ละฟิลด์ 47
4.4 TbDistinct: เก็บค่าที่เป็นไปได้ของแต่ละฟิลด์ 47
4.5 TbCentroid : เก็บค่าจุดศูนย์กลางที่ได้จากการจัดกลุ่มของโปรเจ็กต์นั้นๆ..... 47
5.1 แสดงตาราง Customer ที่นำมาใช้ในการจัดกลุ่มลูกค้า 48

สารบัญภาพ

หน้า

รูปที่

2.1 กระบวนการ Classification..... 7
2.2 ตัวอย่างของ Decision Tree เพื่อวิเคราะห์โอกาสที่ลูกค้าจะเช่าบ้าน.....8
2.3 นิวรอลเน็ตเพื่อวิเคราะห์การเช่าและซื้อบ้านของลูกค้า..... 11
2.4 ตัวอย่าง Clustering 12
3.1 แสดง Leaves ของ Decision tree ในการจัดกลุ่มข้อมูลของลูกค้า 16
3.2 แสดงถึงการนำข้อมูลไปจัดกลุ่ม..... 18
3.3 อธิบายการทำงานของ K-means algorithm 20
3.4 พิกัดของข้อมูล Medicine 23
3.5 พิกัดของ Centroids เริ่มต้น..... 23
3.6 พิกัดของ Centroids ที่ได้จากการคำนวณใน Iteration-1 25
3.7 พิกัดของ Centroids ที่ได้จากการคำนวณใน Iteration-2 26
4.1 แสดง Use Case Diagram ในการทำงานของระบบ.....32
4.2 แสดง Sequence Diagram ของการติดต่อฐานข้อมูล..... 33
4.3 แสดง Sequence Diagram ของการเลือกข้อมูลสำหรับทำไมน์นิ่ง.....34
4.4 แสดง Sequence Diagram ของการแก้ไขข้อมูล..... 35
4.5 แสดง Sequence Diagram ของการแปลงข้อมูลตัวอักษรเป็นตัวเลข..... 35
4.6 แสดง Sequence Diagram ของการแปลงข้อมูล..... 36
4.7 แสดง Sequence Diagram ของการจัดกลุ่มข้อมูล..... 37
4.8 แสดง Sequence Diagram ของการเลือกโมเดลที่ใช้ในการทดสอบ.....37
4.9 แสดง Sequence Diagram ของการทดสอบโมเดล.....39
4.10 แสดง Activity Diagram ของการสร้างโปรเจ็กต์..... 40
4.11 แสดง Activity Diagram ของการทดสอบโมเดล 41
4.12 แสดง Structure Chart ของระบบการจัดกลุ่มข้อมูล.....42
4.13 แสดง Structure Chart ของการสร้างโปรเจ็กต์.....42
4.14 แสดง Structure Chart ของการติดต่อฐานข้อมูล.....43
4.15 แสดง Structure Chart ของการจัดกลุ่มข้อมูล.....44

สารบัญภาพ (ต่อ)

หน้า

รูปที่

4.16 แสดง Structure Chart ของการทดสอบโมเดล.....45
4.17 แสดง Structure Chart ของการทำการทดสอบโมเดล.....46
5.1 แสดงหน้าจอหลักของการทำงาน.....49
5.2 แสดงหน้าจอสำหรับทำการติดต่อฐานข้อมูล.....49
5.3 แสดงตารางทั้งหมดที่มีอยู่ในฐานข้อมูลที่ผู้ใช้เลือก.....50
5.4 แสดงตารางทั้งหมดที่ผู้ใช้เลือก.....51
5.5 แสดงหน้าจอหลักเพื่อเข้าสู่ขั้นตอนการทำไมน์นิ่ง.....51
5.6 แสดงหน้าจอสำหรับเลือกตารางเพื่อใช้ทำไมน์นิ่ง.....52
5.7 แสดงหน้าจอสำหรับเลือกฟิลด์ในการจัดกลุ่มข้อมูล.....52
5.8 แสดงข้อมูลของฟิลด์ที่ถูกเลือกสำหรับทำไมน์นิ่ง.....53
5.9 แสดงหน้าจอสำหรับแก้ไขข้อมูล.....53
5.10 แสดงการแก้ไขข้อมูลในกรณีที่ข้อมูลเป็นตัวเลข.....54
5.11 แสดงการแก้ไขข้อมูลในกรณีที่ข้อมูลเป็นตัวอักษร.....54
5.12 แสดงข้อมูลที่ได้หลังจากการแก้ไข.....55
5.13 แสดงข้อมูลที่ต้องทำการแปลงให้อยู่ในรูปของตัวเลข.....56
5.14 แสดงข้อมูลหลังทำการแปลงข้อมูลแล้ว.....56
5.15 แสดงหน้าจอเพื่อทำการแปลงข้อมูลให้อยู่ในช่วงที่ต้องการ.....57
5.16 แสดงหน้าจอหลังการจัดกลุ่ม.....57
5.17 แสดงกราฟเปรียบเทียบจุดกึ่งกลางของแต่ละกลุ่ม.....58
5.18 แสดงหน้าจอการบันทึกโปรเจ็กต์.....58
5.19 หน้าจอหลักเพื่อเข้าสู่กระบวนการทดสอบโมเดล.....59
5.20 แสดงข้อมูลโมเดลเพื่อให้ผู้ใช้เลือก.....59
5.21 แสดงหน้าจอสำหรับการติดต่อกับฐานข้อมูล.....60
5.22 แสดงข้อมูลที่จะนำมาแม็พกัน.....60
5.23 แสดงหน้าจอการตรวจสอบการแม็พกันของฟิลด์ในตารางกับโมเดล.....61
5.24 แสดงผลลัพธ์ที่ได้หลังจากการทำการทดสอบโมเดล.....62

บทที่ 1
บทนำ

1.1 ความเป็นมาของโครงการ

ในการดำเนินธุรกิจในปัจจุบันมีการแข่งขันกันสูง โดยแต่ละองค์กรก็พยายามหากลยุทธ์ใหม่ๆ ทางการตลาดเพื่อนำมาใช้ปรับปรุงธุรกิจของตนเอง เพื่อให้สามารถแข่งขันกับองค์กรอื่นที่มีแนวทางการตลาดในแนวเดียวกันและสามารถลดความเสี่ยงลงได้ โดยในการวางแผนกลยุทธ์ทางการตลาดนั้นส่วนใหญ่จะได้มาจากการวิเคราะห์ข้อมูลในการดำเนินธุรกิจของบริษัท โดยนำเอาเทคโนโลยีทางด้านดาต้าไมน์นิ่งเข้ามาใช้เพื่อทำการวิเคราะห์ข้อมูลและศึกษาถึงความสัมพันธ์ของข้อมูลในฐานข้อมูล แล้วนำผลที่ได้จากการวิเคราะห์ข้อมูลนั้นไปกำหนดแผนการตลาด ซึ่งในส่วนของกระบวนการทางด้านดาต้าไมน์นิ่งนี้จะเป็นการนำเอาข้อมูลที่มีประโยชน์ที่ซ่อนเร้นอยู่ภายในข้อมูลที่เรามีอยู่มาใช้งาน เพื่อทำให้เกิดประสิทธิภาพสูงสุด และในหลายๆ หน่วยงานได้นำเอาเทคโนโลยีทางด้านดาต้าไมน์นิ่งไปใช้ในกระบวนการการจัดกลุ่มลูกค้า (Clustering) เพื่อหาสินค้าหรือบริการตอบสนองให้แก่ลูกค้าในแต่ละกลุ่มได้ตรงตามเป้าหมายมากที่สุด เช่น การจัดทำโปรโมชั่นให้แก่ลูกค้าในแต่ละกลุ่มเพื่อเพิ่มยอดขายให้สูงขึ้น เป็นต้น โดยอัลกอริทึมที่ศึกษาเพื่อจะใช้ในกระบวนการจัดกลุ่มลูกค้า (Clustering) นี้คือ K-Means Algorithm ซึ่งเป็นอัลกอริทึมหนึ่งที่นิยมใช้ในการจัดกลุ่มข้อมูล จากการจัดกลุ่มข้อมูลลูกค้านี้จะทำให้ช่วยลดเวลาในการทำงานลงและนอกจากนั้นยังช่วยสนับสนุนการตัดสินใจของผู้บริหารอีกด้วย โดยในปัจจุบันได้มีการพัฒนาเครื่องมือที่ช่วยสนับสนุนกระบวนการทางด้านดาต้าไมน์นิ่งมากมาย ไม่ว่าจะเป็น Microsoft SQL Server และอื่นๆ เพื่อที่จะช่วยอำนวยความสะดวกแก่องค์กรที่ต้องการใช้ความสามารถทางด้านดาต้าไมน์นิ่งในการวิเคราะห์ข้อมูลที่ตนมีอยู่ให้เกิดประโยชน์ และใช้ในการวางแผนทางการตลาดขององค์กรต่อไป ดังนั้นจึงได้มีแนวความคิดในการพัฒนาแอพพลิเคชั่นในการจัดกลุ่มข้อมูล เพื่ออำนวยความสะดวกแก่องค์กรที่ต้องการจัดกลุ่มลูกค้าให้มีประสิทธิภาพมากขึ้น

1.2 วัตถุประสงค์ของโครงการ

1.2.1 เพื่อศึกษาวิธีการและขั้นตอนการทำงานของดาต้าไมน์นิ่ง

1.2.2 เพื่อศึกษากระบวนการในการจัดกลุ่มข้อมูลของ K-Means Algorithm

1.2.3 เพื่อนำข้อมูลที่มีอยู่มาใช้ให้เกิดประโยชน์มากที่สุด

1.2.4 เพื่อศึกษาวิธีการสร้างดาต้าไมน์นิ่งโดยใช้ Microsoft SQL Server

1.2.5 เพื่อนำเอาความรู้ในกระบวนการจัดกลุ่มไปประยุกต์ใช้ในการพัฒนาแอพพลิเคชั่น

1.2.6 เพื่อนำข้อมูลที่ได้จากการจัดกลุ่มไปใช้ให้เกิดประโยชน์ต่อองค์กร

1.3 ขอบเขตของโครงการ

โครงการนี้จะเป็นการศึกษาและพัฒนาระบบโดยใช้เทคนิคทางด้านดาต้าไมน์นิ่งเพื่อใช้ในการจัดกลุ่มข้อมูล (Clustering) โดยจะทำการพัฒนาแอพพลิเคชั่นเพื่อเรียกใช้ฟังก์ชั่นของการจัดกลุ่มข้อมูลที่มีใน Microsoft SQL Server ซึ่งเรียกว่า Microsoft Clustering ให้ทำการจัดกลุ่มข้อมูลตามที่ผู้ใช้ต้องการ โดยโปรแกรมที่พัฒนาขึ้นนี้จะช่วยให้ผู้ใช้สามารถนำข้อมูลที่มีอยู่จำนวนมากมาวิเคราะห์หากลุ่มของข้อมูลได้อย่างมีประสิทธิภาพมากขึ้น และนำเอาผลลัพธ์ที่ได้จากการจัดกลุ่มมาแสดงในรูปแบบที่ผู้ใช้สามารถเข้าใจได้ง่าย

1.4 ขั้นตอนการดำเนินงาน

1.4.1 กำหนดหัวข้อ เป้าหมาย และวัตถุประสงค์ ตลอดจนของเขตของโครงการ

1.4.2 ศึกษาทฤษฎี บทความ งานวิจัย และหนังสือที่เกี่ยวข้องกับกระบวนการและเทคนิคทางด้านดาต้าไมน์นิ่ง รวมถึงเทคนิคของการจัดกลุ่มข้อมูล และ K-Means Algorithm

1.4.3 ศึกษาเทคนิคต่างๆ ในการพัฒนาโปรแกรม

1.4.4 ศึกษาเทคนิคการใช้ Analysis Service ใน Microsoft SQL Server

1.4.5 ทำการออกแบบและพัฒนาโปรแกรม

1.4.6 ทำการทดสอบโปรแกรม เพื่อหาข้อผิดพลาดที่เกิดขึ้นและทำการปรับปรุงแก้ไขข้อผิดพลาดที่เกิดขึ้นนั้น

1.4.7 จัดทำเอกสาร

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1.5.1 เพื่อได้เรียนรู้ทฤษฎี เทคนิคและวิธีการของการจัดกลุ่มข้อมูล

1.5.2 เพื่อนำเอาความรู้ที่มีไปประยุกต์ใช้ในการพัฒนาโปรแกรมเพื่อทำการจัดกลุ่มข้อมูลต่างๆ

1.5.3 เพื่อนำเอาข้อมูลที่มีอยู่มาวิเคราะห์และนำไปใช้ให้เกิดประโยชน์มากขึ้น

1.5.4 เพื่อได้เรียนรู้ถึงการทำงานของ Analysis Service ใน Microsoft SQL Server

1.5.5 เพื่อให้การทำงานในองค์กรเร็วขึ้น

บทที่ 2
ดาต้าไมน์นิ่ง

2.1 ดาต้าไมน์นิ่ง (Data Mining)

ดาต้าไมน์นิ่งหรือเรียกอีกอย่างว่าการค้นหาความรู้ในฐานข้อมูล (Knowledge Discovery in Databases – KDD) เป็นเทคนิคเพื่อค้นหารูปแบบ (pattern) บางอย่างจากข้อมูลที่ซ่อนอยู่ในฐานข้อมูลขนาดใหญ่และหาความสัมพันธ์ของข้อมูลที่มีอยู่ให้โดยอัตโนมัติ เพื่อนำข้อมูลที่มีอยู่มาใช้ให้เกิดประโยชน์สูงสุด เช่น การค้นหากฎความสัมพันธ์ (association rules) ของสินค้าในห้างสรรพสินค้า เราอาจจะพบว่าลูกค้า 80 เปอร์เซ็นต์ที่ซื้อเบียร์ จะซื้อผ้าอ้อมเด็กด้วย ซึ่งเป็นข้อมูลให้ทางห้างคิดรายการส่งเสริมการขายใหม่ๆ ได้ หรือธนาคารพบว่า คนทั่วไปที่มีอายุ 20-29 ปี และมีรายได้ในช่วง 20,000-30,000 บาท มักซื้อเครื่องเล่นเอ็มพีสาม ดังนั้นทางธนาคารจึงอาจจะเสนอให้คนกลุ่มนี้ทำบัตรเครดิต โดยแถมเครื่องเล่นดังกล่าว เป็นต้น ซึ่งดาต้าไมน์นิ่งนี้จะอาศัยหลักการทางสถิติ การรู้จำแบบ การเรียนรู้ของเครื่องมาใช้ในการวิเคราะห์ข้อมูลต่างๆ โดยปัจจัยที่สำคัญที่ทำให้ดาต้าไมน์นิ่งได้รับความสนใจก็คือ เป็นการนำเอาข้อมูลที่มีอยู่จำนวนมากมาวิเคราะห์หาความสัมพันธ์และนำไปใช้ให้เกิดประโยชน์แก่องค์กรในการวางแผนกลยุทธ์ทางการตลาดที่จะสามารถแข่งขันกับคู่แข่งได้

2.2 ประเภทข้อมูลที่สามารถทำดาต้าไมน์นิ่ง

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

2. Data Warehouses เป็นการเก็บรวบรวมข้อมูลจากหลายแหล่งมาเก็บไว้ในรูปแบบเดียวกันและรวบรวมไว้ในที่ ๆ เดียวกัน

3. Transactional Database ประกอบด้วยข้อมูลที่แต่ละทรานเซกชันแทนด้วยเหตุการณ์ในขณะใดขณะหนึ่ง เช่น ใบเสร็จรับเงิน จะเก็บข้อมูลในรูป ชื่อลูกค้าและรายการสินค้าที่ลูกค้ารายนั้นซื้อ เป็นต้น

4. Advanced Database เป็นฐานข้อมูลที่จัดเก็บในรูปแบบอื่น ๆ 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 จึงทำให้ผลลัพธ์ที่ได้จาการทำไมน์นิ่งสมเหตุสมผลในช่วงเวลาหนึ่งเท่านั้น ดังนั้นเพื่อให้ได้ผลลัพธ์ที่มีความถูกต้องเหมาะสมอยู่ตลอดเวลาจึงต้องทำไมน์นิ่งใหม่ทุกครั้งในช่วงเวลาที่เหมาะสม

4. ข้อมูลที่มีโครงสร้างซับซ้อน เช่น ข้อมูลรูปภาพ ข้อมูลมัลติมีเดีย ข้อมูลเหล่านี้สามารถนำมาทำไมน์นิ่งได้เช่นกันแต่ต้องใช้เทคนิคการทำดาต้าไมน์นิ่งขั้นสูง

2.4 การเตรียมข้อมูลให้เหมาะสมสำหรับการทำดาต้าไมน์นิ่ง

1. การเลือกข้อมูล (Data selection)

การเลือกข้อมูลเฉพาะที่ต้องการเพื่อที่จะนำมาวิเคราะห์ให้ตรงกับจุดประสงค์ในการทำดาต้าไมน์นิ่งเพื่อทำการแยกข้อมูลที่ไม่ต้องการออกไป ซึ่งจะเป็นการเริ่มต้นของการเตรียมการไมน์นิ่ง การเลือกข้อมูลนั้นจะแตกต่างไปตามวัตถุประสงค์ของแต่ละธุรกิจที่ได้กำหนดเอาไว้ในตอนต้น การเลือกข้อมูลนั้นจำเป็นจะต้องมีความเข้าใจกับชนิดของข้อมูล และประเภทของข้อมูลที่จะต้องนำมาใช้ด้วย เช่นการทำดาต้าไมน์นิ่งประเภทวิเคราะห์ความสัมพันธ์ของข้อมูล (Link analysis) นั้นต้องใช้ข้อมูลประเภททรานเซกชัน เป็นต้น

2. การเตรียมข้อมูล (Data Preprocessing)

ทำการตรวจสอบข้อมูลและแก้ไขเพื่อทำให้ได้ข้อมูลที่มีคุณภาพที่ดี และทำให้ข้อมูลที่ถูกเลือกนั้นถูกต้อง ครบถ้วนตามที่จะต้องนำไปใช้ในการทำดาต้าไมน์นิ่ง เนื่องจากข้อมูลที่ถูกเลือกมาจากกระบวนการเลือกข้อมูลนั้นอาจจะมีบางข้อมูลที่มีจุดบกพร่อง ขาดหายไป หรือข้อมูลที่เก็บไว้นั้นมันล้าสมัย ดังนั้นในขั้นตอนนี้จะต้องทำการพิจารณาเพิ่มเติม 3 ประเด็น ดังนี้

- **การกำจัดค่าของข้อมูลที่ผิดพลาด (Noisy Data)**

ข้อมูลที่มีลักษณะแตกต่างจากข้อมูลที่คาดการณ์เอาไว้ หรือค่าของข้อมูลอาจจะผิดไปจากที่ควรจะเป็นซึ่งอาจจะเกิดจากการป้อนข้อมูลผิด เช่น ข้อมูลอายุเป็น 490 ปี หรือป้อนน้ำหนักเป็น

ข้อมูลติดลบ เป็นต้น ซึ่งข้อมูลที่ผิดนี้อาจเป็นเหตุให้การวิเคราะห์ผิดพลาดได้ ดังนั้นจึงต้องทำการกำจัดข้อมูลที่ผิดพลาดนี้ออกไป

- **การจัดการกับค่าของข้อมูลที่สูญหาย (Missing Value)**

ค่าของข้อมูลที่ไม่ได้ถูกเลือกมาจากขั้นตอนเลือกข้อมูลหรืออาจจะเป็นค่าที่ไม่สมบูรณ์ ที่เราทำการลบออกไประหว่างการกำจัดค่าของข้อมูลที่ผิดพลาด ทำให้ค่าบางค่าอาจสูญหายไปจากความผิดพลาดในการเก็บข้อมูล หรือเกิดจากความผิดพลาดในการบันทึกข้อมูล ซึ่งสามารถแก้ไขได้โดยทำการลบข้อมูลนั้นทิ้งทั้งรายการ หรือทำการบันทึกแทนที่ในค่าที่ขาดหายไปด้วยค่าเฉลี่ย ค่าที่ปรากฏบ่อยๆ หรืออาจบันทึกเป็น "Unknown" เป็นต้น

- **การจัดการข้อมูลที่ไม่ถูกต้อง (Inconsistence Data)**

เนื่องจากข้อมูลที่เลือกมาอาจจะนำมาจากหลายๆ แหล่งมารวมกัน จึงอาจมีข้อมูลที่ไม่ถูกต้องตรงกัน เช่น ชื่อของลูกค้าใน ID เดียวกันอาจจะไม่ตรงกัน สามารถแก้ไขได้โดยพิจารณาว่าลูกค้าเป็นคนเดียวกันหรือไม่ ถ้าใช่ก็ทำการนำข้อมูลที่ได้มีการบันทึกล่าสุดมาแทนค่าในข้อมูลที่เก่ากว่า หรือถ้าไม่ใช่ลูกค้าคนเดียวกันก็ต้องทำการเพิ่ม ID ให้เป็นลูกค้าอีกคนหนึ่ง

3. การแปลงข้อมูล (Data Transformation)

การแปลงข้อมูลจะเป็นการนำข้อมูลที่มีอยู่มาเปลี่ยนแปลงทำให้อยู่ในรูปแบบของข้อมูลที่พร้อมจะนำไปวิเคราะห์กับอัลกอริทึมที่ใช้กับเทคนิคต่างๆ ของดาต้าไมน์นิ่ง เช่น การปรับอัตราส่วนตัวเลขให้อยู่ในช่วง 0-1 เพื่อใช้กับนิวรอลเน็ตเวิร์ก (Neural Network) หรือการแปลงตัวเลขให้เป็นช่วงๆ เพื่อใช้กับอัลกอริทึมของดิซิชั่นทรี (Decision Tree) เป็นต้น โดยสูตรในการคำนวณสำหรับการแปลงข้อมูลจะเป็นไปดังสมการข้างล่างนี้

$$v' = \frac{v - \min_A}{\max_A - \min_A}\left(new\_\max_A - new\_\min_A\right) + new\_\min_A$$

v' = ค่าข้อมูลใหม่ที่ได้หลังจากการแปลง

v = ข้อมูลเดิมที่จะนำมาทำการแปลง

$\min_A$ = ค่าต่ำสุดของข้อมูลเดิมในแอตทริบิวต์ A

$\max_A$ = ค่าสูงสุดของข้อมูลเดิมในแอตทริบิวต์ A

$new\_\min_A$ = ค่าต่ำสุดของข้อมูลใหม่ที่ต้องการทำการแปลงข้อมูลแอตทริบิวต์ A

$new\_\max_A$ = ค่าสูงสุดของข้อมูลใหม่ที่ต้องการทำการแปลงข้อมูลแอตทริบิวต์ A

4. การทำดาต้าไมน์นิ่ง (Data Mining)

คือการประมวลผลข้อมูลตามอัลกอริทึมที่ได้กำหนดเอาไว้

5.. การวิเคราะห์ผลลัพธ์ (Interpretation Evaluation)

หลังจากกระบวนการทำไมน์นิ่งแล้ว ต้องมีการวิเคราะห์ผลลัพธ์ที่ได้ของการประมวลผล ซึ่งจะทำการตีความและประเมินผลลัพธ์ที่ได้จากขั้นตอนการทำดาต้าไมน์นิ่งว่าสามารถนำไปใช้ได้ตามวัตถุประสงค์ที่ต้องการหรือไม่ รวมทั้งเป็นการวิเคราะห์ถึงความถูกต้องของผลที่ได้จากการทำไมน์นิ่ง พิจารณาความเป็นไปได้ของการเกิดผลลัพธ์ดังกล่าว เพราะบางครั้งผลที่ได้จากการทำไมน์นิ่งอาจจะนำไปใช้ประโยชน์ไม่ได้ ดังนั้นจะต้องทำการวิเคราะห์ผลที่ได้ก่อนนำไปใช้งานก่อนเสมอ

2.5 เทคนิคต่างๆ ของดาต้าไมน์นิ่ง

1. Association rule Discovery

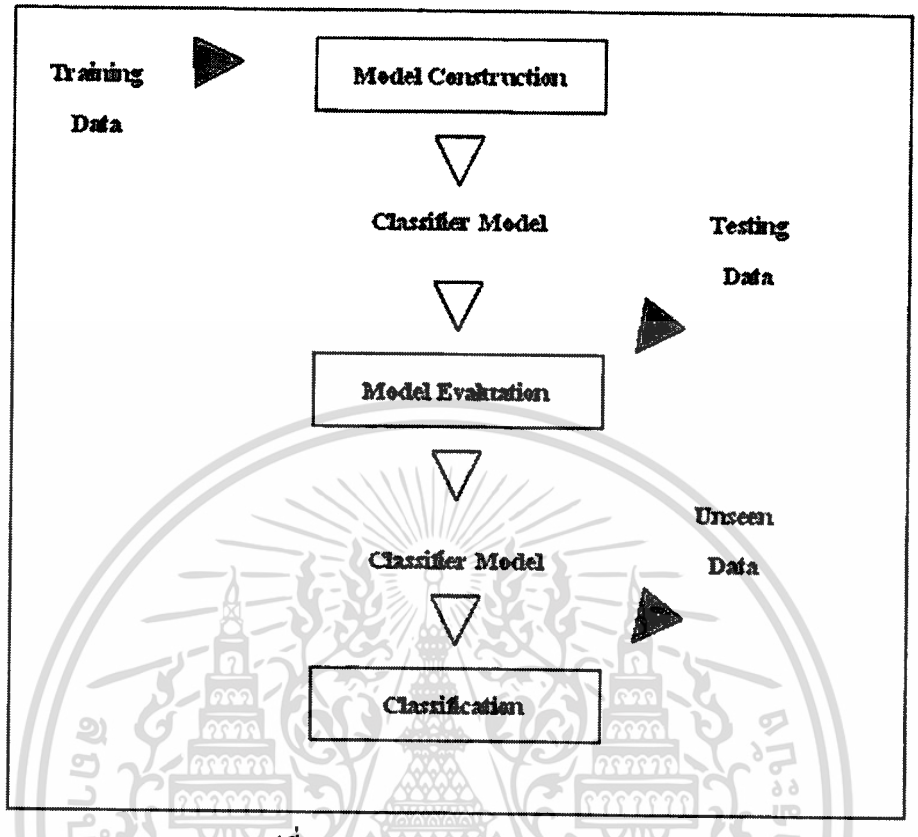
การค้นหาความสัมพันธ์ของข้อมูลจากข้อมูลขนาดใหญ่ที่มีอยู่เพื่อนำไปใช้ในการวิเคราะห์หรือทำนายปรากฏการณ์ต่าง ๆ หรือมากจากการวิเคราะห์การซื้อสินค้าของลูกค้าเรียกว่า " Market Basket Analysis " ซึ่งประเมินจากข้อมูลในตารางที่รวบรวมไว้ ผลการวิเคราะห์ที่ได้จะเป็นคำตอบของปัญหา ซึ่งการวิเคราะห์แบบนี้เป็นการใช้ " กฎความสัมพันธ์ " (Association Rule) เพื่อหาความสัมพันธ์ของข้อมูล

ตัวอย่างการนำเทคนิคนี้ไปประยุกต์ใช้กับงานจริง ได้แก่ ระบบแนะนำหนังสือให้กับลูกค้าแบบอัตโนมัติ ของ Amazon ข้อมูลการสั่งซื้อทั้งหมดของ Amazon ซึ่งมีขนาดใหญ่มากจะถูกนำมาประมวลผลเพื่อหาความสัมพันธ์ของข้อมูล คือ ลูกค้าที่ซื้อหนังสือเล่มหนึ่ง ๆ มักจะซื้อหนังสือเล่มใดพร้อมกันด้วยเสมอ ความสัมพันธ์ที่ได้จากกระบวนการนี้จะสามารถนำไปใช้คาดเดาได้ว่าควรแนะนำหนังสือเล่มใดเพิ่มเติมให้กับลูกค้าที่เพิ่งซื้อหนังสือจากร้าน ตัวอย่างเช่น buys (x , database) -> buys (x , data mining) [80% , 60%] หมายความว่า เมื่อซื้อหนังสือ database แล้วมีโอกาสที่จะซื้อหนังสือ data mining ด้วย 60 % และมีการซื้อทั้งหนังสือ database และหนังสือ data mining พร้อม ๆ กัน 80 %

2. Classification & Prediction

2.1 Classification

เป็นกระบวนการสร้าง Model จัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ ตัวอย่างเช่น จัดกลุ่มนักเรียนว่า ดีมาก ดี ปานกลาง ไม่ดี โดยพิจารณาจากประวัติและผลการเรียน หรือแบ่งประเภทของลูกค้าว่าเชื่อถือได้ หรือไม่โดยพิจารณาจากข้อมูลที่มีอยู่ กระบวนการ Classification นี้แบ่งออกเป็น 3 ขั้นตอน ดังรูปที่ 2.1



รูปที่ 2.1 กระบวนการ Classification

➢ **Model Construction (Learning)**

เป็นขั้นการสร้าง Model โดยการเรียนรู้จากข้อมูลที่ได้กำหนดคลาสไว้เรียบร้อยแล้ว (Training data) ซึ่ง Model ที่ได้อาจแสดงในรูปของ

1. แบบต้นไม้ (Decision Tree)

2. แบบนิวรอลเน็ต (Neural Net)

1. โครงสร้างแบบต้นไม้ของ Decision Tree

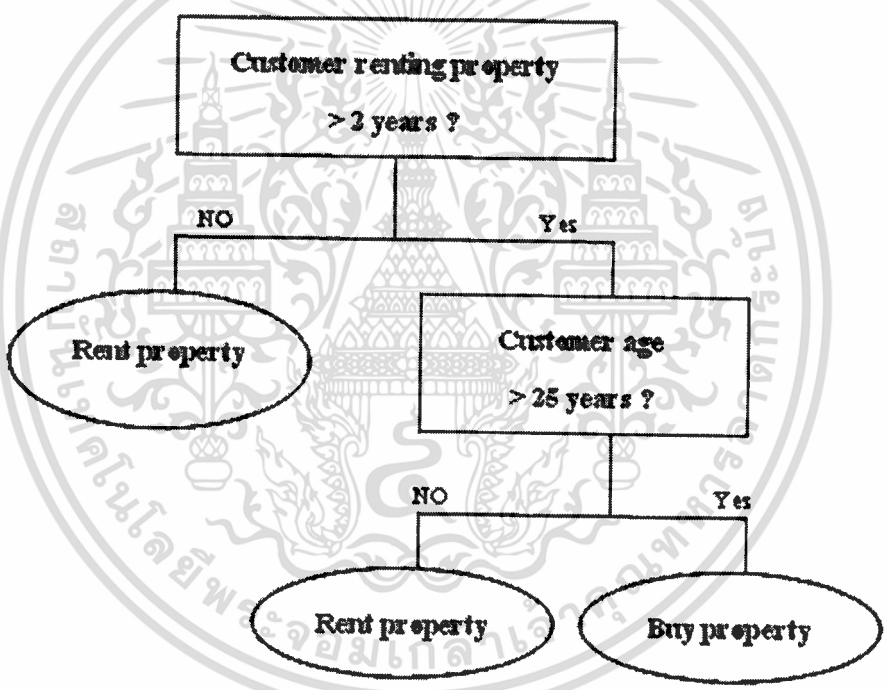
เป็นที่นิยมกันมากเนื่องจากเป็นลักษณะที่คนจำนวนมากคุ้นเคย ทำให้เข้าใจได้ง่าย มีลักษณะเหมือนแผนภูมิองค์กร โดยที่แต่ละโหนดแสดง attribute แต่ละกิ่งแสดงผลในการทดสอบและลีฟโหนดแสดงคลาสที่กำหนดไว้

สมมติว่าบริษัทขนาดใหญ่แห่งหนึ่ง ทำธุรกิจอสังหาริมทรัพย์มีสำนักงานสาขาอยู่ประมาณ 50 แห่ง แต่ละสาขามีพนักงานประจำ เป็นผู้จัดการและพนักงานขาย พนักงานเหล่านี้แต่ละคนจะดูแลอาคารต่าง ๆ หลายแห่งรวมทั้งลูกค้าจำนวนมาก บริษัทจำเป็นต้องใช้ระบบฐานข้อมูลที่กำหนดความสัมพันธ์ระหว่างองค์ประกอบเหล่านี้ เมื่อรวบรวมข้อมูลแบ่งเป็นตารางพื้นฐานต่าง ๆ เช่น ข้อมูลสำนักงานสาขา (Branch) ข้อมูลพนักงาน (Staff) ข้อมูลทรัพย์สิน (Propety) และข้อมูลลูกค้า

(Client) พร้อมทั้งกำหนดความสัมพันธ์ (Relationship) ของข้อมูลเหล่านี้ เช่น ประวัติการเช่าบ้านของลูกค้า (Customer_rental) รายการให้เช่า (Rentals) รายการขายสินทรัพย์ (Sales) เป็นต้น ต่อมาเมื่อมีประชุมกรรมการผู้บริหารของบริษัท ส่วนหนึ่งของรายงานจากฐานข้อมูลสรุปว่า

" 40 % ของลูกค้าที่เช่าบ้านนานกว่าสองปี และมีอายุเกิน 25 ปี จะซื้อบ้านเป็นของตนเอง โดยกรณีเช่นนี้เกิดขึ้น 35 % ของลูกค้าผู้เช่าบ้านของบริษัท"

ดังรูปที่ 2.2 แสดงให้เห็นถึง Decision Tree สำหรับการวิเคราะห์ว่าลูกค้าบ้านเช่าจะมีความสนใจที่จะซื้อบ้านเป็นของตนเองหรือไม่ โดยใช้ปัจจัยในการวิเคราะห์คือ ระยะเวลาที่ลูกค้าได้เช่าบ้านมา และอายุของลูกค้า



รูปที่ 2.2 ตัวอย่างของ Decision Tree เพื่อวิเคราะห์โอกาสที่ลูกค้าจะเช่าบ้าน

จากตัวอย่างพฤติกรรมการเช่าและซื้อบ้านข้างต้น ลองมาตัวอย่างที่เป็นรูปธรรมมากขึ้น ในตาราง Business_Info แสดงถึงรายการทั้งหมดเกี่ยวกับลูกค้าบ้านเช่าของบริษัทโดยมีรายละเอียดเกี่ยวกับอายุ และระยะเวลาเช่า รวมทั้งการซื้อบ้านของลูกค้าแต่ละราย ดังนี้

ตารางที่ 2.1 Business_Info

Age	Rent_Period	Buy
23	3	No
36	1.5	No
20	1.5	No
27	2	Yes
20	1	No
50	2.5	Yes
36	1	No
36	2	Yes
22	2.5	no

SQL สำหรับ Decision Tree ของตัวอย่างนี้แบ่งเป็น 2 ชุด สำหรับปัจจัยแต่ละอย่าง

1. SQL สำหรับ root node ดังนี้

```
SELECT  B.Rent_Period , B.Buy , COUNT(*)
FROM    Business_Info B
WHERE   B.Rent_Period > 2
GROUP BY  B.Rent_Period , B.Buy
```

ผลลัพธ์ของ SQL นี้คือ

ตารางที่ 2.2 ผลลัพธ์ของ SQL สำหรับ root node

Rent_Period	Buy	Yes	No
1	0	2	
1.5	0	2	
2	2	0	
2.5	1	1	
3	0	1	

2. SQL สำหรับ node ที่เป็น child ทางขวาของ root คือ

```
SELECT   B.Age , B.Buy , COUNT(*)
FROM     Business_Info B
WHERE    B.Age > 25
GROUP BY B.Age , B.Buy
```

ผลลัพธ์ของ SQL นี้คือ

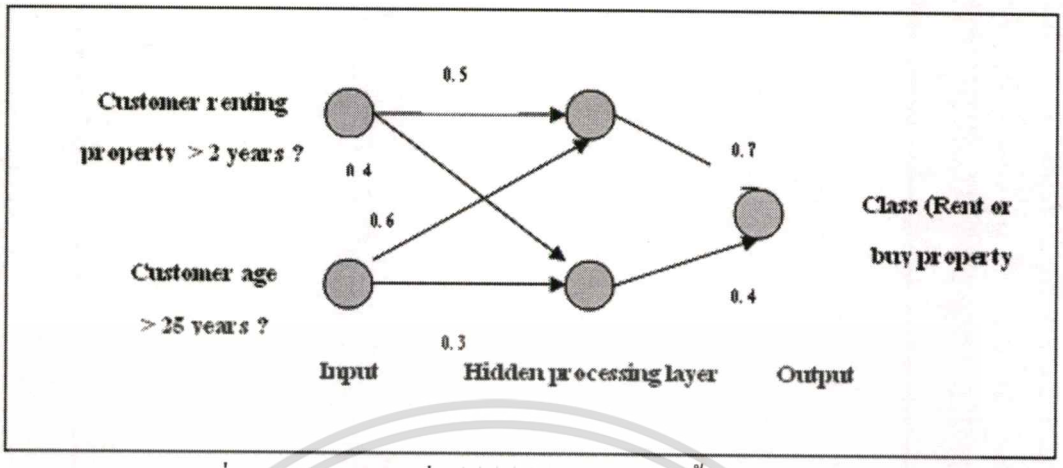
ตารางที่ 2.3 ผลลัพธ์ของ SQL สำหรับ node ที่เป็น child ทางขวาของ root

Rent_Period	Buy	Yes	No
20	0	2	
22	0	1	
23	2	1	
27	1	0	
36	1	2	
50	1	0	

ผลลัพธ์ที่ได้จากแต่ละโหนดของ Decision Tree เรียกว่า AVC set (Attribute Value, Class label) จากตัวอย่างข้างต้นจะเห็นว่ามี 2 AVC sets เพื่อใช้ในการจัดกลุ่มลูกค้า

2. นิวรอลเน็ต หรือ นิวรอลเน็ตเวิร์ก (Neural Net)

เป็นเทคโนโลยีที่มีที่มาจากงานวิจัยด้านปัญญาประดิษฐ์ Artificial Intelligence:AI เพื่อใช้ในการคำนวณค่าฟังก์ชันจากกลุ่มข้อมูล วิธีการของ นิวรอลเน็ต (แท้จริงต้องเรียกให้เต็มว่า Artificial Neural Networks หรือ ANN) เป็นวิธีการที่ให้เครื่องเรียนรู้จากตัวอย่างต้นแบบ แล้วฝึก (Train) ให้ระบบได้รู้จักที่จะคิดแก้ปัญหาที่กว้างขึ้นได้ ในโครงสร้างของนิวรอลเน็ตจะประกอบด้วยโหนด (node) สำหรับ Input – Output และการประมวลผล กระจายอยู่ในโครงสร้างเป็นชั้น ๆ ได้แก่ Input layer, Output layer และ Hidden layers การประมวลผลของนิวรอลเน็ตจะอาศัยการส่งการทำงานผ่านโหนดต่าง ๆ ใน layer เหล่านี้ สำหรับตัวอย่างรูปที่ 2.3 เป็นการวิเคราะห์แบบเดียวกับรูปที่ 2.2 ในโครงสร้างแบบนิวรอลเน็ต



รูปที่ 2.3 นิวรอลเน็ตเพื่อวิเคราะห์การเช่าและซื้อบ้านของลูกค้า

➢ **Model Evaluation (Accuracy)**

เป็นขั้นการประมาณความถูกต้องโดยอาศัยข้อมูลที่ใช้ทดสอบ (Testing data) ซึ่งคลาสที่แท้จริงของข้อมูลที่ใช้ทดสอบนี้จะถูกนำมาเปรียบเทียบกับคลาสที่หามาได้จาก Model เพื่อทดสอบความถูกต้อง

➢ **Model Usage (Classification)**

เป็น Model สำหรับใช้ข้อมูลที่ไม่เคยเห็นมาก่อน (Unseen data) โดยจะทำการกำหนดคลาสให้กับ Object ใหม่ที่ได้มา หรือทำนายค่าออกมาตามที่ต้องการ

2.2 Prediction

เป็นการทำนายหาค่าที่ต้องการจากข้อมูลที่มีอยู่ ตัวอย่างเช่น หายอดขายของเดือนถัดไปจากข้อมูลที่มีอยู่ หรือทำนายโรคจากอาการของคนไข้ในอดีต เป็นต้น

3. Database clustering หรือ Segmentation

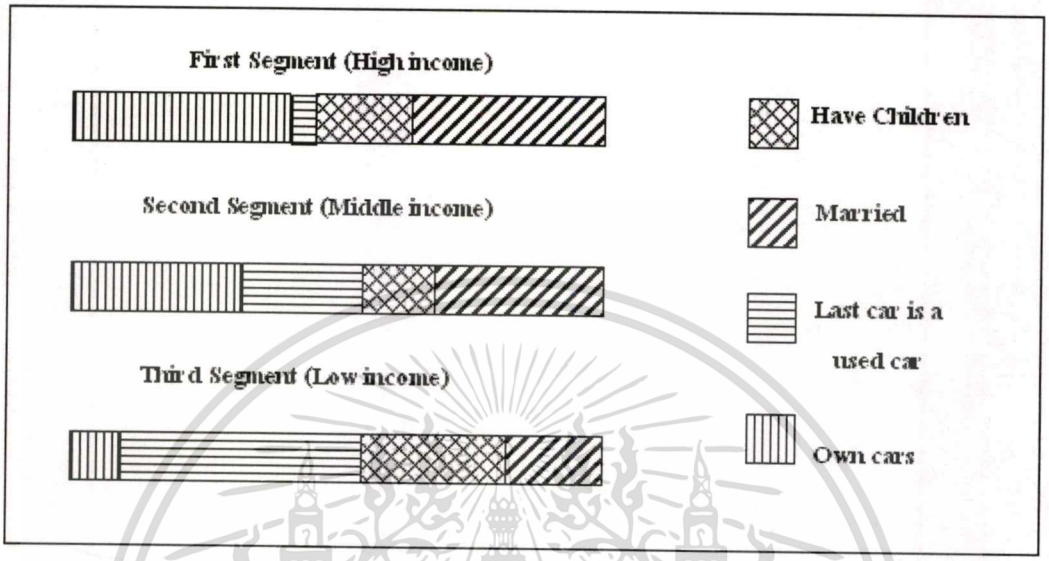
เทคนิคการลดขนาดของข้อมูลด้วยการรวมกลุ่มตัวแปรที่มีลักษณะเดียวกันไว้ด้วยกัน ตัวอย่างเช่น บริษัทจำหน่ายรถยนต์ได้แยกกลุ่มลูกค้าออกเป็น 3 กลุ่ม คือ

1. กลุ่มผู้มีรายได้สูง (>$80,000)
2. กลุ่มผู้มีรายได้ปานกลาง ($25,000 to $ 80,000)
3. กลุ่มผู้มีรายได้ต่ำ (Less than $25,000)

และภายในแต่ละกลุ่มยังแยกออกเป็น

- Have Children
- Married
- Last car is a used car

- Own cars



รูปที่ 2.4 ตัวอย่าง Clustering

จากข้อมูลข้างต้นทำให้ทางบริษัทรู้ว่าเมื่อมีลูกค้าเข้ามาที่บริษัทควรจะเสนอขายรถประเภทใด เช่น ถ้าเป็นกลุ่มผู้มีรายได้สูงควรจะเสนอรถใหม่ เป็นรถครอบครัวขนาดใหญ่พอสมควร แต่ถ้าเป็นผู้มีรายได้ค่อนข้างต่ำควรเสนอรถมือสอง ขนาดค่อนข้างเล็ก

4. Deviation Detection

เป็นกรรมวิธีในการหาค่าที่แตกต่างไปจากค่ามาตรฐาน หรือค่าที่คาดคิดไว้ว่าต่างไปมากน้อยเพียงใด โดยทั่วไปมักใช้วิธีการทางสถิติ หรือการแสดงให้เห็นภาพ (Visualization) สำหรับเทคนิคนี้ใช้ในการตรวจสอบ ลายเซ็นปลอม หรือบัตรเครดิตปลอม รวมทั้งการตรวจหาจุดบกพร่องของชิ้นงานในโรงงานอุตสาหกรรม

5. Link Analysis

จุดมุ่งหมายของ Link Analysis คือ การสร้าง Link ที่รียกว่า “associations” ระหว่าง Record เดียว หรือ กลุ่มของ Record ในฐานข้อมูล Link analysis สามารถแบ่งออกเป็น 3 ชนิด คือ

- Associations discovery
- Sequential pattern discovery
- Similar time sequence discovery

2.6 การประยุกต์ใช้งานดาต้าไมน์นิ่ง

- ธุรกิจค้าปลีกสามารถใช้งานดาต้าไมน์นิ่งในการพิจารณาหากลยุทธ์ให้เป็นที่สนใจกับผู้บริโภคในรูปแบบต่าง ๆ เช่น ที่ว่างในชั้นวางของจะจัดการอย่างไรถึงจะเพิ่มยอดขายได้ เช่นที่ Midas ซึ่งเป็นผู้แทนจำหน่ายอะไหล่สำหรับอุตสาหกรรมรถยนต์ งานที่ต้องทำคือการจัดการกับข้อมูลที่ได้รับจากสาขาทั้งหมด ซึ่งจะต้องทำการรวบรวมและวิเคราะห์อย่างทันท่วงที
- กิจการโทรคมนาคม เช่นที่ Bouygues Telecom ได้นำมาใช้ตรวจสอบการโกงโดยวิเคราะห์รูปแบบการใช้งานของสมาชิกลูกข่ายในการใช้งานโทรศัพท์ เช่น คาบเวลาที่ใช้จุดหมายปลายทาง ความถี่ที่ใช้ ฯลฯ และคาดการณ์ข้อบกพร่องที่เป็นไปได้ในการชำระเงิน เทคนิคนี้ยังได้ถูกนำมาใช้กับลูกค้าโทรศัพท์เคลื่อนที่ซึ่งระบบสามารถตรวจสอบได้ว่าที่ใดที่เสี่ยงที่จะสูญเสียลูกค้าสูงในการแข่งขัน France Telecom ได้ค้นหาวิธีรวมกลุ่มผู้ใช้ให้เป็นหนึ่งเดียวด้วยการสร้างแรงดึงดูดในเรื่องค่าใช้จ่ายและพัฒนาเรื่องความจงรักภักดีต่อตัวสินค้า
- การวิเคราะห์ผลิตภัณฑ์ เก็บรวบรวมลักษณะและราคาของผลิตภัณฑ์ทั้งหมดสร้างโมเดลด้วยเทคนิคดาต้าไมน์นิ่งและใช้โมเดลในการทำนายราคาผลิตภัณฑ์ตัวอื่น ๆ
- การวิเคราะห์ลูกค้า
  - ช่วยแบ่งกลุ่มและวิเคราะห์ลูกค้าเพื่อที่จะผลิตและเสนอสินค้าได้ตรงตามกลุ่มเป้าหมายแต่ละกลุ่ม
  - ทำนายว่าลูกค้าคนใดจะเลิกใช้บริการจากบริษัทภายใน 6 เดือนหน้า
- การวิเคราะห์การขาย
  - ช่วยในการโฆษณาสินค้าได้อย่างเหมาะสมและตรงตามเป้าหมาย
  - ช่วยในการวางสินค้าในชั้นได้อย่างเหมาะสม

บทที่ 3

Customer Segmentation

3.1 ความหมายของ Customer Segmentation

Segmentation เป็นการนำข้อมูลมาแบ่งเป็นกลุ่มซึ่งสมาชิกภายในกลุ่มจะมีลักษณะที่คล้ายกัน และแต่ละกลุ่มที่ได้ก็จะต้องมีความแตกต่างกัน โดยที่ไม่รู้ล่วงหน้าว่าจะมีทั้งหมดกี่กลุ่ม ถ้าข้อมูลมีลักษณะคล้ายกันก็จะจัดไว้ในกลุ่มเดียวกัน ยกตัวอย่างของข้อมูลลูกค้าในองค์กร เช่น การแบ่งเพศหญิงชายก็ถือว่าเป็นการจัดกลุ่ม ซึ่งสมาชิกที่อยู่ภายในกลุ่มนั้นอาจจะมีความคล้ายกันในหลายเหตุผลที่แตกต่างออกไป เช่น อาจจะมีความคล้ายกันในเรื่องของรายได้ หรืออาจจะมีความคล้ายกันในเรื่องของความคิดและพฤติกรรม เป็นต้น ซึ่งในการจัดกลุ่มข้อมูลนั้นจะทำให้เข้าใจความต้องการของลูกค้ายิ่งขึ้น และสามารถกำหนดเป้าหมายทางการตลาดได้ชัดเจนมากขึ้น ในการจัดกลุ่มข้อมูลนั้นเราอาจจะใช้รูปแบบของข้อมูลทางสถิติ (Statistical) หรือเทคนิคทางด้านดาต้าไมน์นิ่ง โดยแบ่งออกเป็น 2 อย่างคือ การทำนายการจัดกลุ่มข้อมูล (Predictive segmentation) และการจัดกลุ่มข้อมูล (Clustering)

หลักสำคัญที่ได้จากการจัดกลุ่มของข้อมูลที่แท้จริงนั้นจะต้องเป็นไปตามนี้

1. Collectively exhaustive คือข้อมูลในฐานข้อมูลทุกตัวจะต้องถูกนำไปจัดกลุ่ม
2. Mutually exclusive ข้อมูลของลูกค้าแต่ละคนในฐานข้อมูลจะสามารถถูกนำไปจัดกลุ่มได้เพียงครั้งเดียว คือจะไม่สามารถนำข้อมูลตัวเดียวกันไปจัดไว้ในกลุ่มอื่นได้ถ้าข้อมูลนั้นถูกจัดกลุ่มแล้ว เช่น เมื่อเราต้องการที่จะทำการจัดกลุ่มลูกค้าโดยแบ่งเป็น 3 กลุ่มด้วยกัน คือ
    1) ลูกค้าที่เป็นผู้ทำงานแล้ว
    2) ลูกค้าที่เป็นนักศึกษาระดับปริญญาตรี
    3) ลูกค้าที่มีการศึกษาต่ำกว่าปริญญาตรี

โดยมีลูกค้าคนหนึ่งซึ่งมีรหัสลูกค้าเป็น 47066111 โดยเราได้แบ่งลูกค้าคนนี้ไปอยู่ในกลุ่มที่ 2 คือกลุ่มลูกค้าที่เป็นนักศึกษาระดับปริญญาตรีแล้ว เราก็ไม่สามารถที่จะนำลูกค้าคนนี้ไปจัดให้อยู่ในกลุ่มอื่นๆ อีกได้ ซึ่งก็คือกลุ่มที่ 1 และกลุ่มที่ 3

ในการสร้างกลุ่มข้อมูลนั้นเราจะใช้วิธีการตัดสินใจโดยใช้โครงสร้างแบบต้นไม้ (Decision tree methods) และใช้หลักการของการจัดกลุ่มข้อมูล (Clustering approaches) ซึ่งภายในเทคนิคนี้ก็

ยังมีความแตกต่างที่สำคัญอย่างหนึ่งคือ การจัดกลุ่มข้อมูลนั้นเป็น Supervised learning หรือเป็น Unsupervised learning

- Supervised learning เป็นเทคนิคในการสร้างกลุ่มข้อมูลตามจุดมุ่งหมายที่วางเอาไว้ ตัวอย่างเช่นการสร้างกลุ่มลูกค้าที่มีคุณภาพสูงกับลูกค้าที่มีคุณภาพต่ำ
- Unsupervised learning เป็นเทคนิคการจัดกลุ่มที่ดึงเอาข้อมูลที่เราสนใจที่มีลักษณะเหมือนกันออกมาเพื่อใช้ช่วยในการจัดกลุ่มข้อมูล

เราสามารถนำข้อมูลของลูกค้าที่ได้ มาวิเคราะห์เป็น 3 ลักษณะ ดังนี้

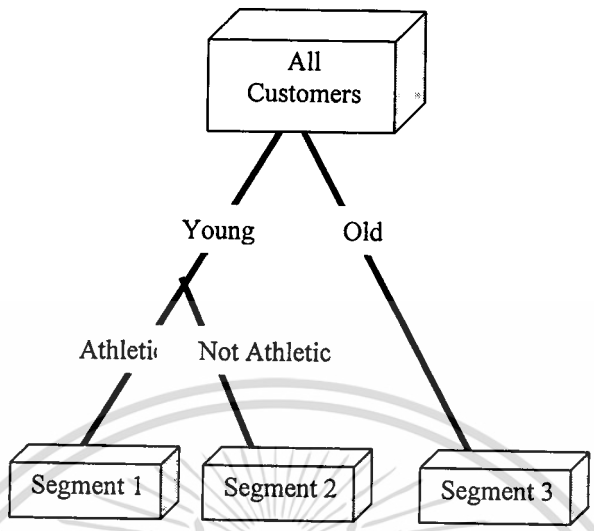
1. Demographic segmentation เป็นการจัดกลุ่มข้อมูลโดยอาศัยข้อมูลของลูกค้ามาจัดกลุ่ม เช่นศึกษาว่าที่อยู่ของลูกค้าอยู่ที่ไหน, สุขภาพของลูกค้าเป็นอย่างไร, ลูกค้ามีการศึกษาเป็นเช่นไร และอื่นๆ โดยอย่างน้อยที่สุดการรู้ถึงที่อยู่ของลูกค้าก็จะสามารถติดต่อถึงลูกค้าเหล่านั้นได้โดยผ่านช่องทางการสื่อสารทางการตลาด

2. Psychographic segmentation เป็นการจัดกลุ่มข้อมูลโดยใช้พฤติกรรมหรือความคิดของลูกค้าเป็นหลัก

3. Targeted segmentation ถูกใช้เมื่อเราตั้งเป้าหมายเอาไว้แล้ว ตัวอย่างเช่น การจัดกลุ่มของลูกค้าที่มีเป้าหมายคือเป็นลูกค้าที่ใช้สินค้าขององค์กรจำนวนมาก เราสามารถจัดกลุ่มลูกค้าโดยใช้ตัวสินค้าที่เราสามารถแข่งขันกับคู่แข่งมาเป็นตัวแบ่งกลุ่มลูกค้าได้

ดาต้าไมน์นิ่งที่ใช้ในการจัดกลุ่มข้อมูลนั้นมีหลายวิธี วิธีแรกคือมันจะใช้ในการกำหนดกลุ่มลูกค้าซึ่งได้มาจากการทำนายพฤติกรรมของลูกค้า ตัวอย่างเช่น Leaf nodes ของ Decision tree (ต้นไม้ในการตัดสินใจ) สามารถถูกมองเป็นกลุ่มแต่ละกลุ่มได้ ซึ่งแต่ละกลุ่มก็จะถูกกำหนดโดยลักษณะเฉพาะที่แน่นอนของลูกค้าและสำหรับลูกค้าทุกๆ คนก็จะมีลักษณะเฉพาะนั้น

ถ้า Decision tree ถูกใช้ในการสร้างกลุ่มข้อมูล ดังนั้นเราจะแน่ใจได้ว่าข้อมูลที่ได้จะเป็น Mutually exclusive และ Collectively exhaustive คือข้อมูลของลูกค้าแต่ละคนในฐานข้อมูลจะสามารถถูกนำไปจัดกลุ่มได้เพียงครั้งเดียวและข้อมูลทุกตัวจะต้องถูกนำไปจัดกลุ่ม ในรูปที่ 3.1 จะแสดงให้เห็นถึง Leaves ของ Decision tree ที่ใช้ในการจัดกลุ่มข้อมูลของลูกค้า



รูปที่ 3.1 แสดง Leaves ของ Decision tree ในการจัดกลุ่มข้อมูลของลูกค้า

ไม่ได้เฉพาะแค่ Decision tree เท่านั้นที่ใช้ในการจัดกลุ่มข้อมูล แต่เรายังสามารถใช้เทคนิคอื่นของดาต้าไมน์นิ่งในการวิเคราะห์เพื่อจัดกลุ่มข้อมูลของลูกค้าในฐานข้อมูลได้ โดยแต่ละกลุ่มก็จะประกอบด้วยข้อมูลของลูกค้าที่มีลักษณะที่เหมือนกัน

กระบวนการของการจัดกลุ่มข้อมูลเข้าไปไว้ในกลุ่มของข้อมูลที่มีความคล้ายคลึงกันนั้นจะเรียกว่า Clustering (การจัดกลุ่มข้อมูล) กลุ่ม (Cluster) จะเก็บข้อมูลที่มีลักษณะคล้ายกับข้อมูลอื่นที่อยู่ภายในกลุ่มเดียวกัน และข้อมูลที่มีลักษณะต่างกันก็จะถูกจัดไว้ในกลุ่มอื่น ซึ่งกลุ่มของข้อมูลแต่ละกลุ่มจะถูกเก็บรวบรวมไว้เป็นกลุ่มใหญ่เพียงกลุ่มเดียวในหลายๆ แอพพลิเคชั่น

โดยการจัดกลุ่มข้อมูลนั้นเป็นสิ่งหนึ่งที่เราสามารถระบุความหนาแน่น และความเบาบางของข้อมูลในบริเวณนั้นได้ ซึ่งจะทำให้เราทราบถึงรูปแบบการกระจายของข้อมูลทั้งหมดที่มีอยู่และรู้ถึงความสัมพันธ์ระหว่างคุณสมบัติของข้อมูลเหล่านั้นได้ด้วย

ในทางธุรกิจ การจัดกลุ่มของข้อมูลนั้นสามารถช่วยในการจัดกลุ่มลูกค้าของบริษัทและจัดกลุ่มการซื้อสินค้าของลูกค้าที่มีลักษณะแตกต่างกันอีกด้วย

ในการจัดกลุ่มข้อมูลนี้เป็นเครื่องมืออันหนึ่งที่จะใช้ช่วยให้เราเข้าใจถึงการกระจายของข้อมูลหรือให้เรามองเห็นลักษณะเฉพาะของแต่ละกลุ่ม และมุ่งเจาะลึกไปที่กลุ่มที่เราสนใจจะศึกษา ในอีกทางเลือกหนึ่งคือการจัดกลุ่มข้อมูลนี้อาจจะถูกใช้ในขั้นตอนการทำงานของอัลกอริทึมบางอย่าง เช่นการแสดงลักษณะเฉพาะของข้อมูลและการแยกประเภทของข้อมูลออกเป็นกลุ่มๆ เพื่อจะใช้ในการค้นหากลุ่มของข้อมูล

ซึ่งการจัดกลุ่มข้อมูลนั้นกลายมาเป็นหัวข้อที่สำคัญในงานวิจัยด้านดาต้าไมน์นิ่ง

ในสาขาทางสถิติ การวิเคราะห์กลุ่มนั้นมีการศึกษากันอย่างแพร่หลายมาหลายปีแล้ว โดยจะมุ่งเน้นศึกษาในเรื่อง Distance-based cluster analysis ซึ่งเครื่องมือที่ใช้ในการจัดกลุ่ม (วิเคราะห์กลุ่ม) ก็จะมี K-means, K-mediods และวิธีอื่นๆ อีกหลายวิธีซึ่งถูกสร้างเป็น Software packages การจัดกลุ่มของข้อมูลนั้นเป็นตัวอย่างของ Unsupervised learning ซึ่งไม่เหมือนกับการจัดแยกประเภทของกลุ่มข้อมูล (Classification) คือ Clustering และ Unsupervised learning ไม่ได้มีการกำหนดกลุ่มของข้อมูลไว้ก่อนล่วงหน้า จากเหตุผลนี้จะทำให้การจัดกลุ่มของข้อมูล (Clustering) จึงเป็นรูปแบบของการเรียนรู้โดยการสังเกต (Learning by observation) ซึ่ง Concept ในการจัดกลุ่มข้อมูลนั้นจะประกอบด้วย 2 องค์ประกอบหลักคือ

1) เป็นการค้นหากลุ่มของข้อมูลที่เหมาะสมออกมาจัดเป็นกลุ่ม

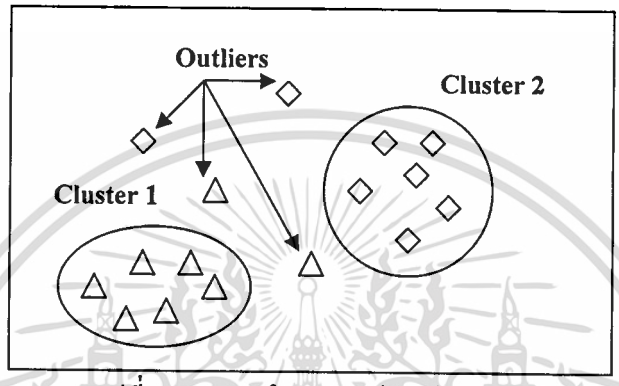
2) อธิบายลักษณะเฉพาะของกลุ่มแต่ละกลุ่มซึ่งได้จากการจัดแบ่งประเภทของกลุ่ม

สิ่งที่ควรคำนึงถึงเมื่อจะทำการสร้างอัลกอริทึมในการจัดกลุ่มข้อมูล มีหลักดังต่อไปนี้

- Scalability : คืออัลกอริทึมที่สร้างขึ้นมานั้นจะต้องสามารถรองรับกับข้อมูลที่มีปริมาณมากได้ หรือรองรับข้อมูลที่เพิ่มขึ้นในอนาคตได้
- Ability to deal with different types of attributes : คืออัลกอริทึมที่สร้างมานั้นควรมีความสามารถในการจัดการกับชนิดของข้อมูลที่แตกต่างกันได้
- Discovery of clusters with arbitrary shape : สามารถใช้กับการจัดกลุ่มข้อมูลที่มีลักษณะหลายรูปแบบได้
- Minimal requirements for domain knowledge to determine input parameters : คือการที่สามารถใส่ Input ของจำนวนกลุ่มที่มีขนาดแน่นอนเข้าไปในการทำงานของอัลกอริทึมของการจัดกลุ่มข้อมูล หรือผู้ใช้สามารถที่จะ Input ค่าของจำนวนกลุ่มที่จะใช้ในการจัดกลุ่มที่ต้องการลงไปได้
- Ability to deal with noisy data : เป็นการจัดการกับข้อมูลที่เราไม่สนใจหรือไม่ต้องการได้
- Insensitivity to the order of input records : ควรพัฒนาอัลกอริทึมที่ไม่ขึ้นอยู่กับลำดับของข้อมูล
- High dimensionality : ความสามารถของอัลกอริทึมที่สามารถจัดการกับข้อมูลที่มีหลายๆ มิติได้
- Constraint-based clustering : การจัดกลุ่มข้อมูลที่มีการกำหนดข้อบังคับต่างๆ เอาไว้
- Interpretability and usability : สามารถอธิบายขั้นตอนหรือขบวนการในการทำงานของอัลกอริทึม และสามารถนำอัลกอริทึมนั้นไปใช้งานได้จริง

3.2 Clustering

Clustering เป็นการแบ่งส่วนของข้อมูลเข้าไปในกลุ่มของ Object ที่เหมือนกัน โดยข้อมูลที่อยู่ภายในกลุ่มเดียวกันจะเหมือนกันและจะต่างจากกลุ่มอื่น โดยที่ไม่มีการกำหนดกลุ่มที่ใช้ในการแบ่งเอาไว้ล่วงหน้า



รูปที่ 3.2 แสดงถึงการนำข้อมูลไปจัดกลุ่ม

โดยแอพพลิเคชั่นที่ใช้สำหรับวิเคราะห์กลุ่มนั้นมีมากมาย ไม่ว่าจะเป็น Market/Customer segmentation, Pattern recognition, Biological studies และ Web document classification ซึ่งการจัดกลุ่มนั้นจะอยู่ในงานวิจัยทางด้านดาต้าไมน์นิ่งและเราสามารถแบ่งประเภทของอัลกอริทึมที่ใช้ในการจัดกลุ่มได้ดังนี้ คือ Partitioning, Hierarchical, Density-based, และ Model-based methods ซึ่งอัลกอริทึมที่จะทำการศึกษานี้ก็จะเป็น K-means algorithm ของ Partitioning method โดยจะกล่าวในหัวข้อต่อไป

อัลกอริทึมที่ใช้ในการจัดกลุ่มนั้นจะพยายามค้นหาข้อมูลของกลุ่มตามธรรมชาติ ซึ่งก็หมายความว่ามันจะหาว่าข้อมูลที่กำลังจะทำการจัดกลุ่มนั้นมีลักษณะเหมือนกับข้อมูลในกลุ่มใดมากที่สุด แล้วมันก็จะเอาข้อมูลนั้นไปจัดไว้ในกลุ่มที่มีลักษณะเหมือนกันที่สุด ซึ่งอัลกอริทึมสำหรับจัดกลุ่มนั้นจะหา Centroid ของกลุ่มข้อมูลเหล่านั้น ซึ่งค่า Centroid นี้จะเป็นค่า Mean ของกลุ่ม

โดยทั่วไปแล้วจะใช้ Distance function ซึ่งเป็นฟังก์ชันที่ใช้กำหนดความคล้ายคลึงกันในกลุ่มของข้อมูล โดยพิจารณาจากระยะระหว่างข้อมูล 2 ตัวซึ่งจะอยู่ในรูปของ Metric ซึ่งตัวอย่างของ Distance function ก็คือ Manhattan และ Euclidian distance functions

กำหนดให้ 2 p-dimensional data object คือ

$i = (x_{i1}, x_{i2}, ..., x_{ip})$ และ

$j = (x_{j1}, x_{j2}, ..., x_{jp})$

ซึ่ง Distance function จะถูกนิยามได้ดังนี้

Euclidian Distance Function:

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

ซึ่งส่วนมากจะนิยมใช้ Euclidean distance function เป็นเครื่องมือวัดระยะห่าง (Distance measure)

ข้อดีของการจัดกลุ่ม คือ ทำให้ข้อมูลที่อยู่ในกลุ่มเดียวกันมีความคล้ายกัน และสามารถค้นหารูปแบบที่แอบซ่อนอยู่ได้

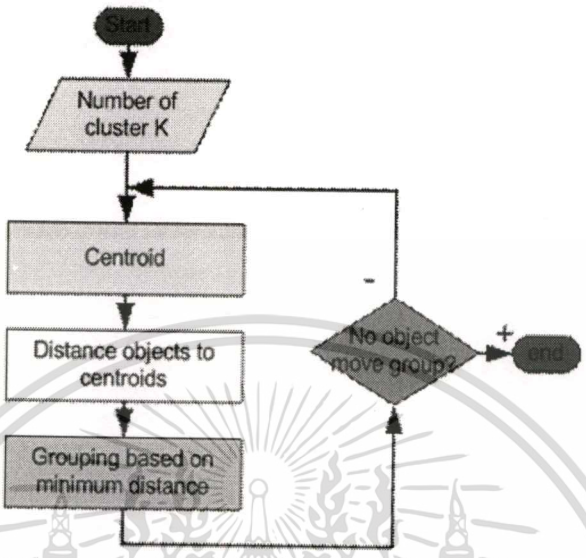
3.3 ความหมายของ K-means algorithm

K-means algorithm เป็นวิธีหนึ่งที่นิยมใช้ในการจัดกลุ่มของข้อมูล โดยแอพพลิเคชั่นทางด้านดาต้าไมน์นิ่งเชิงพาณิชย์เกือบทั้งหมดได้รวมเอา Clustering algorithm ไว้ด้วย ซึ่งความสามารถหลักของ K-means algorithm ก็คือการกำหนดจำนวนกลุ่มที่ต้องการใช้ในการจัดกลุ่มได้ก่อนล่วงหน้าซึ่งจำนวนกลุ่มที่กำหนดนี้จะแทนด้วยตัวแปร k ตัวอย่างเช่นถ้าต้องการจะจัดกลุ่มข้อมูลเป็น 10 กลุ่ม ดังนั้น k=10 ผลลัพธ์สุดท้ายของการจัดกลุ่มขึ้นอยู่กับค่า k ที่เลือก

เราสามารถกำหนดค่า k ให้เท่ากับ 1 ก็ได้ แต่ผลลัพธ์ที่ได้นั้นจะไม่มีความหมาย เพราะว่าข้อมูลทุกๆตัวจะถูกจัดกลุ่มเข้าไปไว้ในกลุ่มเพียงกลุ่มเดียวเท่านั้น ในอีกกรณีคือการกำหนดให้ค่า k เท่ากับจำนวนข้อมูลทั้งหมดที่จะนำไปจัดกลุ่มซึ่งในกรณีนี้ผลลัพธ์ที่ได้ก็จะไม่มีความหมายเช่นเดียวกัน เพราะว่าการจัดกลุ่มในกรณีนี้ผลลัพธ์ที่ได้ก็เหมือนกับข้อมูลก่อนการจัดกลุ่ม เพราะฉะนั้นการกำหนดจำนวนของกลุ่มที่จะใช้ในการจัดกลุ่มที่เป็นไปได้นั้นก็จะขึ้นอยู่กับการกำหนดค่าให้แก่ k จึงไม่มีกฎในการเลือกค่า k และเป็นเรื่องที่ดีถ้ามีการเปลี่ยนค่าของตัวแปรที่ใช้ในการจัดกลุ่ม

K-means algorithm เป็นหนึ่งในกลุ่มของอัลกอริทึมที่ถูกเรียกว่า Partitioning methods ซึ่งเป็นการสร้างกลุ่มของข้อมูลโดยกำหนดค่า Mean ที่อยู่ตรงกลางให้กับแต่ละกลุ่ม

อธิบายการทำงานของ K-means algorithm



รูปที่ 3.3 อธิบายการทำงานของ K-means algorithm

Step 1 เริ่มด้วยการเลือกค่าของ k ซึ่ง k คือจำนวนของกลุ่มที่จะจัดกลุ่ม

Step 2 ใส่ค่าเริ่มต้น (Centroid) ที่จะใช้ในการแบ่งข้อมูลให้กับกลุ่ม k กลุ่ม ซึ่งอาจจะได้จากการสุ่ม หรือใช้ข้อมูล k ตัวแรกในการจัดกลุ่ม จากนั้นกำหนดให้ข้อมูลแต่ละตัวที่เหลือ (n-k) เข้าไปไว้ในกลุ่มที่มีค่าใกล้เคียงกับค่า Centroid ในกลุ่มนั้นมากที่สุด โดยจะได้จากการคำนวณหาค่า Distance ซึ่งจะจัดข้อมูลไว้ในกลุ่มที่คำนวณได้ค่า Distance น้อยที่สุด หลังจากที่กำหนดข้อมูลเข้าไปไว้ในกลุ่มเรียบร้อยแล้ว ก็ให้ทำการคำนวณค่า Centroid ของแต่ละกลุ่มใหม่

Step 3 เอาข้อมูลแต่ละตัวไปคำนวณหาค่า Distance โดยเปรียบเทียบกับค่า Centroid ของแต่ละกลุ่ม ถ้าข้อมูลนั้นมีค่าใกล้เคียงกับค่า Centroid ในกลุ่มอื่นมากกว่า ก็ให้ย้ายข้อมูลไปไว้ในกลุ่มนั้น แล้วให้ทำการปรับค่า Centroid ของกลุ่มที่มีการเปลี่ยนแปลงข้อมูลในกลุ่มอีกครั้ง

Step 4 ทำซ้ำในขั้นตอนที่ 3 จนกระทั่งไม่มีการเคลื่อนย้ายข้อมูล หรือจนกระทั่งไม่มีการเปลี่ยนแปลงค่าของ Centroid ในแต่ละกลุ่ม

ถ้าจำนวนของข้อมูลน้อยกว่าจำนวนของกลุ่ม ดังนั้นเราจะกำหนดให้ข้อมูลแต่ละตัวเป็น Centroid ของกลุ่ม ซึ่งแต่ละ Centroid จะมีหมายเลขกลุ่ม ถ้าจำนวนของข้อมูลมากกว่าจำนวนของกลุ่ม ข้อมูลแต่ละตัวก็จะถูกคำนวณหาค่า Distance เทียบกับ Centroid ทุกตัวที่มี และจะจัดข้อมูลนั้นเข้าไปไว้ในกลุ่มที่มีค่า Distance ต่ำที่สุด

เราจะไม่ทราบตำแหน่งที่แน่นอนของ Centroid ดังนั้นจำเป็นที่จะต้องปรับค่าตำแหน่งของ Centroid ใหม่เมื่อมีการเปลี่ยนแปลงของข้อมูล แล้วก็จะกำหนดข้อมูลทุกๆ ตัวให้กับ Centroid ใหม่นี้

ประสิทธิภาพในการทำงานของ K-means algorithm คือ O(knt)

ซึ่ง k คือจำนวนของกลุ่ม

n เป็นจำนวนของ object ทั้งหมด

t คือจำนวนรอบที่ algorithm ทำงาน

อธิบายการทำงานของ direct k-means algorithm โดยสมมติให้จำนวนของกลุ่มที่จะใช้ในการจัดกลุ่มของ k-means เท่ากับ k โดยสุ่มค่าขึ้นมา $k : (w_1, \ldots, w_k)$ ค่าจากทั้งหมด $n : (i_1, \ldots, i_n)$ ในขั้นแรกซึ่งจะได้

$w_j = i_l, j \in \{1, \ldots, k\}, l \in \{1, \ldots, n\}$

<u>การทำงานของ Direct k-means clustering algorithm</u>

function Direct-k-means()

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

Each cluster C_j is associated with prototype w_j

Repeat

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

Assign i_l to the cluster C_{j^*} with nearest prototype w_{j^*}.

(i.e., $| i_l - w_{j^*} | \leq | i_l - w_j |, j \in \{1, \ldots, k\}$)

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

Update the prototype w_j to be the

centroid of all samples currently

in C_j, so that $w_j = \sum_{i_l \in C_j} i_l / | C_j |$

Compute the error function:

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

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

จากอัลกอริทึมข้างบนนี้จะแสดงการทำงานของ Direct k-means clustering algorithm ซึ่ง C_j คือกลุ่มที่ j ซึ่งค่าของมันจะเป็น Disjoint subset ของข้อมูลที่นำเข้าสู่อัลกอริทึม โดยที่คุณภาพของการจัดกลุ่มนี้จะถูกกำหนดโดย error function ตามสมการข้างล่างนี้

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

การเลือกค่า k ที่เหมาะสมนั้นคือปัญหาซึ่งจะขึ้นอยู่กับกรณีที่จะทำการพิจารณา โดยทั่วไปแล้วผู้ใช้จะทดลองสุ่มค่า k ขึ้นมาหลายๆ ค่า สมมติว่ามีอยู่ n รูปแบบ และมี d dimension เพราะฉะนั้นเวลาที่ใช้ในการคำนวณของ Direct k-means algorithm ในแต่ละรอบนั้นจะประกอบด้วย 3 ส่วนหลักๆ ดังนี้

1. เวลาที่ใช้ในการทำงานของ loop for แรกในรูปที่ 6 จะเป็น O(nkd)
2. เวลาที่ใช้ในการคำนวณหาค่า centroid ซึ่งอยู่ใน loop for ที่ 2 จะเป็น O(nd)
3. ส่วนเวลาที่ใช้ในการคำนวณ error function จะเป็น O(nd)

จำนวนของรอบการทำงานทั้งหมดนั้นมีช่วงกว้างโดยเริ่มจากค่าน้อยๆ ไปจนถึงหลายพัน ซึ่งจะขึ้นอยู่กับจำนวนของ Pattern, จำนวนของกลุ่ม และการกระจายของข้อมูลที่นำเข้า ด้วยเหตุนี้การทำงานด้วยวิธี K-means จึงสามารถคำนวณได้ดี

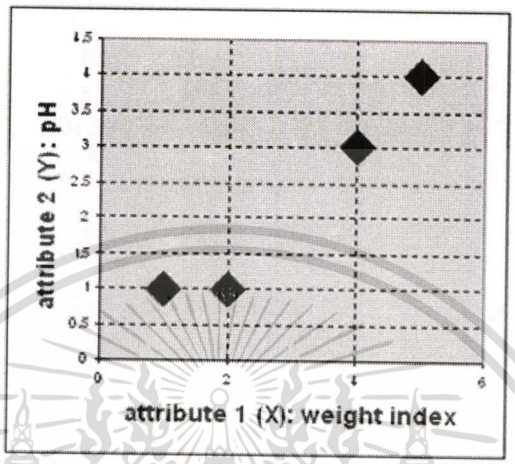
3.4 ตัวอย่างการจัดกลุ่มของข้อมูล

สมมติเรามี Objects อยู่ 4 Objects ซึ่งแต่ละ Object มี 2 Attributes/Features ดังแสดงในตารางข้างล่างนี้ จุดประสงค์คือต้องการจัดกลุ่มยา (Medicine) เป็น 2 กลุ่ม (k=2) โดยใช้ค่า ph และ weight index ในการจัดกลุ่มของยา

ตารางที่ 3.1 แสดงตัวอย่างของข้อมูล Medicine ที่จะใช้ในการจัดกลุ่ม

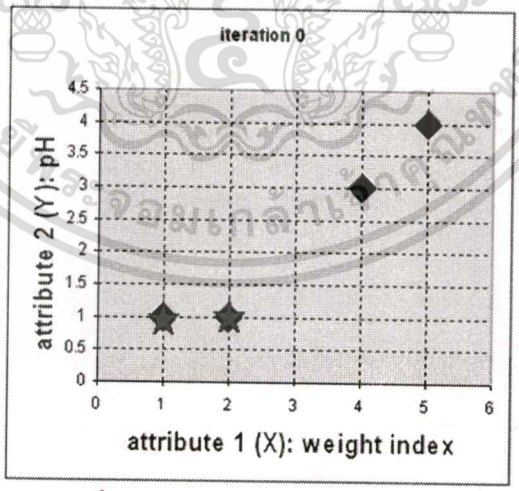
Object	attribute 1 (X): weight index	attribute 2 (Y): pH
Medicine A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4

ตัวยา (Medicine) แต่ละตัวจะแทน 1 จุด ด้วย 2 attributes (X, Y) ซึ่งแสดงเป็นพิกัดดังรูปข้างล่างนี้



รูปที่ 3.4 พิกัดของข้อมูล Medicine

Step 1 กำหนดค่าของ Centroids เริ่มต้น สมมติเราใช้ Medicine A และ Medicine B เป็น Centroids เริ่มต้น เราจะให้ c_1 และ c_2 แทนด้วยพิกัดของ Centroids ดังนั้น $c_1 = (1, 1)$ และ $c_2 = (2, 1)$



รูปที่ 3.5 พิกัดของ Centroids เริ่มต้น

Step 2 เป็นการหาค่า Distance หรือระยะระหว่าง Object แต่ละตัวเปรียบเทียบกับค่า Centroids ในแต่ละกลุ่ม ซึ่งเราจะใช้ Euclidean distance ในการหา ดังนั้นเราจะได้ Distance matrix ที่ Iteration 0 เป็น

$$\mathbf{D}^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad \begin{matrix} \mathbf{c}_1 = (1,1) \text{ group - 1} \\ \mathbf{c}_2 = (2,1) \text{ group - 2} \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \begin{matrix} X \\ Y \end{matrix}$$

แต่ละ Column ใน Distance matrix แสดงถึง Object แถวแรกของ Distance matrix เป็นระยะระหว่าง Object แต่ละ Object เปรียบเทียบกับ Centroid ตัวแรก และแถวที่สองก็คือระยะระหว่างแต่ละ Object เปรียบเทียบกับ Centroid ตัวที่ 2 ตัวอย่างเช่น Distance หรือระยะระหว่าง Medicine C = (4, 3) กับ Centroid ตัวแรก ($c_1 = (1, 1)$) คือ $\sqrt{(4-1)^2 + (3-1)^2} = 3.61$ และระยะระหว่าง Centrois ตัวที่ 2 ($c_2 = (2, 1)$) คือ $\sqrt{(4-2)^2 + (3-1)^2} = 2.83$ เป็นต้น

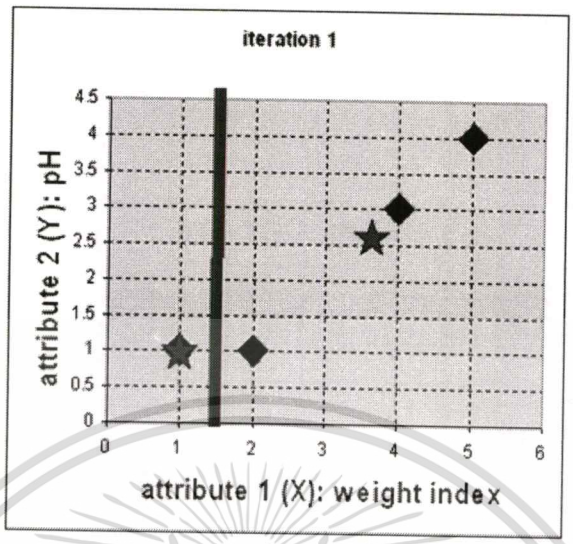
Step 3 เป็นการจัดกลุ่ม Objects คือเรากำหนดแต่ละ Object ไปไว้ในกลุ่มที่มี Distance ที่น้อยที่สุด ดังนั้น Medicine A จะถูกกำหนดให้อยู่ในกลุ่มที่ 1, Medicine B อยู่ในกลุ่ม 2, Medicine C อยู่ในกลุ่ม 2 และ Medicine D ก็อยู่ในกลุ่ม 2 เช่นเดียวกัน ซึ่ง object ที่อยู่ในกลุ่มใดนั้นแทนด้วย 1 ดังแสดงใน Group matrix ด้านล่างนี้

$$\mathbf{G}^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \quad \begin{matrix} \text{group - 1} \\ \text{group - 2} \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

Step 4 ใน Iteration-1 นี้จะเป็นการหาค่า Centroids ใหม่ ซึ่งเมื่อเราทราบสมาชิกในแต่ละกลุ่มแล้ว ในขั้นตอนนี้เราจะทำการคำนวณหาค่า Centroid ของแต่ละกลุ่มใหม่ ซึ่งในกลุ่มที่ 1 นั้นมีสมาชิกเพียงหนึ่งตัว ดังนั้นค่า Centroid ก็ยังคงเป็นตัวเดิมคือ $c_1 = (1, 1)$ ในกลุ่มที่ 2 ซึ่งมีสมาชิดอยู่ 3 ตัว ดังนั้นค่า Centroid คือการหาค่าเฉลี่ยของพิกัดระหว่างสมาชิก 3 ตัว เพราะฉะนั้นจะได้

$$c_2 = \left(\frac{2+4+5}{3}, \frac{1+3+4}{3}\right) = \left(\frac{11}{3}, \frac{8}{3}\right)$$



รูปที่ 3.6 พิกัดของ Centroids ที่ได้จากการคำนวณใน Iteration-1

Step 5 ใน Iteration-1, จะเป็นการคำนวณหา Distance ของทุกๆ Object เปรียบเทียบกับ Centroids ใหม่ คล้ายใน Step ที่ 2 ซึ่ง Distance matrix ใน Iteration 1 คือ

$$\mathbf{D}^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \quad \begin{matrix} \mathbf{c}_1 = (1,1) \text{ group - 1} \\ \mathbf{c}_2 = (11/3,8/3) \text{ group - 2} \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \begin{matrix} X \\ Y \end{matrix}$$

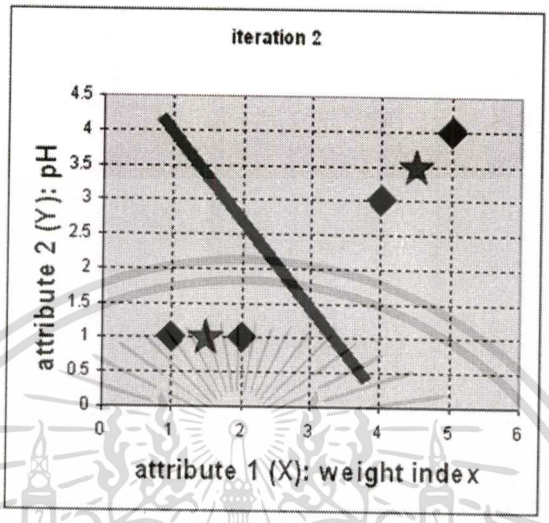
Step 6 Iteration-1, จะเป็นการจัดกลุ่มของ Object เหมือนใน Step 3 ซึ่งเราจะทำการกำหนดแต่ละ Object เข้าไปในกลุ่มที่มี Distance น้อยสุด จาก Distance matrix ที่ได้จากการคำนวณใหม่นี้เราจะย้าย Medicine B ไปไว้ในกลุ่ม 1 ในขณะที่ Object ตัวอื่นๆ ก็ยังอยู่ในกลุ่มเหมือนเดิม ซึ่งจะได้ Group matrix ดังแสดงข้างล่างนี้

$$\mathbf{G}^1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad \begin{matrix} \text{group - 1} \\ \text{group - 2} \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

Step 7 Iteration 2, เป็นการหาค่า Centroids ใหม่ ซึ่งเราจะทำซ้ำ Step 4 เพื่อคำนวณหาพิกัดของ Centroids ใหม่ โดยจะพิจารณาจากการจัดกลุ่มในรอบที่แล้ว ซึ่งทั้งในกลุ่มที่ 1 และ 2 ต่างก็มีสมาชิก 2 ตัว ดังนั้น Centroids ใหม่ที่ได้คือ

$$c_1 = \left(\frac{1+2}{2}, \frac{1+1}{2}\right) = \left(1\frac{1}{2}, 1\right) \text{ และ } c_2 = \left(\frac{4+5}{2}, \frac{3+4}{2}\right) = \left(4\frac{1}{2}, 3\frac{1}{2}\right)$$



รูปที่ 3.7 พิกัดของ Centroids ที่ได้จากการคำนวณใน Iteration-2

Step 8 Iteration-2, ทำซ้ำใน Step 2 อีกครั้ง เราจะได้ Distance matrix ใหม่ใน Iteration 2 เป็นดังนี้

$$\mathbf{D}^2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix}$$

$\mathbf{c}_1 = (3/2,1)$ group - 1

$\mathbf{c}_2 = (9/2,7/2)$ group - 2

A B C D

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix}$$

X

Y

Step 9 Iteration-2, เป็นการกำหนดแต่ละ Object เข้าไปอยู่ในกลุ่มที่มี Distance น้อยที่สุด

$$\mathbf{G}^2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

group - 1

group – 2

A B C D

ผลลัพธ์ที่ได้ คือ $\mathbf{G}^2 = \mathbf{G}^1$ ซึ่งการเปรียบเทียบการจัดกลุ่มของรอบสุดท้ายและในรอบนี้แสดงให้เห็นว่า Object ไม่ได้มีการเปลี่ยนกลุ่ม ด้วยเหตุนี้การทำงานของ K-means algorithm ได้

มาถึงจุดที่ไม่มีการเปลี่ยนแปลงแล้ว ดังนั้นเราจะเอากลุ่มที่ได้จากการจัดกลุ่มครั้งสุดท้ายมาเป็นผลลัพธ์ ดังแสดงในตารางข้างล่างนี้

ตารางที่ 3.2 แสดงผลลัพธ์ที่ได้จากการจัดกลุ่ม

Object	Feature 1 (X): weight index	Feature 2 (Y): pH	Group (result)
Medicine A	1	1	1
Medicine B	2	1	1
Medicine C	4	3	2
Medicine D	5	4	2

3.5 จุดอ่อนของ K-means algorithm

1. ใช้ได้เพียงข้อมูลที่มีรูปแบบเป็นตัวเลขหรือมีการแปลงเป็นตัวเลขแล้วเท่านั้น
2. จำเป็นต้องระบุจำนวนของกลุ่ม (k) ที่ต้องการจะจัดกลุ่มก่อนการจัดกลุ่มเสมอ
3. ทำงานผิดพลาดกับข้อมูลที่เป็น Noisy และ Outliers

บทที่ 4
การวิเคราะห์และออกแบบระบบ

4.1 ความต้องการของระบบ

4.1.1 การสร้างโปรเจ็กต์

1. ฟังก์ชันการเลือกข้อมูล

ฟังก์ชันการเลือกข้อมูลเป็นขั้นตอนแรกของการเตรียมข้อมูลสำหรับทำดาต้าไมน์นิ่ง โดยผู้ใช้งานสามารถเลือกฐานข้อมูลจาก Microsoft SQL Server 2000 และเลือกตารางที่ต้องการใช้งานมาจริงๆ เพียงหนึ่งตาราง จากนั้นก็ให้ทำการเลือกฟิลด์ที่จะใช้สำหรับทำดาต้าไมน์นิ่งตามต้องการ โดยฟิลด์นั้นจะเป็นฟิลด์ที่อยู่ในตารางที่ทำการเลือกไปแล้วนั้น

2. ฟังก์ชันการแก้ไขข้อมูล

ฟังก์ชันการแก้ไขข้อมูลนี้เป็นขั้นตอนที่สอง โดยจะเป็นส่วนของการนำข้อมูลจากฟิลด์ที่เลือกมาทำการแก้ไขในกรณีที่ฟิลด์มีค่าเป็นค่าว่าง เพื่อที่จะเพิ่มประสิทธิภาพของการทำดาต้าไมน์นิ่งให้ได้ผลดียิ่งขึ้น

3. ฟังก์ชันการแปลงหรือปรับเปลี่ยนข้อมูล

สำหรับฟังก์ชันการแปลงหรือปรับเปลี่ยนข้อมูลนี้จะเป็นขั้นตอนที่สาม โดยจะเป็นการนำเอาค่าซึ่งเป็นตัวอักษรมาทำการแปลงให้เป็นตัวเลข จากนั้นก็นำเอาค่าข้อมูลทั้งหมดซึ่งได้ถูกแปลงเป็นตัวเลขแล้วนั้นมาทำการแปลงให้อยู่ในช่วงที่ต้องการอีกที เพื่อที่จะทำให้ผลลัพธ์จากการทำดาต้าไมน์นิ่งของข้อมูลในแต่ละแอตทริบิวต์มีค่าไม่แตกต่างกันมากเกินไป

4. ฟังก์ชันการจัดกลุ่มข้อมูล

ฟังก์ชันนี้เป็นขั้นตอนที่สี่ซึ่งเป็นขั้นตอนสำหรับการทำดาต้าไมน์นิ่งของการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม K-means ซึ่งผู้ใช้สามารถจะกำหนดจำนวนกลุ่มเพื่อจะทำการจัดกลุ่มได้ตามต้องการ แต่จำนวนกลุ่มที่กำหนดนั้นจะต้องไม่เกินจำนวนเรคอร์ดทั้งหมดที่มีอยู่ในฐานข้อมูลซึ่งได้จากการแก้ไขข้อมูลแล้วนั้น จากนั้นโปรแกรมก็จะทำการจัดกลุ่มข้อมูลตามจำนวนกลุ่มที่กำหนดนั้นและแสดงผลลัพธ์ออกมาโดยจะแบ่งเป็นสองส่วน คือ ส่วนแรกจะเป็นส่วนของจุดศูนย์กลางของกลุ่มข้อมูล และส่วนที่สองจะแสดงสมาชิกต่างๆ ที่อยู่ในแต่ละกลุ่มนั้น

5. ฟังก์ชันการแสดงผลการจัดกลุ่มด้วยกราฟ

ฟังก์ชันนี้จะเป็นการทำงานหลังจากที่ทำการจัดกลุ่มข้อมูลแล้ว โดยจะแสดงจุดศูนย์กลางของกลุ่มข้อมูลในแต่ละกลุ่มเปรียบเทียบกันด้วยกราฟเส้น

6. ฟังก์ชันการบันทึกข้อมูล

ฟังก์ชันนี้จะเป็นการนำเอาข้อมูลที่ได้หลังจากการจัดกลุ่มมาจัดเก็บลงฐานข้อมูล Microsoft SQL Server 2000 เพื่อนำข้อมูลไปใช้ในการทดสอบโมเดลต่อไป

4.1.2 การทดสอบโมเดล

1. ฟังก์ชันการเลือกโมเดลที่จะใช้สำหรับทดสอบข้อมูล

เป็นฟังก์ชันสำหรับเลือกโมเดลที่ได้จากการสร้างโปรเจ็กต์เพื่อนำมาใช้สำหรับทดสอบข้อมูลที่เราต้องการ

2. ฟังก์ชันการเลือกข้อมูลที่จะใช้สำหรับทดสอบ

ฟังก์ชันนี้จะเป็นการเลือกข้อมูลจากฐานข้อมูลที่จะนำมาทดสอบ และเลือกตารางที่ต้องการใช้ทดสอบมาหนึ่งตาราง

3. ฟังก์ชันการแม็พข้อมูล

ฟังก์ชันการแม็พข้อมูลนี้จะเป็นการเลือกฟิลด์จากตารางในฐานข้อมูลที่เราเลือก เพื่อนำมาแม็พเข้ากับฟิลด์ที่มีอยู่ในโมเดลที่เลือกในขั้นแรกของการทดสอบโมเดล โดยฟิลด์ที่นำมาแม็พกันนั้นจะต้องมีค่าของข้อมูลที่เหมือนหรือคล้ายกัน จากนั้นโปรแกรมจะทำการตรวจสอบความเป็นไปได้ในการแม็พกันของข้อมูลให้ ถ้าแม็พกันไม่ได้ผู้ใช้ก็จะต้องทำการเลือกฟิลด์จากตารางข้อมูลที่จะใช้ทดสอบใหม่

4. ฟังก์ชันการแก้ไขและแปลงค่าของข้อมูล

ฟังก์ชันนี้จะทำการแก้ไขข้อมูลที่มีค่าเป็นค่าว่าง โดยจะทำการลบเรคอร์ดนั้นทิ้งไป จากนั้นก็จะนำข้อมูลที่ได้จากการแก้ไขแล้วนั้นไปทำการแปลงให้อยู่ในช่วงของค่าต้องการ

5. ฟังก์ชันการทดสอบโมเดล

ฟังก์ชันนี้จะทำงานหลังจากที่ทำการแปลงข้อมูลเรียบร้อยแล้ว โดยมันจะเอาข้อมูลทั้งหมดไปทำการเปรียบเทียบระยะห่างกับจุดศูนย์กลางในแต่ละกลุ่มของข้อมูลที่ได้จากโมเดล

6. ฟังก์ชันการแสดงผลจากการทดสอบโมเดล

ฟังก์ชันนี้จะเป็นการแสดงผลลัพธ์ออกมาเป็นสองส่วนคือ ส่วนแรกจะแสดงค่าจุดศูนย์กลางในแต่ละกลุ่มของข้อมูลที่ได้มาจากโมเดลซึ่งเลือกมาทำการทดสอบ และส่วนที่สองจะแสดงให้เห็นว่าข้อมูลแต่ละเรคอร์ดนั้นอยู่ในกลุ่มใด

4.2 องค์ประกอบของระบบงาน

ระบบการจัดกลุ่มข้อมูลถูกพัฒนาขึ้นมาโดยใช้อัลกอริทึม K-means ซึ่งประกอบด้วยส่วนประกอบหลักๆ อยู่ 3 ส่วน ดังนี้

4.2.1 ส่วนนำข้อมูลเข้า

ส่วนนี้จะเป็นการนำเอาข้อมูลเข้ามาเพื่อนำไปใช้สำหรับทำการจัดกลุ่มข้อมูล โดยจะแบ่งเป็น 2 ส่วนคือ ส่วนของการสร้างโปรเจ็กต์และส่วนของการทดสอบโมเดล โดยสามารถแบ่งได้ดังนี้

1. ส่วนของการสร้างโปรเจ็กต์

1.1 ฐานข้อมูลที่จะใช้เพื่อดึงข้อมูลมาใช้ โดยฐานข้อมูลที่เลือกเข้ามานั้นจะต้องเป็นฐานข้อมูล Microsoft SQL Server 2000

1.2 ตารางข้อมูลโดยเลือกจากฐานข้อมูลที่เลือกไว้ในข้อที่ 1.1

1.3 ฟิลด์ข้อมูล โดยเลือกจากตารางที่เลือกไว้ในข้อที่ 1.2

1.4 จำนวนกลุ่มที่จะใช้ในการจัดกลุ่ม

2. ส่วนของการทดสอบโมเดล

2.1 โมเดลที่จะนำไปใช้ในการทดสอบข้อมูล

2.2 ฐานข้อมูลที่ต้องการนำมาทดสอบ โดยฐานข้อมูลที่เลือกเข้ามานั้นจะต้องเป็นฐานข้อมูล Microsoft SQL Server 2000

2.3 ตารางข้อมูล โดยเลือกจากฐานข้อมูลที่เลือกไว้ในข้อที่ 2.2

2.4 ฟิลด์ข้อมูลที่จะนำไปแม็พกับฟิลด์ของโมเดล โดยเลือกจากตารางที่เลือกไว้ในข้อที่ 2.3

4.2.2 ส่วนวิเคราะห์และประมวลผล

ขั้นตอนนี้เป็นการประมวลผลข้อมูลที่นำเข้ามา โดยการประมวลผลนั้นจะแบ่งเป็น 2 ส่วนคือ ส่วนของการเตรียมข้อมูล และส่วนของการทำไมน์นิ่ง ดังต่อไปนี้

1. ส่วนของการเตรียมข้อมูล

ในส่วนของการเตรียมข้อมูลนี้จะเป็นการเตรียมข้อมูลเพื่อจะนำไปใช้ในขั้นของการทำไมน์นิ่ง โดยเราสามารถแบ่งเป็น 2 ส่วนคือ การเตรียมข้อมูลสำหรับการสร้างโปรเจ็กต์ และส่วนของการเตรียมข้อมูลสำหรับการทดสอบโมเดล โดยมีขั้นตอนดังต่อไปนี้

<u>การเตรียมข้อมูลสำหรับการสร้างโปรเจ็กต์</u>

- เลือกฟิลด์ข้อมูลที่ต้องการทำการจัดกลุ่ม
- ทำการแก้ไขข้อมูลในกรณีค่าข้อมูลนั้นเป็นค่าว่าง โดยอาจจะทำการลบเรคอร์ดที่มีค่าว่างนั้นทิ้งไป หรืออาจจะทำการแทนค่าว่างด้วยค่า Mean ของแอตทริบิวต์นั้น เป็นต้น

- ทำการแปลงข้อมูลที่เป็นตัวอักษรให้เป็นตัวเลข
- ทำการแปลงค่าของข้อมูลให้อยู่ในช่วงที่ต้องการเพื่อทำให้การทำงานของอัลกอริทึมมีประสิทธิภาพมากขึ้น

<u>การเตรียมข้อมูลสำหรับการทดสอบโมเดล</u>

- เลือกโมเดลที่ต้องการนำไปใช้ทำการทดสอบข้อมูล
- เลือกฟิลด์จากตารางข้อมูลที่ต้องการนำไปใช้ทำการทดสอบ
- ทำการแม็พฟิลด์จากโมเดลที่จะนำไปทดสอบกับฟิลด์ที่ทำการเลือก ว่าสามารถแม็พกันได้หรือไม่
- ทำการแก้ไขข้อมูลที่มีค่าเป็นค่าว่าง โดยการลบเรคอร์ดที่มีค่าเป็นค่าว่างนั้นทิ้ง
- ทำการแปลงข้อมูลที่ได้หลังจากการแก้ไขแล้วให้อยู่ในช่วงที่ต้องการ

2. ส่วนการทำไมน์นิ่ง

ในส่วนการทำไมน์นิ่งนี้จะเป็นการจัดกลุ่มของข้อมูลตามการทำงานของอัลกอริทึม K-means โดยเป็นการหาความเหมือนกันของข้อมูล ซึ่งพิจารณาจากระยะห่างจากจุดศูนย์กลางของกลุ่ม โดยจะแบ่งเป็น 2 ส่วนคือ ส่วนของการทำไมน์นิ่งสำหรับการสร้างโปรเจ็กต์ และส่วนของการทำไมน์นิ่งสำหรับการทดสอบโมเดล ซึ่งมีขั้นตอนดังต่อไปนี้

<u>การทำไมน์นิ่งสำหรับการสร้างโปรเจ็กต์</u>

1. ทำการกำหนดจุดศูนย์กลางของกลุ่มข้อมูลเริ่มต้นตามจำนวนกลุ่มที่จะใช้สำหรับทำการจัดกลุ่ม โดยในที่นี้เราจะกำหนดให้ข้อมูลตั้งแต่เรคอร์ดแรกเป็นจุดศูนย์กลางเริ่มต้นไปจนครบจำนวนกลุ่มข้อมูลที่จะใช้สำหรับทำการจัดกลุ่ม
2. ทำการคำนวณหาค่า Distance หรือระยะห่างของข้อมูลแต่ละตัวเปรียบเทียบกับจุดศูนย์กลางในแต่ละกลุ่ม
3. ทำการกำหนดกลุ่มของข้อมูล โดยได้จากการนำเอาค่า Distance ที่หาได้ในแต่ละกลุ่มมาเปรียบเทียบกัน โดยจัดข้อมูลเข้าไปไว้ในกลุ่มที่มีค่า Distance ที่น้อยที่สุด
4. ทำการคำนวณหาค่าจุดศูนย์กลางของกลุ่มใหม่ แล้วดูว่าค่าของจุดศูนย์กลางในแต่ละกลุ่มที่หาได้นั้นมีการเปลี่ยนแปลงไปจากเดิมหรือไม่ ถ้าไม่เปลี่ยนก็เป็นการเสร็จสิ้นกระบวนการทำงาน แต่ถ้าเปลี่ยนก็ให้ย้อนกลับไปทำตั้งแต่ข้อที่ 2 ใหม่

<u>การทำไมน์นิ่งสำหรับการทดสอบโมเดล</u>

1. ทำการคำนวณหาค่า Distance หรือระยะห่างของข้อมูลแต่ละตัวเปรียบเทียบกับจุดศูนย์กลางในแต่ละกลุ่ม (โดยจุดศูนย์กลางของข้อมูลได้มาจากโมเดลที่จะนำไปใช้ในการทดสอบข้อมูล)

2. ทำการกำหนดกลุ่มของข้อมูล โดยได้จากการนำเอาค่า Distance ที่หาได้ในแต่ละกลุ่มมาเปรียบเทียบกัน โดยจัดข้อมูลเข้าไปไว้ในกลุ่มที่มีค่า Distance ที่น้อยที่สุด

4.2.3 ส่วนของการแสดงผล

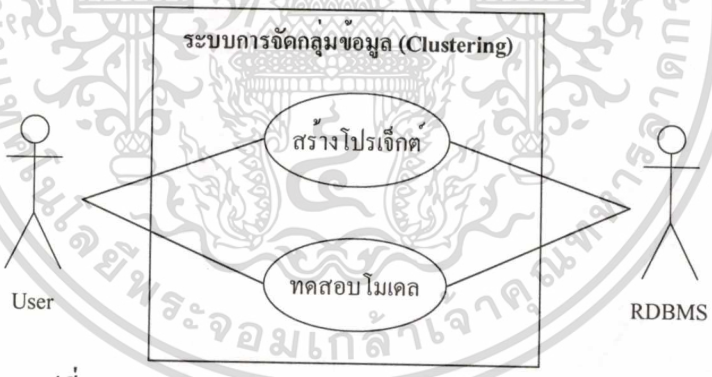
หลังจากที่ได้ทำการจัดกลุ่มแล้วจะทำการแสดงผลข้อมูลออกมาทางหน้าจอ โดยที่ข้อมูลที่นำมาแสดงนั้นจะคล้ายกันสำหรับการแสดงผลหลังการสร้างโปรเจ็กต์ และการแสดงผลหลังการทดสอบโมเดล โดยมีขั้นตอนดังนี้

- แสดงจุดศูนย์กลางของกลุ่ม
- แสดงข้อมูลในแต่ละกลุ่ม

4.3 ขั้นตอนการทำงานของระบบ

4.3.1 Use Case Diagram

ในการออกแบบระบบจะใช้ Use Case Diagram อธิบายการทำงานของระบบการจัดกลุ่มข้อมูล ซึ่งจะแสดงให้เห็นในรูปที่ 4.1



รูปที่ 4.1 แสดง Use Case Diagram ในการทำงานของระบบ

4.3.2 Sequence Diagram

Sequence Diagram สามารถแบ่งพิจารณาได้ 2 ส่วนหลักๆ ดังนี้

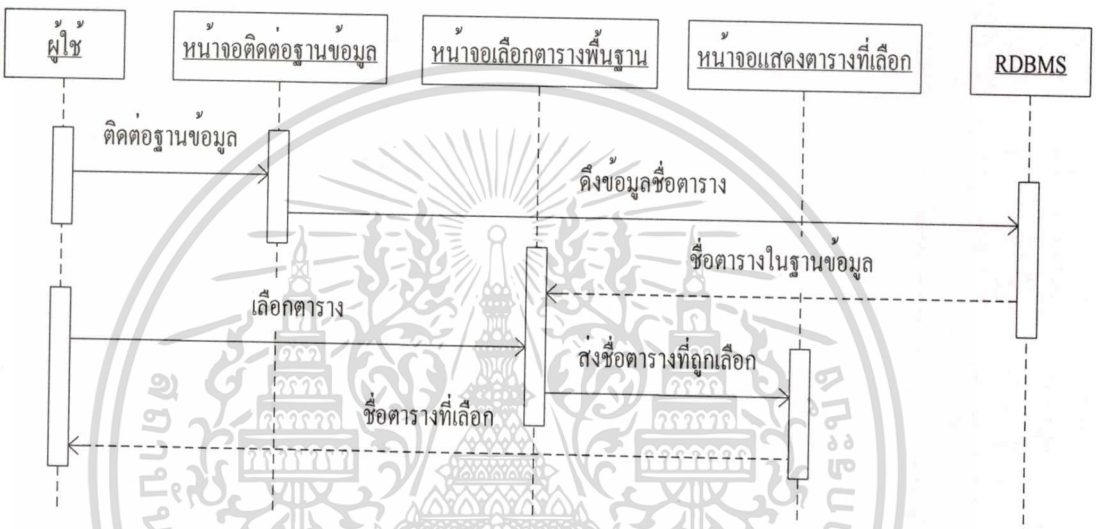
4.3.2.1 Sequence Diagram ของการสร้างโปรเจ็กต์

1) Sequence Diagram ของการติดต่อฐานข้อมูล

- ผู้ใช้จะต้องทำการติดต่อไปยังฐานข้อมูล โดยการเรียกใช้หน้าจอติดต่อฐานข้อมูลแล้วพิมพ์ชื่อเซิฟเวอร์และชื่อฐานข้อมูลที่ต้องการจะติดต่อ
- จากนั้นหน้าจอติดต่อฐานข้อมูลจะทำการติดต่อไปยัง RDBMS ตามชื่อเซิฟเวอร์และชื่อฐานข้อมูลที่ผู้ใช้ระบุเอาไว้ เพื่อทำการดึงข้อมูลชื่อตารางที่มีในฐานข้อมูลนั้นออกมา

- แล้วตัว RDBMS จะทำการแสดงชื่อตารางที่มีอยู่ในฐานข้อมูลทั้งหมดไปยังหน้าขอเลือกตารางพื้นฐาน

- ให้ผู้ใช้ทำการเลือกตารางที่ต้องการใช้จากหน้าจอเลือกตารางพื้นฐาน แล้วข้อมูลชื่อตารางที่ถูกเลือกจะถูกส่งไปแสดงที่หน้าจอแสดงตารางที่เลือก เพื่อแสดงตารางที่ผู้ใช้ทำการเลือกอีกครั้ง



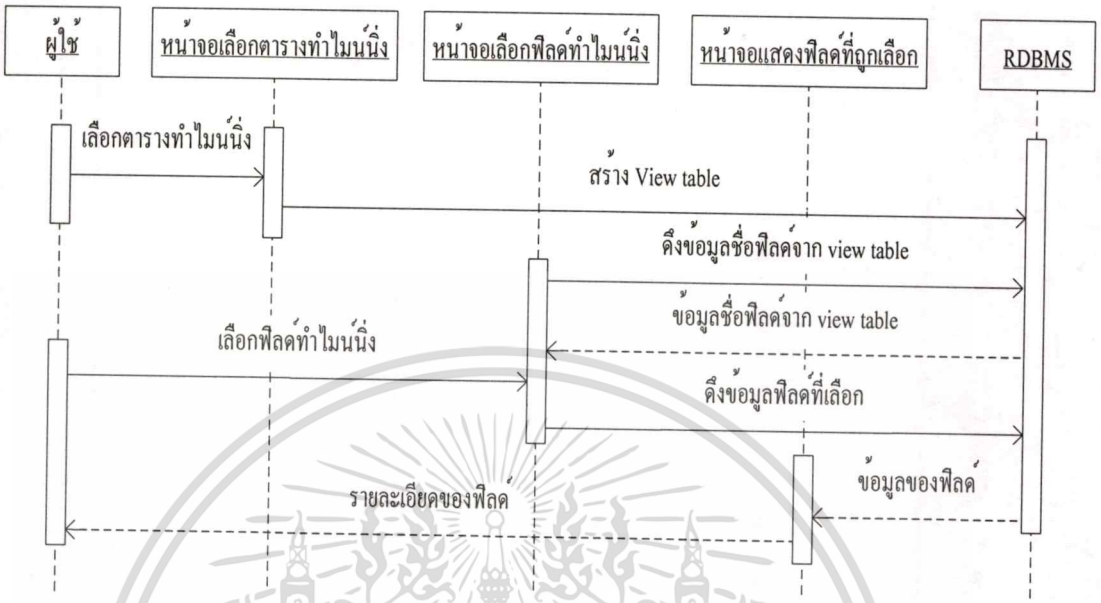
รูปที่ 4.2 แสดง Sequence Diagram ของการติดต่อฐานข้อมูล

2) Sequence Diagram ของการเลือกข้อมูลสำหรับทำไมน์นิ่ง

- ให้ผู้ใช้ทำการเลือกตารางที่ต้องการทำไมน์นิ่งที่หน้าจอเลือกตารางทำไมน์นิ่ง จากนั้นหน้าจอเลือกตารางทำไมน์นิ่งจะทำการติดต่อไปยัง RDBMS เพื่อทำการสร้าง View table

- ที่หน้าจอเลือกฟิลด์ทำไมน์นิ่งจะแสดงข้อมูลชื่อฟิลด์ที่มีทั้งหมดใน View table เพื่อให้ผู้ใช้ทำการเลือกฟิลด์ที่ต้องการใช้ในการทำไมน์นิ่ง

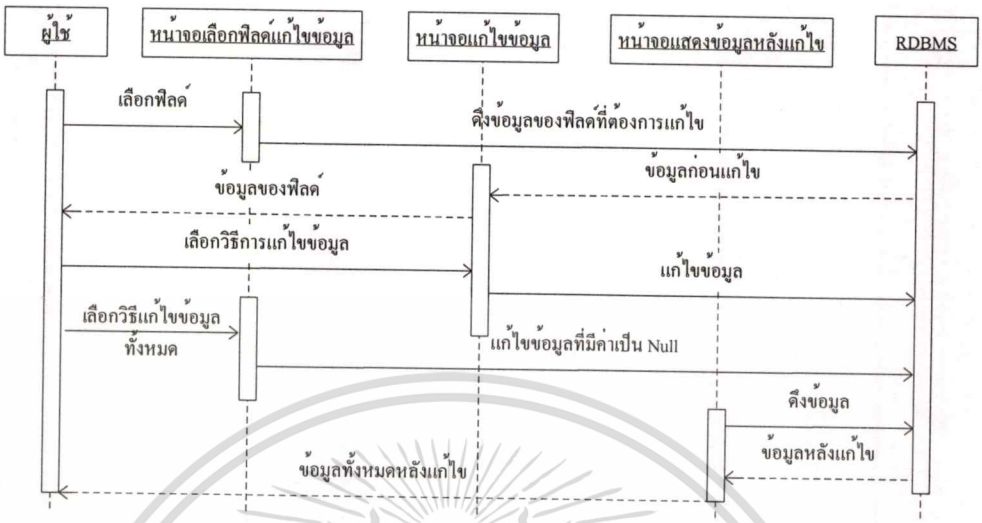
- ให้ผู้ใช้ทำการเลือกฟิลด์ที่ต้องการใช้สำหรับทำไมน์นิ่งที่หน้าจอเลือกฟิลด์ทำไมน์นิ่ง แล้วที่หน้าจอนี้จะทำการติดต่อไปยัง RDBMS เพื่อดึงข้อมูลของฟิลด์ที่จะใช้ทำไมน์นิ่งออกมาแสดงที่หน้าจอแสดงฟิลด์ที่ถูกเลือก เพื่อให้ผู้ใช้ตรวจสอบฟิลด์ที่เลือกไปว่าถูกต้องหรือไม่ และเพื่อให้ทราบถึงชนิดของข้อมูลในแต่ละฟิลด์นั้น



รูปที่ 4.3 แสดง Sequence Diagram ของการเลือกข้อมูลสำหรับทำไมน์นิ่ง

3) Sequence Diagram ของการแก้ไขข้อมูล

- ให้ผู้ใช้ทำการเลือกฟิลด์ที่ต้องการแก้ไขที่หน้าจอเลือกฟิลด์แก้ไขข้อมูล จากนั้นที่หน้าจอเลือกฟิลด์แก้ไขข้อมูลจะทำการดึงข้อมูลของฟิลด์ที่ต้องการแก้ไขมาจาก RDBMS
- ตัว RDBMS จะแสดงข้อมูลของฟิลด์ที่ผู้ใช้ต้องการแก้ไขไปที่หน้าจอแก้ไขข้อมูล
- จากนั้นให้ผู้ใช้ทำการเลือกวิธีที่ต้องการแก้ไขข้อมูลที่หน้าจอแก้ไขข้อมูล แล้วหน้าจอแก้ไขข้อมูลจะทำการแก้ไขข้อมูลไปยัง RDBMS ซึ่งในกรณีเช่นนี้จะเป็นการแก้ไขข้อมูลทีละฟิลด์
- แต่ถ้าผู้ใช้ต้องการแก้ไขข้อมูลทั้งหมดในคราวเดียวก็สามารถเลือกวิธีการแก้ไขข้อมูลทั้งหมดที่หน้าจอเลือกฟิลด์แก้ไขข้อมูลได้เลย แล้วมันจะทำการแก้ไขข้อมูลที่มีค่าเป็นค่า Null ในฐานข้อมูลให้
- ข้อมูลที่ได้หลังจากการแก้ไขจะแสดงที่หน้าจอแสดงข้อมูลหลังการแก้ไข



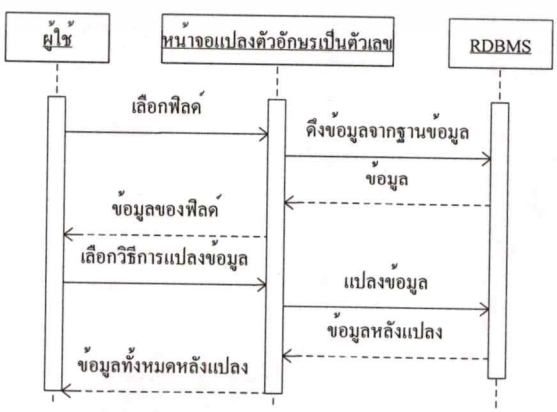
รูปที่ 4.4 แสดง Sequence Diagram ของการแก้ไขข้อมูล

4) Sequence Diagram ของการแปลงข้อมูลตัวอักษรเป็นตัวเลข

- ให้ผู้ใช้ทำการเลือกฟิลด์ที่ต้องการแปลงที่หน้าจอแปลงตัวอักษรเป็นตัวเลข จากนั้นที่หน้าจอแปลงตัวอักษรเป็นตัวเลขจะดึงข้อมูลของฟิลด์มาจากฐานข้อมูลเพื่อมาแสดงที่หน้าจอแปลงตัวอักษรเป็นตัวเลข

- ให้ผู้ใช้เลือกวิธีการแปลงข้อมูลที่หน้าจอแปลงตัวอักษรเป็นตัวเลขเพื่อทำการแปลงข้อมูล แล้วที่หน้าจอแปลงตัวอักษรเป็นตัวเลขจะสั่งให้ RDBMS ทำการแก้ไขข้อมูลในฐานข้อมูล

- ผลลัพธ์จากการแปลงจะถูกแสดงที่หน้าจอแปลงตัวอักษรเป็นตัวเลข



รูปที่ 4.5 แสดง Sequence Diagram ของการแปลงข้อมูลตัวอักษรเป็นตัวเลข

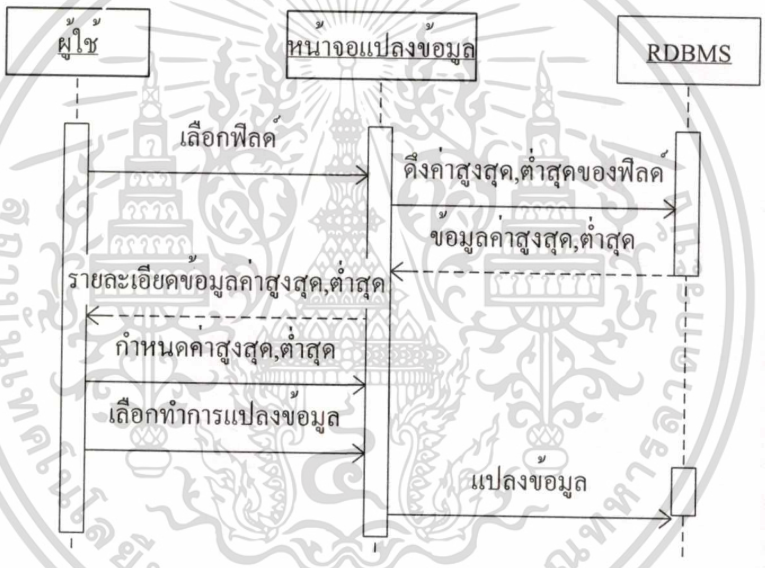
5) Sequence Diagram ของการแปลงข้อมูล

- ให้ผู้ใช้ทำการเลือกฟิลด์ที่ต้องการแปลงที่หน้าจอแปลงข้อมูล

- หน้าจอแปลงข้อมูลจะทำการดึงค่าสูงสุด, ต่ำสุดของฟิลด์จาก RDBMS ขึ้นมาแสดงที่หน้าจอแปลงข้อมูล

- ให้ผู้ใช้ทำการกำหนดค่าสูงสุด, ต่ำสุดที่ต้องใช้ในการแปลงข้อมูล แล้วเลือกทำการแปลงข้อมูลที่หน้าจอแปลงข้อมูล

- หน้าจอแปลงข้อมูลทำการแปลงข้อมูลไปยัง RDBMS



รูปที่ 4.6 แสดง Sequence Diagram ของการแปลงข้อมูล

6) Sequence Diagram ของการจัดกลุ่มข้อมูล

- ให้ผู้ใช้ทำการกำหนดจำนวนกลุ่มที่ต้องการจัดกลุ่มที่หน้าจอการจัดกลุ่มข้อมูล

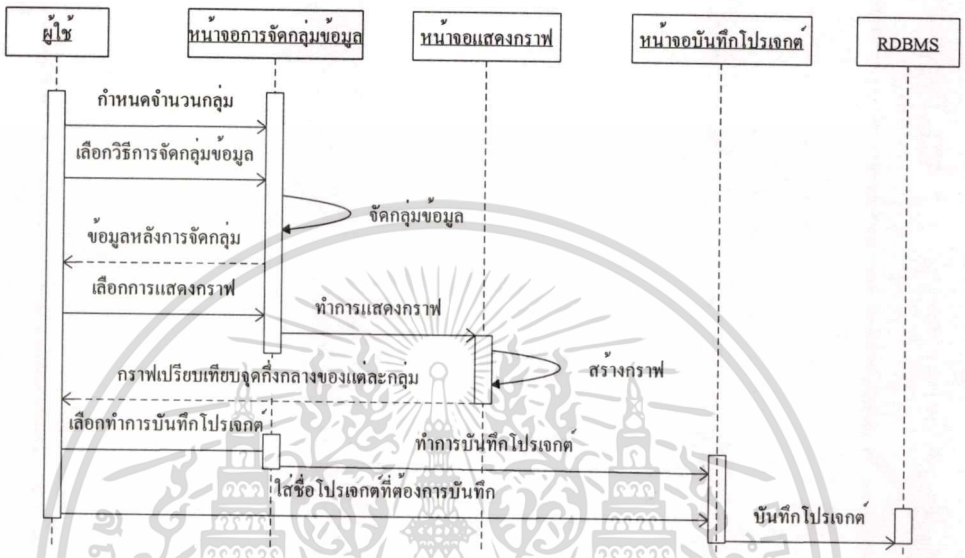
- จากนั้นให้ผู้ใช้ทำการเลือกวิธีการที่จะใช้ในการจัดกลุ่มที่หน้าจอการจัดกลุ่มข้อมูล

- หน้าจอการจัดกลุ่มข้อมูลจะแสดงผลของการจัดกลุ่มให้ผู้ใช้ทราบ

- ในกรณีที่ผู้ใช้ต้องการดูกราฟก็ทำได้โดยการเลือกการแสดงกราฟที่หน้าจอการจัดกลุ่มข้อมูล จากนั้นหน้าจอการจัดกลุ่มจะสั่งให้หน้าจอแสดงกราฟทำการแสดงกราฟให้แก่ผู้ใช้

- ให้ผู้ใช้ทำการบันทึกโปรเจ็กต์โดยเลือกทำการบันทึกโปรเจ็กต์ที่หน้าจอการจัดกลุ่มข้อมูล

- จากนั้นให้ผู้ใช้ใส่ชื่อโปรเจ็กต์ที่ต้องการบันทึกในหน้าจอบันทึกโปรเจ็กต์เพื่อทำการบันทึก แล้วหน้าจอบันทึกโปรเจ็กต์จะทำการบันทึกโปรเจ็กต์นี้เข้าไปใน RDBMS



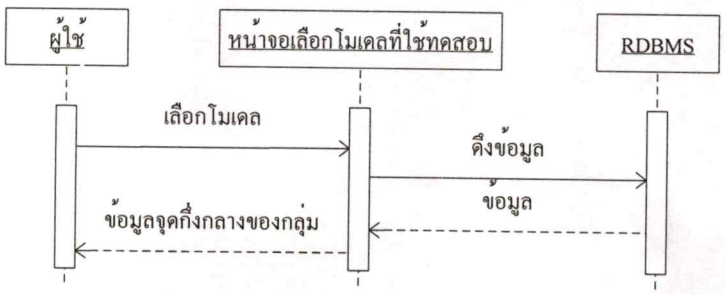
รูปที่ 4.7 แสดง Sequence Diagram ของการจัดกลุ่มข้อมูล

4.3.2.2 Sequence Diagram ของการสร้างทดสอบโมเดล

1) Sequence Diagram ของการเลือกโมเดลที่ใช้ในการทดสอบ

- ให้ผู้ใช้ทำการเลือกโมเดลที่ต้องการใช้ในการทดสอบที่หน้าจอเลือกโมเดลที่ใช้ทดสอบ

- ที่หน้าจอเลือกโมเดลที่ใช้ทดสอบจะทำการดึงข้อมูลจากฐานข้อมูลมาแสดงให้ผู้ใช้ดู



รูปที่ 4.8 แสดง Sequence Diagram ของการเลือกโมเดลที่ใช้ในการทดสอบ

2) Sequence Diagram ของการทดสอบโมเดล

- ให้ผู้ใช้ทำการติดต่อฐานข้อมูลที่หน้าจอติดต่อฐานข้อมูลเพื่อทำการดึงข้อมูลชื่อตารางออกมาจาก RDBMS

- หน้าจอติดต่อฐานข้อมูลจะทำการดึงข้อมูลชื่อตารางจาก RDBMS มาแสดงที่หน้าจอแม็พข้อมูล

- ให้ผู้ใช้ทำการเลือกตารางที่ต้องการแม็พข้อมูลที่หน้าจอแม็พข้อมูล

- หน้าจอแม็พข้อมูลจะทำการดึงข้อมูลชื่อฟิลด์จาก RDBMS มาแสดงที่หน้าจอแม็พข้อมูล

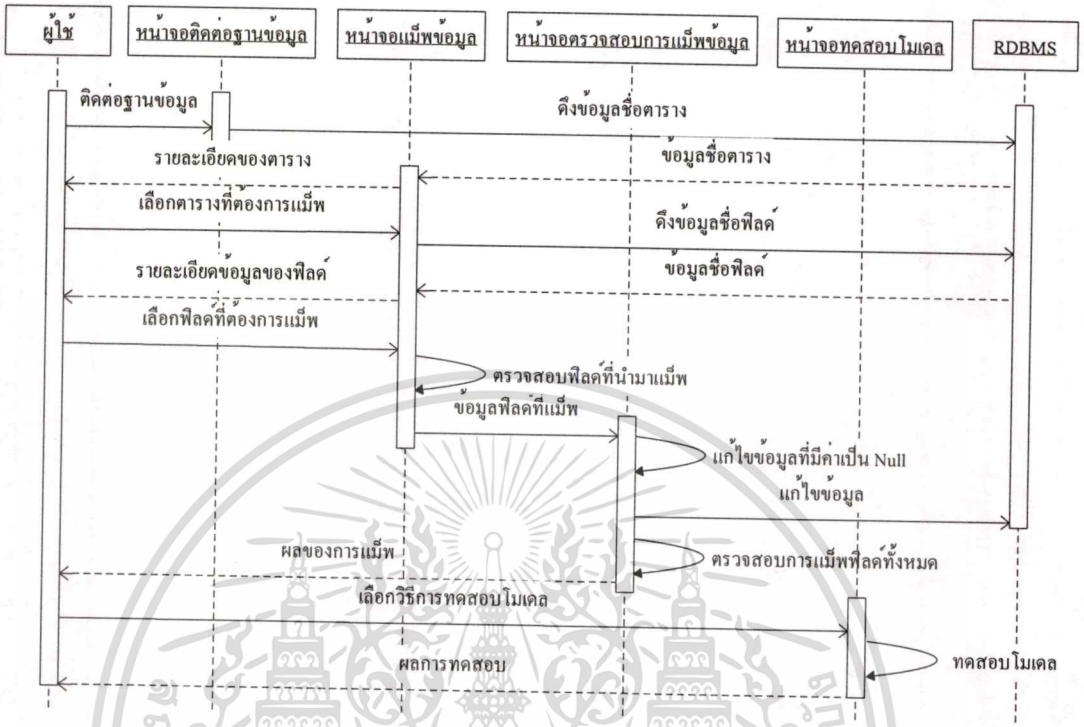
- จากนั้นให้ผู้ใช้ทำการเลือกฟิลด์ที่ต้องการแม็พเพื่อใช้ในการทดสอบโมเดลที่หน้าจอแม็พข้อมูล

- หน้าจอแม็พข้อมูลจะทำการตรวจสอบฟิลด์ที่นำมาแม็พว่าสามารถแม็พกันได้หรือไม่

- หน้าจอแม็พข้อมูลจะส่งข้อมูลฟิลด์ที่แม็พไปที่หน้าจอตรวจสอบการแม็พข้อมูลแล้วที่หน้าจอตรวจสอบการแม็พข้อมูลจะทำการแก้ไขข้อมูลที่มีค่าเป็น Null ใน RDBMS

- จากนั้นหน้าจอตรวจสอบการแม็พข้อมูลจะทำการตรวจสอบการแม็พฟิลด์ทั้งหมดว่าสามารถแม็พกันได้หรือไม่ แล้วจะแสดงผลของการแม็พให้ผู้ใช้งานทราบว่าข้อมูลสามารถแม็พกันได้หรือไม่

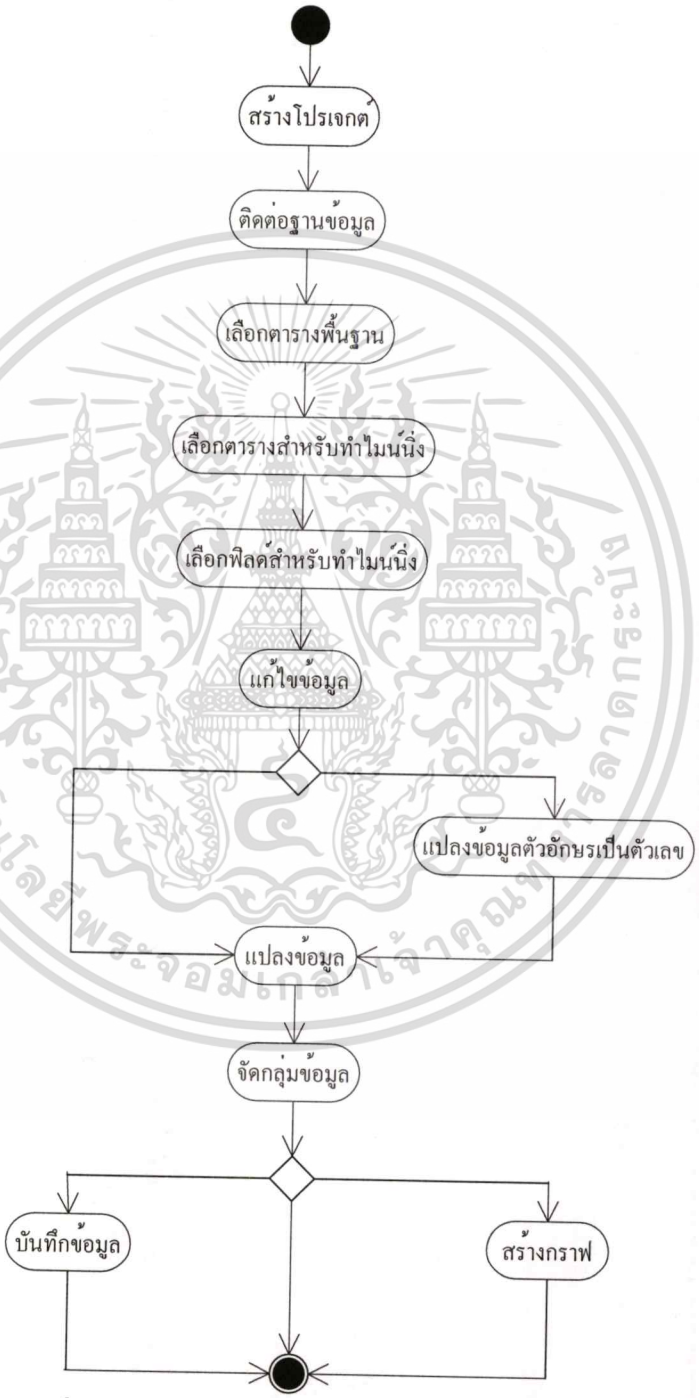
- ผู้ใช้ทำการเลือกวิธีการมดสอบโมเดลที่หน้าจอทดสอบโมเดล แล้วหน้าจอทดสอบโมเดลจะทำการทดสอบโมเดลแล้วแสดงผลลัพธ์ที่ได้จากการทดสอบให้แก่ผู้ใช้ดู



รูปที่ 4.9 แสดง Sequence Diagram ของการทดสอบโมเดล

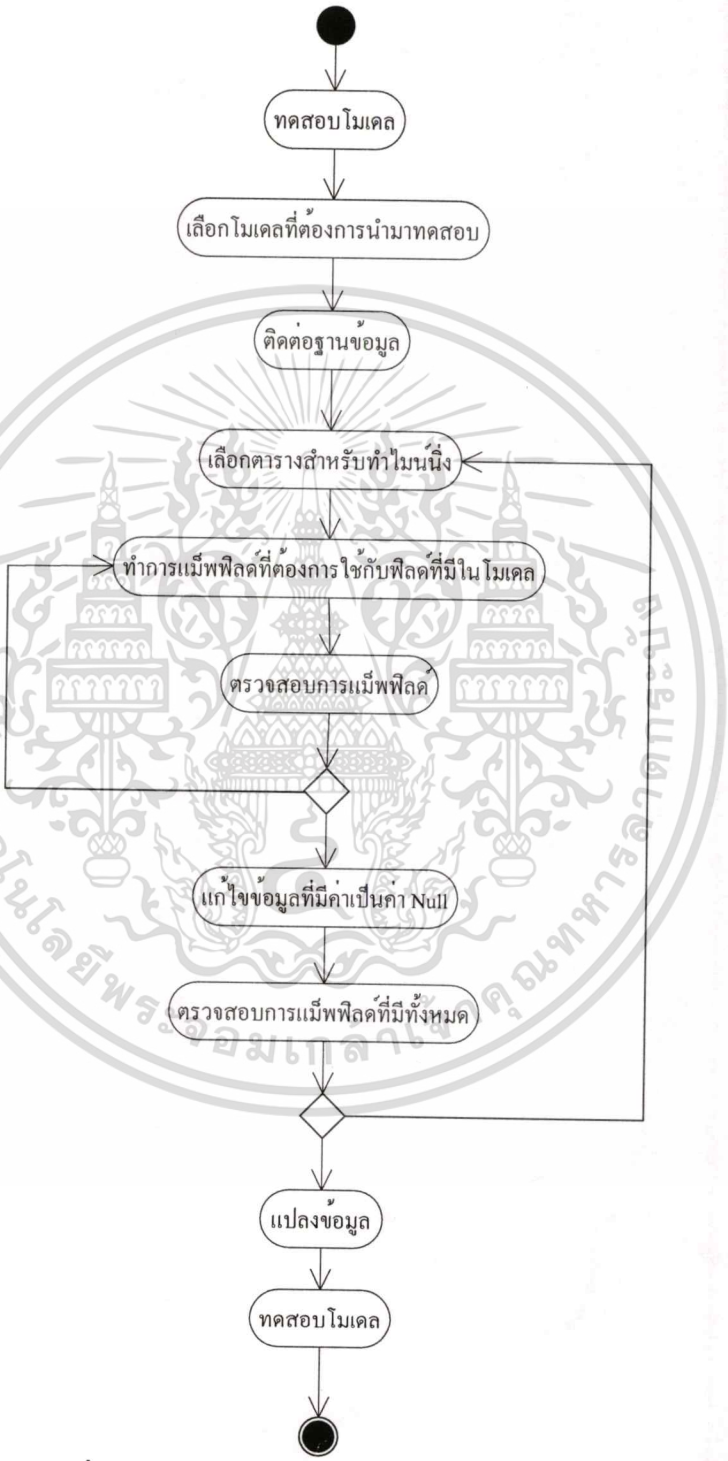
4.3.3 Activity Diagram

1) Activity diagram ของการสร้างโปรเจ็กต์



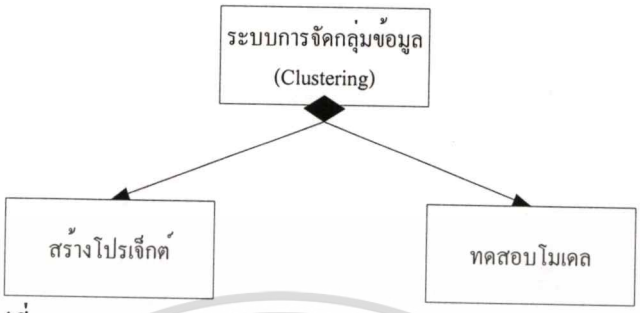
รูปที่ 4.10 แสดง Activity Diagram ของการสร้างโปรเจ็กต์

2) Activity diagram ของการทดสอบโมเดล



รูปที่ 4.11 แสดง Activity Diagram ของการทดสอบโมเดล

4.3.4 Structure Chart

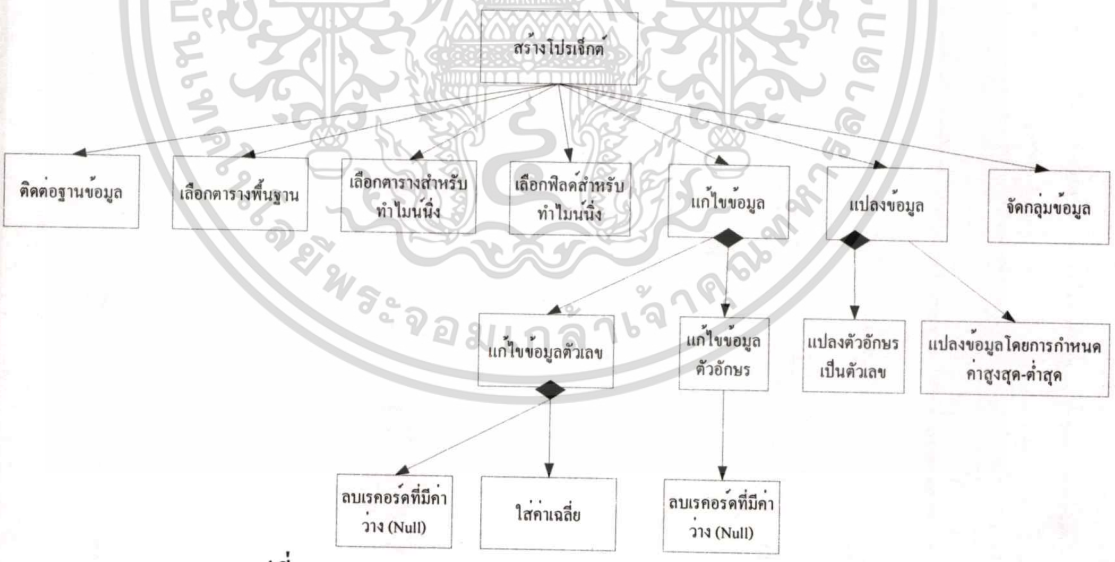


รูปที่ 4.12 แสดง Structure Chart ของระบบการจัดกลุ่มข้อมูล

โครงสร้างหลักของระบบการจัดกลุ่มข้อมูล (Clustering) จะประกอบด้วย 2 ส่วนหลักๆ คือ

1. สร้างโปรเจ็กต์
2. ทดสอบโมเดล

โดยส่วนของการสร้างโปรเจ็กต์จะมี module ย่อยๆ ดังรูปที่ 4.13 ข้างล่างนี้



รูปที่ 4.13 แสดง Structure Chart ของการสร้างโปรเจ็กต์

ซึ่งใน Structure Chart ของการสร้างโปรเจ็กต์นี้จะประกอบด้วย 7 Module หลักๆ ดังต่อไปนี้

1. ติดต่อฐานข้อมูล เพื่อติดต่อกับฐานข้อมูลที่ผู้ใช้ต้องการใช้สำหรับสร้างโปรเจ็กต์
2. เลือกตารางพื้นฐาน คือระบบจะแสดงชื่อตารางที่มีในฐานข้อมูลที่ผู้ใช้เลือก แล้วหลังจากนั้นก็ให้ผู้ใช้เลือกตารางพื้นฐานสำหรับนำไปใช้งานในขั้นถัดไป

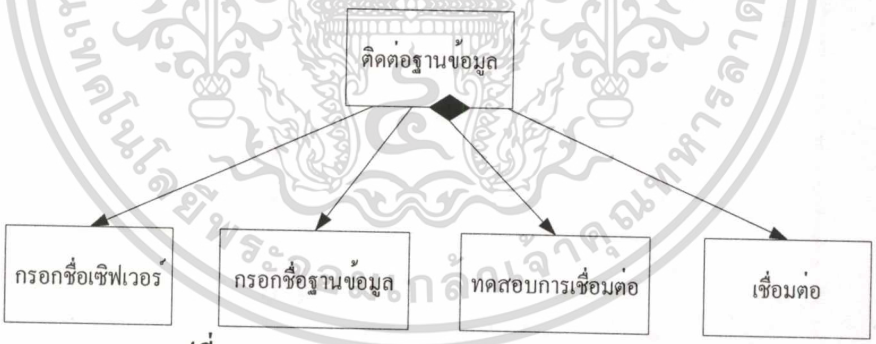
3. เลือกตารางสำหรับทำไมน์นิ่ง คือให้ผู้ใช้ทำการเลือกตารางที่จะใช้ทำไมน์นิ่งจากตารางพื้นฐานที่เลือกในขั้นที่แล้วมา 1 ตาราง

4. เลือกฟิลด์สำหรับทำไมน์นิ่ง โดยหลังจากที่ผู้ใช้ได้เลือกตารางสำหรับใช้ทำไมน์นิ่งแล้ว ระบบจะแสดงชื่อฟิลด์ทั้งหมดที่มีในตารางนั้น แล้วให้ผู้ใช้เลือกฟิลด์ที่ต้องการจะใช้ทำไมน์นิ่ง เพื่อทำงานในขั้นถัดไป

5. แก้ไขข้อมูล ถ้าชนิดของฟิลด์เป็นตัวเลขก็จะเลือก module แก้ไขข้อมูลตัวเลข ซึ่งการแก้ไขข้อมูลที่เป็นตัวเลขนั้นทำได้โดยการลบเรคอร์ดที่มีค่าว่าง (Null) หรือใส่ค่าเฉลี่ย แต่ถ้าชนิดของฟิลด์เป็นตัวอักษรก็จะเลือก module ของการแก้ไขข้อมูลตัวอักษรซึ่งทำได้โดยลบเรคอร์ดที่มีค่าว่าง (Null) ทำเช่นนี้จนครบทุกฟิลด์

6. แปลงข้อมูล ถ้าฟิลด์ที่เลือกมีชนิดของฟิลด์เป็นตัวอักษรก็ให้ทำการแปลงตัวอักษรเป็นตัวเลข โดยระบบจะแสดงค่าที่ไม่ซ้ำกันในฟิลด์นั้นๆ แล้วจะแทนค่าตัวเลขที่กำหนดโดยระบบให้กับแต่ละค่าของฟิลด์นั้น หลังจากนั้นผู้ใช้ก็จะต้องทำการแปลงค่าข้อมูลทุกฟิลด์อีกครั้ง โดยผู้ใช้จะต้องกำหนดค่าสูงสุดและต่ำสุดของฟิลด์ที่ต้องการจะแปลง

7. จัดกลุ่มข้อมูล



รูปที่ 4.14 แสดง Structure Chart ของการติดต่อฐานข้อมูล

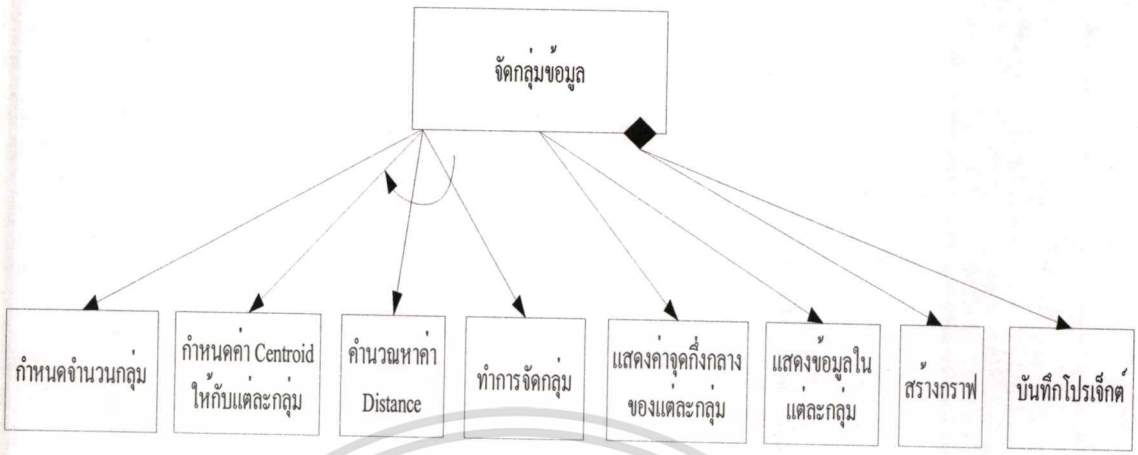
ใน Structure Chart ของการติดต่อฐานข้อมูลนี้จะประกอบด้วย module ย่อย 4 module คือ

1. กรอกชื่อเซิฟเวอร์ ให้ผู้ใช้ทำการกรอกชื่อเซิฟเวอร์ที่ต้องการติดต่อ

2. กรอกชื่อฐานข้อมูล คือหลังจากที่กรอกชื่อเซิฟเวอร์แล้วผู้ใช้จะต้องกรอกชื่อฐานข้อมูลที่ต้องการใช้สำหรับสร้างโปรเจ็กต์หรือทดสอบโมเดลด้วย

3. ทดสอบการเชื่อมต่อ ในกรณีที่ผู้ใช้ต้องการทดสอบว่าการเชื่อมต่อฐานข้อมูลนั้นสำเร็จหรือไม่ก็ทำได้โดยการเลือก module นี้เพื่อทดสอบการเชื่อมต่อ

4. เชื่อมต่อ จะเป็นการเชื่อมต่อไปยังฐานข้อมูลและเซิฟเวอร์ที่ผู้ใช้ต้องการ



รูปที่ 4.15 แสดง Structure Chart ของการจัดกลุ่มข้อมูล

จากรูปข้างต้นแสดงให้เห็นถึง Structure Chart ของการจัดกลุ่มข้อมูล โดยจะมี module ย่อย 8 module ดังต่อไปนี้

1. กำหนดจำนวนกลุ่ม คือผู้ใช้จะต้องทำการกำหนดจำนวนกลุ่มที่จะใช้สำหรับทำการจัดกลุ่ม ในที่นี้สมมติให้เป็น K กลุ่ม

2. กำหนดค่า Centroid ให้กับแต่ละกลุ่ม โดยถ้าหากว่าเป็นการทำงานในรอบแรกจะกำหนดค่า Centroid เริ่มต้นโดยกำหนดให้ใช้เป็น K ตัวแรก แต่ถ้าเป็นการทำงานในรอบถัดๆ ไปจะใช้ค่า mean ในแต่ละกลุ่มเป็นค่า Centroid

3. คำนวณหาค่า Distance

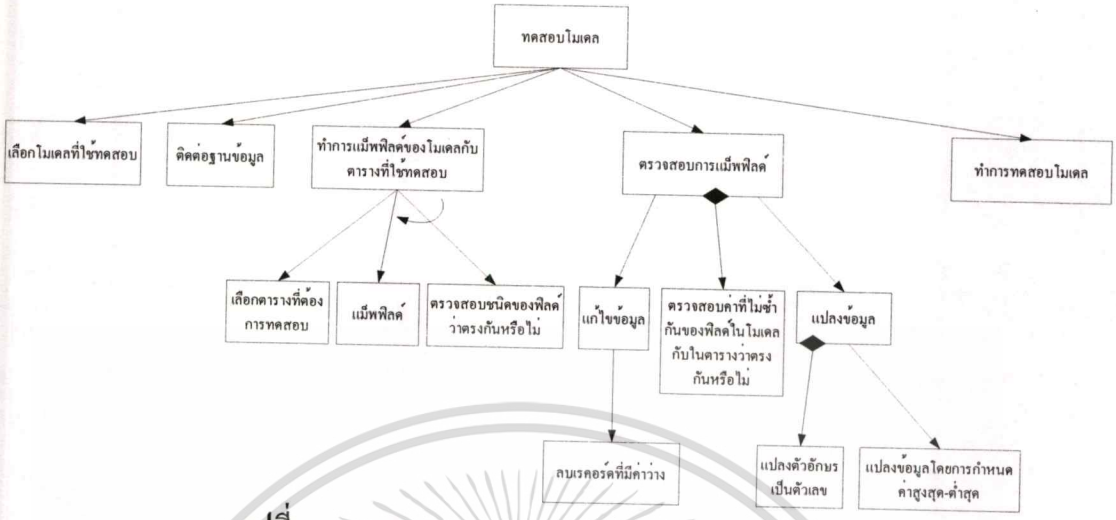
4. ทำการจัดกลุ่ม โดยจัดข้อมูลแต่ละตัวเข้าไปไว้ในกลุ่มที่มีค่าใกล้เคียงกับค่า Centroid ของกลุ่มนั้นมากที่สุด ซึ่งจะต้องทำให้ครบทุกตัวแล้วจะกลับไปเริ่มทำงานในข้อ 2 ใหม่ จนกระทั่งข้อมูลไม่มีการเปลี่ยนกลุ่มหรือค่า Centroid ไม่เปลี่ยน

5. แสดงค่าจุดกึ่งกลางของแต่ละกลุ่ม โดยระบบจะแสดงค่าจุดกึ่งกลางหรือค่า Centroid ของแต่ละกลุ่มให้ผู้ใช้ดู

6. แสดงข้อมูลในแต่ละกลุ่ม ซึ่งจะแสดงว่าในแต่ละกลุ่มมีสมาชิกเป็นอะไรบ้าง

7. สร้างกราฟ แสดงการเปรียบเทียบค่าจุดกึ่งกลางของแต่ละกลุ่ม

8. บันทึกโปรเจ็กต์ เพื่อนำไปใช้ในขั้นของการทดสอบโมเดล



รูปที่ 4.16 แสดง Structure Chart ของการทดสอบโมเดล

Structure Chart ของการทดสอบโมเดลนี้จะมี module ย่อยด้วยกัน 5 module หลักๆ ดังนี้

1. เลือกโมเดลที่ใช้ทดสอบ คือเลือก โมเดลที่จะนำไปใช้ในการทดสอบโมเดล

2. ติดต่อฐานข้อมูล เหมือนกับการติดต่อฐานข้อมูลในส่วนของการสร้างโปรเจ็กต์ เพื่อติดต่อเข้าไปในฐานข้อมูลที่ต้องการดึงเอาข้อมูลไปทดสอบ

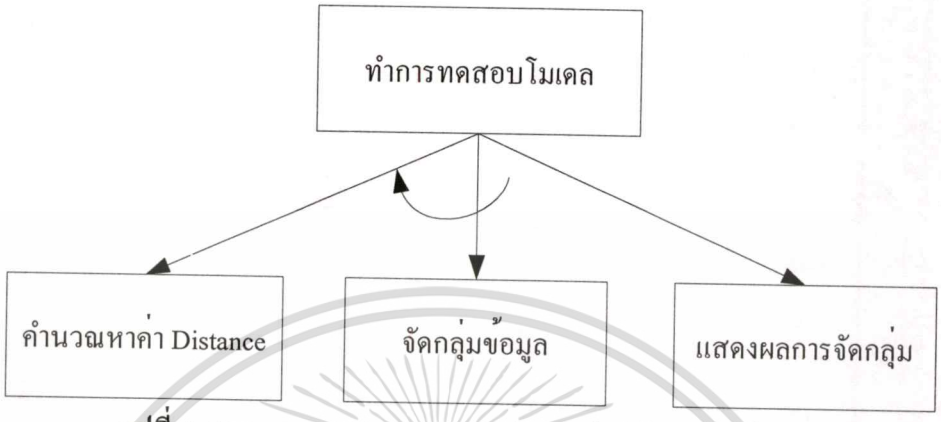
3. ทำการแม็พฟิลด์ของโมเดลกับตารางที่ใช้ทดสอบ แบ่งเป็น 3 module ย่อย ดังต่อไปนี้

- เลือกตารางที่ต้องการทดสอบ เพื่อเลือกตารางที่จะนำไปใช้ในการทดสอบกับโมเดล
- แม็พฟิลด์ โดยการเลือกฟิลด์ที่มีในโมเดลกับฟิลด์ที่มีในตาราง
- ตรวจสอบชนิดของฟิลด์ว่าตรงกันหรือไม่ ระบบทำการตรวจสอบว่าชนิดของฟิลด์ที่เลือกในโมเดลกับในตารางตรงกันหรือไม่ ถ้าไม่ตรงก็ให้ทำการเลือกฟิลด์เพื่อทำการแม็พใหม่ แต่ถ้าชนิดของฟิลด์ตรงกันก็ให้ทำการแม็พฟิลด์ถัดไป ทำจนกระทั่งครบทุกฟิลด์ในโมเดล

4. ตรวจสอบการแม็พฟิลด์ โดยในส่วนนี้จะประกอบด้วย 3 ส่วนหลักๆ ดังนี้

- แก้ไขข้อมูลโดยการลบเรคอร์ดที่มีค่าว่าง (Null) ทิ้งไป
- ตรวจสอบค่าที่ไม่ซ้ำกันของฟิลด์ในโมเดลกับในตารางว่าตรงกันหรือไม่ ในกรณีที่ชนิดของข้อมูลเป็นตัวอักษรจึงจะทำการตรวจสอบ โดยตรวจว่าค่าที่ไม่ซ้ำกันของฟิลด์ในโมเดลกับในตารางที่ใช้ทำการทดสอบว่ามันตรงกันหรือไม่ ถ้าไม่ตรงก็ให้ทำการเลือกตารางเพื่อทำการแม็พฟิลด์ใหม่ แต่ถ้าตรงกันก็ไปทำงานยังขั้นตอนถัดไป
- แปลงข้อมูล ถ้าชนิดของฟิลด์เป็นตัวอักษรก็ให้ทำการแปลงตัวอักษรเป็นตัวเลข โดยระบบจะแทนค่าตัวเลขให้กับค่าที่ไม่ซ้ำกันของฟิลด์ตามค่าที่มีในโมเดลที่นำมาทดสอบ และระบบจะทำการแปลงค่าสูงสุด-ต่ำสุดให้กับฟิลด์ของตารางที่จะใช้ทดสอบตามค่าสูงสุด-ต่ำสุดของฟิลด์ที่แม็พกับโมเดลนั้นๆ

5. ทดสอบโมเดล



รูปที่ 4.17 แสดง Structure Chart ของการทำการทดสอบโมเดล

โดยใน Structure Chart ของการทำการทดสอบโมเดลนี้จะแบ่งเป็น 3 module ดังนี้

1. คำนวณหาค่า Distance โดยการหาค่า Distance ของข้อมูลแต่ละตัวเปรียบเทียบกับค่า Centroid ในแต่ละกลุ่มของโมเดลที่นำมาทดสอบ

2. จัดกลุ่มข้อมูล โดยจัดข้อมูลนั้นไปไว้ในกลุ่มที่ใกล้เคียงกับกลุ่มนั้นมากที่สุด คือจัดไปไว้ในกลุ่มที่มีค่า Distance ที่น้อยที่สุด

3. แสดงผลการจัดกลุ่ม คือแสดงผลการทดสอบโมเดลว่าข้อมูลที่ได้จะอยู่กลุ่มใด

4.4 Data Dictionary

ตารางที่ 4.1 RuleName : เป็นตารางซึ่งเก็บชื่อโปรเจ็กต์

Attribute Name	Contents	Type	Size	Key
Rule_id	หมายเลขของโปรเจ็กต์	int	4	
Rule_Name	ชื่อโปรเจ็กต์	varchar	50	

ตารางที่ 4.2 FieldName : เก็บชื่อและชนิดของฟิลด์ที่ถูกใช้ในโปรเจ็กต์นั้นๆ

Attribute Name	Contents	Type	Size	Key
Rule_id	หมายเลขของโปรเจ็กต์	int	4	
Field_Name	ชื่อฟิลด์	varchar	50	
Field_Type	ชนิดของฟิลด์	varchar	50	

ตารางที่ 4.3 TbMinMax : เก็บค่าน้อยสุดและมากสุดที่ใช้ทำการแปลงข้อมูลของแต่ละฟิลด์

Attribute Name	Contents	Type	Size	Key
Rule_id	หมายเลขของโปรเจ็กต์	int	4	
Field_Name	ชื่อฟิลด์	varchar	50	
T_Min	ค่าน้อยสุดที่ใช้สำหรับแปลงข้อมูล	real	4	
T_Max	ค่ามากสุดที่ใช้สำหรับแปลงข้อมูล	real	4	

ตารางที่ 4.4 TbDistinct: เก็บค่าที่เป็นไปได้ของแต่ละฟิลด์

Attribute Name	Contents	Type	Size	Key
Rule_id	หมายเลขของโปรเจ็กต์	int	4	
Field_Name	ชื่อฟิลด์	varchar	50	
Distinct_value	ค่าที่เป็นไปได้ของฟิลด์	varchar	50	

ตารางที่ 4.5 TbCentroid : เก็บค่าจุดศูนย์กลางที่ได้จากการจัดกลุ่มของโปรเจ็กต์นั้นๆ

Attribute Name	Contents	Type	Size	Key
Rule_id	หมายเลขของโปรเจ็กต์	int	4	
Field_Name	ชื่อฟิลด์	varchar	50	
group_id	หมายเลขกลุ่ม	int	4	
centroid_value	ค่าจุดศูนย์กลางที่หาได้	real	4	
Data_value	ค่าของข้อมูลชนิดตัวอักษรที่เป็นไปได้ของกลุ่ม	varchar	50	

บทที่ 5
การประยุกต์ใช้ดาต้าไมน์นิ่งเพื่อทำการจัดกลุ่มข้อมูล

5.1 กำหนดวัตถุประสงค์

เนื่องจากในปัจจุบันธุรกิจที่เป็นการให้บริการเกี่ยวกับโทรศัพท์เคลื่อนที่นั้น จำเป็นที่จะต้องนำเอาข้อมูลของการใช้โทรศัพท์ของลูกค้ามาทำการวิเคราะห์เพื่อจัดทำแผนกลยุทธ์ทางการตลาด ที่จำเป็นในการกำหนดโปรโมชั่นให้สามารถตอบสนองแก่ลูกค้าได้ และเพื่อดึงดูดความสนใจของลูกค้าให้หันมาใช้บริการของตนให้มากยิ่งขึ้น ซึ่งจะส่งผลให้บริษัทนั้นสามารถแข่งขันกับคู่แข่งอื่นๆ ได้ แต่เนื่องจากข้อมูลที่ถูกนำมาพิจารณานั้นมีจำนวนมาก ซึ่งจะต้องใช้เวลาในการวิเคราะห์ข้อมูลที่มีจำนวนมากนั้นเป็นเวลานาน ดังนั้นจึงมีแนวความคิดที่จะนำเอาดาต้าไมน์นิ่งมาประยุกต์ใช้เพื่อแก้ปัญหานี้ โดยมีวัตถุประสงค์เพื่อทำการจัดกลุ่มข้อมูลการใช้โทรศัพท์ของลูกค้าออกเป็นกลุ่มๆ ซึ่งข้อมูลที่เราจะนำเอามาแบ่งกลุ่มลูกค้านี้จะประกอบด้วย เพศของลูกค้า, อายุของลูกค้า, บริการที่ใช้, อายุการใช้งาน (โดยนับตั้งแต่วันแรกที่ใช้จนถึงปัจจุบัน มีหน่วยเป็นวัน), จำนวนครั้งการติดต่อกับบริษัท จากนั้นก็จะนำเอาผลของการจัดกลุ่มไปใช้ในการกำหนดโปรโมชั่นต่อไป

5.2 การใช้งานระบบการจัดกลุ่มข้อมูล

ในการประยุกต์ใช้งานระบบนี้เราจะดึงเอาข้อมูลการใช้โทรศัพท์ของลูกค้าออกมาจากตาราง Customer ซึ่งจัดเก็บอยู่ในฐานข้อมูล Microsoft SQL Server 2000 ชื่อ Clustering เพื่อนำมาทำการวิเคราะห์และใช้ในการจัดกลุ่มลูกค้า โดยข้อมูลที่อยู่ในตาราง Customer จะแสดงให้เห็นในตารางที่ 5.1 ดังต่อไปนี้

ตารางที่ 5.1 แสดงตาราง Customer ที่นำมาใช้ในการจัดกลุ่มลูกค้า

ชื่อฟิลด์	ประเภทของฟิลด์
อายุของลูกค้า (Age)	float
เพศของลูกค้า (Sex)	varchar
บริการที่ใช้ (Services)	varchar
อายุการใช้งาน (Time_Services)	float
จำนวนครั้งการติดต่อกับบริษัท (Num_Of_Services)	float

โดยในระบบเราสามารถแบ่งการทำงานออกเป็น 2 ส่วนหลักๆ ดังต่อไปนี้

5.2.1 ส่วนของการสร้างโปรเจ็กต์

โดยในส่วนนี้เราจะอธิบายการทำงานเป็นขั้นตอนได้ดังนี้

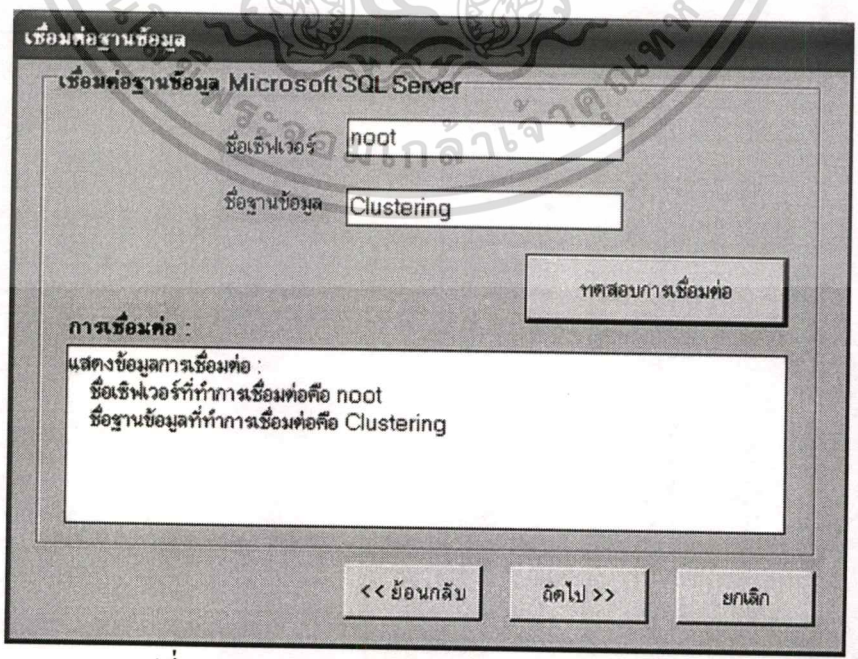
1. เข้าสู่เมนูหลัก

เป็นขั้นตอนเริ่มต้นในการสร้างโปรเจ็กต์โดยเมื่อเปิดโปรแกรมขึ้นมาจะแสดงให้หน้าจอดังรูปที่ 5.1

รูปที่ 5.1 แสดงหน้าจอหลักของการทำงาน

2. ทำการติดต่อฐานข้อมูล

ให้เราทำการคลิกที่ปุ่มสร้างโปรเจ็กต์ ในหน้าจอหลัก จากนั้นจะเข้าสู่หน้าจอเพื่อทำการติดต่อฐานข้อมูลดังรูปที่ 5.2

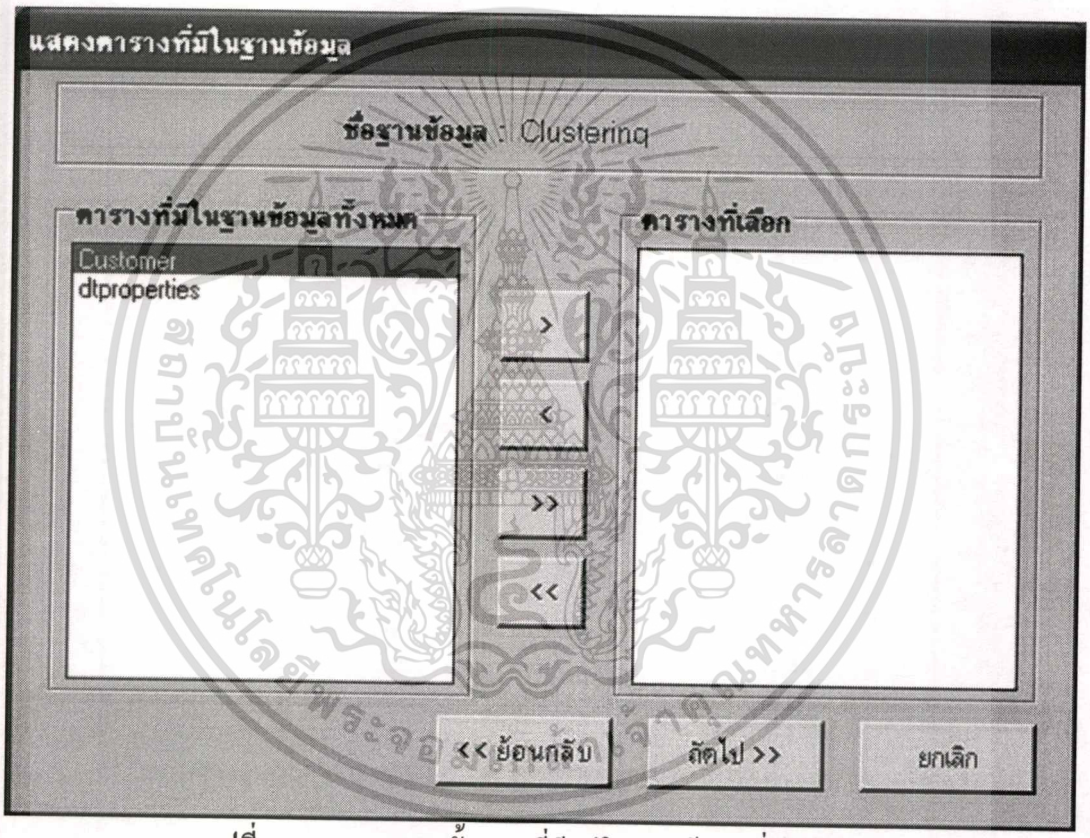


รูปที่ 5.2 แสดงหน้าจอสำหรับทำการติดต่อฐานข้อมูล

เมื่อผู้ใช้อยู่ที่หน้าจอนี้ให้ทำการพิมพ์ชื่อเซิฟเวอร์และชื่อฐานข้อมูลที่ต้องการจะติดต่อไป แล้วให้คลิกที่ปุ่มถัดไป

3. ทำการเลือกตารางที่ต้องการใช้สร้างโปรเจ็กต์

หลังจากที่ผู้ใช้ทำการติดต่อฐานข้อมูลแล้วก็จะแสดงหน้าจอสำหรับให้ผู้ใช้เลือกตาราง โดยในหน้าจอนี้จะแสดงตารางทั้งหมดที่มีอยู่ในฐานข้อมูลที่ผู้ใช้เลือกมาในขั้นตอนก่อนหน้านี้ ซึ่งแสดงดังรูปที่ 5.3

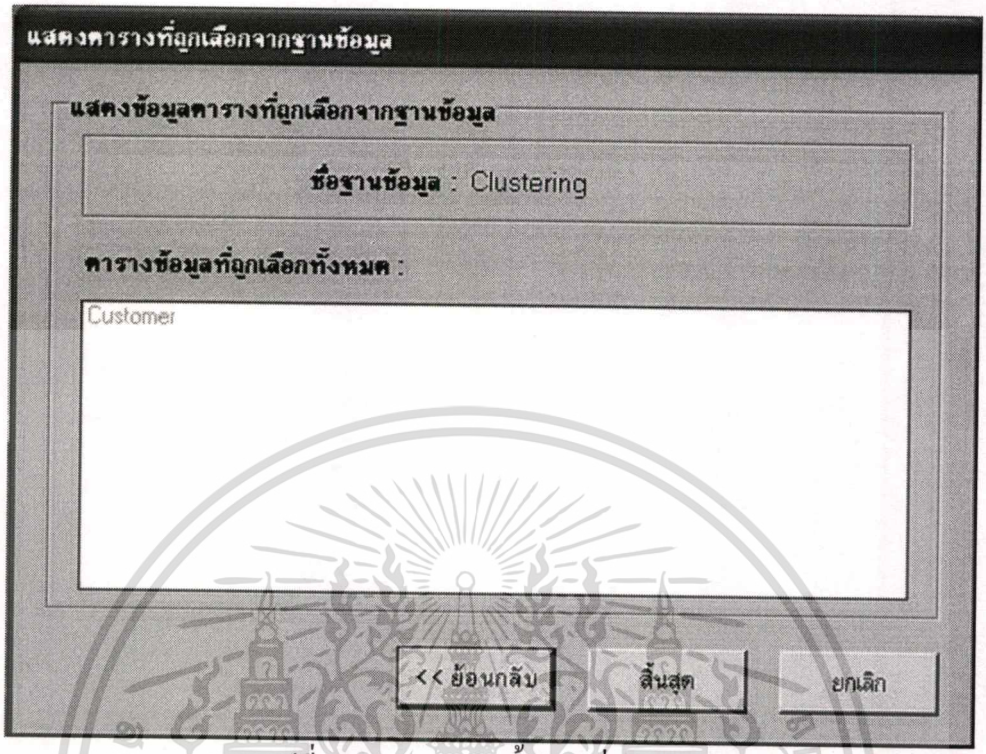


รูปที่ 5.3 แสดงตารางทั้งหมดที่มีอยู่ในฐานข้อมูลที่ผู้ใช้เลือก

จากนั้นให้ผู้ใช้เลือกตารางที่ต้องการใช้ เมื่อผู้ใช้เลือกเสร็จก็ให้กดปุ่มถัดไป

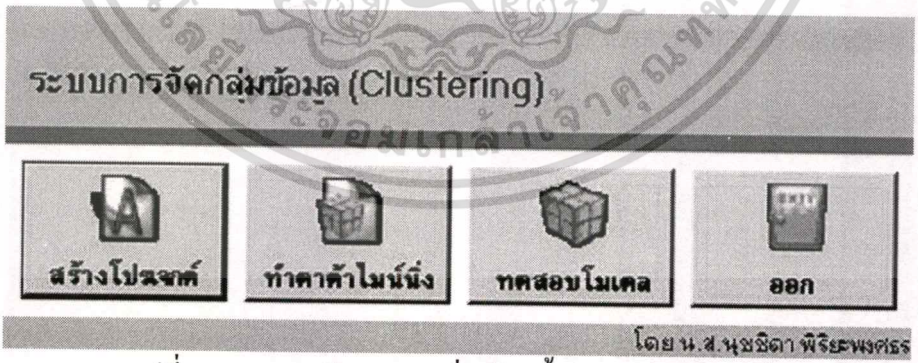
4. แสดงตารางที่ถูกเลือก

ในหน้าจอนี้จะเป็นการแสดงตารางที่ผู้ใช้เลือกมาในขั้นตอนของการเลือกตาราง ดังรูปที่ 5.4



รูปที่ 5.4 แสดงตารางทั้งหมดที่ผู้ใช้เลือก

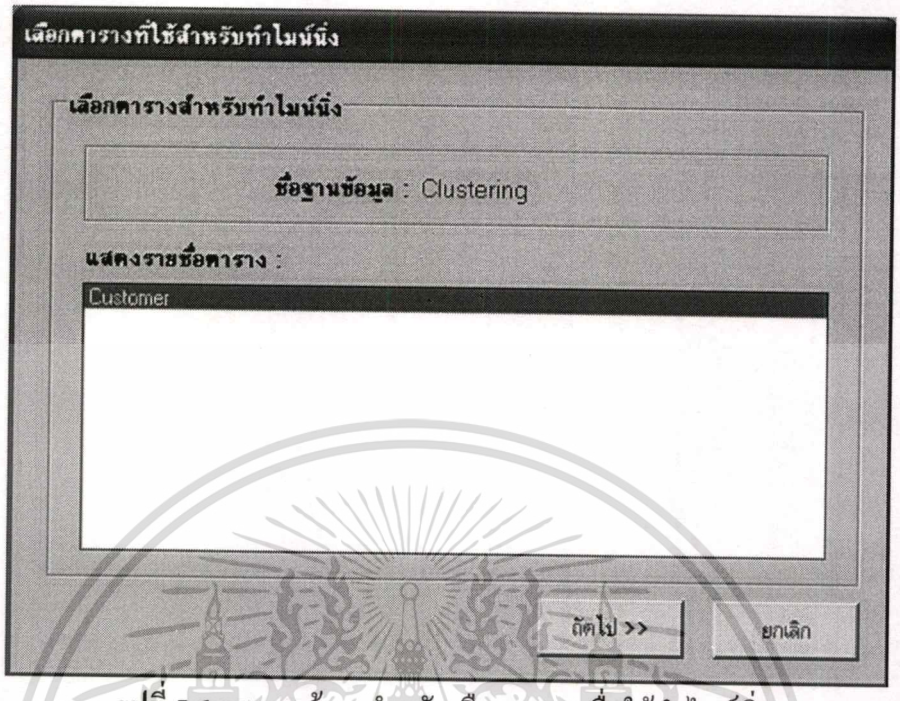
จากนั้นให้ผู้ใช้คลิกที่ปุ่มสิ้นสุด เพื่อเข้าสู่ขั้นตอนการทำไมน์นิ่งโดยให้ผู้ใช้คลิกที่ปุ่ม ทำดาต้าไมน์นิ่ง ดังรูปที่ 5.5



รูปที่ 5.5 แสดงหน้าจอหลักเพื่อเข้าสู่ขั้นตอนการทำไมน์นิ่ง

5. ทำการเลือกตารางที่ต้องการใช้ทำไมน์นิ่ง

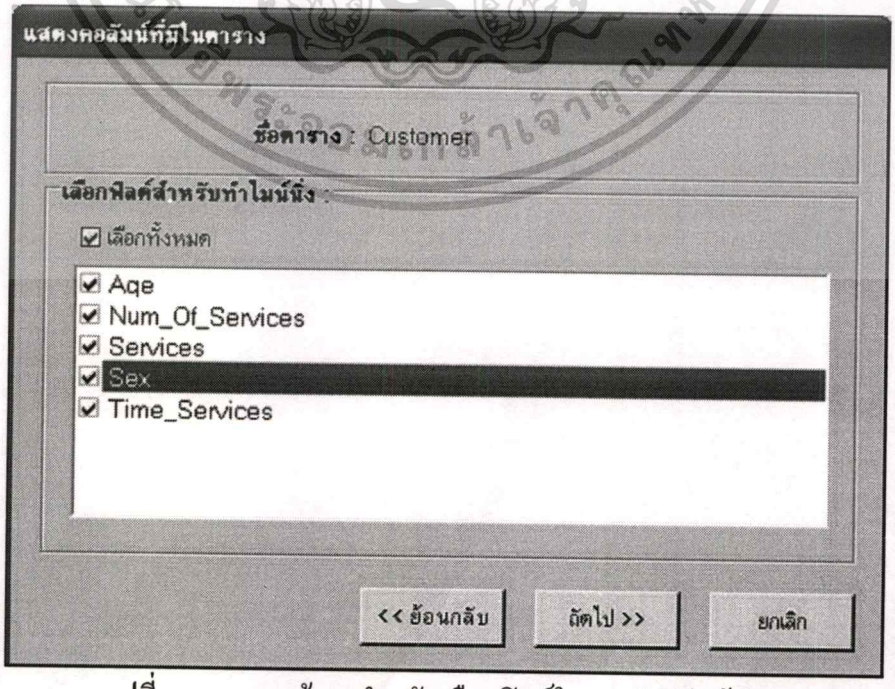
หลังจากที่ผู้ใช้คลิกที่ปุ่มทำดาต้าไมน์นิ่งที่หน้าจอหลัก ก็จะเข้าสู่หน้าจอให้เลือกตารางสำหรับทำไมน์นิ่ง โดยผู้ใช้สามารถเลือกได้มาหนึ่งตารางดังแสดงในรูปที่ 5.6 หลังจากที่ทำการเลือกตารางแล้วนั้นก็ให้คลิกที่ปุ่ม ถัดไปเพื่อทำงานในขั้นถัดไป



รูปที่ 5.6 แสดงหน้าจอสำหรับเลือกตารางเพื่อใช้ทำไมน์นิ่ง

6. เลือกฟิลด์ที่ต้องการใช้ทำไมน์นิ่ง

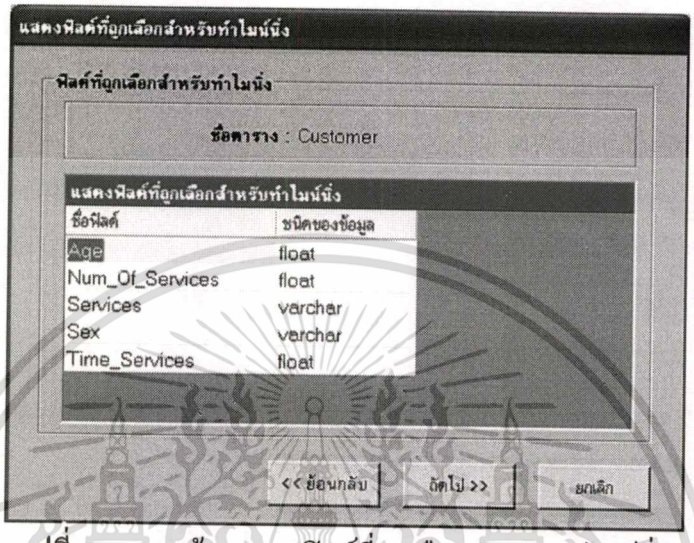
ในหน้าจอนี้จะให้ผู้ใช้ทำการเลือกฟิลด์ที่ต้องการใช้ในการจัดกลุ่มข้อมูล ดังแสดงในรูปที่ 5.7 หลังจากเลือกฟิลด์ที่ต้องการเสร็จก็ให้ผู้ใช้คลิกที่ปุ่มถัดไป เพื่อทำงานในหน้าถัดไป



รูปที่ 5.7 แสดงหน้าจอสำหรับเลือกฟิลด์ในการจัดกลุ่มข้อมูล

7. แสดงฟิลด์ที่ถูกเลือกสำหรับทำไมน์นิ่ง

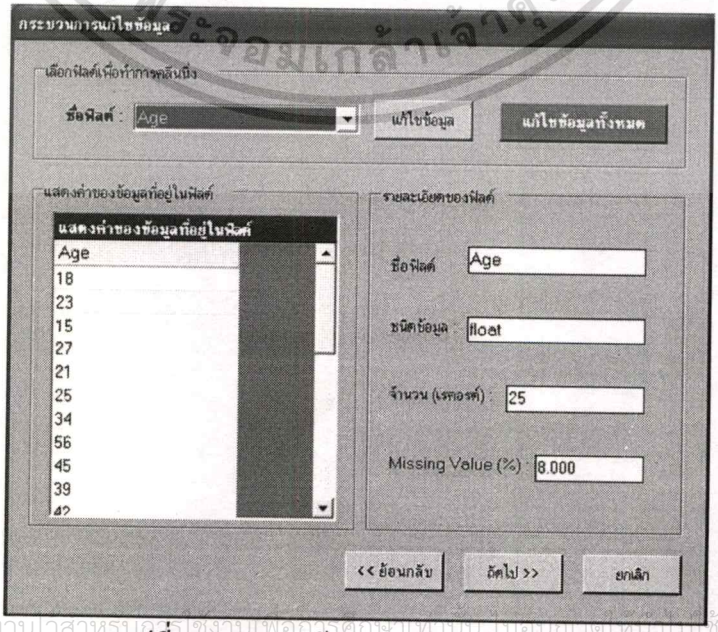
หน้าจอนี้จะแสดงข้อมูลของฟิลด์ที่ถูกเลือกเพื่อใช้ในการทำไมน์นิ่ง ดังแสดงในรูปที่ 5.8



รูปที่ 5.8 แสดงข้อมูลของฟิลด์ที่ถูกเลือกสำหรับทำไมน์นิ่ง

8. ทำการแก้ไขข้อมูล

ในขั้นตอนนี้จะเป็นการแก้ไขข้อมูลในกรณีที่ข้อมูลมีค่าเป็นค่าว่าง โดยผู้ใช้สามารถแก้ไขข้อมูลทีละตัวด้วยการคลิกที่ปุ่มแก้ไขข้อมูล หรือแก้ไขข้อมูลทั้งหมดโดยคลิกที่ปุ่มแก้ไขข้อมูลทั้งหมด ดังรูปที่ 5.9 นี้

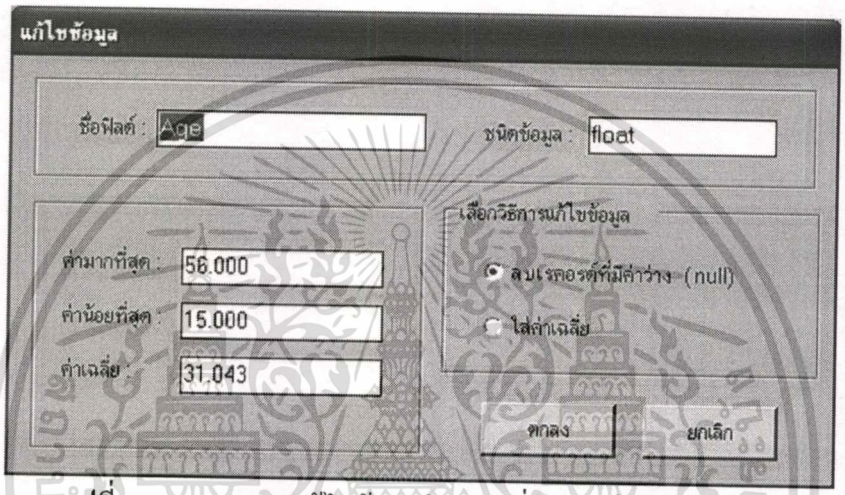


รูปที่ 5.9 แสดงหน้าจอสำหรับแก้ไขข้อมูล

โดยในการแก้ไขข้อมูลนั้นสามารถพิจารณาได้เป็น 2 กรณี ดังนี้

8.1 แก้ไขข้อมูลในกรณีที่ชนิดของข้อมูลเป็นตัวเลขดังแสดงในรูปที่ 5.10 โดยมีวิธีการแก้ไขดังนี้

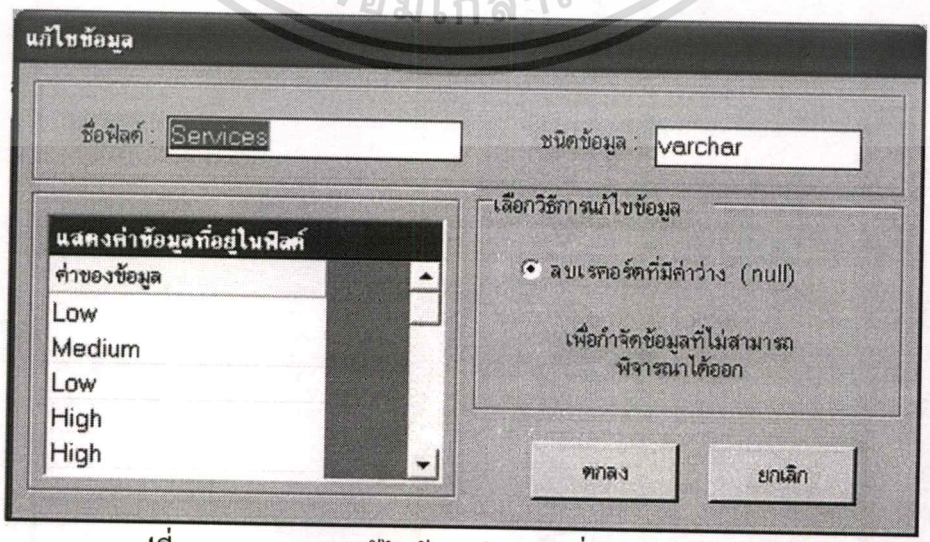
- แก้ไขได้ด้วยการลบเรคอร์ดที่มีค่าว่าง (null)
- แก้ไขด้วยการใส่ค่าเฉลี่ยลงไปในเรคอร์ดที่เป็นค่าว่าง



รูปที่ 5.10 แสดงการแก้ไขข้อมูลในกรณีที่ข้อมูลเป็นตัวเลข

8.2 แก้ไขข้อมูลในกรณีที่ชนิดของข้อมูลเป็นตัวอักษรดังแสดงในรูปที่ 5.11 โดยมีวิธีการแก้ไขดังนี้

- แก้ไขได้ด้วยการลบเรคอร์ดที่มีค่าว่าง (null)

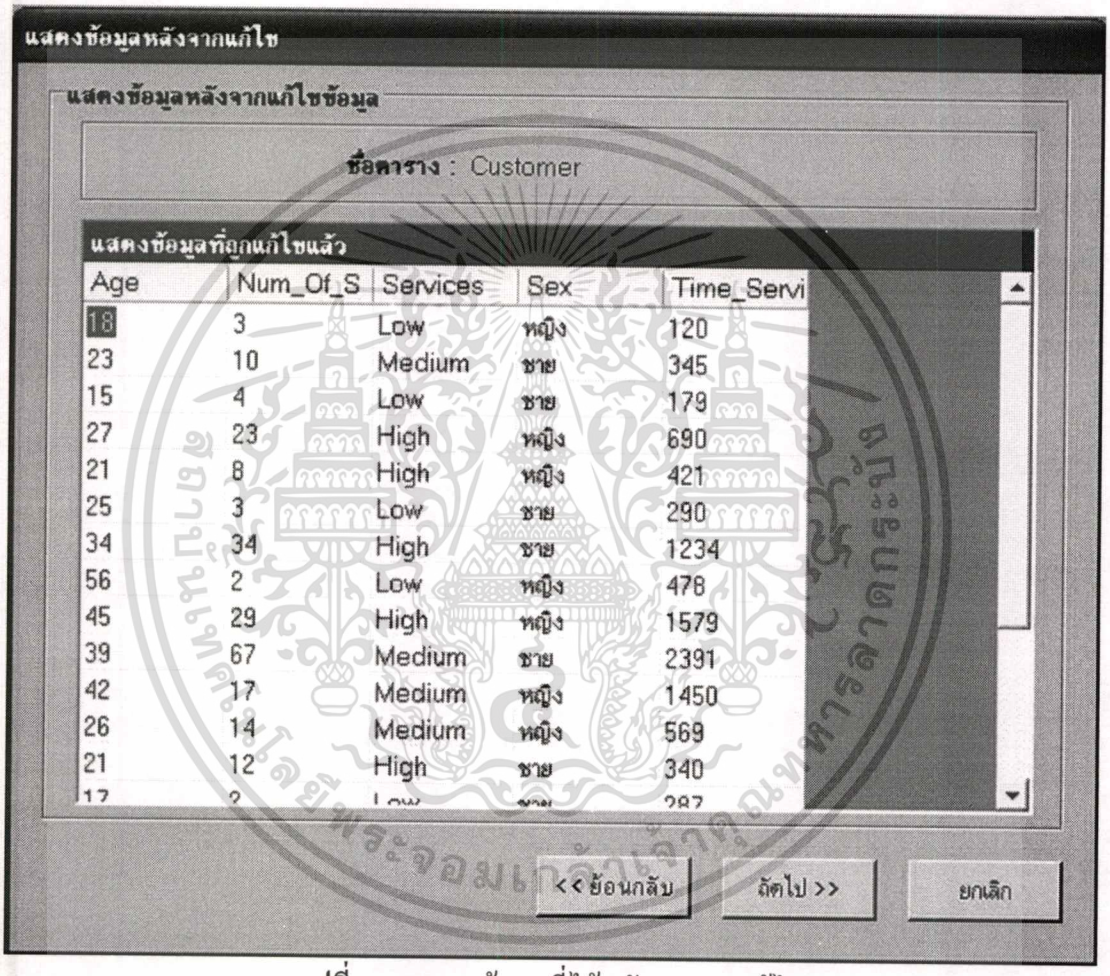


รูปที่ 5.11 แสดงการแก้ไขข้อมูลในกรณีที่ข้อมูลเป็นตัวอักษร

เมื่อทำการแก้ไขข้อมูลเสร็จแล้วก็ให้ผู้ใช้คลิกที่ปุ่มถัดไป เพื่อไปทำงานยังขั้นตอนถัดไป

9. แสดงข้อมูลหลังการแก้ไข

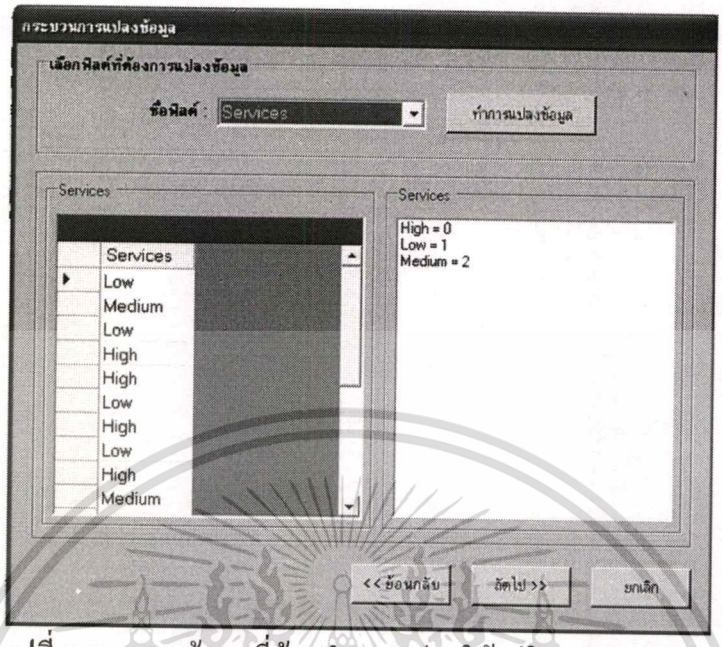
หน้าจอนี้จะแสดงข้อมูลทั้งหมดที่มีอยู่ในฐานที่ได้ทำการแก้ไขเรียบร้อยแล้ว ดังแสดงในรูปที่ 5.12 จากนั้นก็ให้ผู้ใช้คลิกที่ปุ่มถัดไป



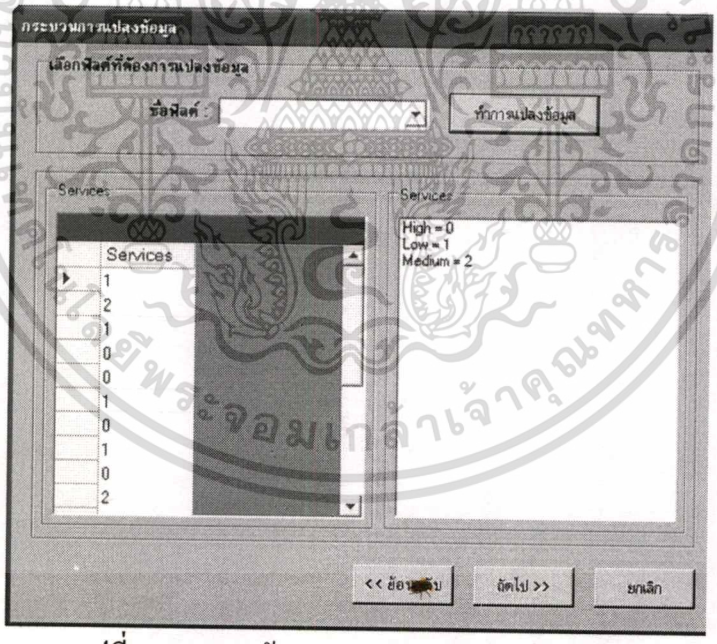
รูปที่ 5.12 แสดงข้อมูลที่ได้หลังจากการแก้ไข

10. ทำการแปลงข้อมูลจากตัวอักษรให้เป็นตัวเลข

ในขั้นตอนนี้จะเป็นการแปลงข้อมูลซึ่งมีชนิดของข้อมูลเป็นตัวอักษรให้กลายเป็นตัวเลข เพื่อนำไปใช้ในการคำนวณในขั้นถัดไป ดังแสดงในรูปที่ 5.13 โดยให้ผู้ใช้เลือกฟิลด์แล้วคลิกที่ปุ่มทำการแปลงข้อมูล เพื่อทำการแปลงข้อมูลให้เป็นตัวเลข โดยในรูปที่ 5.14 จะแสดงข้อมูลหลังจากการแปลง เมื่อข้อมูลทุกตัวถูกแปลงเสร็จก็ให้ผู้ใช้คลิกปุ่มถัดไป เพื่อทำงานในขั้นต่อไป



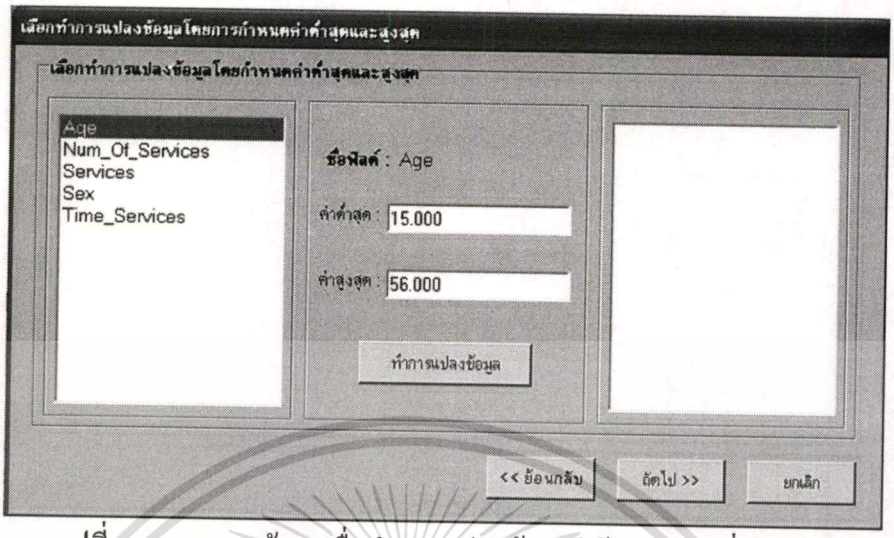
รูปที่ 5.13 แสดงข้อมูลที่ต้องทำการแปลงให้อยู่ในรูปของตัวเลข



รูปที่ 5.14 แสดงข้อมูลหลังทำการแปลงข้อมูลแล้ว

11. ทำการแปลงข้อมูลทุกฟิลด์ให้อยู่ในช่วงที่ต้องการ

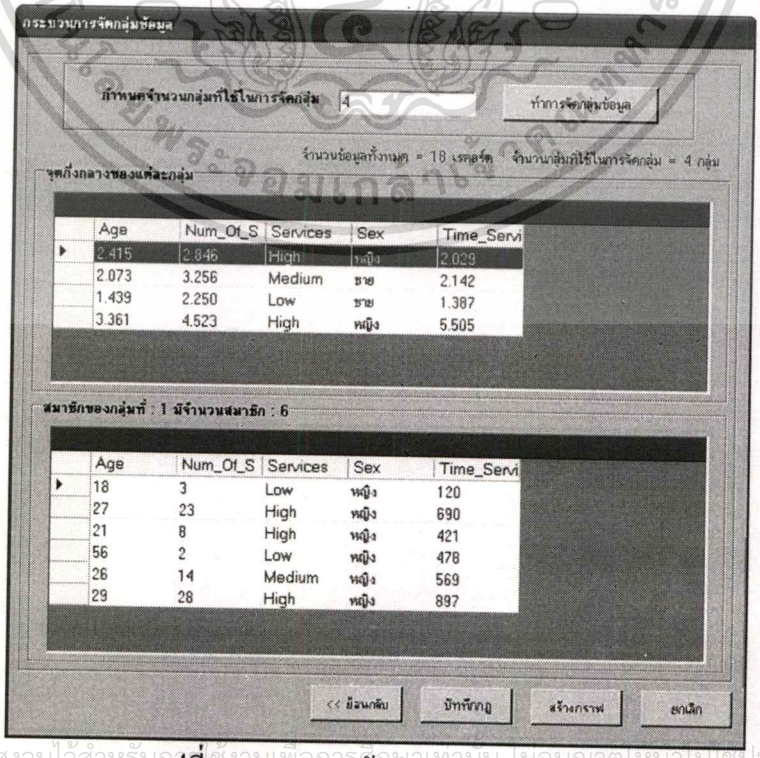
ขั้นตอนนี้เป็นขั้นตอนของการแปลงข้อมูลทุกฟิลด์ให้อยู่ในช่วงที่ต้องการ โดยการกำหนดค่าต่ำสุดและค่าสูงสุดที่ต้องการทำการแปลงแล้วคลิกที่ปุ่มทำการแปลงข้อมูล ดังแสดงในรูปที่ 5.15 โดยผู้ใช้จะต้องทำการแปลงค่าของข้อมูลทุกฟิลด์ให้อยู่ในช่วงที่ต้องการ จากนั้นก็ให้คลิกที่ปุ่มถัดไป



รูปที่ 5.15 แสดงหน้าจอเพื่อทำการแปลงข้อมูลให้อยู่ในช่วงที่ต้องการ

12. ทำการจัดกลุ่มข้อมูล

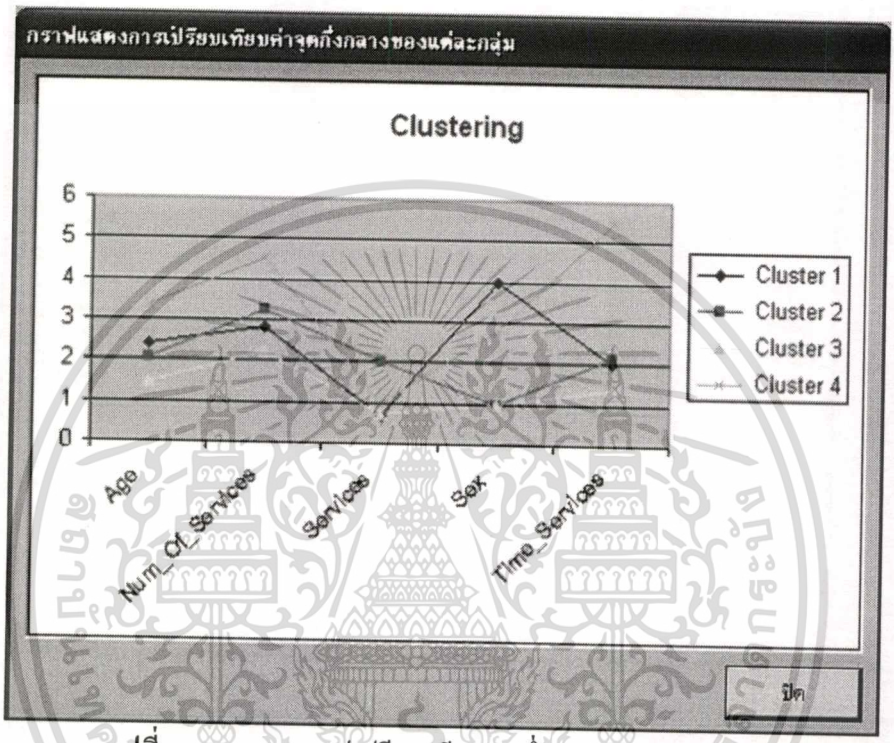
ในขั้นตอนนี้ผู้ใช้จะต้องทำการกำหนดจำนวนกลุ่มข้อมูลที่ต้องใช้ในการจัดกลุ่มข้อมูลก่อน แล้วคลิกที่ปุ่มทำการจัดกลุ่มข้อมูล เพื่อทำการจัดกลุ่มข้อมูลตามจำนวนกลุ่มที่กำหนด เมื่อผู้ใช้ต้องการดูว่าข้อมูลในกลุ่มนี้มีสมาชิกเป็นตัวใดบ้างสามารถดูได้ด้วยการคลิกที่เรคอร์ดซึ่งแสดงจุดกึ่งกลางของแต่ละกลุ่ม แล้วผลลัพธ์จะปรากฏในส่วนของสมาชิกของกลุ่ม ดังแสดงในรูปที่ 5.16



รูปที่ 5.16 แสดงหน้าจอหลังการจัดกลุ่ม

13. ทำการสร้างกราฟเปรียบเทียบจุดกึ่งกลางของแต่ละกลุ่ม

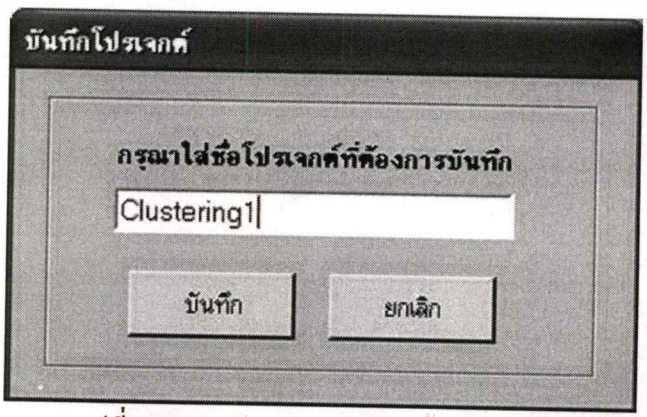
ทำได้โดยการคลิกที่ปุ่มสร้างกราฟเพื่อทำการดูกราฟเปรียบเทียบจุดกึ่งกลางของแต่ละกลุ่ม ซึ่งจะแสดงให้เห็นดังรูปที่ 5.17



รูปที่ 5.17 แสดงกราฟเปรียบเทียบจุดกึ่งกลางของแต่ละกลุ่ม

14. ทำการบันทึกโปรเจ็กต์

ในขั้นตอนนี้จะเป็นการบันทึกโปรเจ็กต์เพื่อนำไปใช้ในขั้นของการทดสอบโมเดล ดังแสดงในรูปที่ 5.18 จากนั้นให้ผู้ใช้ใส่ชื่อโปรเจ็กต์ตามต้องการแล้วคลิกที่ปุ่มตกลง



รูปที่ 5.18 แสดงหน้าจอการบันทึกโปรเจ็กต์

5.2.2 ส่วนของการทดสอบโมเดล

ในส่วนนี้จะเป็นส่วนของการนำเอาข้อมูลมาทดสอบกับโมเดลที่เราสร้างเอาไว้ในส่วนของการสร้างโปรเจ็กต์ โดยมีขั้นตอนดังนี้

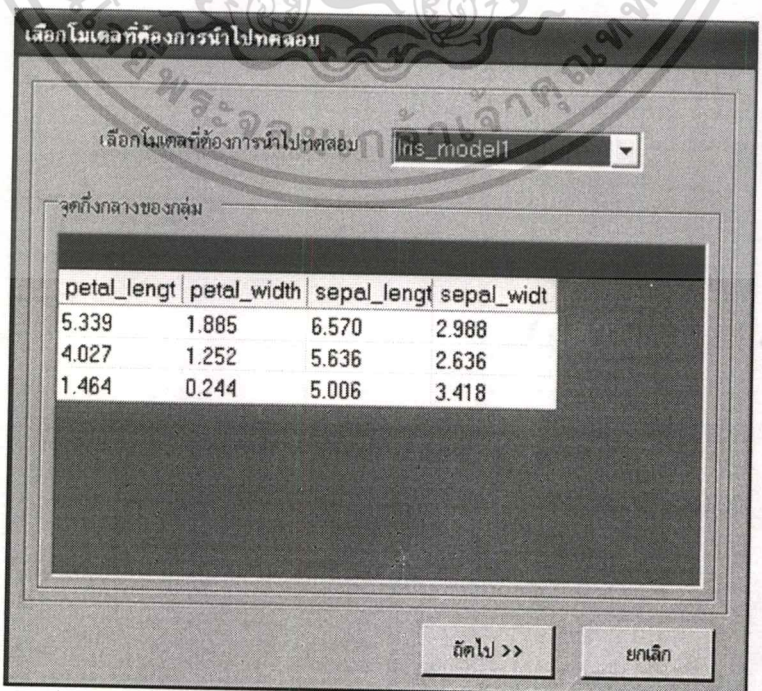
1. เข้าสู่กระบวนการทดสอบโมเดล

ในขั้นตอนนี้ทำได้โดยการคลิกที่ปุ่มทดสอบโมเดลซึ่งอยู่ในหน้าจอหลัก ดังแสดงในรูปที่ 5.19

รูปที่ 5.19 หน้าจอหลักเพื่อเข้าสู่กระบวนการทดสอบโมเดล

2. เลือกโมเดลที่ต้องการนำไปทดสอบ

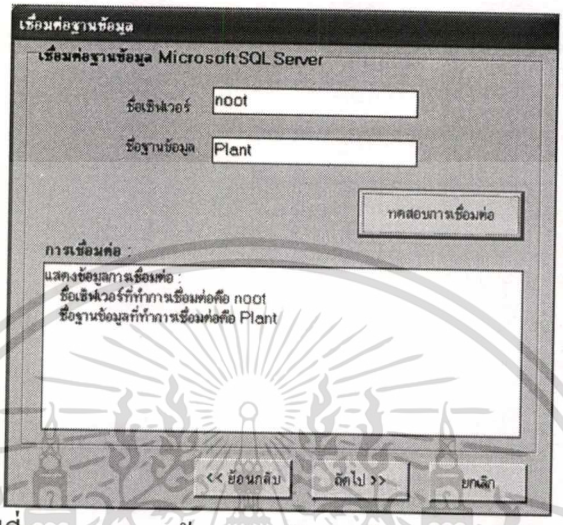
ในขั้นตอนนี้จะให้ผู้ใช้เลือกโมเดลที่ต้องการนำมาใช้ในการทดสอบข้อมูล ดังแสดงในรูปที่ 5.20



รูปที่ 5.20 แสดงข้อมูลโมเดลเพื่อให้ผู้ใช้เลือก

3. ทำการติดต่อฐานข้อมูลเพื่อเลือกข้อมูลมาใช้ในการทดสอบ

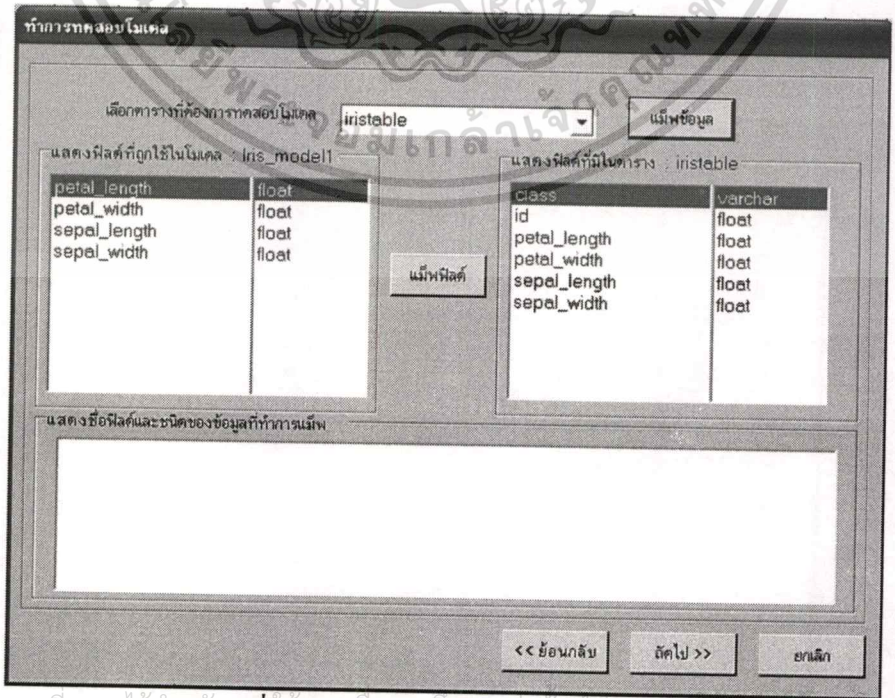
ขั้นตอนนี้จะเป็นการเลือกฐานข้อมูลที่ต้องการนำมาทดสอบ ดังแสดงในรูปที่ 5.21



รูปที่ 5.21 แสดงหน้าจอสำหรับการติดต่อกับฐานข้อมูล

4. ทำการเลือกตารางที่ต้องการนำมาทดสอบ

เป็นการเลือกตารางเพื่อนำมาใช้ในการทดสอบโมเดลหลังจากนั้นให้ผู้ใช้คลิกที่ปุ่มแม็พข้อมูลเพื่อแสดงฟิลด์ที่จะต้องทำการแม็พ ดังแสดงในรูปที่ 5.22

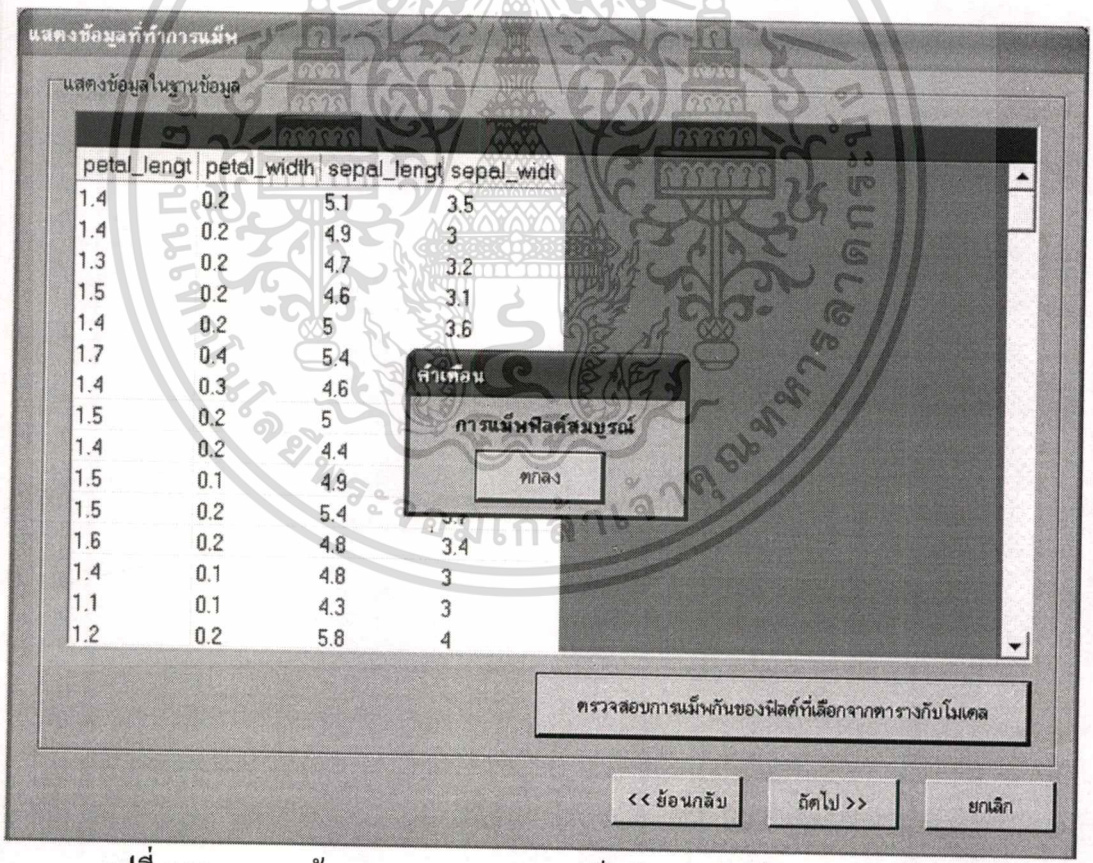


รูปที่ 5.22 แสดงข้อมูลที่จะนำมาแม็พกัน

หลังจากนั้นให้ผู้ใช้เลือกฟิลด์ในส่วนของฟิลด์ที่ถูกใช้ในกฎและฟิลด์ที่มีในตารางเพื่อนำมาแม็พกัน โดยการคลิกที่ปุ่มแม็พฟิลด์ โดยฟิลด์ที่นำมาแม็พกันนั้นค่าของข้อมูลที่อยู่ในฟิลด์จะต้องมีความคล้ายคลึงกัน จึงจะสามารถทำการแม็พกันได้ เมื่อผู้ใช้ทำการแม็พฟิลด์ครบทุกฟิลด์ในส่วนของฟิลด์ที่อยู่ในกฎแล้วก็ให้คลิกที่ปุ่มถัดไป เพื่อทำงานในหน้าถัดไป

5. ทำการตรวจสอบการแม็พกันของฟิลด์

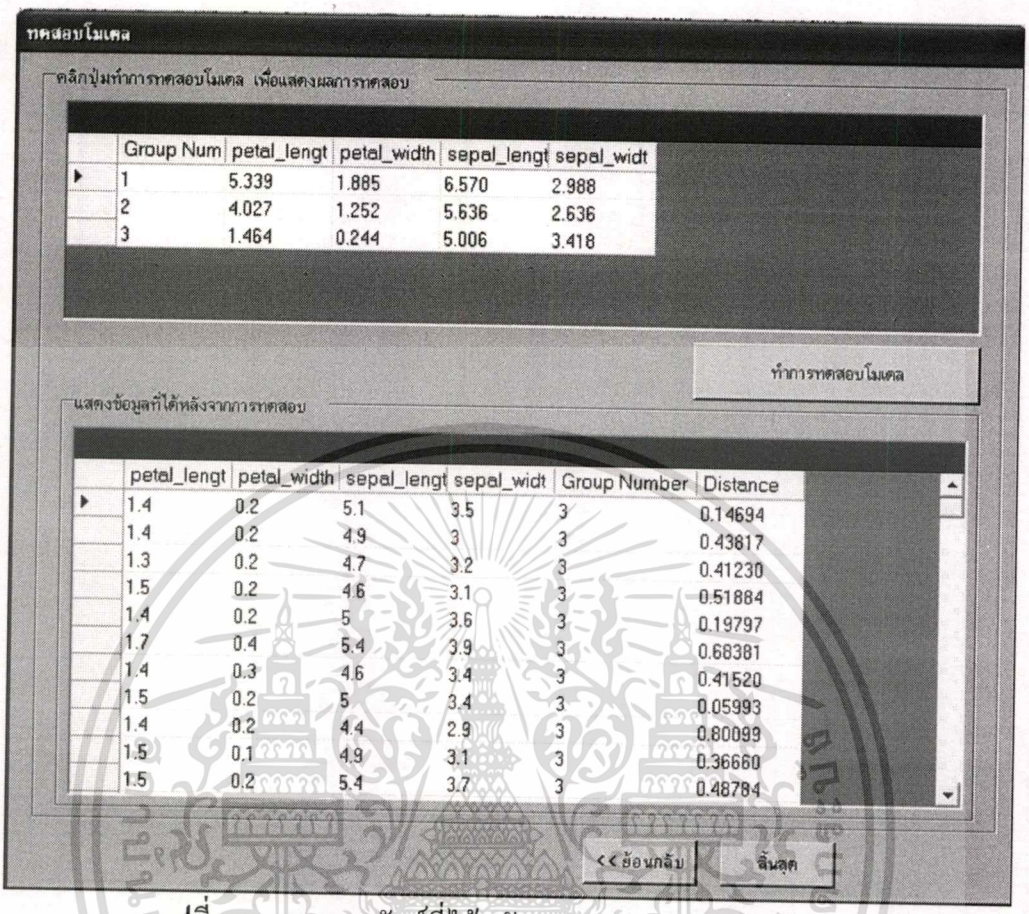
โดยให้ผู้ใช้คลิกที่ปุ่มตรวจสอการแม็พกันของฟิลด์ที่เลือกจากตารางกับโมเดล เพื่อทำการตรวจสอบการแม็พกันของฟิลด์ ถ้าหากว่าค่าของข้อมูลของฟิลด์ที่นำมาแม็พกันมีความคล้ายคลึงกันก็แสดงว่าการแม็พฟิลด์นั้นสมบูรณ์ แล้วให้ผู้คลิกที่ปุ่มถัดไปเพื่อไปทำงานยังขั้นตอนต่อไป แต่ถ้าไม่สามารถแม็พฟิลด์ได้ก็ให้ผู้ใช้กลับไปทำการแม็พฟิลด์ใหม่อีกครั้ง ดังแสดงในรูปที่ 5.23



รูปที่ 5.23 แสดงหน้าจอการตรวจสอบการแม็พกันของฟิลด์ในตารางกับโมเดล

6. ทำการทดสอบโมเดล

ในขั้นตอนนี้จะเป็นการทดสอบโมเดล โดยนำเอาข้อมูลที่เลือกมามาทำการจัดกลุ่มข้อมูล ซึ่งผู้ใช้จะต้องคลิกที่ปุ่มทำการทดสอบโมเดลเพื่อทำการจัดกลุ่มข้อมูล ดังแสดงในรูปที่ 5.24



รูปที่ 5.24 แสดงผลลัพธ์ที่ได้หลังจากการทำการทดสอบโมเดล

จากผลการจัดกลุ่มข้อมูลสามารถสรุปได้ดังนี้

กลุ่มที่ 1 จะมีสมาชิกทั้งหมด 3 เรคอร์ด โดยสมาชิกที่อยู่ในกลุ่มนี้จะเป็นเพศหญิง โดยใช้บริการแบบ Low และ Medium

กลุ่มที่ 2 มีสมาชิกด้วยกันทั้งหมด 4 เรคอร์ด โดยเป็นเพศชายและใช้บริการแบบ Medium

กลุ่มที่ 3 จะประกอบด้วยสมาชิกทั้งหมด 5 เรคอร์ด โดยเป็นเพศชายและใช้บริการแบบ Low กับ High

กลุ่มที่ 4 ประกอบด้วยสมาชิกด้วยกันทั้งหมด 6 เรคอร์ด โดยเป็นเพศหญิงทั้งหมดและใช้บริการแบบ High กับ Low

บทที่ 6
สรุปผลการศึกษาและข้อเสนอแนะ

6.1 สรุปผลการศึกษา

การทำดาต้าไมน์นิ่งเป็นกระบวนการในการค้นหาข้อมูลที่ซ้อนเร้นอยู่ภายในฐานข้อมูล และนำข้อมูลที่มีอยู่มาทำให้เกิดประโยชน์ จากนั้นจะนำข้อมูลที่ได้จากการทำดาต้าไมน์นิ่งไปใช้ช่วยในการวางแผนธุรกิจและช่วยในการสนับสนุนการตัดสินใจเกี่ยวกับการทำธุรกิจต่างๆ โดยในขั้นตอนของการทำดาต้าไมน์นิ่งจะต้องทำการกำหนดวัตถุประสงค์ในการทำดาต้าไมน์นิ่งก่อน หลังจากนั้นก็ทำการเตรียมข้อมูล ซึ่งในขั้นตอนของการเตรียมข้อมูลนี้จะประกอบด้วยการเลือกข้อมูลเข้ามาเพื่อใช้ทำไมน์นิ่ง จากนั้นก็ทำการแก้ไขข้อมูลในกรณีที่ข้อมูลนั้นมีค่าเป็นค่าว่าง แล้วทำการแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมสำหรับใช้ในการทำไมน์นิ่ง หลังจากเสร็จสิ้นขบวนการเตรียมข้อมูลแล้วก็เข้าสู่กระบวนการทำไมน์นิ่ง หลังจากนั้นจะนำผลลัพธ์ที่ได้จากการทำไมน์นิ่งไปใช้ให้เกิดประโยชน์กับธุรกิจต่อไป

หลังจากที่ได้ทำการศึกษาหลักการดังการแล้วจึงมีแนวคิดในการพัฒนาเครื่องมือที่ใช้ช่วยในการจัดกลุ่มข้อมูล (Clustering) โดยใช้อัลกอริทึม K-means ซึ่งเป็นอัลกอริทึมหนึ่งที่นำมาใช้ในการจัดกลุ่มข้อมูล โดยข้อมูลที่จะใช้ในอัลกอริทึมนี้จะต้องเป็นข้อมูลที่เป็นตัวเลขเท่านั้น ดังนั้นหลังจากขั้นตอนของการแก้ไขข้อมูลแล้วถ้ามีฟิลด์ใดมีชนิดของข้อมูลเป็นตัวเลข เราจะต้องทำการแปลงให้อยู่ในรูปของตัวเลขให้หมด เพื่อจะสามารถนำไปใช้คำนวณในอัลกอริทึมได้

6.2 ข้อเสนอแนะ

1. ระบบที่พัฒนาขึ้นสามารถใช้ได้กับฐานข้อมูล Microsoft SQL Server เท่านั้น
2. ระบบที่พัฒนาขึ้นจะใช้เวลาในการประมวลผลแตกต่างกัน ขึ้นอยู่กับจำนวนข้อมูลในฐานข้อมูล, ความเร็วของ CPU, ขนาดของหน่วยความจำ เป็นต้น

บรรณานุกรม

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

กิตติ ภักดีวัฒนะกุล และ กิตตพงษ์ กลมกล่อม. 2544. **UML วิเคราะห์และออกแบบระบบเชิงวัตถุ.** กรุงเทพฯ: เคทีพี คอมพ์ แอนด์ คอนซัลท์.

ธาริน สิทธิธรรมชารี. 2537. **Microsoft Visual Basic .Net ฉบับสมบูรณ์.** กรุงเทพฯ:ซัคเซส มีเดีย.

ศุภชัย สมพานิช. 2546. **สร้างระบบงานฐานข้อมูลด้วย Visual Basic .NET.** กรุงเทพฯ: Dev Book.

สมพร จิวรสกุล. 2545. **คู่มือการติดตั้งและใช้งาน Microsoft SQL Server 2000 ฉบับสมบูรณ์.** กรุงเทพฯ: Infopress Developer Book.

Ching-Yi Wu, Suan Zhu, Rohit Thadani, Kevin Kirkpatrick and Christopher Swope. 2005. **DATA MINING.** [Online]. Available: http://userfs.cec.wustl.edu/~cse530/2004/DataMIning.ppt

Eric Williams. 2004. **Customer relationship management.** [Online]. Available: http://searchcrm.techtarget.com/sDefinition/0,,sid11_gci213567,00.html

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

Jiawei Han & 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/

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

ประวัติผู้เขียน

ชื่อ : นางสาวนุชชิดา พิริยะพงศธร

วันเดือนปีเกิด : 30 สิงหาคม 2524

สถานที่เกิด : โรงพยาบาลสัมประสิทธิประสงค์ จังหวัดอุบลราชธานี

ประวัติการศึกษา :

มัธยมต้น : โรงเรียนลือคำหาญ

มัธยมปลาย : โรงเรียนนารีนุกูล

ปริญญาตรี : มหาวิทยาลัยขอนแก่น คณะวิทยาศาสตร์ วิทยาการสารสนเทศ