# การพัฒนาระบบดาต้าไมน์นิ่งโดยใช้อัลกอริทึม CLARANS

# DEVELOPMENT OF DATA MINING USING CLARANS ALGORITHM

โดย

สิริภูมิ สุดดีพงศ์

SIRIPOOM SUDDEEPHONG

อาจารย์ที่ปรึกษา

รศ.ดร.วรพจน์ กรีสุระเดช



รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ภาคเรียนที่ 2 ปีการศึกษา 2549

# DEVELOPMENT OF DATA MINING USING CLARANS ALGORITHM

**SIRIPOOM SUDDEEPHONG**

**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**

**2/ 2006**

| | |
|---|---|
| **หัวข้อ** | การพัฒนาระบบดาต้าไมน์นิ่งโดยใช้อัลกอริทึม CLARANS |
| **นักศึกษา** | นายสิริภูมิ สุดดีพงศ์ |
| **รหัสนักศึกษา** | 48066424 |
| **ปริญญา** | วิทยาศาสตรมหาบัณฑิต |
| **สาขาวิชา** | เทคโนโลยีสารสนเทศ |
| **แขนงวิชา** | วิทยาการสารสนเทศ |
| **ปีการศึกษา** | 2549 |
| **อาจารย์ที่ปรึกษา** | รศ.ดร. วรพจน์ กรีสุระเดช |

## บทคัดย่อ

ข้อมูลข่าวสารเป็นสิ่งที่มีประโยชน์ต่อการทำธุรกิจเป็นอย่างมากในโลกอันทันสมัย ข้อมูลได้มีการพัฒนาเพิ่มขึ้นอยู่ตลอดเวลา ปัญหาพิเศษนี้จึงได้นำความรู้ทางด้านดาต้าไมน์นิ่ง มาใช้เป็นแนวทางในการวิเคราะห์ข้อมูลข่าวสารที่มีอยู่ และนำข้อมูลที่ผ่านการวิเคราะห์มาใช้ให้เกิดประโยชน์ต่อธุรกิจได้อย่างแท้จริง

จุดมุ่งหมายของปัญหาพิเศษนี้เพื่อการพัฒนาระบบดาต้าไมน์นิ่งโดยนำอัลกอริทึม CLARANS มาใช้ในการแบ่งกลุ่มข้อมูล โดยคาดว่าจะสามารถนำระบบที่ได้พัฒนานี้ไปใช้แบ่งกลุ่มข้อมูลได้อย่างสะดวกรวดเร็ว และง่ายยิ่งขึ้น

| | |
|---|---|
| **Title** | Development of Data Mining using CLARANS Algorithm |
| **Student** | Mr. Siripoom Suddeephong |
| **Student ID.** | 48066424 |
| **Degree** | Master of Science |
| **Programme** | Information Science |
| **Academic Year** | 2006 |
| **Advisor** | Assoc. Prof. Dr. Worapoj Kreesuradej |


## ABSTRACT

Information is the useful for business in modern lifestyle. New information are consistently added. This special project bring the data mining knowledge to analyze the information we have and bring it to useful for business.

This data mining system use CLARANS algorithm to cluster information and expect that can use by convenience and simple.

# กิตติกรรมประกาศ

ขอกราบขอบพระคุณบิดา มารดาที่ได้ทำการสนับสนุนทางด้านกำลังใจ ความเข้าใจ และทุนทรัพย์จนสามารถทำให้ข้าพเจ้าทำงานสำเร็จได้ด้วยดี

ขอขอบพระคุณรศ.ดร.วรพจน์ กรีสุระเดช ซึ่งเป็นอาจารย์ผู้รับผิดชอบในการพัฒนาระบบงานนี้ ที่กรุณาให้คำแนะนำอันเกิดประโยชน์แก่ข้าพเจ้า และเป็นที่ปรึกษาในการแก้ปัญหาต่างๆ ที่เกิดขึ้นจนผ่านพ้นไปได้ด้วยดี รวมทั้งยังเป็นผู้ตรวจสอบความความถูกต้องของเนื้อหา และรูปแบบของปัญหาพิเศษฉบับนี้ให้สมบูรณ์ยิ่งขึ้น

ขอขอบคุณพี่อาทิตยา เชื้อจันอัด ที่เขียนโปรแกรมขั้นการเตรียมข้อมูลได้ดี และสวยมากๆ ทำให้สามารถนำมาพัฒนาระบบดาต้าไมน์นิ่งได้ง่ายกว่าการเริ่มทำตั้งแต่ต้น

ขอขอบคุณเพื่อนๆ IS 19.1 รวมทั้งพี่ๆ ทุกคน ที่คอยให้คำปรึกษาแก่ข้าพเจ้าด้วยความเต็มใจ คอยให้กำลังใจในยามที่ข้าพเจ้าย่อท้อ และช่วยเหลือในยามที่ข้าพเจ้าติดปัญหาต่างๆ ตลอดการทำงาน

ขอขอบพระคุณผู้ที่บริจาคพื้นที่การศึกษาในเขตลาดกระบัง ทำให้ก่อเกิดเป็นสถานศึกษาอันมีชื่อเสียงให้ข้าพเจ้าได้รู้จัก มิฉะนั้นแล้วข้าพเจ้าคงไม่ได้มาทำงานชิ้นนี้ ไม่ได้มาอยู่ที่นี่ และไม่ได้มาเจอสิ่งดีดีที่เกิดขึ้นในชีวิตไม่ว่าจะเป็น อาจารย์ เพื่อน พี่ ธรรมชาติแวดล้อม และผู้คนในเขตลาดกระบัง

สิริภูมิ สุดดีพงศ์

# สารบัญ

หน้า

# สารบัญ(ต่อ)

หน้า

# สารบัญ(ต่อ)

หน้า

# สารบัญตาราง

ตารางที่ หน้า

# สารบัญรูป

# บทที่ 1
# บทนำ

## 1.1 ความสำคัญและที่มาของปัญหา

เนื่องจากในปัจจุบันเป็นยุคที่มีข้อมูลข่าวสารอยู่เป็นจำนวนมาก ดังนั้นผู้ที่สามารถนำข้อมูลข่าวสารมาใช้ประโยชน์ได้มากก็จะมีโอกาสเป็นผู้สำเร็จในองค์กรธุรกิจมากกว่าผู้อื่น และจากการเพิ่มขึ้นอย่างรวดเร็วของข้อมูล ทำให้การขยายขนาดของฐานข้อมูลเป็นไปอย่างรวดเร็ว ทำให้การวิเคราะห์ฐานข้อมูลที่มีขนาดใหญ่นั้นเป็นไปได้ยากขึ้นเรื่อยๆ ดังนั้นเพื่อที่จะให้เกิดความสะดวกในการวิเคราะห์ข้อมูล และสามารถทำความเข้าใจในข้อมูลได้ง่ายยิ่งขึ้น จึงได้นำความรู้ในเรื่องของดาต้าไมน์นิ่ง(Data Mining) ซึ่งเป็นเทคนิคในการวิเคราะห์ข้อมูล โดยเฉพาะการวิเคราะห์ข้อมูลที่มีปริมาณมากจนเกินความสามารถที่จะนำมาวิเคราะห์ได้ด้วยคน มาช่วยในการวิเคราะห์ข้อมูลที่เป็นอัตโนมัติ ทำให้เกิดความสะดวกต่อการวิเคราะห์ข้อมูลมากยิ่งขึ้น

## 1.2 วัตถุประสงค์ของการพัฒนาระบบงาน

วัตถุประสงค์ของการพัฒนาระบบงาน มีดังต่อไปนี้

1) เพื่อให้สามารถนำดาต้าไมน์นิ่งในเรื่องของการแบ่งกลุ่มฐานข้อมูล(Database Segmentation) มาใช้ในการจัดกลุ่มข้อมูลได้

2) เพื่อให้ข้อมูลที่ผ่านกระบวนการของระบบที่พัฒนาขึ้นมา สามารถนำไปใช้เป็นข้อมูลอย่างหนึ่งในการสนับสนุนการตัดสินใจทางธุรกิจได้เป็นอย่างดี

3) เพื่อให้เกิดความสะดวกในการวิเคราะห์ข้อมูลมากยิ่งขึ้น

4) สามารถนำไปใช้กับฐานข้อมูลที่มีขนาดใหญ่ได้

## 1.3 ขอบเขตของการพัฒนาระบบงาน

เป็นการศึกษาดาต้าไมน์นิ่งในเรื่องของการแบ่งกลุ่มฐานข้อมูล ซึ่งมีขอบเขตดังนี้

1) DBMS ที่ใช้ในการพัฒนาระบบคือ Microsoft SQL Server 2000

2) ใช้โปรแกรม Microsoft Visual Studio .NET 2003 ในการพัฒนาระบบ

3) ระบบงานนี้เป็นการพัฒนาระบบต่อมาจากส่วนการเตรียมข้อมูล โดยเพิ่มส่วนในการจัดกลุ่มข้อมูลเข้าไป

4) ใช้อัลกอริทึม CLARANS ในการจัดกลุ่มฐานข้อมูล

## 1.4 ขั้นตอนในการดำเนินงาน

ขั้นตอนในการดำเนินงาน มีดังต่อไปนี้

1) ศึกษาความรู้เกี่ยวกับดาต้าไมน์นิ่งที่ใช้เทคนิคการแบ่งกลุ่มฐานข้อมูล

2) ศึกษาขั้นตอนวิธีการของอัลกอริทึม CLARANS เพื่อใช้ในการจัดกลุ่มฐานข้อมูล

3) กำหนดแหล่งข้อมูลที่จะนำมาใช้ในการวิเคราะห์จัดกลุ่มข้อมูล

4) เขียนโปรแกรมเพื่อการพัฒนาระบบดาต้าไมน์นิ่งโดยใช้อัลกอริทึม CLARANS

5) ทดสอบและปรับปรุงโปรแกรมเพื่อให้ใช้งานได้อย่างถูกต้องสมบูรณ์

6) วิเคราะห์และประเมินผลที่ได้จากการพัฒนาระบบ

## 1.5 ประโยชน์ที่คาดว่าจะได้รับ

ประโยชน์ที่คาดว่าจะได้รับ มีดังต่อไปนี้

1) ทำให้ได้รับความรู้ และความเข้าใจถึงขั้นตอนและวิธีการต่างๆ ของการทำดาต้าไมน์นิ่งได้อย่างดี

2) ได้ระบบดาต้าไมน์นิ่งที่สามารถนำไปจัดกลุ่มข้อมูลในฐานข้อมูลได้

3) มีความสะดวกสบายในการวิเคราะห์ข้อมูลมากกว่าการวิเคราะห์ข้อมูลด้วยคน

4) สามารถนำข้อมูลที่ผ่านกระบวนการของระบบที่พัฒนาขึ้นมาไปใช้งานจริงได้

# บทที่ 2
# ทฤษฎี และหลักการที่เกี่ยวข้อง

## 2.1 ดาต้าไมน์นิ่ง(Data Mining)

ดาต้าไมน์นิ่งเป็นวิธีการที่นำมาใช้ในการวิเคราะห์ข้อมูล เพื่อค้นหาแนวโน้มหรือรูปแบบความสัมพันธ์ของข้อมูลซึ่งมีอยู่จริงในฐานข้อมูล และรูปแบบเหล่านั้นได้ถูกซ่อนไว้ภายในข้อมูลจำนวนมากที่มีอยู่ ดาต้าไมน์นิ่งจะทำการสำรวจและวิเคราะห์ข้อมูลให้อยู่ในรูปแบบที่มีความหมายและอยู่ในรูปของกฎ โดยความสัมพันธ์ที่ได้จะแสดงให้เห็นถึงความรู้ต่างๆ ที่มีประโยชน์ในฐานข้อมูล สามารถที่จะนำข้อมูลเหล่านี้ไปใช้เป็นแนวทางในการตัดสินใจเพื่อก่อให้เกิดผลดีในองค์กรธุรกิจต่อไปได้

## 2.2 ขั้นตอนการทำดาต้าไมน์นิ่ง

เพื่อที่จะให้ได้ข้อมูลหลังจากการทำดาต้าไมน์นิ่งออกมามีประสิทธิภาพแล้ว ควรที่จะปฏิบัติตามขั้นตอนในการทำดาต้าไมน์นิ่ง ซึ่งแบ่งออกเป็น 5 ขั้นตอน คือ

### 2.2.1 กำหนดวัตถุประสงค์ทางธุรกิจ(Business Objective Determination)

เป็นการนิยามถึงจุดประสงค์ หรือความต้องการของธุรกิจที่ต้องการหา เช่น ทำอย่างไรถึงจะสามารถขายสินค้าให้ลูกค้าได้มากที่สุด หรือถ้ามองลึกลงไปอีกจะได้ว่า ลูกค้าประเภทไหนที่ชอบซื้อสินค้า A เพื่อที่จะให้ได้ข้อมูลตามที่ต้องการ ต้องมีข้อมูลที่อธิบายเกี่ยวกับลูกค้าที่ชอบซื้อสินค้า A จากนั้นจึงจะสามารถเตรียมข้อมูลสำหรับทำดาต้าไมน์นิ่งได้

ในการกำหนดวัตถุประสงค์นั้นจะต้องเข้าใจถึงปัญหาของงานเป็นอย่างดี เพราะถ้ากำหนดวัตถุประสงค์ไม่ชัดเจนแล้ว อาจทำให้ผลลัพธ์ที่ได้เกิดความคลุมเครือ ไม่สามารถที่จะนำไปใช้ได้ อาจทำให้ต้องกลับมาเริ่มต้นใหม่อีกครั้ง ซึ่งจะทำให้เสียเวลาเป็นอย่างมาก

### 2.2.2 การรวบรวมและเตรียมข้อมูล(Data Gathering and Preparation)

ข้อมูลที่จะใช้ในการทำดาต้าไมน์นิ่งนั้นจะต้องเตรียมให้อยู่ในรูปแบบที่เหมาะสมก่อนเสมอ ซึ่งเป็นขั้นตอนที่มีความสำคัญมาก และต้องใช้เวลาในการทำงานมากกว่าขั้นตอนอื่นๆ ของขั้นตอนการทำดาต้าไมน์นิ่ง ซึ่งในขั้นตอนนี้จะแบ่งออกเป็นขั้นตอนย่อยอีก คือ

**2.2.2.1 การคัดเลือกข้อมูล(Data Selection)**

เป็นการวิเคราะห์แล้วเลือกสิ่งที่สนใจ และมีความสำคัญออกมาจากข้อมูลที่มีอยู่ และเลือกข้อมูลให้ตรงจุดที่สามารถนำมาแก้ปัญหาได้ ซึ่งจะขึ้นอยู่กับวัตถุประสงค์ทางธุรกิจที่ได้กำหนดไว้ในตอนแรก และต้องมีการจัดรูปแบบของข้อมูลให้พร้อมต่อการใช้งาน โดยอาจจะมีการกำหนดชนิด ค่า รูปแบบ และลักษณะที่ชัดเจนเอาไว้ และบางส่วนของข้อมูลที่นำมาใช้ อาจจะต้องมีการแก้ไขให้สามารถนำมาวิเคราะห์ได้ง่ายขึ้น เช่น จากฟิลด์ "DATE_OF_BIRTH" แก้ไขเป็น "AGE" เป็นต้น

**2.2.2.2 การทำข้อมูลให้สมบูรณ์(Data Cleansing)**

ข้อมูลที่ได้มานั้นจะเป็นข้อมูลที่ยังไม่สมบูรณ์ ที่จะสามารถนำไปใช้ผ่านกระบวนการดาต้าไมน์นิ่งได้ จึงต้องมีการจัดการข้อมูลให้สามารถนำไปใช้ได้ก่อน โดยการเตรียมข้อมูลเบื้องต้นมีวิธีการดังนี้

1) เลือกเฉพาะคอลัมน์สำคัญที่คาดว่าจะสามารถนำมาใช้ประโยชน์ได้ และเป็นคอลัมน์ที่มีข้อมูลอยู่ค่อนข้างครบถ้วน

2) สำหรับคอลัมน์ที่มีค่าในทุกๆ แถวเป็นค่าเดียวกัน จะเป็นข้อมูลที่ไม่สามารถแยกความแตกต่างของแต่ละแถวได้เลย ดังนั้นในการทำดาต้าไมน์นิ่งจะไม่สามารถใช้ประโยชน์จากคอลัมน์นี้ได้ ดังนั้นจึงไม่นำคอลัมน์นี้มาพิจารณา

3) คอลัมน์ที่มีค่าที่ไม่ซ้ำกันเลย เช่น หมายเลขสมาชิก ชื่อสมาชิก และหมายเลขโทรศัพท์ เป็นต้น ข้อมูลเหล่านี้ไม่สามารถหาแถวที่มีข้อมูลสัมพันธ์กันได้เลย ในการทำดาต้าไมน์นิ่งจะไม่สามารถนำข้อมูลเหล่านี้มาใช้ประโยชน์ได้ ดังนั้นควรกำจัดคอลัมน์ที่มีข้อมูลไม่ซ้ำกันเลยออก แต่ข้อมูลเหล่านี้อาจให้รายละเอียดบางอย่างที่น่าสนใจกับเราได้ เช่น ข้อมูลหมายเลขลูกค้าสามารถแบ่งลูกค้าเก่ากับลูกค้าใหม่ได้ ข้อมูลหมายเลขโทรศัพท์สามารถบอกตำแหน่งที่อยู่แบบคร่าวๆ ได้ เป็นต้น

4) แก้ไขข้อมูลให้ถูกต้องสมบูรณ์ ได้แก่ การแก้ไขค่าว่างของข้อมูล ซึ่งสามารถแก้ไขได้หลายวิธี เช่น แก้ไขโดยกำจัดข้อมูลที่ในแถวเป็นค่าว่าง(NULL)

5) ปรับเปลี่ยนข้อมูลให้มีค่าเหมาะสมกับการตัดสินใจ เช่น ข้อมูลที่อยู่ของแต่ละคนอาจจะไม่ซ้ำกันเลย ดังนั้นจึงต้องปรับเปลี่ยนข้อมูลให้อยู่ในรูปแบบที่จะสามารถนำไปใช้ได้ เช่น เปลี่ยนให้เป็น Bangkok และ Non-Bangkok อย่างใดอย่างหนึ่ง เป็นต้น

6) การจัดกลุ่มข้อมูลเพื่อลดการกระจาย(Binning Data) เช่น ถ้าเกรดในแต่ละวิชามีเกรด {A, B+, B, C+, C, D+ ,D, F} เพื่อลดการกระจายของข้อมูลเกรด จะสามารถจัดกลุ่มเกรดออกเป็นสามกลุ่ม ได้ดังนี้คือ เกรด {A, B+, B} เป็น High เกรด {C+, C} เป็น Medium และเกรด {D+, D, F} เป็น Low

**2.2.2.3 การปรับเปลี่ยนรูปแบบข้อมูล(Data Transformation)**

เป็นการแปลงข้อมูล รวบรวม และจัดสรรข้อมูล ให้อยู่ในรูปแบบที่เหมาะสมพร้อมที่จะนำไปใช้ในการวิเคราะห์ได้ ซึ่งจะส่งผลให้ข้อมูลที่ผ่านการแปลงข้อมูลมามีคุณภาพมากขึ้น และช่วยให้สอดคล้องกับเทคนิคที่จะนำไปใช้งานได้มากขึ้นด้วย เช่น การแปลงข้อมูลที่อยู่ในรูปของตัวอักษรให้เป็นตัวเลข เป็นต้น

**2.2.3 การทำดาต้าไมน์นิ่ง(Data Mining)**

เป็นการนำอัลกอริทึมของดาต้าไมน์นิ่งมากลั่นกรองข้อมูล เพื่อหารูปแบบความสัมพันธ์และสร้างแบบจำลองที่จะนำมาทำนาย โดยต้องเลือกอัลกอริทึมให้เหมาะสมกับข้อมูลที่ได้เตรียมมา ซึ่งการทำดาต้าไมน์นิ่งมีอยู่ด้วยกัน 4 ชนิด คือ

**2.2.3.1 Clustering Model**

เป็นการแบ่งกลุ่มของฐานข้อมูลให้อยู่เป็นคลัสเตอร์(cluster) ย่อยๆ โดยที่แต่ละคลัสเตอร์ย่อยนั้น จะประกอบไปด้วยออบเจ็กต์(object) ที่เหมือนกัน ซึ่งในการแบ่งกลุ่มข้อมูลนั้นไม่สามารถที่จะกำหนดได้ว่าข้อมูลควรที่จะอยู่ในกลุ่มใด แต่จะเป็นการกำหนดกลุ่มจากธรรมชาติของข้อมูลเองมากกว่า

**2.2.3.2 Predictive Model**

เป็นการสร้างแบบจำลองเพื่อทำนายหาค่าความเป็นไปได้ โดยใช้หลักการสังเกตจากข้อมูลที่มีอยู่ โดยแบบจำลองในการพัฒนาจะแบ่งออกเป็นสองช่วงด้วยกัน คือ

- Training เป็นช่วงในการสร้างแบบจำลองขึ้นมาใหม่โดยใช้ข้อมูลในอดีต และใช้ข้อมูลในปริมาณมาก

- Testing เป็นการตรวจสอบประสิทธิภาพของแบบจำลองใหม่ที่สร้างขึ้นมา ซึ่งใช้ข้อมูลในปริมาณที่ไม่มากนัก

สำหรับเทคนิคที่มีในวิธีนี้มีอยู่สองเทคนิค คือ

- Classification เป็นการทำนายว่าสิ่งที่พิจารณาควรที่จะอยู่ภายในกลุ่มใด ซึ่งสามารถแบ่งกลุ่มได้อย่างชัดเจน

- Value prediction เป็นการทำนายค่าที่เป็นตัวเลขค่าความต่อเนื่องของข้อมูล

**2.2.3.3 Link Analysis**

เป็นการศึกษาถึงความสัมพันธ์ของข้อมูล เพื่อที่จะดูว่ากลุ่มของข้อมูลมีความสัมพันธ์กันในลักษณะใด ซึ่งความสัมพันธ์นี้ถูกเรียกว่า Associations Link Analysis เป็นวิธีที่นิยมเป็นอย่างมากในการนำมาวิเคราะห์เพื่อหาความสัมพันธ์ในการซื้อสินค้าของลูกค้า ซึ่งเทคนิคที่มีในวิธีนี้มีสามเทคนิค คือ Associations discovery Sequential pattern discovery และ Similar time sequence discovery

**2.2.3.4 Deviation Detection**

เป็นวิธีการในการวิเคราะห์หาสิ่งที่มีความแตกต่างในข้อมูล การตรวจจับสิ่งผิดปกติต่างๆ เป็นกรรมวิธีในการหาค่าที่แตกต่างไปจากค่ามาตรฐานหรือค่าที่คาดคิดไว้ ว่าต่างไปมากน้อยเพียงใด ซึ่งเป็นการสรุปข้อมูลออกมาในรูปแบบการแสดงผลทางกราฟิก โดยใช้เทคนิคทางสถิติหรือการแสดงให้เห็นภาพ เทคนิคนี้ถูกใช้ในงานทางด้านการตรวจสอบการปลอมลายเซ็น หรือบัตรเครดิตปลอม รวมทั้งการตรวจหาจุดบกพร่องของชิ้นงานในโรงงานอุตสาหกรรม

**2.2.4 การวิเคราะห์ผลลัพธ์ที่ได้(Analysis of Results)**

เป็นการนำผลลัพธ์ที่ได้จากการทำดาต้าไมน์นิ่ง มาวิเคราะห์ตีความเพื่อหารูปแบบความสัมพันธ์ที่ซ่อนอยู่ โดยถ้าผลลัพธ์ที่ได้ออกมาไม่ตรงกับวัตถุประสงค์ที่วางไว้ ให้ย้อนกลับไปทำใหม่

**2.2.5 การนำความรู้มาใช้(Assimilation of Knowledge)**

เป็นการนำผลลัพธ์ที่ได้หลังจากการวิเคราะห์แล้ว มารวมกับผลการทดสอบอื่นๆ ที่เกี่ยวข้อง เพื่อนำมาวิเคราะห์ให้ได้บทสรุปที่ต้องการ และสามารถนำไปใช้ต่อไปได้

# บทที่ 3

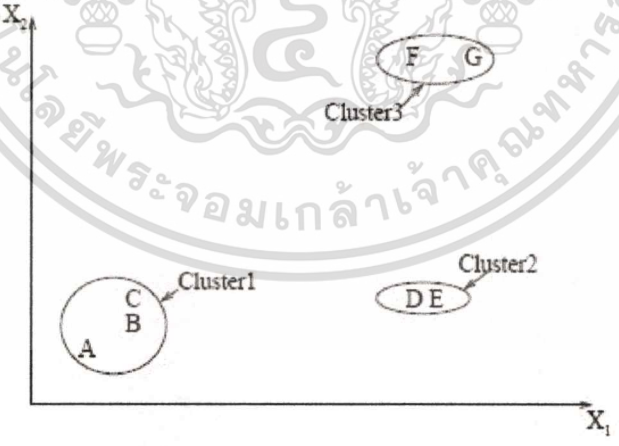
# Database Segmentation

## 3.1 ประเภทของการแบ่งกลุ่มฐานข้อมูล(Database Segmentation)

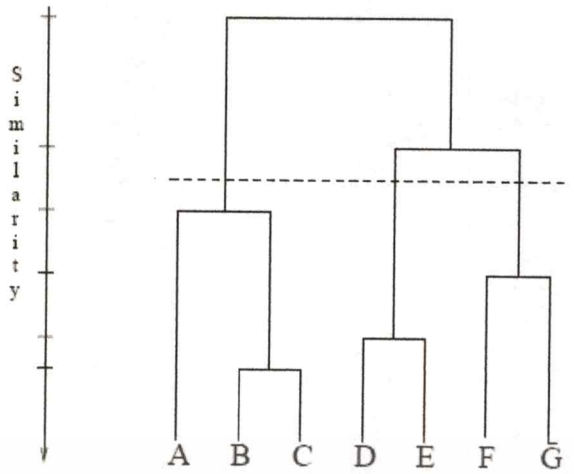
การแบ่งกลุ่มฐานข้อมูลนั้นเรียกได้อีกอย่างหนึ่งว่า Data Clustering ซึ่งมีอยู่ด้วยกันหลายวิธีการ และในแต่ละวิธีการมีอัลกอริทึมที่ใช้อยู่มากมายหลายประเภทด้วยกัน การเลือกอัลกอริทึมที่จะนำมาใช้งานขึ้นอยู่กับชนิดของข้อมูลที่มีอยู่ และนอกจากนั้นยังขึ้นอยู่กับจุดประสงค์ของงานนั้นอีกด้วย ประเภทของการแบ่งกลุ่มฐานข้อมูลมีอยู่ด้วยกันหลายประเภทดังนี้

### 3.1.1 วิธีการแบบ Hierarchical Clustering Algorithms

วิธีการนี้เป็นการจัดกลุ่มข้อมูล โดยให้ข้อมูลแยกออกเป็นลำดับชั้นจากข้อมูลที่มีอยู่ ตามภาพประกอบตามในรูปที่ 3.1 ให้แต่ละออบเจ็กต์เป็นอักษร A ถึง G อยู่ในสามคลัสเตอร์ดังรูป และผลที่ได้จากวิธีการจัดกลุ่มข้อมูลโดยใช้โครงสร้างลำดับชั้นนี้จะใช้เดนโดแกรม(dendrogram) แทนระดับของกลุ่ม และรูปแบบที่มีความสัมพันธ์กัน โดยที่จุดเจ็ดจุดที่คล้ายกันในรูปที่ 3.1 นั้นได้แสดงไว้ในรูปที่ 3.2 และเดนโดแกรมสามารถแตกเป็นระดับชั้นที่ต่างกันเพื่อแสดงคลัสเตอร์ที่ต่างกันของข้อมูลนั้นได้ด้วย




**รูปที่ 3.1** จุดที่ตกบนสามคลัสเตอร์




**รูปที่ 3.2** เดนโดแกรมที่ได้มาจาก single-link

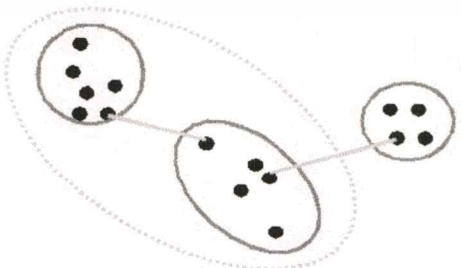
สำหรับรูปแบบในการจัดกลุ่มของวิธีการนี้มีอยู่ด้วยกันสองลักษณะ คือ

1) Agglomerative เป็นการจัดกลุ่มแบบ bottom-up คือเริ่มจากแต่ละออบเจ็กต์ที่อยู่ในกลุ่มคนละกลุ่มกัน แล้วรวมกลุ่มแต่ละกลุ่มเข้าด้วยกันจนกระทั่งสิ้นสุดเงื่อนไข หรือรวมแต่ละกลุ่มเข้าด้วยกันหมดแล้ว

2) Divisive เป็นการจัดกลุ่มแบบ top-down คือเริ่มจากคลัสเตอร์ที่มีออบเจ็กต์อยู่หลายๆ ออบเจ็กต์ในคลัสเตอร์นั้น แล้วแบ่งกลุ่มให้มีขนาดเล็กลงเรื่อยๆ จนสิ้นสุดเงื่อนไข หรือเหลือเพียงออบเจ็กต์เดียว

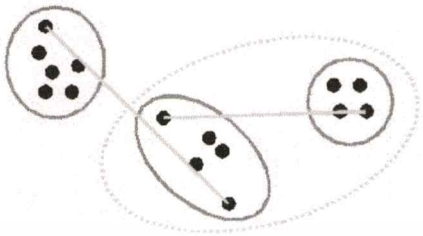
และในวิธีการนี้จะใช้การแสดงข้อมูลในลักษณะของการบอกระยะห่างระหว่างคู่ของออบเจ็กต์หรือคลัสเตอร์ โดยการวัดระยะห่างนั้นแบ่งออกได้เป็น 3 รูปแบบ คือ

1) Single-link คือวัดระยะห่างระหว่างสองคลัสเตอร์ แล้วเลือกใช้ระยะห่างของคู่ออบเจ็กต์ที่สั้นที่สุด ในทุกๆ คู่ของออบเจ็กต์ เป็นวิธีที่เร็วที่สุดแต่ได้ข้อมูลที่มีความสัมพันธ์กันน้อยที่สุด ตามรูปที่ 3.3



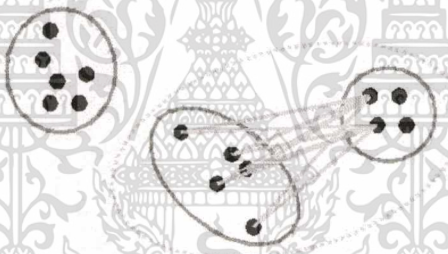
**รูปที่ 3.3** Single-link

2) Complete-link คือ วัดระยะห่างระหว่างสองคลัสเตอร์ แล้วเลือกใช้ระยะห่างของคู่ออบเจ็กต์ที่ยาวที่สุด ในทุกๆ คู่ของออบเจ็กต์ วิธีนี้จะได้กลุ่มของคลัสเตอร์ที่มีความสัมพันธ์กันมาก แต่ว่าหาได้ช้า ตามรูปที่ 3.4



**รูปที่ 3.4** Complete-link

3) Average-link คือวัดระยะห่างระหว่างสองคลัสเตอร์ แล้วเลือกใช้ระยะห่างเฉลี่ยของทุกออบเจ็กต์ วิธีนี้จะได้กลุ่มของคลัสเตอร์ที่มีความสัมพันธ์กันมากแต่แต่ว่าหาได้ช้าเหมือนกับ Complete-link ตามรูปที่ 3.5



**รูปที่ 3.5** Average-link

### 3.1.2 วิธีการแบบ Partitional Algorithms

หลักการของวิธีพาร์ทิชันอัลกอริทึม(Partition Algorithms) นี้คือ ถ้ามีข้อมูลอยู่ n ตัว จะแบ่งข้อมูลออกเป็น k คลัสเตอร์ โดยที่ค่า k มีค่าน้อยกว่าหรือเท่ากับ n ซึ่งแต่ละพาร์ทิชันนั้นก็คือตัวแทนของคลัสเตอร์

ในพาร์ทิชันอัลกอริทึมนั้น มีอัลกอริทึมอยู่หลายๆ วิธี วิธีที่นิยมใช้กันมีดังนี้

#### 3.1.2.1 K-Means Clustering Algorithm

จะทำการจัดกลุ่มโดยอาศัยระยะห่างระหว่างจุดเทียบกับจุดศูนย์กลางของคลัสเตอร์ เป็นตัวหาว่าจุดนั้นควรจะอยู่ในคลัสเตอร์ใด ในการตัดสินใจเลือกค่า k ให้เหมาะสมได้นั้นอาจจะต้องใช้วิธีทดสอบค่า k ไปเรื่อยๆ แล้วประเมินหาค่า k ที่เหมาะสมจากหลักการที่ว่า ระยะห่างระหว่างออบเจ็กต์ในคลัสเตอร์เดียวกันควรจะน้อยที่สุด และระยะห่างระหว่างออบเจ็กต์ที่อยู่ต่างคลัสเตอร์กันควรจะมากที่สุด

**3.1.2..2 K-Medoids**

วิธีการนี้มีอีกชื่อหนึ่งว่า Partitioning Around Medoids (PAM) โดยในแต่ละคลัสเตอร์จะแสดงโดยออบเจ็กต์หนึ่งที่อยู่ใกล้กับศูนย์กลางของคลัสเตอร์

K-Means กับ K-Medoids นั้นจะแตกต่างกันในการกำหนดจุดศูนย์กลาง โดยที่ K-Means จะใช้ค่าเฉลี่ยของระยะห่างระหว่างออบเจ็กต์ในการกำหนดจุดศูนย์กลาง ซึ่งค่าเฉลี่ยจะไม่ได้เป็นค่าที่แสดงถึงตัวออบเจ็กต์ใดๆ เลย แต่ในวิธีการของ K-Medoids จะอาศัยการหา medoid จากออบเจ็กต์ที่อยู่ในคลัสเตอร์นั้นๆ

**3.1.2.3 Clustering Large Applications(CLARA)**

เป็นอัลกอริทึมที่พัฒนามาจาก K-Means และ K-Medoids เพื่อให้เหมาะสมกับข้อมูลที่มีปริมาณมาก โดยจะมีวิธีการคิดตามพื้นฐานของ K-Means และ k-Medoid เป็นหลักและอาศัยหลักการทางสถิติเข้ามาช่วย คือจะทำการสุ่มกลุ่มของออบเจ็กต์ขึ้นมาจากจำนวนออบเจ็กต์ทั้งหมดแล้วนำออบเจ็กต์ที่สุ่มมาไปทำตามอัลกอริทึมของ K-Medoids ซึ่งจะเป็นการลดจำนวนครั้งในการทำงานลงไปได้มาก

**3.1.2.4 Clustering Large Applications based on Randomized Search (CLARANS)**

จะใช้การค้นหาแบบสุ่มเลือกโหนดขึ้นมา เพื่อเปรียบเทียบไปเรื่อยๆ จนกระทั่งได้โหนดเป็นที่น่าพอใจตามต้องการ โดยที่จะไม่ใช้การค้นหาแบบระเอียด โดยจะแสดงถึงรายละเอียดในหัวข้อนี้อีกครั้งหนึ่ง ในบทที่ 4

**3.1.3 วิธีการแบบ Density-based**

มีแนวความคิดที่ว่าจะปล่อยให้คลัสเตอร์ขยายไปเรื่อยๆ ตนกว่าความหนาแน่น(density) หรือก็คือจำนวนของออบเจ็กต์ ของคลัสเตอร์ที่อยู่ใกล้กันมีค่ามากกว่าค่าที่กำหนดเอาไว้ในตอนแรก ซึ่งจะสามารถช่วยให้กำจัดข้อมูลที่ไม่เกี่ยวข้องออกไปได้ อัลกอริทึมที่น่าสนใจในวิธีนี้ได้แก่ DBSCAN OPTICS และ DENCLUE

**3.1.4 วิธีการแบบ Grid-based**

วิธีนี้จะทำการ Quantize ที่ว่างของออบเจ็กต์ลงในเซลล์จำนวนหนึ่ง ซึ่งอยู่ในรูปของโครงสร้างแบบกริด(Grid) ซึ่งในทุกๆ ขั้นตอนของการแบ่งกลุ่มฐานข้อมูล จะถูกจัดการลงบนโครงสร้างแบบกริด มีข้อดีคือ ใช้เวลาในการทำงานไม่มากนัก โดยจะไม่ขึ้นกับจำนวนของข้อมูลแต่จะขึ้นกับจำนวนของเซลล์ในแต่ละส่วนในที่ว่างของ Quantize อัลกอริทึมที่น่าสนใจในวิธีนี้ได้แก่ STING CLIQUE และ Wave cluster

## 3.2 อัลกอริทึมของการจัดกลุ่มที่ใช้วิธีการค้นหาแบบสุ่ม(CLARANS)

อัลกอริทึมของการจัดกลุ่มที่ใช้วิธีการค้นหาแบบสุ่ม(CLARANS) จะอยู่ในรูปของกระบวนการในการหา medoid ออกมาจำนวน k medoid โดยเป็นการหาออกมาจากกราฟ โดยลักษณะของกราฟมีดังนี้

### 3.2.1 ลักษณะของกราฟ

ให้มี $n$ ออบเจ็กต์ กราฟนี้แทนเป็น $G_{n,k}$ แต่ละโหนดจะถูกแทนค่าด้วยเซต(set) ของ $k$ ออบเจ็กต์ $\{O_{m_1},...,O_{m_k}\}$ โดยที่ $O_{m_1},...,O_{m_k}$ เป็น medoid ที่ถูกเลือก ดังนั้นเซตของโหนดในกราฟจะเป็นเซต $\{O_{m_1},...,O_{m_k} \mid O_{m_1},...,O_{m_k}$ เป็นออบเจ็กต์ในเซตข้อมูล(data set)$\}$

โหนดสองโหนดจะเป็นเพื่อนบ้าน(neighbor) กันถ้าในเซตของสองโหนดนั้นมีออบเจ็กต์แตกต่างกันแค่ตัวเดียว หรือจะหมายความว่า โหนดสองโหนด $S_1=\{O_{m_1},...,O_{m_k}\}$ และ $S_2=\{O_{w_1},...,O_{w_k}\}$ จะเป็นเพื่อนบ้านกันก็ต่อเมื่อ ขนาดของเซตจากการอินเตอเซคชัน (intersection) กันระหว่าง $S_1$ กับ $S_2$ มีค่าเป็น $k$ - 1 นั่นก็คือ $|S_1 \cap S_2| = k-1$ เราจะเห็นได้ว่าแต่ละโหนดจะมีเพื่อนบ้าน $k(n - k)$ ตัว ดังนั้นแต่ละโหนดสามารถนำมากำหนดค่า cost ที่อธิบายถึงผลรวมของความแตกต่างระหว่างทุกๆ ออบเจ็กต์ และ medoid ของคลัสเตอร์ของมัน

อัลกอริทึม CLARANS จะไม่ทำการตรวจสอบในทุกๆ เพื่อนบ้านของโหนด แต่จะหาที่กราฟต้นแบบ $G_{n,k}$ เลย โดยที่ CLARANS จะทำการสุ่มเลือกโหนดใดๆ ที่อยู่ภายในกราฟ และทำการสุ่มเลือกโหนดที่อยู่ใกล้เคียงกันขึ้นมาอีกหนึ่งโหนด โดยถ้า cost ของโหนดใกล้เคียงที่ถูกเลือกขึ้นมามีค่าน้อยกว่าโหนดที่เลือกอยู่ในตอนแรก อัลกอริทึม CLARANS จะเลือกใช้โหนดที่อยู่ใกล้เคียงโหนดต่อๆ ไปขึ้นมา เพื่อจะทำการเปรียบเทียบไปเรื่อยๆ โดยจะทำการเลือกสุ่มโหนดใกล้เคียงไปเรื่อยๆ จนกระทั่งได้โหนดที่พอใจแล้ว หรืออาจมีการกำหนดถึงจำนวนที่มากที่สุดของโหนดที่ต้องการจะตรวจสอบเอาไว้ก่อน ต่อมาโหนดที่ได้ถูกเลือกไว้จะมีการระบุว่าเป็น local minimum เอาไว้

ในการหาผลลัพธ์ที่ดีที่สุดอัลกอริทึม CLARANS จะทำขั้นตอนต่างๆ ในการค้นหา local minimum ซ้ำอีกครั้งจากโหนดตอนเริ่มต้นที่ต่างกันออกไปตามจำนวนครั้งที่กำหนด จากนั้นโหนดกับค่า cost ที่ต่ำที่สุดที่ถูกเลือกมานั้น จะถือว่าเป็น cluster สุดท้าย โดยให้ maxneighbor เป็นจำนวนที่มากที่สุดของโหนดที่อยู่ใกล้เคียงเพื่อการตรวจสอบ และให้ numlocal เป็นจำนวนของ local minima ที่ได้ โดยที่ CLARANS จะวาดตัวอย่างของเพื่อนบ้านในแต่ละขั้นตอนของการหาเลย นี่ถือเป็นข้อดีที่ไม่จำกัดการค้นหาอยู่ที่บริเวณบริเวณเดียว ขั้นตอนในของอัลกอริทึม CLARANS มีดังนี้

**3.2.2 ขั้นตอนของอัลกอริทึม CLARANS**

รูปแบบของอัลกอริทึมของ CLARANS มีดังนี้

1) ใส่ค่าพารามิเตอร์จำนวนคลัสเตอร์ *numlocal* และ *maxneighbor* ในตอนเริ่มต้นให้ i เท่ากับ 1 และให้ *mincost* เป็นค่ามากๆ

2) ให้ *current* เป็นโหนดที่สุ่มขึ้นในกราฟ $G_{n,k}$ ซึ่งในโหนดนี้จะมีจำนวนข้อมูลเท่ากับจำนวนคลัสเตอร์ที่ใส่เข้ามา

3) ให้ *j* เท่ากับ 1

4) พิจารณาสุ่มโหนดเพื่อนบ้าน(*S*) ของ *current* แล้วคำนวณหาค่า cost ของทั้งสองโหนด

5) ถ้า S มีค่า cost ต่ำกว่าค่า cost ของ *current* ให้ตั้ง *current* เป็น *S* และกลับไปที่ขั้นที่ 3

6) ถ้าค่า cost ของ S ไม่ต่ำกว่าค่า cost ของ *current* ให้เพิ่มค่า j ขึ้นอีก 1 หากค่า $j \leq$ *maxneighbor* กลับไปที่ขั้นที่ 4

7) แต่ถ้า $j >$ *maxneighbor* เปรียบเทียบค่า cost ของ *current* กับ *mincost* ถ้าค่า cost ของ *current* มีค่าน้อยกว่าให้ตั้งค่า *mincost* เท่ากับค่า cost ของ *current* และตั้ง *current* ให้เป็น *bestnode*

8) เพิ่มค่า *i* ขึ้น 1 ถ้า $i >$ *numlocal* จะได้ค่าที่ต้องการคือ *bestnode* ถ้า $i \leq$ *numlocal* ให้กลับไปที่ขั้นที่ 2

ขั้นที่ 3 ถึง 6 เป็นการค้นหาโหนดที่มีค่า cost ต่ำ แต่ถ้าโหนดปัจจุบัน(current node) ได้ถูกนำไปเทียบกับค่าที่มากที่สุดของเพื่อนบ้านของโหนดแล้ว(*maxneighbor*) ยังคงมีค่า cost ต่ำที่สุดบริเวณโหนดปัจจุบันนั้นถือว่าเป็น local minimum แล้วในขั้นที่ 7 ค่า cost ของบริเวณดังกล่าวจะถูกนำมาเปรียบเทียบกับค่า cost ที่ต่ำที่สุดที่ได้รับมาอยู่แล้ว ค่า cost ที่ต่ำกว่าจากทั้งสองอันจะถูกเก็บอยู่ในค่า *mincost* หลังจากนั้นจะทำซ้ำเพื่อหา local minima อื่นๆ จนกระทั่งพบค่า *numlocal* ของบริเวณนั้น

สำหรับการหาค่า cost ระหว่างออบเจ็กต์สองออบเจ็กต์ที่ใช้ในระบบนี้ จะหาได้จากระยะห่างยูคลีเดียน(Euclidean distance) โดยที่ระยะห่างยูคลีเดียนระหว่างออบเจ็กต์สองออบเจ็กต์ให้เป็น $(x_1, x_2, \ldots x_n)$ และ $(y_1, y_2, \ldots y_n)$ จะนิยามได้ว่า

$$\text{Euclidean Distance} = \sqrt{|x_1 - y_1|^2 + |x_2 - y_2|^2 + \ldots + |x_n - y_n|^2}$$

จากขั้นตอนดังกล่าวสามารถเขียนเป็น pseudocode ได้ดังนี้

```
Input numCluster numlocal and maxneighbor
Set mincost to large number
For i = 1 to numlocal
        randomly select current in graph
        set j = 1
        Repeat
                randomly select neighbor S of current
                if cost(S) < cost(current )
                        set current = S
                        set j = 1
                else j++
        Until j > maxneighbor
        compare cost(current) with mincost
        If cost(current) < mincost
                mincost = cost(current)
                set bestnode = current
        End If
End For
Return bestnode
```

ประสิทธิภาพของ CLARANS จะขึ้นกับค่าพารามิเตอร์สองตัวคือ *maxneighbor* และ *numlocal* เป็นสำคัญ ถ้าทั้งสองพารามิเตอร์มีค่ามาก ก็จะมีความเป็นไปได้ว่ากลุ่มคลัสเตอร์จะมีค่า cost ที่ต่ำ ซึ่งก็คือมีความสัมพันธ์กันของข้อมูลที่ดี และได้ข้อมูลที่มีประสิทธิภาพสูง

# บทที่ 4
# วิธีการดำเนินงาน

## 4.1 ระบบงาน

การทำงานของระบบจัดกลุ่มข้อมูลที่พัฒนาขึ้น จะประกอบไปด้วยส่วนหลักๆ สามส่วน คือ ส่วนข้อมูลเข้า(Input) ส่วนประมวลผล(Process) และส่วนแสดงผล(Output) ซึ่งมีรายละเอียดดังต่อไปนี้

### 4.1.1 ส่วนข้อมูลเข้า

เป็นข้อมูลเข้าที่จะนำมาใช้ในการจัดกลุ่มข้อมูล โดยข้อมูลนำเข้าที่ใช้คือ ไฟล์ฐานข้อมูลที่ประกอบไปด้วยข้อมูลที่ต้องการจัดกลุ่ม โดยฐานข้อมูลที่เลือกเข้ามานั้นจะต้องเป็นฐานข้อมูลจาก Microsoft SQL Server

### 4.1.2 ส่วนประมวลผล

ส่วนประมวลผลจะแบ่งออกเป็นสองส่วนด้วยกัน คือส่วนการเตรียมข้อมูล และส่วนการทำดาต้าไมน์นิ่ง

#### 4.1.2.1 ส่วนการเตรียมข้อมูล

เป็นขั้นตอนในการเตรียมข้อมูลเพื่อจะนำไปใช้ต่อในขั้นตอนการทำไมน์นิ่ง โดยจะประกอบไปด้วย

1) การเลือกข้อมูล(Data Selection) เมื่อได้ทำการเลือกไฟล์ฐานข้อมูลที่ต้องการได้แล้ว ต้องมาเลือกตาราง และฟิลด์ของข้อมูลที่ต้องการในการทำไมน์นิ่งต่อไป

2) การทำข้อมูลให้สมบูรณ์(Data Cleansing) เมื่อทำการเลือกข้อมูลเสร็จแล้ว ต้องทำการแก้ไขค่าของข้อมูลต่างๆ เช่น ค่าว่าง โดยวิธีต่างๆ เพื่อให้ข้อมูลที่จะนำเข้าการทำดาต้าไมนิ่งนั้นมีประสิทธิภาพ

3) การปรับเปลี่ยนรูปแบบข้อมูล(Data Transformation) เป็นขั้นตอนในการเปลี่ยนรูปแบบของข้อมูลให้อยู่ในรูปแบบที่เหมาะสมต่อวิธีการทำดาต้าไมน์นิ่งที่ต้องการ เช่น ปรับเปลี่ยนข้อมูลของ Numerical ให้อยู่ในช่วงหนึ่งๆ หรือการปรับเปลี่ยนข้อมูลที่อยู่ในรูปของข้อความตัวอักษรไปเป็นข้อมูลที่เป็นตัวเลข เพื่อให้ข้อมูลมีความเหมาะสมต่อการทำไมน์นิ่ง

4) การสำรวจข้อมูล(Data Exploration) เป็นการสำรวจข้อมูลขั้นสุดท้าย ก่อนที่จะนำไปทำการดาต้าไมน์นิ่ง ซึ่งจะแสดงผลในรูปแบบของกราฟแท่ง และกราฟวงกลม

**4.1.2.2 ส่วนการทำดาต้าไมน์นิ่ง**

เมื่อข้อมูลผ่านส่วนการเตรียมข้อมูลมาแล้ว โปรแกรมระบบจะทำการประมวลผลจัดกลุ่มข้อมูล เพื่อให้ได้ข้อมูลตามจำนวนกลุ่มที่ต้องการ

**4.1.3 ส่วนแสดงผล**

เมื่อโปรแกรมทำการประมวลผลเรียบร้อยแล้ว จะแสดงผลลัพธ์ที่ได้จากการประมวลผลออกมาทางจอภาพของคอมพิวเตอร์

## 4.2 ขั้นตอนการทำงานของระบบ

การทำงานของระบบในแต่ละขั้นตอนมีการทำงานดังต่อไปนี้

**4.2.1 การเลือกข้อมูล(Data Selection)**

1) เลือกไฟล์จากฐานข้อมูล โดยระบุชื่อของเซิร์ฟเวอร์ และชื่อฐานข้อมูลที่ต้องการติดต่อ

2) ระบุวิธีการเลือกข้อมูล

- เลือกข้อมูลจากหนึ่งตาราง
- เลือกข้อมูลจากหลายตาราง

3) เลือกแอตทริบิวต์(Attribute) ที่ต้องการนำมาทำดาต้าไมน์นิ่ง

- ข้อมูลจากหนึ่งตาราง เลือกแอตทริบิวต์ของตารางที่เลือกได้เลย
- ข้อมูลจากหลายตาราง ให้ใช้คำสั่ง SQL ในการเลือกแอตทริบิวต์

4) กดปุ่ม execute เพื่อทำงานในขั้นต่อไป

**4.2.2 การทำข้อมูลให้สมบูรณ์(Data Cleaning)**

1) ถ้าเลือก Auto Cleaning จะเป็นการลบลบเรคอร์ดที่มีค่าเป็น Null ออกทุกๆ เรคอร์ด และจะเป็นการไปยังขั้นตอนต่อไป

2) เลือกแอตทริบิวต์ที่ต้องการแก้ไข แล้วเลือกวิธีในการแก้ไขข้อมูล

สำหรับข้อมูลที่เป็น Categorical มีสามทางเลือกคือ

- เติมค่าฐานนิยม (Mode)
- เติม Unknown
- ลบเรคอร์ดที่มีค่าเป็น Null

สำหรับข้อมูลที่เป็น Numerical มีสามทางเลือกคือ

- เติมค่าเฉลี่ย
- เติมค่าที่ต้องการ
- ลบเรคอร์ดที่มีค่าเป็น Null

3) เลือกแอตทริบิวต์ที่ต้องการแก้ไขให้ครบทุกๆ แอตทริบิวต์ ถ้าแอตทริบิวต์ไหนที่ไม่มีค่าว่างจะปรากฏปุ่ม No Missing ขึ้นมา ให้กดที่ปุ่มนั้น เมื่อทำครบทุกแอตทริบิวต์แล้วจะไปยังขั้นตอนต่อไปทันที

**4.2.3 การปรับเปลี่ยนรูปแบบข้อมูล(Data Transformation)**

แบ่งข้อมูลออกเป็นสองแบบ คือข้อมูลที่เป็น Categorical กับข้อมูลที่เป็น Numerical

1) ข้อมูลที่เป็น Categorical สามารถเลือกรูปแบบการแปลงข้อมูลได้สองวิธี คือ One of N Coding และแปลงข้อมูลให้อยู่ในรูปของตัวเลข หากไม่ต้องการแปลงค่าให้เลือกแอตทริบิวต์แล้วเลือกที่ No Transform

2) ข้อมูลที่เป็น Numerical สามารถเลือกรูปแบบการแปลงข้อมูลได้สามวิธี คือ Normalization Construct Attribute และ Numerical to Categorical หากไม่ต้องการแปลงค่าให้เลือกแอตทริบิวต์แล้วเลือกที่ No Transform

เมื่อทำครบแล้วให้ไปยังขั้นตอนสำรวจข้อมูล หรือขั้นการจัดกลุ่มข้อมูลต่อไปได้เลย

**4.2.4 การสำรวจข้อมูล(Data Exploration)**

เลือกที่แอตทริบิวต์ที่ต้องการดูข้อมูล ระบบจะแสดงข้อมูลโดยที่

1) ข้อมูลที่เป็น Categorical ระบบจะเลือกค่าที่แตกต่างกัน(distinct) ออกมา แล้วนับจำนวนของแต่ละค่ามาแสดงผลในรูปแบบของกราฟแท่ง และกราฟวงกลม

2) ข้อมูลที่เป็น Numerical ระบบจะแบ่งข้อมูลออกเป็น 10 ช่วง แล้วนับจำนวนข้อมูลของแต่ละช่วงมาแสดงผลในรูปแบบของกราฟแท่ง และกราฟวงกลม

**4.2.5 การจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม CLARANS**

1) ใส่จำนวนคลัสเตอร์ที่ต้องการจัดกลุ่ม ค่า Numlocal และค่า Maxneighbor

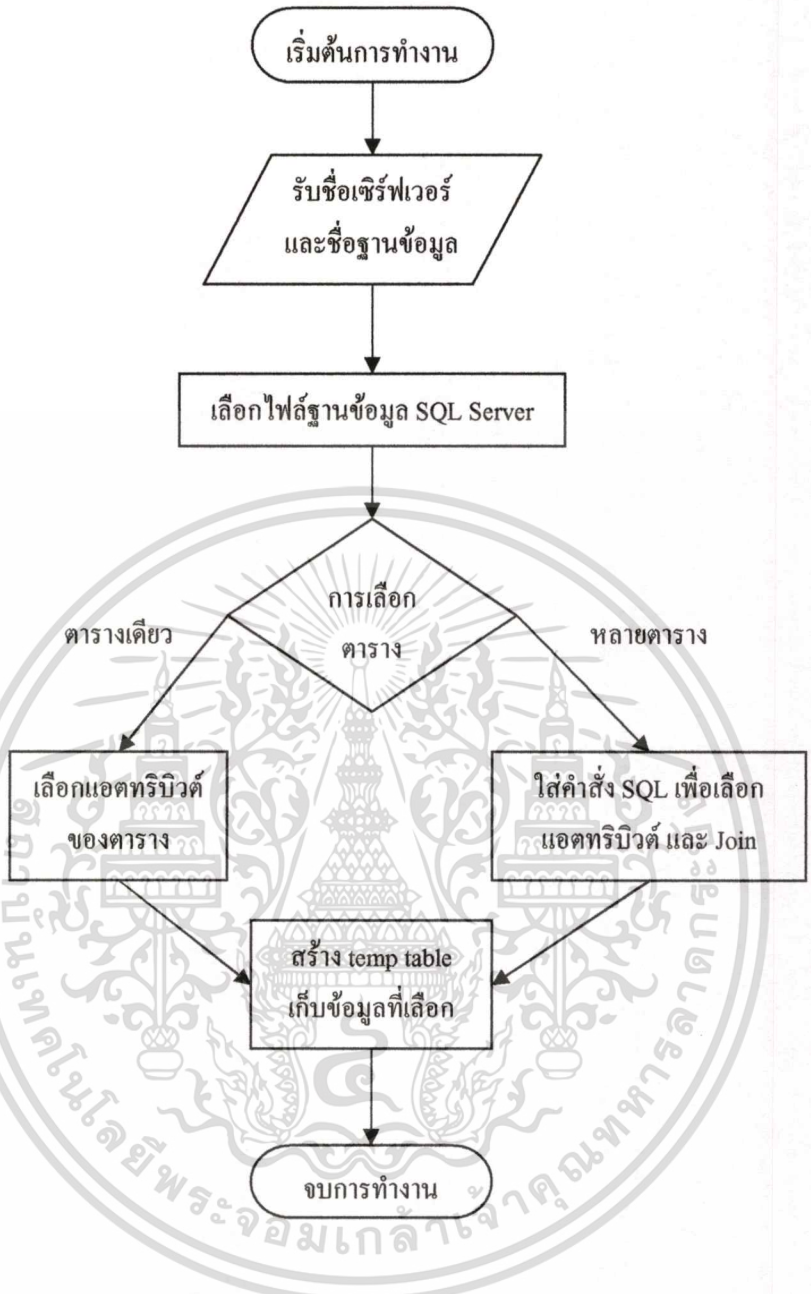
2) กดปุ่ม execute เพื่อประมวลผลจัดกลุ่มข้อมูล

3) จะแสดงผลการจัดกลุ่มที่ได้

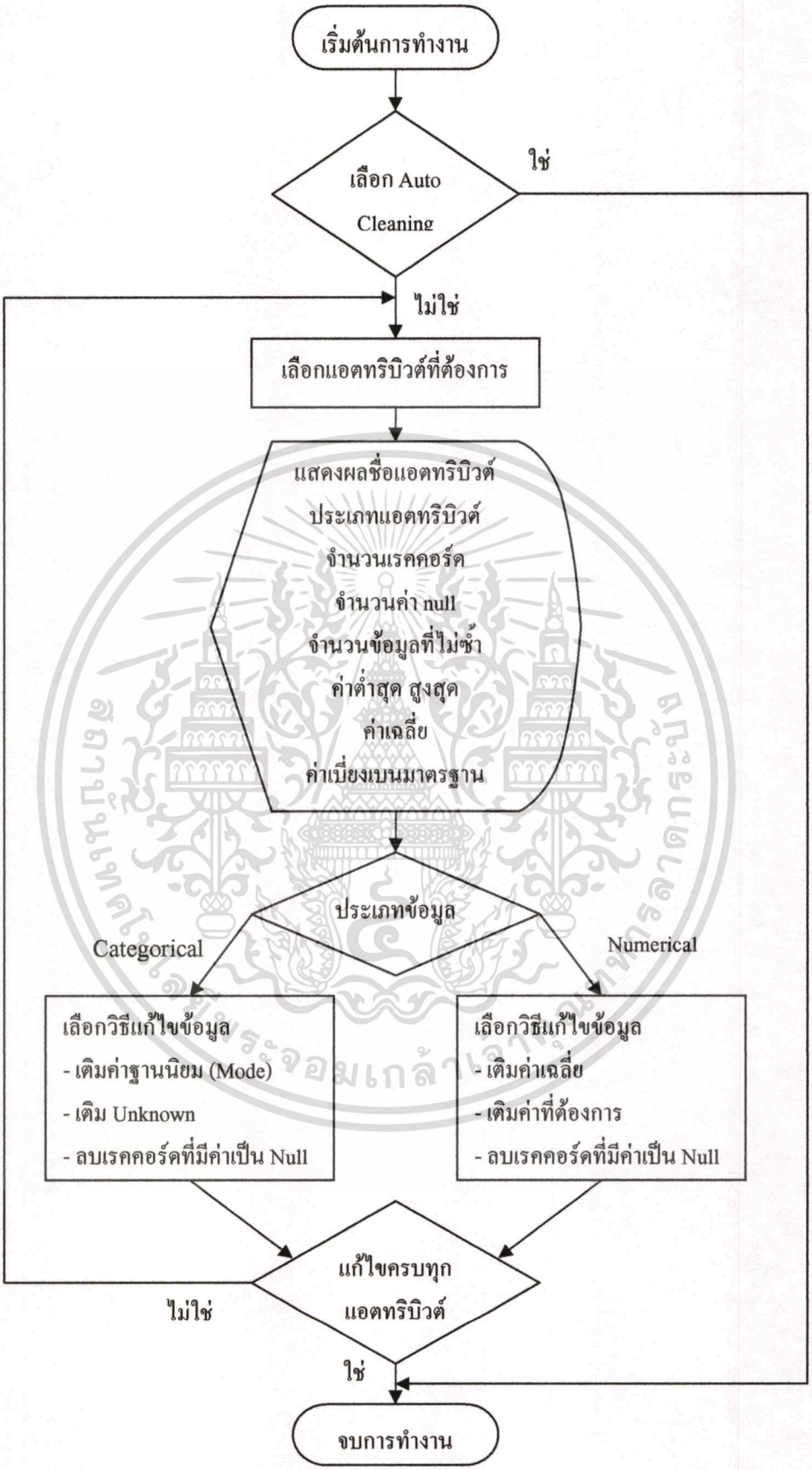
4) จบการทำงาน

## 4.3 แผนผังการทำงานของระบบ

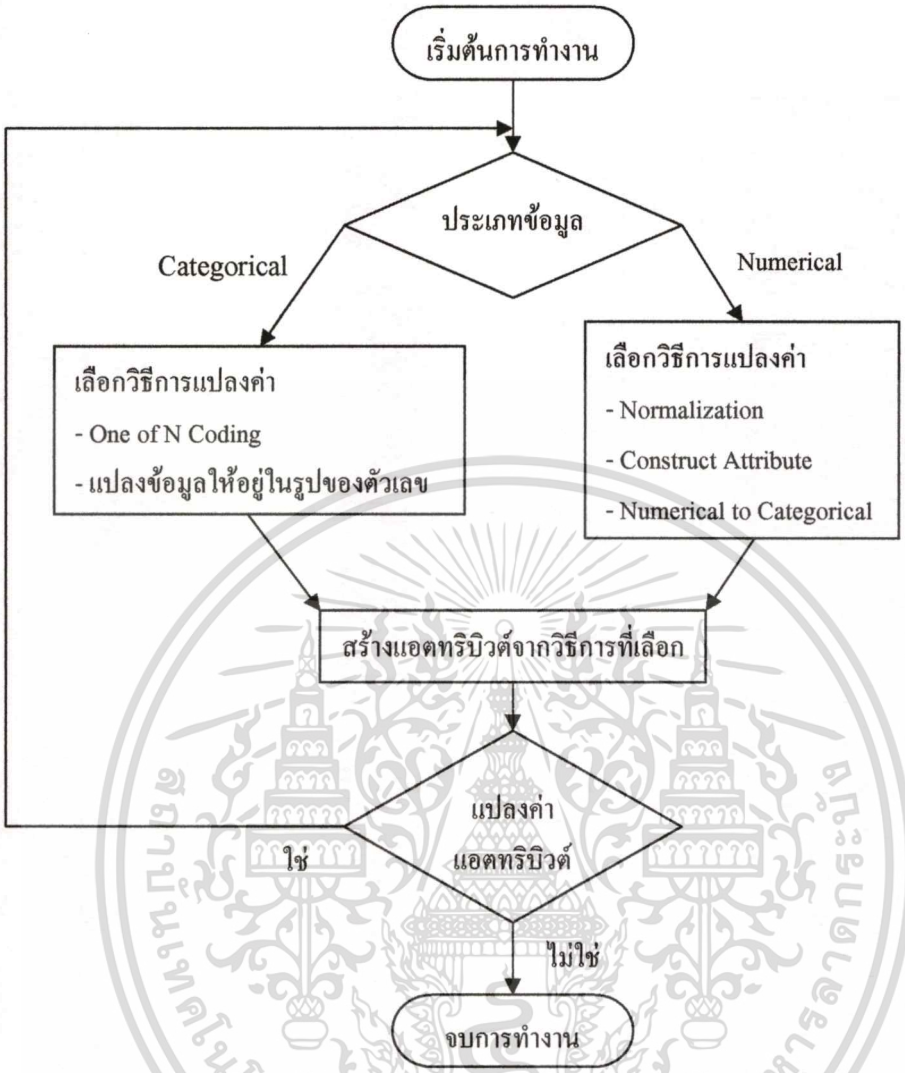

**รูปที่ 4.1** ผังงานแสดงขั้นตอนการทำงานหลักของระบบ



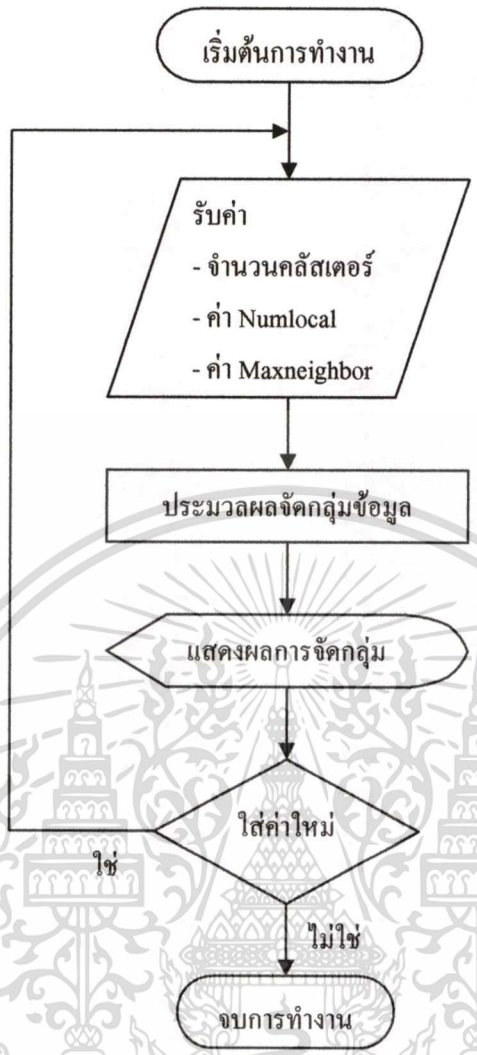

**รูปที่ 4.2** ผังงานย่อยแสดงขั้นตอนการเลือกข้อมูล(Data Selection)




**รูปที่ 4.3** ผังงานย่อยแสดงขั้นตอนแก้ไขข้อมูลให้สมบูรณ์(Data Cleaning)




**รูปที่ 4.4** ผังงานย่อยแสดงขั้นตอนการแปลงข้อมูลให้อยู่ในรูปที่เหมาะสม(Data Transformation)




**รูปที่ 4.5** ผังงานย่อยแสดงขั้นตอนการจัดกลุ่มข้อมูลด้วยอัลกอริทึม CLARANS

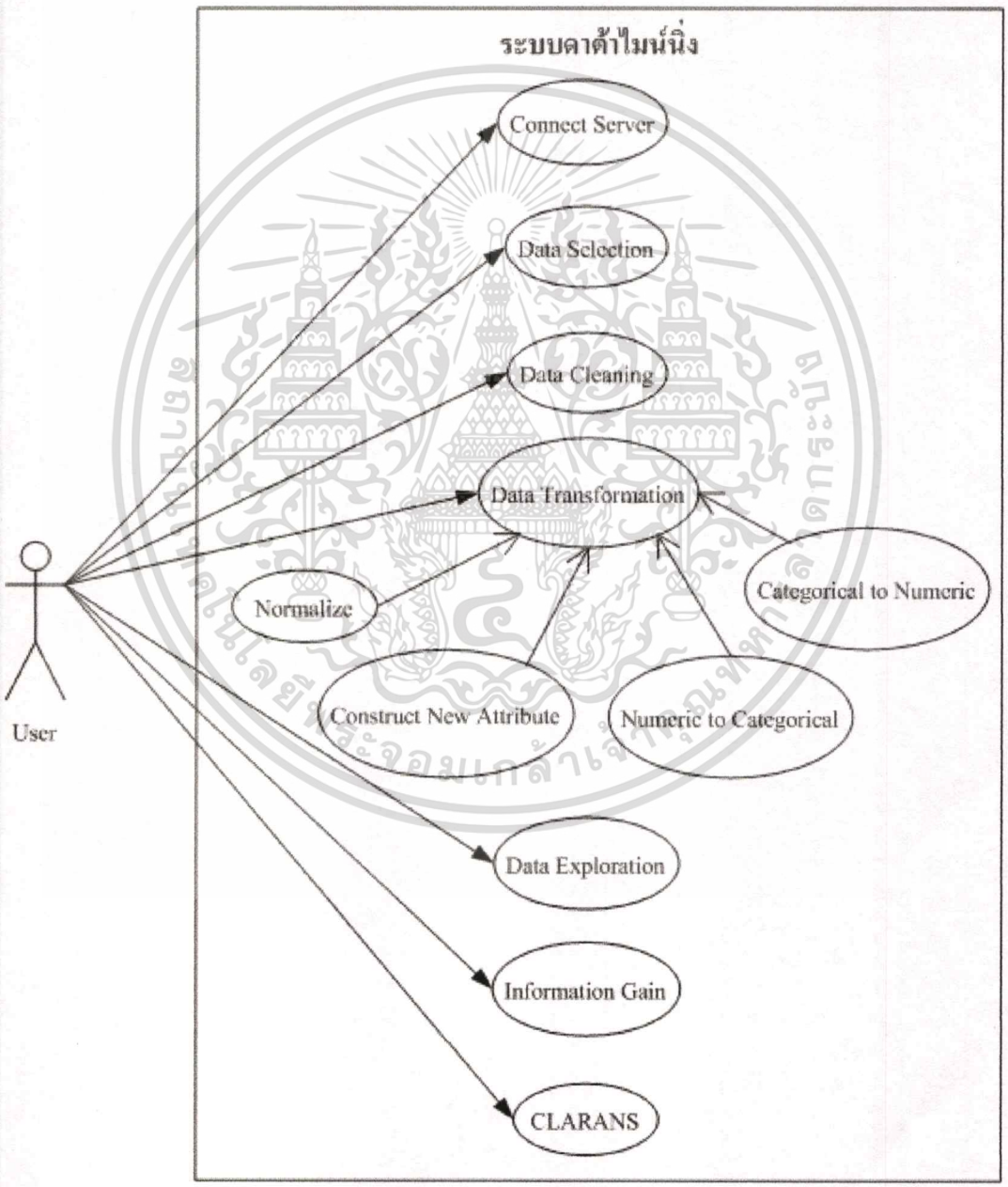
**4.4 แบบจำลองเชิงแนวคิดของระบบงาน(Conceptual Model)**

ออกแบบแบบจำลองใน 3 มุมมอง ดังนี้

**4.4.1 มุมมองเชิงฟังก์ชันของระบบ(Functional View)**

**4.4.1.1 ยูสเคสไดอะแกรม(Use Case Diagram)**

แสดงการทำงานของระบบและความสัมพันธ์ระหว่างแอคเตอร์(Actor) และยูสเคส ดังรูปที่ 4.5 ประกอบด้วยแอคเตอร์ คือ ผู้ใช้(User) และมีความสัมพันธ์กับ 9 ยูสเคส ดังนี้




**รูปที่ 4.6** แสดงยูสเคสของระบบดาต้าไมน์นิ่ง

**4.4.1.2 คำอธิบายยูสเคส(Use Case Description)**

จากยูสเคสในรูปที่ 4.5 อธิบายได้ดังนี้

**ตารางที่ 4.1** คำอธิบายยูสเคส Connect Server

| Use Case Name | Connect Server | |
|---|---|---|
| Brief Description | เมื่อเริ่มต้นเข้าสู่ระบบผู้ใช้ ต้องทำการเชื่อมต่อกับฐานข้อมูลก่อน | |
| Preconditions | ผู้ใช้มีความต้องการเชื่อมต่อกับฐานข้อมูลเพื่อการทำไมน์นิ่ง | |
| Postconditions | เชื่อมต่อกับฐานข้อมูล | |
| Actors | ผู้ใช้ | |
| Triggering events | ผู้ใช้ใส่ชื่อเซิร์ฟเวอร์ และชื่อฐานข้อมูล | |
| Input | ชื่อเซิร์ฟเวอร์ และชื่อฐานข้อมูล | |
| Output | สถานะการระเชื่อมต่อกับฐานข้อมูล | |
| Flow of Events | Actor | System |
| | - ผู้ใช้ใส่ชื่อเซิร์ฟเวอร์ และชื่อฐานข้อมูล<br>- เลือกทำการเชื่อมต่อใหม่ได้ตามต้องการ | - ตรวจสอบชื่อเซิร์ฟเวอร์ และชื่อฐานข้อมูล<br>- ตอบรับหรือปฏิเสธการเชื่อมต่อ |
| Exception condition | - หากไม่ใส่ชื่อเซิร์ฟเวอร์ หรือชื่อฐานข้อมูล ระบบจะแจ้งข้อผิดพลาด<br>- ระบบไม่สามารถเชื่อมต่อกับฐานข้อมูลได้ หากไม่มีไฟล์ฐานข้อมูลนั้น<br>- มีการแจ้งข้อผิดพลาด | |

ก่อนจะใช้งานระบบได้ในส่วนอื่นต่อไป ผู้ใช้จะต้องทำการเชื่อมต่อเข้าสู่ระบบฐานข้อมูลก่อน โดยการกรอกชื่อเซิร์ฟเวอร์ และชื่อฐานข้อมูลที่ต้องการ หากชื่อเซิร์ฟเวอร์ และชื่อฐานข้อมูลถูกต้อง ระบบก็จะทำการเชื่อมต่อเข้ากับฐานข้อมูลได้

**ตารางที่ 4.2** คำอธิบายยูสเคส Data Selection

| Use Case Name | Data Selection | |
|---|---|---|
| Brief Description | ให้ผู้ใช้เลือกตาราง และแอตทริบิวต์ที่ต้องการทำไมน์นิ่ง | |
| Preconditions | ระบบต้องเชื่อมต่อกับฐานข้อมูลแล้ว | |
| Postconditions | แสดงรายชื่อแอตทริบิวต์ที่เลือกจากตารางนั้น | |
| Actors | ผู้ใช้ | |
| Triggering events | ผู้ใช้เลือกตาราง และแอตทริบิวต์ที่ต้องการ | |
| Input | ชื่อแอตทริบิวต์ หรือคำสั่ง SQL | |
| Output | ชื่อแอตทริบิวต์ที่เลือก | |
| Flow of Events | Actor | System |
| | - ผู้ใช้เลือกตาราง แล้วเลือกชื่อแอตทริบิวต์ที่ต้องการ<br>- ผู้ใช้เลือกว่าจะเลือกข้อมูลจากตารางเดียว หรือจากหลายตาราง<br>- หากเลือกข้อมูลจากหลายตาราง ผู้ใช้ต้องใส่คำสั่ง SQL ในการเลือกแอตทริบิวต์ | - ตรวจสอบว่าผู้ใช้เลือกจากตารางเดียวหรือหลายตาราง<br>- ตรวจสอบคำสั่ง SQL<br>- ตอบรับหรือปฏิเสธการเชื่อมต่อ |
| Exception condition | - มีการแจ้งข้อผิดพลาด | |

เมื่อเชื่อมต่อกับฐานข้อมูลแล้ว ระบบจะแสดงตาราง และแอตทริบิวต์ที่มีในฐานข้อมูลนั้น ผู้ใช้สามารถเลือกตาราง และแอตทริบิวต์ที่ต้องการ เพื่อนำมาใช้ในการวิเคราะห์ข้อมูลในขั้นต่อๆ ไป ผู้ใช้ควรจะกำหนดวัตถุประสงค์ของการวิเคราะห์ข้อมูลไว้ล่วงหน้า เพื่อสามารถเลือกข้อมูลที่ต้องการได้อย่างถูกต้อง ไม่ต้องกลับมาเริ่มเชื่อมต่อกับฐานข้อมูลใหม่อีกรอบหนึ่ง

**ตารางที่ 4.3** คำอธิบายยูสเคส Data Cleaning

| Use Case Name | Data Cleaning | |
|---|---|---|
| Brief Description | ให้ผู้ใช้เลือกวิธีในการทำให้ข้อมูลสมบูรณ์ | |
| Preconditions | ผู้ใช้ต้องเลือกแอตทริบิวต์ที่ต้องการทำไมน์นิ่งแล้ว | |
| Postconditions | ทุกแอตทริบิวต์ผ่านการทำให้ข้อมูลสมบูรณ์ ไม่มีค่าว่างเหลืออยู่ | |
| Actors | ผู้ใช้ | |
| Triggering events | ผู้ใช้เลือกแอตทริบิวต์แล้วเลือกวิธีการที่ต้องการ | |
| Input | ชื่อแอตทริบิวต์ และวิธีการที่เลือกไว้ | |
| Output | แอตทริบิวต์ถูกทำให้ไม่มีค่าว่าง | |
| Flow of Events | Actor | System |
| | - ผู้ใช้เลือกชื่อแอตทริบิวต์ที่ต้องการ<br>- ผู้ใช้เลือกวิธีการในการทำให้ข้อมูลสมบูรณ์ | - แสดงรายละเอียดของข้อมูลในแอตทริบิวต์ที่ผู้ใช้เลือก<br>- ตรวจสอบว่าข้อมูลในแอตทริบิวต์เป็น Categorical หรือ Numerical<br>- แสดงวิธีการทำให้ข้อมูลสมบูรณ์ตามชนิดของข้อมูล<br>- ประมวลผลตามวิธีการที่ผู้ใช้เลือก |
| Exception condition | - หากทำไม่ครบทุกแอตทริบิวต์ระบบจะไม่ทำงานต่อในขั้นต่อไป | |

เมื่อเลือกแอตทริบิวต์ที่ต้องการแล้ว ผู้ใช้จะต้องทำข้อมูลสมบูรณ์ โดยเลือกวิธีที่ต้องการตามแต่ละประเภทของข้อมูลดังที่ได้อธิบายไว้ในรายงาน

**ตารางที่ 4.4** คำอธิบายยูสเคส Data Transformation

| Use Case Name | Data Transformation | |
|---|---|---|
| Brief Description | ให้ผู้ใช้เลือกวิธีการปรับเปลี่ยนข้อมูลที่ต้องการ | |
| Preconditions | ข้อมูลต้องผ่านการทำให้ข้อมูลสมบูรณ์แล้ว | |
| Postconditions | ปรับเปลี่ยนรูปแบบของข้อมูลให้เหมาะสมต่อการทำไมน์นิ่ง | |
| Actors | ผู้ใช้ | |
| Triggering events | ผู้ใช้เลือกวิธีการเปลี่ยนรูปแบบข้อมูล | |
| Input | วิธีการปรับเปลี่ยนข้อมูล | |
| Output | รายละเอียดของวิธีการที่เลือก | |
| Flow of Events | Actor | System |
| | - ผู้ใช้เลือกวิธีการปรับเปลี่ยนข้อมูล | - ตรวจสอบวิธีการปรับเปลี่ยนข้อมูล<br>- ตรวจสอบว่าข้อมูลในแอตทริบิวต์เป็น Categorical หรือ Numerical<br>-แสดงรายละเอียดของแต่ละวิธีการ |
| Exception condition | - มีการแจ้งข้อผิดพลาด | |

เมื่อผ่านการทำข้อมูลสมบูรณ์แล้ว ผู้ใช้จะต้องทำการปรับเปลี่ยนรูปแบบของข้อมูลให้เหมาะสมต่อวิธีการในการวิเคราะห์ข้อมูลที่ต้องการ เช่น การแปลงตัวอักษรเป็นตัวเลข สำหรับการพัฒนาระบบนี้ใช้อัลกอริทึม CLARANS ในการวิเคราะห์แล้วจัดกลุ่มข้อมูล ซึ่งจะเหมาะสมกับข้อมูลที่เป็นประเภทตัวเลขเท่านั้น ดังนั้นหากต้องการวิเคราะห์ข้อมูลโดยใช้อัลกอริทึม CLARANS จะต้องทำการแปลงข้อมูลประเภทตัวอักษรให้เป็นประเภทตัวเลขก่อน

**ตารางที่ 4.5** คำอธิบายยูสเคส Normalize

| Use Case Name | Normalize | |
|---|---|---|
| Brief Description | เมื่อผู้ใช้ต้องการปรับเปลี่ยนข้อมูลแบบ Normalize | |
| Preconditions | ข้อมูลต้องผ่านการทำให้ข้อมูลสมบูรณ์แล้ว | |
| Postconditions | ปรับเปลี่ยนข้อมูลที่เป็น Numerical แบบ Normalize | |
| Actors | ผู้ใช้ | |
| Triggering events | ผู้ใช้เลือกวิธีการปรับเปลี่ยนข้อมูลแบบ Normalize | |
| Input | - วิธีการในการทำ Normalize<br>- ค่า Min และ Max หากเลือกวิธีการ Min-Max Normalization | |
| Output | แอตทริบิวต์ที่ปรับเปลี่ยนข้อมูลแล้ว | |
| Flow of Events | Actor | System |
| | - เลือกแอตทริบิวต์ที่ต้องการแปลงค่า<br>- เลือกวิธีการ Normalize<br>- หากเลือกวิธีการ Min-Max Normalization ให้ใส่ค่า Min และ Max ที่ต้องการ | - แสดงรายละเอียดของแอตทริบิวต์ที่ผู้ใช้เลือก<br>- ตรวจสอบค่า Min และ Max<br>- ประมวลผลแปลงข้อมูล<br>-แสดงรายละเอียดของแอตทริบิวต์ที่ผ่านการปรับเปลี่ยนข้อมูลแล้ว |
| Exception condition | - ถ้าผู้ใช้ไม่ระบุค่า Min และค่า Max ระบบจะใส่ค่า 0 และ 1 ตามลำดับ | |

เป็นวิธีการในการปรับเปลี่ยนข้อมูลให้อยู่ในช่วง Min Max ตามแต่ที่ผู้ใช้กำหนด ซึ่งจะใช้กับข้อมูลประเภทตัวเลข

**ตารางที่ 4.6** คำอธิบายยูสเคส Construct New Attribute

| Use Case Name | Construct New Attribute | |
|---|---|---|
| Brief Description | เมื่อผู้ใช้ต้องการปรับเปลี่ยนข้อมูลแบบ Construct New Attribute | |
| Preconditions | ข้อมูลต้องผ่านการทำให้ข้อมูลสมบูรณ์แล้ว | |
| Postconditions | เพิ่มแอตทริบิวต์ใหม่ตามสูตรคำนวณที่ต้องการ | |
| Actors | ผู้ใช้ | |
| Triggering events | ผู้ใช้เลือกวิธีการปรับเปลี่ยนข้อมูลแบบ Construct New Attribute | |
| Input | ชื่อแอตทริบิวต์ใหม่ และค่าการคำนวณ | |
| Output | แอตทริบิวต์ที่ปรับเปลี่ยนข้อมูลแล้ว | |
| Flow of Events | Actor | System |
| | - เลือกแอตทริบิวต์ที่ต้องการแปลงค่า<br>- ผู้ใช้ใส่ชื่อแอตทริบิวต์ใหม่<br>- ใส่สูตรคณิตศาสตร์ตามที่ต้องการ | - แสดงรายละเอียดของแอตทริบิวต์ที่ผู้ใช้เลือก<br>- ตรวจสอบชื่อแอตทริบิวต์ไม่ให้ซ้ำกับที่มีอยู่<br>- ประมวลผลตามสูตรคำนวณที่ผู้ใช้ใส่<br>- แสดงรายละเอียดของแอตทริบิวต์ที่ผ่านการปรับเปลี่ยนข้อมูลแล้ว |
| Exception condition | - ถ้าผู้ใช้ไม่ระบุชื่อแอตทริบิวต์ระบบจะตั้งชื่อให้อัตโนมัติ<br>- หากใส่สูตรคณิตศาสตร์ที่ระบบไม่สามารถคำนวณได้ ระบบจะแจ้งข้อผิดพลาด | |

เป็นวิธีการในการปรับเปลี่ยนข้อมูลโดยการใส่สูตรคณิตศาสตร์ ตามแต่ที่ผู้ใช้กำหนดเพื่อสร้างแอตทริบิวต์ใหม่ขึ้นมา ผู้ใช้สามารนำแอตทริบิวต์ต่างๆ กันมาใส่สูตรคณิตศาสตร์ร่วมกันได้ ซึ่งจะใช้กับข้อมูลประเภทตัวเลข

**ตารางที่ 4.7** คำอธิบายยูสเคส Numeric to Categorical

| Use Case Name | Numeric to Categorical | |
|---|---|---|
| Brief Description | เมื่อผู้ใช้ต้องการปรับเปลี่ยนข้อมูลแบบ Numeric to Categorical | |
| Preconditions | ข้อมูลต้องผ่านการทำให้ข้อมูลสมบูรณ์แล้ว | |
| Postconditions | ปรับเปลี่ยนข้อมูลที่เป็น Numerical ไปเป็น Categorical | |
| Actors | ผู้ใช้ | |
| Triggering events | ผู้ใช้เลือกวิธีการปรับเปลี่ยนข้อมูลแบบ Numeric to Categorical | |
| Input | ช่วงของตัวเลข จำนวนกลุ่ม และระดับของข้อมูล | |
| Output | แอตทริบิวต์ที่ปรับเปลี่ยนข้อมูลแล้ว | |
| Flow of Events | Actor | System |
| | - เลือกแอตทริบิวต์ที่ต้องการแปลงค่า<br>- ผู้ใช้ใส่ช่วงตัวเลขของข้อมูลที่ต้องการ<br>- ใส่จำนวนกลุ่มที่ต้องการ<br>- ใส่ระดับของข้อมูลที่ต้องการ | - แสดงค่าน้อยที่สุด และมากที่สุดของข้อมูลในแอตทริบิวต์<br>- ตรวจสอบช่วงตัวเลขใหม่ และระดับข้อมูลที่ผู้ใช้ใส่<br>- ประมวลผลแปลงค่าข้อมูล<br>- แสดงรายละเอียดของแอตทริบิวต์ที่ผ่านการปรับเปลี่ยนข้อมูลแล้ว |
| Exception condition | - ถ้าผู้ใช้ไม่ระบุช่วงตัวเลขระบบจะแจ้งข้อผิดพลาด | |

เป็นวิธีการในการปรับเปลี่ยนข้อมูล โดยการแปลงข้อมูลที่อยู่ในรูปของตัวเลขให้ไปอยู่ในรูปของตัวอักษรตามที่ผู้ใช้กำหนด

**ตารางที่ 4.8** คำอธิบายยูสเคส Categorical to Numeric

| Use Case Name | Categorical to Numeric | |
|---|---|---|
| Brief Description | เมื่อผู้ใช้ต้องการปรับเปลี่ยนข้อมูลแบบ Categorical to Numeric | |
| Preconditions | ข้อมูลต้องผ่านการทำให้ข้อมูลสมบูรณ์แล้ว | |
| Postconditions | ปรับเปลี่ยนข้อมูลที่เป็น Categorical ไปเป็น Numerical | |
| Actors | ผู้ใช้ | |
| Triggering events | ผู้ใช้เลือกวิธีการปรับเปลี่ยนข้อมูลแบบ Categorical to Numeric | |
| Input | ประเภทในการแปลงค่า | |
| Output | แอตทริบิวต์ที่ปรับเปลี่ยนข้อมูลแล้ว | |
| Flow of Events | Actor | System |
| | - เลือกแอตทริบิวต์ที่ต้องการแปลงค่า<br>- เลือกประเภทในการแปลงค่า | - แสดงรายละเอียดของแอตทริบิวต์ที่ผู้ใช้เลือก<br>- ตรวจสอบประเภทในการแปลงค่า<br>- ประมวลผลแปลงค่าข้อมูล<br>- แสดงรายละเอียดของแอตทริบิวต์ที่ผ่านการปรับเปลี่ยนข้อมูลแล้ว |
| Exception condition | - ถ้าผู้ใช้ไม่ระบุวิธีการปรับเปลี่ยนข้อมูลระบบจะแจ้งข้อผิดพลาด<br>- ถ้าผู้ใช้ไม่เลือกแอตทริบิวต์แล้วเลือกปรับเปลี่ยนข้อมูลระบบจะแจ้งข้อผิดพลาด | |

เป็นวิธีการในการปรับเปลี่ยนข้อมูลโดยการแปลงข้อมูลที่อยู่ในรูปของตัวอักษรให้ไปอยู่ในรูปของตัวเลข

**ตารางที่ 4.9** คำอธิบายยูสเคส Data Exploration

| Use Case Name | Data Exploration | |
|---|---|---|
| Brief Description | เมื่อผู้ใช้ต้องการสำรวจดูข้อมูล | |
| Preconditions | ข้อมูลต้องผ่านการทำให้ข้อมูลสมบูรณ์แล้ว | |
| Postconditions | แสดงรายละเอียดของข้อมูล | |
| Actors | ผู้ใช้ | |
| Triggering events | ผู้ใช้เลือกแอตทริบิวต์ที่ต้องการสำรวจข้อมูล | |
| Input | แอตทริบิวต์ที่ผู้ใช้เลือก | |
| Output | ข้อมูลของแอตทริบิวต์ กราฟแท่ง กราฟวงกลม | |
| Flow of Events | Actor | System |
| | - เลือกแอตทริบิวต์ที่ต้องการสำรวจข้อมูล<br>- เลือกประเภทการแสดงกราฟ | - แสดงรายละเอียดของแอตทริบิวต์ที่ผู้ใช้เลือก<br>- แสดงรูปแบบของกราฟตามรูปแบบที่ผู้ใช้เลือก |
| Exception condition | - ถ้าผู้ใช้ไม่ระบุวิธีการแสดงกราฟ ระบบจะแสดงกราฟเป็นแบบ 2 มิติ | |

เป็นส่วนที่ใช้ในการตรวจสอบข้อมูลของข้อมูล เช่น ค่าเฉลี่ย ค่า min/max จำนวนของข้อมูล เป็นต้น โดยระบบจะแสดงออกมาในรูปแบบของกราฟตามแต่ที่ผู้ใช้ต้องการ

**ตารางที่ 4.10** คำอธิบายยูสเคส Information Gain

| Use Case Name | Information Gain | |
|---|---|---|
| Brief Description | ใช้การหาค่า Gain เพื่อนำไปใช้ในการไมน์นิ่ง | |
| Preconditions | ข้อมูลต้องผ่านการปรับเปลี่ยนข้อมูลเป็น Categorical แล้ว | |
| Postconditions | เก็บข้อมูลลงฐานข้อมูล | |
| Actors | ผู้ใช้ | |
| Triggering events | ผู้ใช้เลือกค่า Gain จากข้อมูลที่ได้ | |
| Input | แอตทริบิวต์ที่เป็น Categorical | |
| Output | ตารางเก็บค่า Gain | |
| Flow of Events | Actor | System |
| | - เลือกแอตทริบิวต์ที่ต้องการหาค่า Gain<br>- เลือกค่า Gain ที่ต้องการ<br>- ใส่ชื่อฐานข้อมูลที่ต้องการ<br>- เลือกบันทึกลงฐานข้อมูล | - แสดงค่า Entropy ของแอตทริบิวต์ที่เลือก<br>- ประมวลผลหาค่า Gain<br>- บันทึกค่า Gain ลงฐานข้อมูล |
| Exception condition | - ถ้าผู้ใช้ไม่เลือกแอตทริบิวต์ระบบจะแจ้งเตือนให้เลือกแอตทริบิวต์<br>- ถ้าผู้ใช้ไม่ใส่ชื่อฐานข้อมูลระบบจะตั้งชื่อฐานข้อมูลให้อัตโนมัติ | |

เป็นส่วนที่ใช้ในการหาค่า Information gain เพื่อใช้ในการเลือก attribute ในแต่ละ node เพื่อไปใช้ทำดาต้าไมน์นิ่งต่อไป ซึ่ง attribute ตัวใดที่มีค่า Information gain สูงสุดหรือว่ามีค่า entropy น้อยจะถูกเลือกให้เป็น attribute ของ node นั้น และ attribute ตัวนี้จะลดจำนวนข้อมูลที่จะใช้ในการสร้าง tree

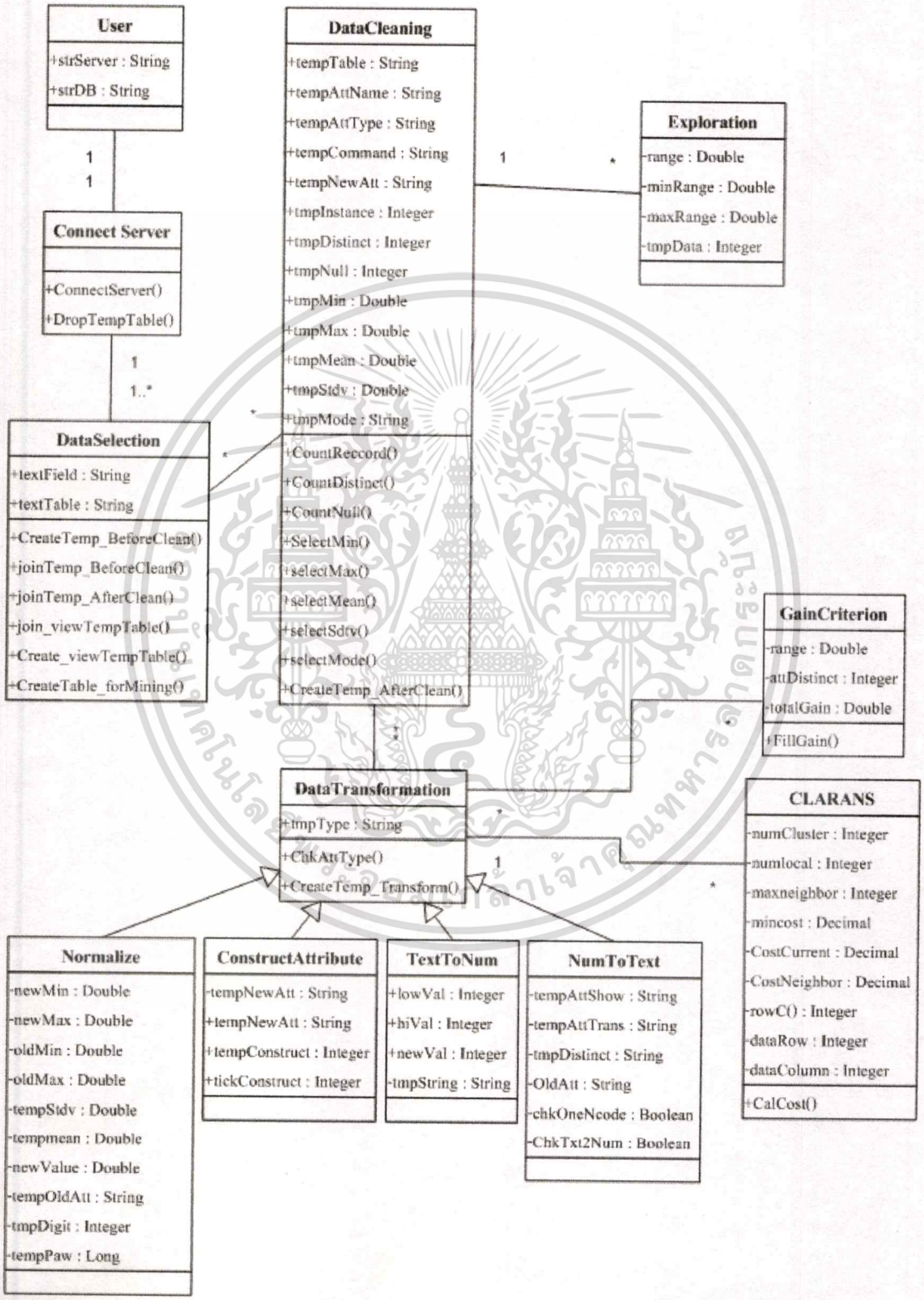
**ตารางที่ 4.11** คำอธิบายยูสเคส CLARANS

| Use Case Name | CLARANS | |
|---|---|---|
| Brief Description | ใช้จัดกลุ่มข้อมูลด้วยอัลกอริทึม CLARANS | |
| Preconditions | ข้อมูลต้องผ่านการปรับเปลี่ยนข้อมูลเป็น Numerical แล้ว | |
| Postconditions | แสดงผลการจัดกลุ่มของข้อมูล | |
| Actors | ผู้ใช้ | |
| Triggering events | ผู้ใช้เลือกจัดกลุ่มข้อมูล | |
| Input | จำนวนคลัสเตอร์ ค่า numlocal และค่า maxneighbor | |
| Output | ข้อมูลที่ผ่านการจัดกลุ่มแล้ว | |
| Flow of Events | Actor | System |
| | - ผู้ใช้ใส่จำนวนคลัสเตอร์ ค่า numlocal และ maxneighbor<br>- เลือกประมวลผล และดูการจัดกลุ่มที่ได้ | - แสดงรายละเอียดของข้อมูลในฐานข้อมูลก่อนการจัดกลุ่ม<br>- ประมวลผลจัดกลุ่มข้อมูลตามค่าที่ผู้ใช้ใส่<br>- แสดงผลของการจัดกลุ่มข้อมูล |
| Exception condition | - ถ้าผู้ใช้ไม่ระบุจำนวนคลัสเตอร์ ค่า numlocal และ maxneighbor ระบบจะแจ้งข้อผิดพลาด<br>- ถ้าผู้ใช้ระบุจำนวนคลัสเตอร์ ค่า numlocal และ maxneighbor เป็น 0 ระบบจะแจ้งข้อผิดพลาด | |

เป็นส่วนในการวิเคราะห์จัดกลุ่มข้อมูลโดยใช้อัลกอริทึม CLARANS โดยผู้ใช้จะต้องใส่จำนวนคลัสเตอร์ที่ต้องการ รวมไปถึงค่า numlocal และค่า maxneighbor ด้วย ระบบจะทำการจัดกลุ่มข้อมูลแล้วแสดงจุดศูนย์กลางคลัสเตอร์ แสดงข้อมูลและจำนวนข้อมูลที่อยู่ในคลัสเตอร์ แสดงค่าเฉลี่ยของข้อมูลในคลัสเตอร์ และแสดงค่า cost เฉลี่ยทั้งภายในคลัสเตอร์และระหว่างคลัสเตอร์

### 4.4.2 มุมมองเชิงโครงสร้างของระบบ(Structural View)

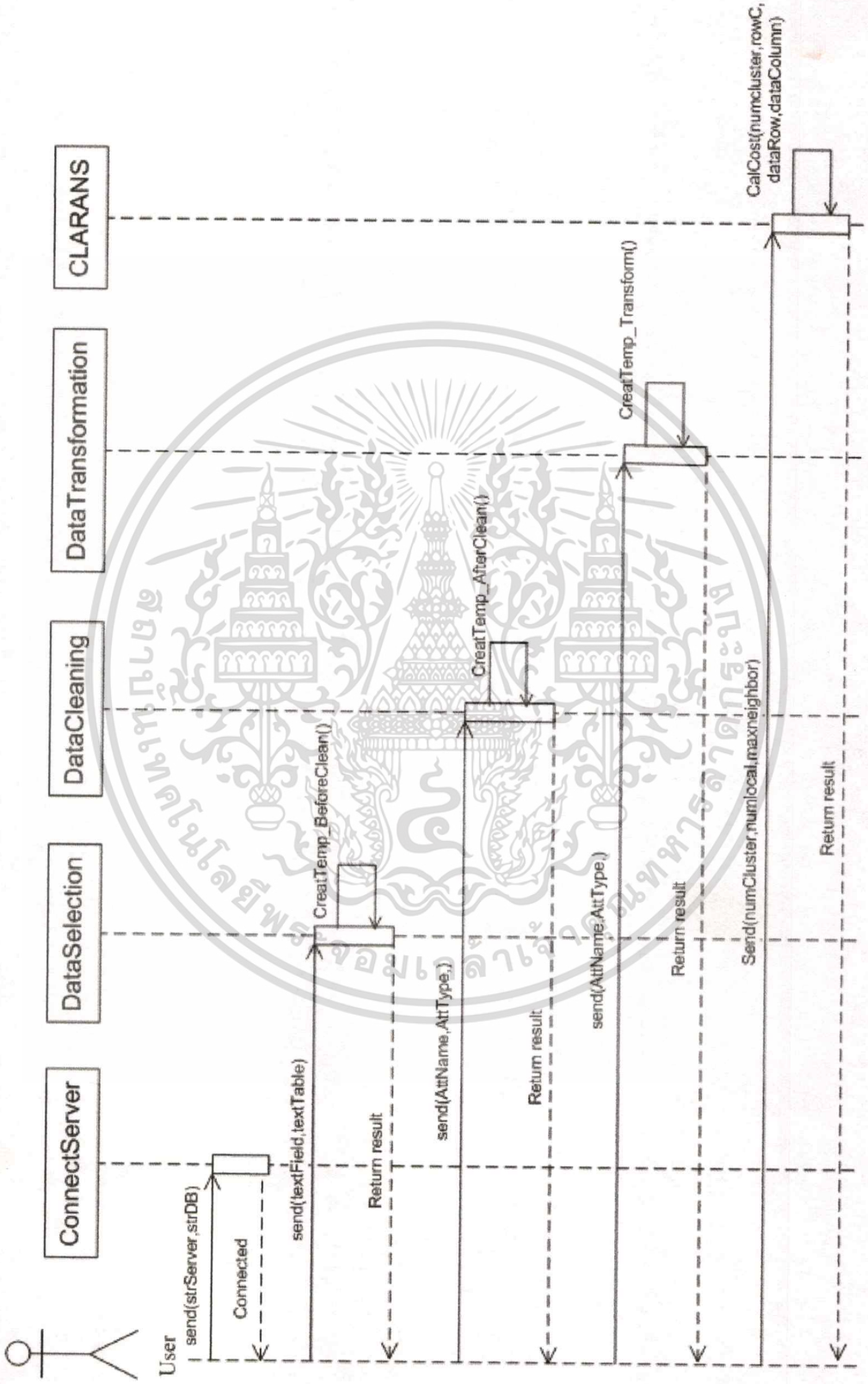
แสดงเป็นคลาสไดอะแกรม(Class Diagram) ดังรูปที่ 4.7




**รูปที่ 4.7** แสดงคลาสไดอะแกรมของระบบดาต้าไมน์นิ่ง

**4.4.3 มุมมองเชิงพฤติกรรมของระบบ(Behavioral View)**

แสดงเป็นซีเควนต์ ไดอะแกรม(Sequence Diagram) ดังรูปที่ 4.8




**รูปที่ 4.8** แสดงซีเควนต์ไดอะแกรมของการจัดกลุ่มข้อมูล

# บทที่ 5
# การประยุกต์การใช้โปรแกรม

## 5.1 เครื่องมือที่ใช้ในการพัฒนาระบบ

โปรแกรมภาษาที่ใช้ในการพัฒนาระบบงาน คือ Microsoft Visual Basic .NET 2003 ซึ่งเป็นเครื่องมือที่ใช้สามารถใช้พัฒนาระบบงานบนระบบปฏิบัติการ Windows ได้ ส่วนระบบฐานข้อมูลที่ใช้ คือ Microsoft SQL Server 2000 ในการติดต่อกับตัวโปรแกรมภาษา Microsoft Visual Basic .NET 2003

## 5.2 ขั้นตอนและรายละเอียดในกานใช้งาน

แสดงถึงขั้นตอนวิธีการใช้งานของโปรแกรม และรายละเอียดของโปรแกรมซึ่งมีดังต่อไปนี้

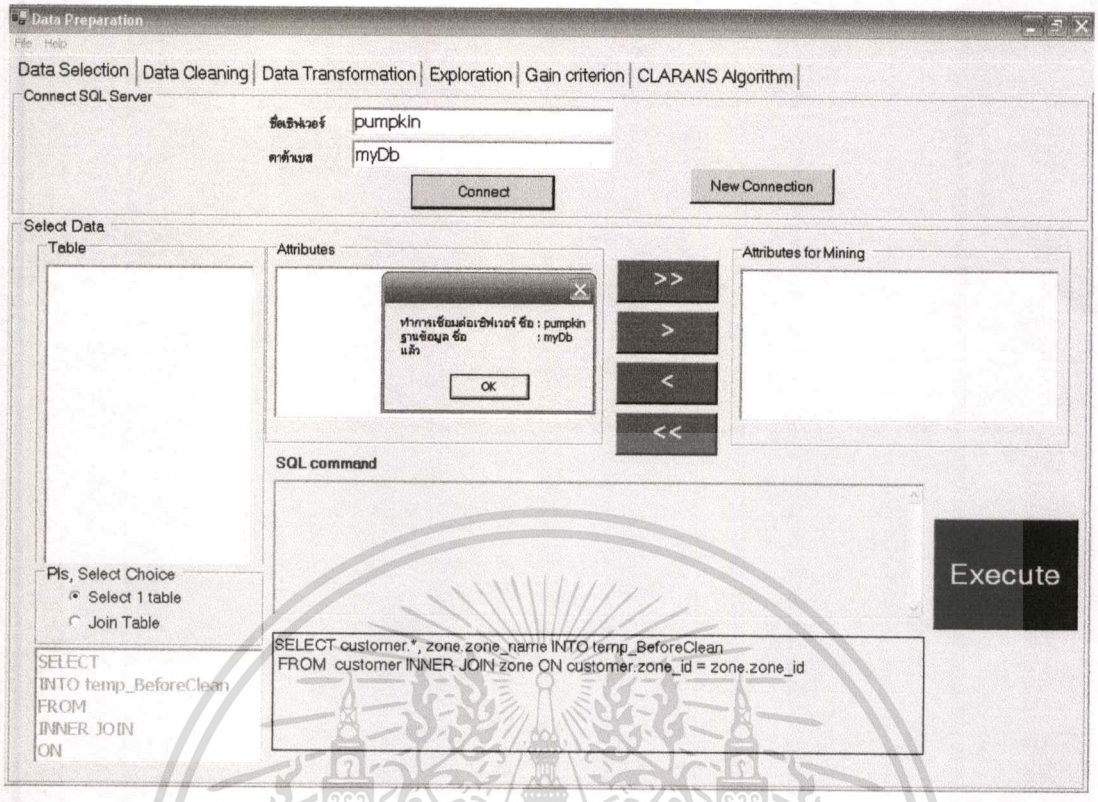
### 5.2.1 การติดต่อกับฐานข้อมูล

เมื่อเข้าสู่โปรแกรมในขั้นตอนแรกจะต้องติดต่อกับฐานข้อมูลก่อน ซึ่งฐานข้อมูลที่ใช้ในการติดต่อ คือ Microsoft SQL Server 2000 โดย เมื่อเปิดโปรแกรมขึ้นมา ผู้ใช้ต้องกรอกข้อมูลชื่อเซิร์ฟเวอร์ และชื่อดาต้าเบส ที่ต้องการติดต่อ แล้วกดปุ่ม Connect เพื่อเชื่อมต่อกับฐานข้อมูลที่ได้ระบุไว้ ระบบจะแสดงข้อความให้ทราบว่าได้เชื่อมต่อเรียบร้อยแล้ว ดังในรูปที่ 5.1

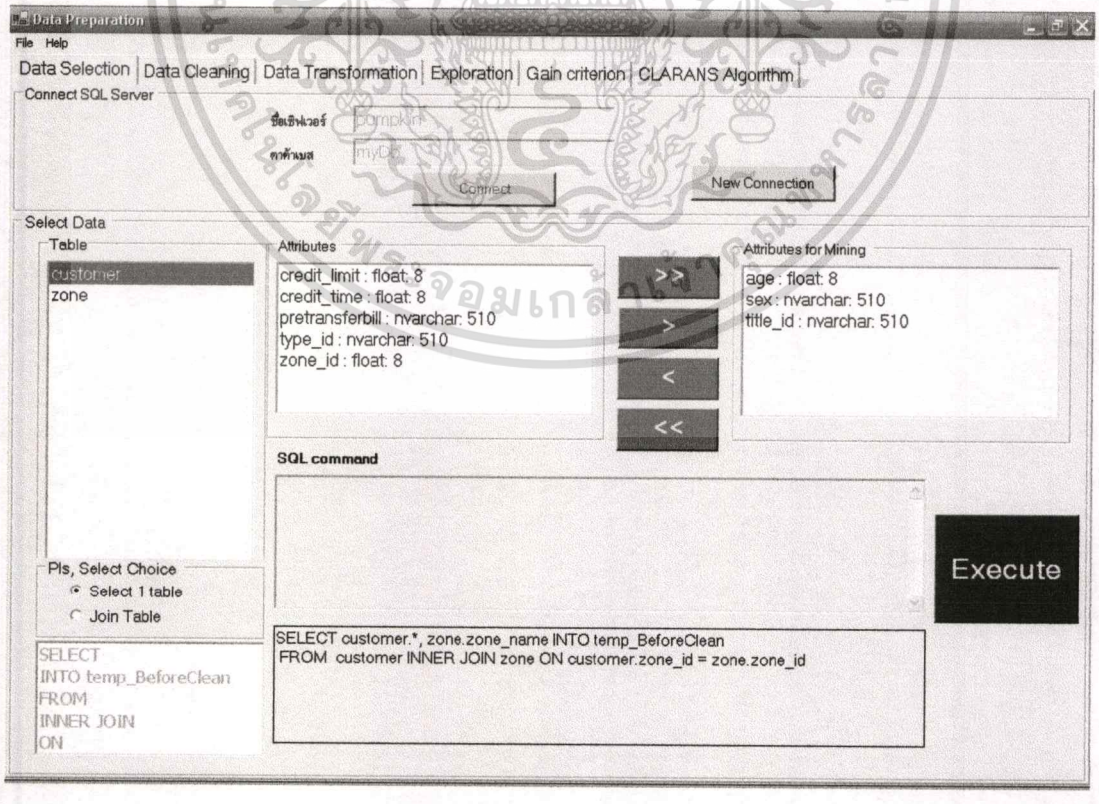
### 5.2.2 การเลือกข้อมูล(Data Selection)

เมื่อติดต่อกับฐานข้อมูลแล้วหน้าจอจะแสดงรายชื่อของตารางในช่อง Table เมื่อเราเลือกที่ชื่อตาราง ในช่อง Attributes จะปรากฏรายชื่อแอตทริบิวต์ของตารางนั้น

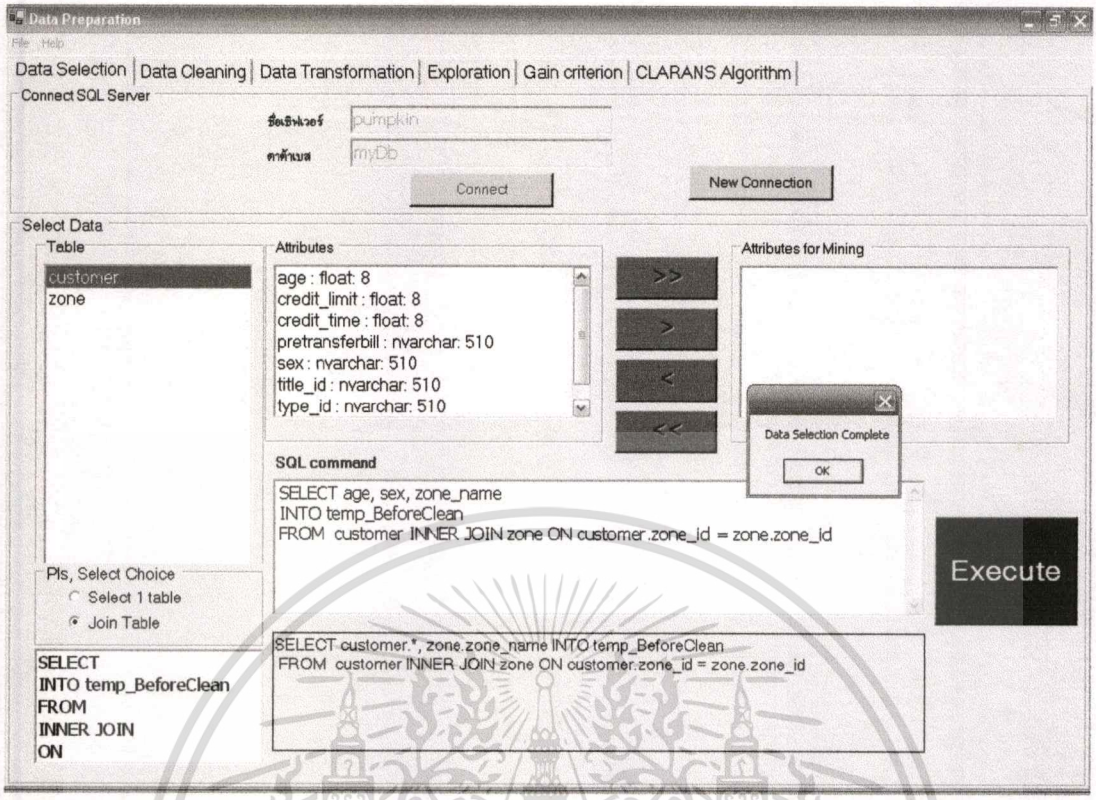
สำหรับการเลือกแอตทริบิวต์ที่ต้องการนำไปจัดกลุ่มนั้นจะแบ่งออกเป็นการเลือกข้อมูลจากตารางเดียวดังรูปที่ 5.2 และการเลือกข้อมูลจากหลายตารางดังรูปที่ 5.3 โดยการเลือกข้อมูลจากหลายตารางผู้ใช้จะต้องใส่คำสั่งของ SQL ลงในช่อง SQL Command เพื่อเลือกข้อมูลที่ต้องการ ซึ่งจะต้องทราบถึงความสัมพันธ์ของข้อมูลในแต่ละตารางด้วย สำหรับทั้งสองวิธีนั้นผู้ใช้สามารถเลือกได้เพียงวิธีเดียวเท่านั้น




**รูปที่ 5.1** หน้าจอแสดงการติดต่อกับฐานข้อมูล




**รูปที่ 5.2** หน้าจอแสดงการเลือกข้อมูลจากตารางเดียว




**รูปที่ 5.3** หน้าจอแสดงการเลือกข้อมูลจากหลายตาราง

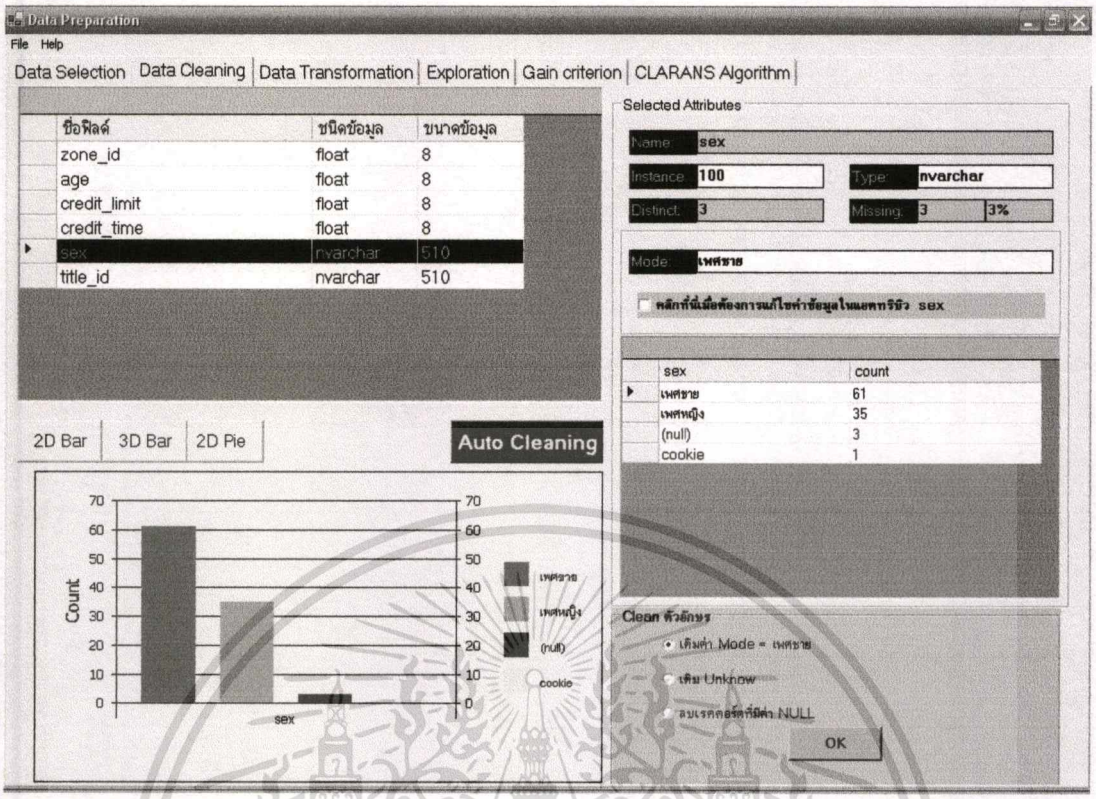
### 5.2.3 การทำข้อมูลให้สมบูรณ์(Data Cleansing)

เมื่อเลือกข้อมูลที่ต้องการเสร็จแล้ว จะต้องทำการแก้ไขข้อมูลให้สมบูรณ์ เช่น ลบค่าว่างออก เป็นต้น ถ้ากดปุ่ม Auto Cleaning จะเป็นการลบลบเรคอร์ดที่มีค่าเป็น Null ออกทุกเรคอร์ด และจะเป็นการไปยังขั้นตอนต่อไป

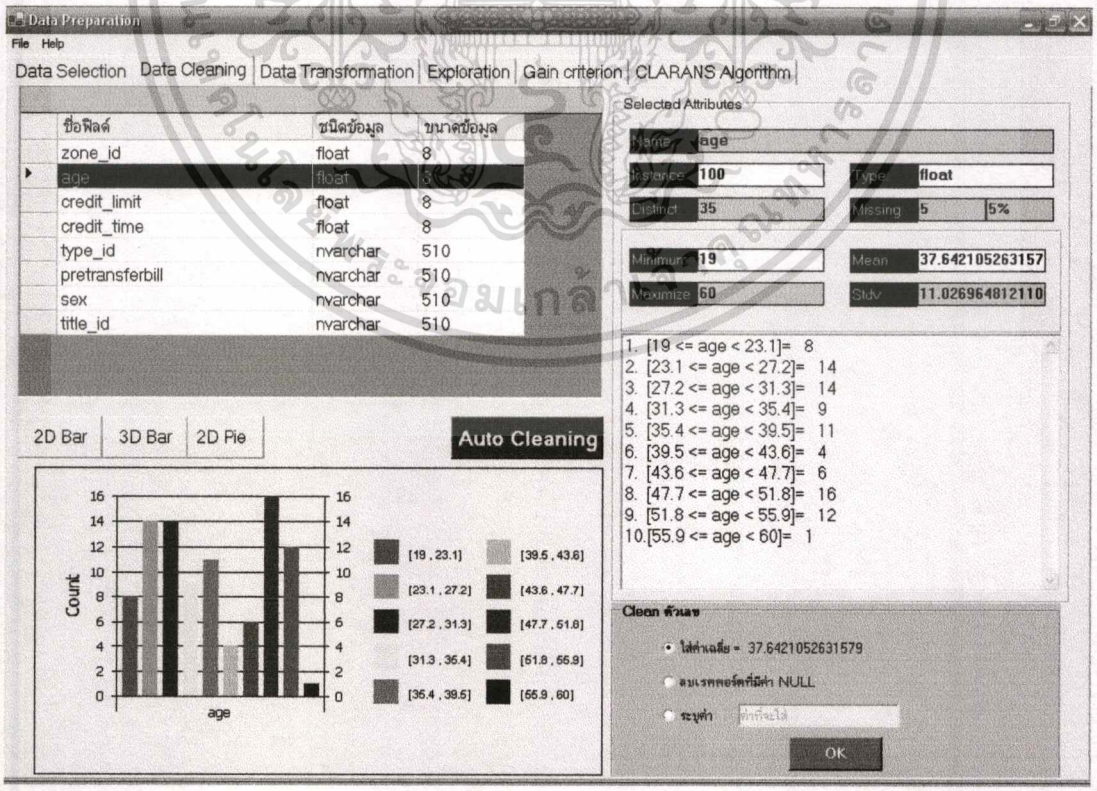
การทำข้อมูลให้สมบูรณ์จะแบ่งข้อมูลออกเป็นสองแบบ คือ ข้อมูลที่เป็น Categorical และข้อมูลที่เป็น Numerical

1) ข้อมูลที่เป็น Categorical จะปรากฏวิธีการทำข้อมูลให้สมบูรณ์ที่ช่อง “Clean ตัวอักษร” ขึ้นมาให้เลือกสามวิธีคือ เติมค่า Mode เติม Unknown และลบเรคอร์ดที่มีค่า NULL ให้เลือกวิธีใดวิธีหนึ่งที่ต้องการแล้วกดปุ่ม OK ให้ทำการแก้ไขข้อมูลให้ครบทุกแอตทริบิวต์ ดังรูปที่ 5.4

2) ข้อมูลที่เป็น Numerical จะปรากฏวิธีการทำข้อมูลให้สมบูรณ์ที่ช่อง “Clean ตัวเลข” ขึ้นมาให้เลือกสามวิธีคือ ใส่ค่าเฉลี่ย ลบเรคอร์ดที่มีค่า NULL และระบุค่า ให้เลือกวิธีใดวิธีหนึ่งที่ต้องการแล้วกดปุ่ม OK ทำการแก้ไขข้อมูลให้ครบทุกแอตทริบิวต์ ดังรูปที่ 5.5




**รูปที่ 5.4** หน้าจอแสดงการทำข้อมูลให้สมบูรณ์สำหรับข้อมูลที่เป็น Categorical




**รูปที่ 5.5** หน้าจอแสดงการทำข้อมูลให้สมบูรณ์สำหรับข้อมูลที่เป็น Numerical

**5.2.4 การปรับเปลี่ยนรูปแบบข้อมูล(Data Transformation)**

เป็นขั้นตอนในการเปลี่ยนรูปแบบของข้อมูลเดิมให้อยู่ในรูปแบบที่เหมาะสมต่อวิธีการทำดาต้าไมน์นิ่งที่เลือกใช้ซึ่งในระบบงานนี้ใช้อัลกอริทึม CLARANS ในการจัดกลุ่มข้อมูล ซึ่งรูปแบบของข้อมูลที่เหมาะสมต่อการจัดกลุ่มในวิธีนี้คือข้อมูลต้องเป็น Numerical ทั้งหมด

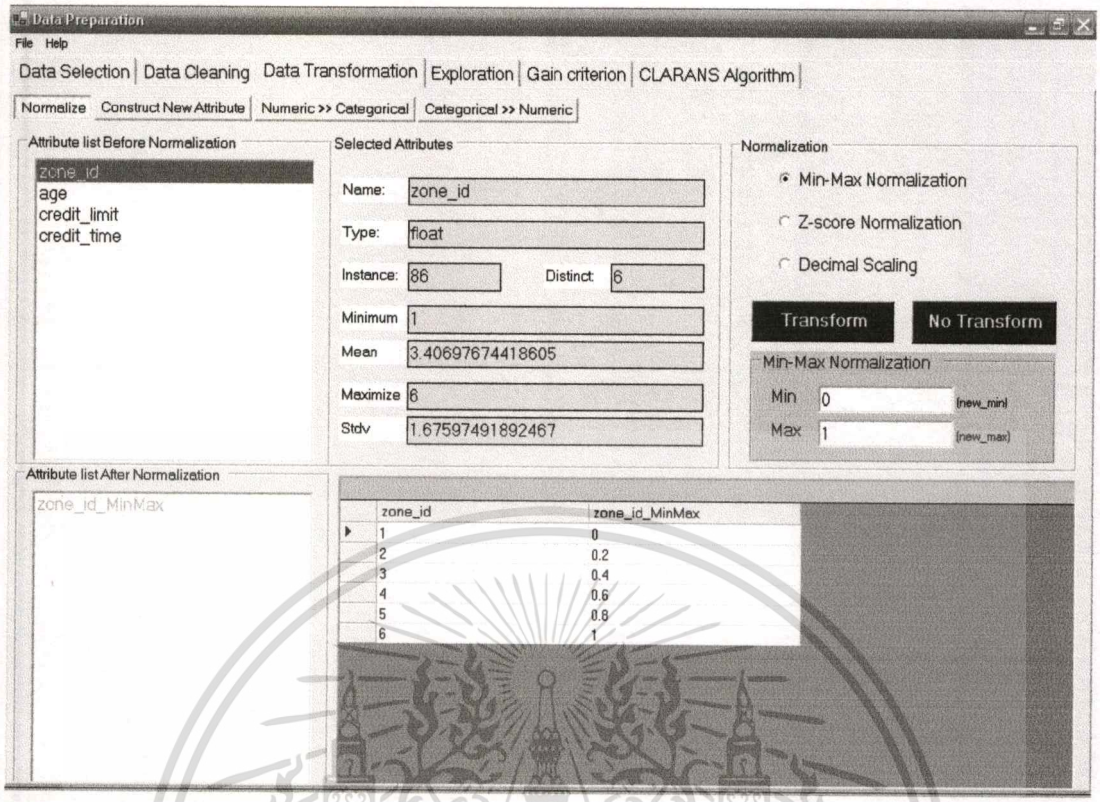
สำหรับขั้นตอนในการปรับเปลี่ยนรูปแบบข้อมูลนั้นแบ่งออกเป็นหลายวิธีดังต่อไปนี้

1) Normalization เป็นการแปลงค่าข้อมูลในแอตทริบิวต์ให้มีค่าอยู่ในขอบเขตที่กำหนด โดยวิธีการ Normalize จะมีอยู่สามวิธีคือ Min-Max Normalization Z-score Normalization และ Decimal Scaling ซึ่งหนึ่งแอตทริบิวต์สามารถเลือกแปลงค่าได้หลายวิธี ดังรูปที่ 5.6

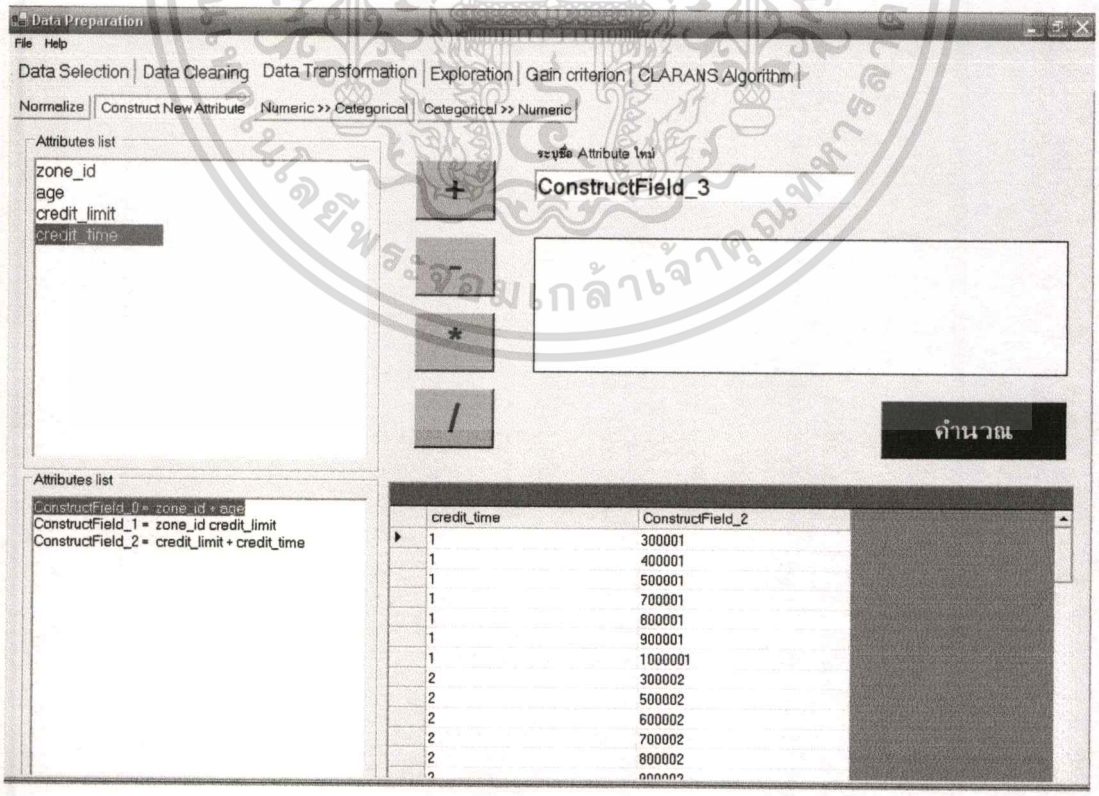
2) Construct New Attribute เป็นการแปลงค่าข้อมูลในแอตทริบิวต์ ให้เป็นแอตทริบิวต์ใหม่ ซึ่งมีค่ามาจากสูตรคำนวณที่ผู้ใช้ต้องการ โดยที่ผู้ใช้สามารถตั้งชื่อแอตทริบิวต์ใหม่ได้ และใส่สูตรคำนวณที่ช่องว่างทางด้านขวา ดังรูปที่ 5.7

3) Numerical to Categorical เป็นการแปลงค่าข้อมูลในแอตทริบิวต์ที่มีข้อมูลเป็นตัวเลข ให้เป็นข้อมูลที่มีค่าเป็นตัวอักษร หรือข้อความ โดยผู้ใช้สามารถกำหนดช่วงของข้อมูล และจำนวนกลุ่มใหม่ได้ ดังรูปที่ 5.8

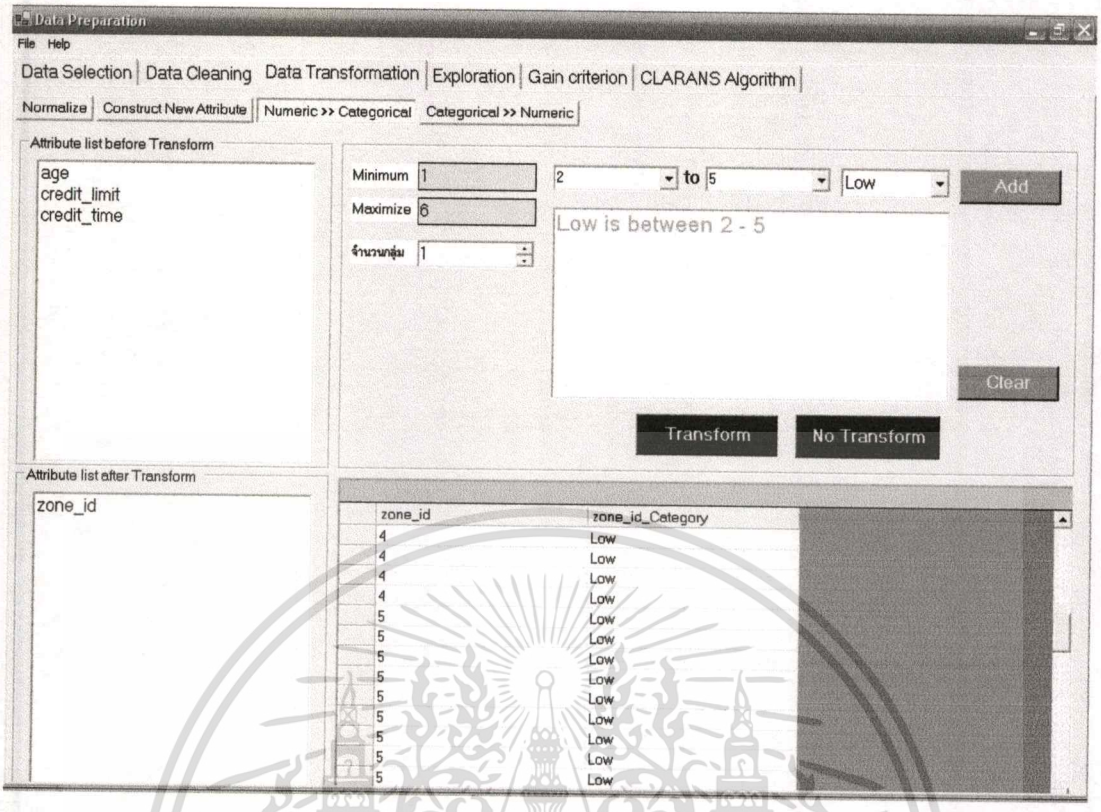
3) Categorical to Numerical เป็นการแปลงค่าข้อมูลในแอตทริบิวต์ที่มีข้อมูลเป็นตัวอักษร หรือข้อความ ให้เป็นข้อมูลที่มีค่าเป็นตัวเลข ซึ่งมีวิธีการแปลงอยู่ 2 วิธี คือ One of N Coding และแปลงตัวอักษรเป็นตัวเลขซึ่งจะเป็นตัวเลขที่ไล่เรียงเป็นลำดับไป ดังรูปที่ 5.9



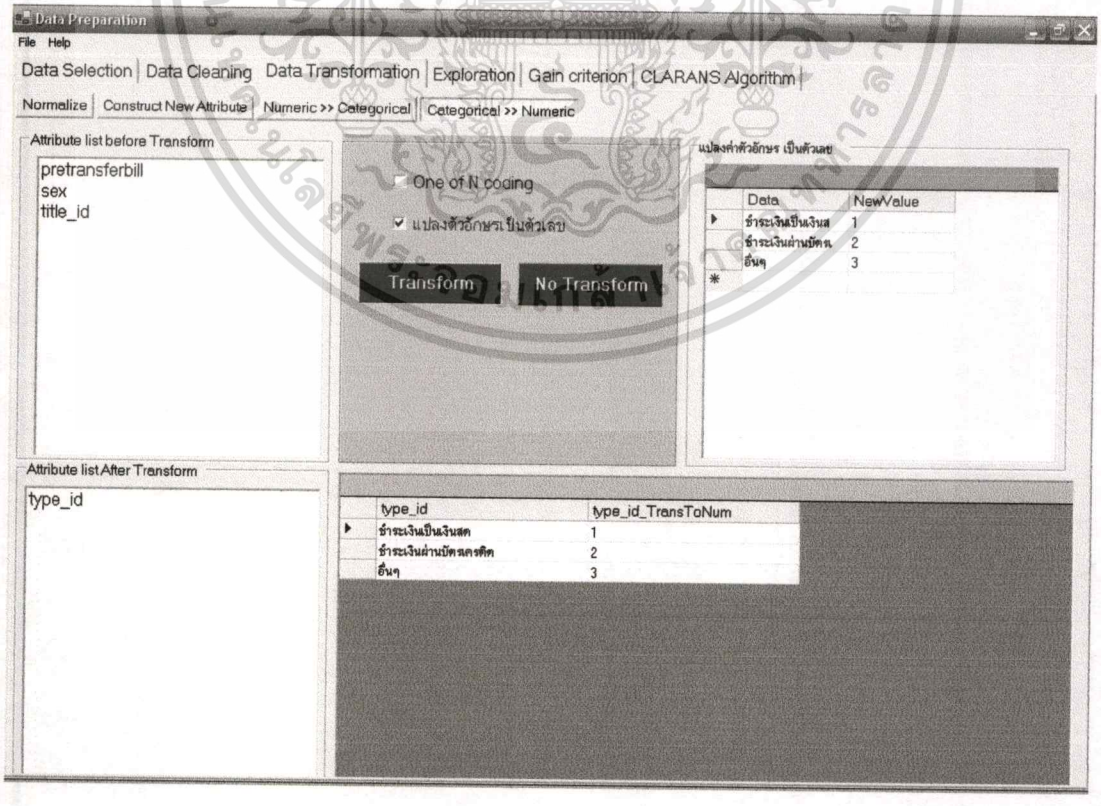

**รูปที่ 5.6** หน้าจอแสดงการปรับเปลี่ยนรูปแบบข้อมูลด้วยวิธี Normalization




**รูปที่ 5.7** หน้าจอแสดงการปรับเปลี่ยนรูปแบบข้อมูลด้วยวิธี Construct New Attribute



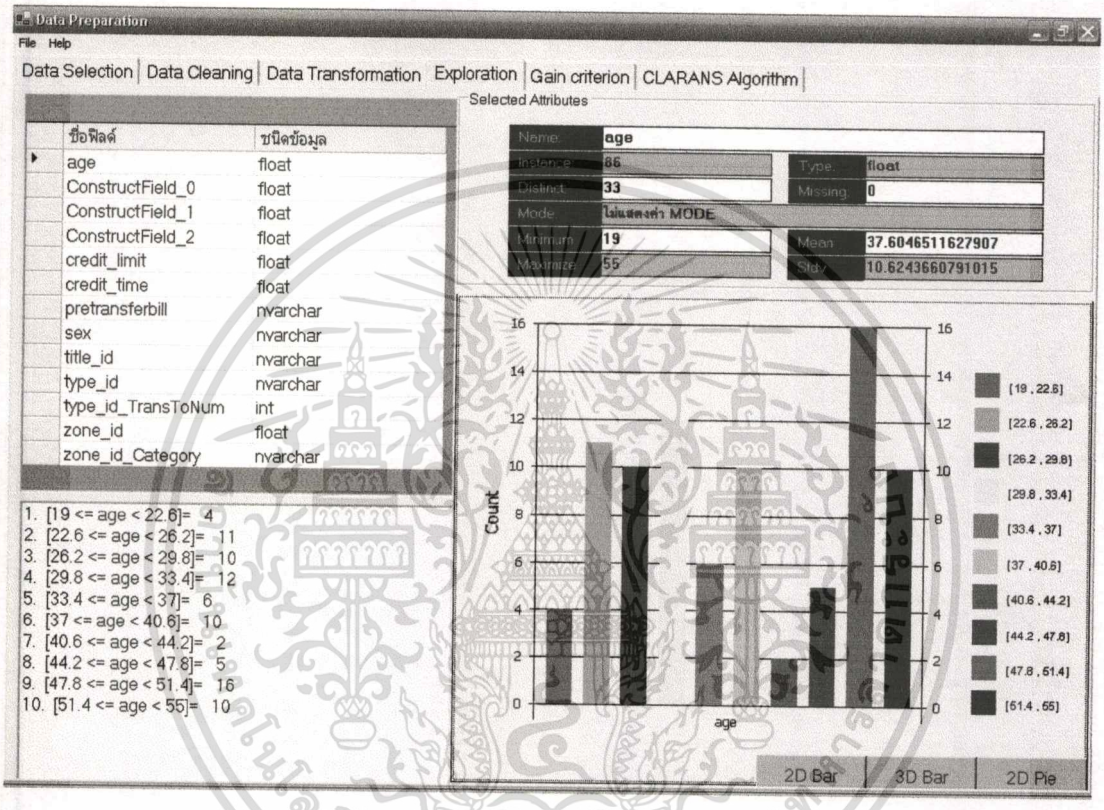

**รูปที่ 5.8** หน้าจอแสดงการปรับเปลี่ยนรูปแบบข้อมูลด้วยวิธี Numerical to Categorical




**รูปที่ 5.9** หน้าจอแสดงการปรับเปลี่ยนรูปแบบข้อมูลด้วยวิธี Categorical to Numerical

### 5.2.5 การสำรวจข้อมูล(Data Exploration)

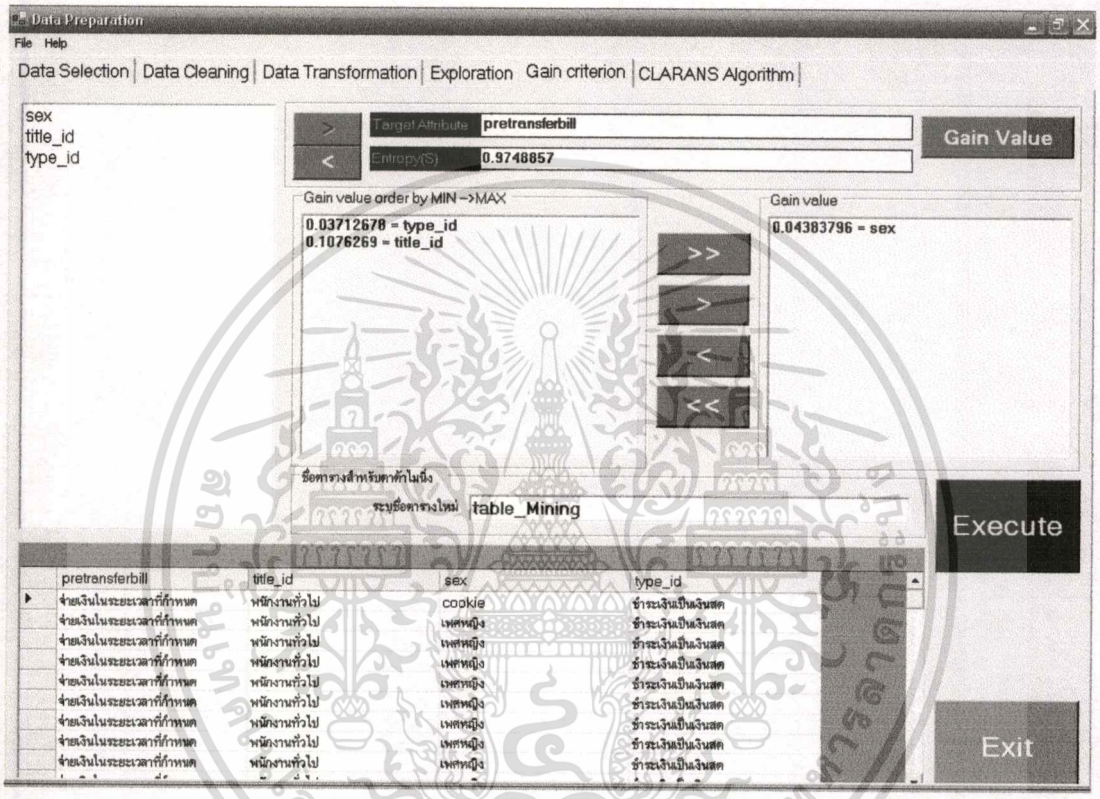
เป็นการสำรวจข้อมูลก่อนที่จะนำไปทำการคาต้าไมน์นิ่ง เมื่อผู้ใช้เลือกที่แอตทริบิวต์ระบบจะแสดงจำนวนเรคอร์ด ประเภทของแอตทริบิวต์ ค่าสูงสุด ค่าต่ำสุด ค่าเฉลี่ย และค่าเบี่ยงเบนมาตรฐาน ผู้ใช้สามารถเลือกรูปแบบการแสดงกราฟได้ 3 แบบคือ แบบกราฟแท่ง 2 มิติ กราฟแท่ง 3 มิติ และแบบวงกลม ดังรูปที่ 5.10




**รูปที่ 5.10** หน้าจอแสดงการสำรวจข้อมูล

### 5.2.6 ส่วนการหาค่า Gain criterion

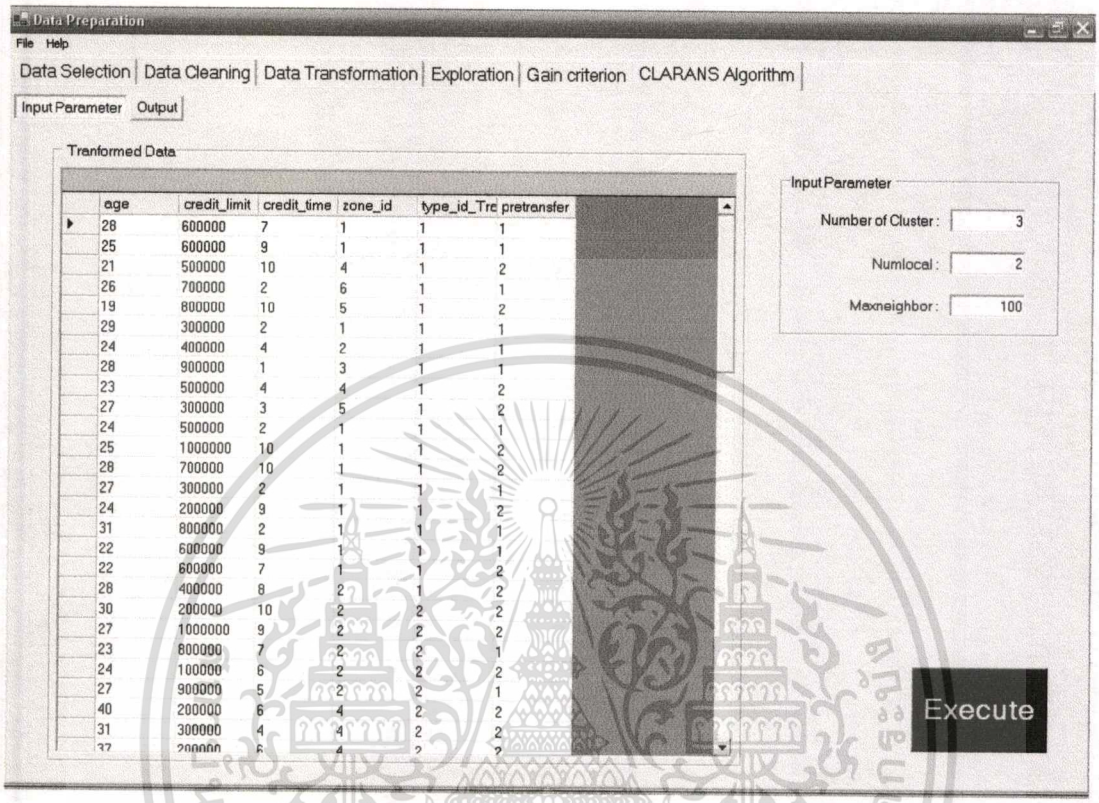
การหาค่า Gain criterion จะใช้หาค่าได้กับข้อมูลที่เป็น Categorical เท่านั้น เมื่อผู้ใช้เลือกแอตทริบิวต์ที่ต้องการ ระบบจะคำนวณค่า Entropy แล้วแสดงให้ทราบ ผู้ใช้สามารถนำค่า Gain ที่ได้มาพิจารณาประกอบในการเลือกข้อมูลเพื่อนำไปใช้ในการทำดาต้าไมนิ่งต่อไป โดยสามารถสร้างตารางใหม่เพื่อเก็บข้อมูลไว้ได้ด้วย ดังรูปที่ 5.11




**รูปที่ 5.11** หน้าจอแสดงส่วนการหาค่า Gain criterion

### 5.2.7 การจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม CLARANS

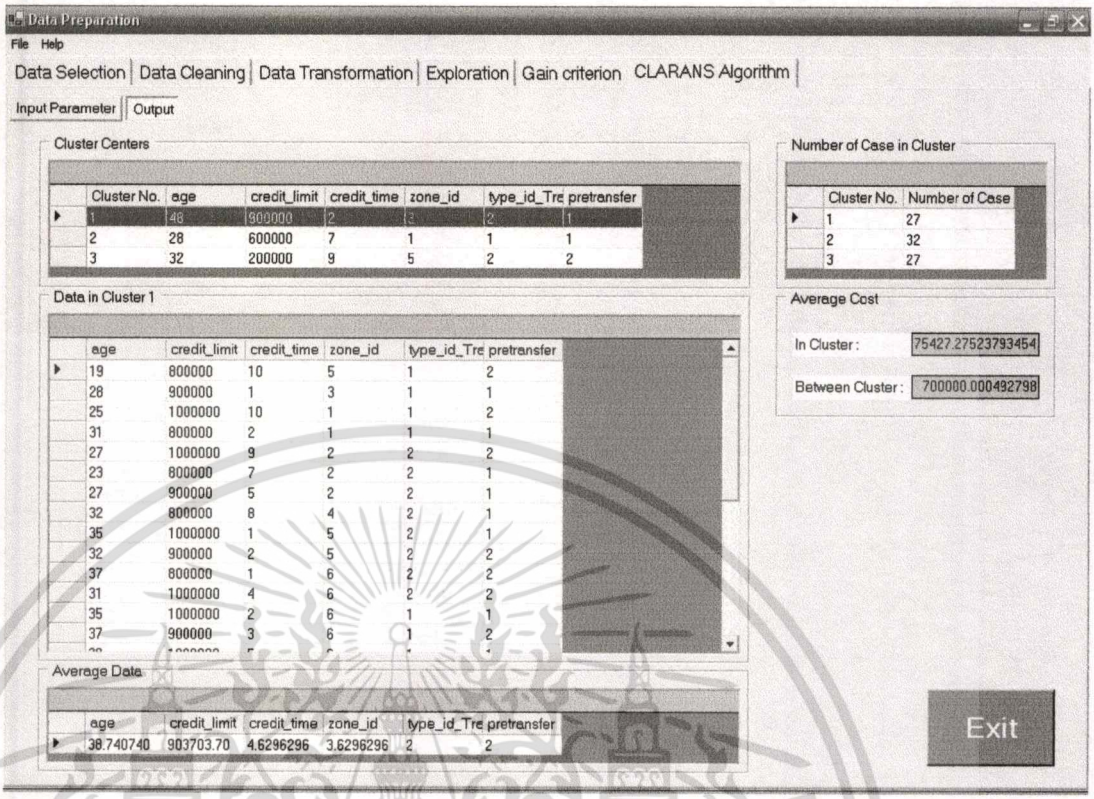
หลังจากผ่านการเตรียมข้อมูลมาแล้ว ระบบจะแสดงหน้าจอดังรูปที่ 5.12




**รูปที่ 5.12** หน้าจอแสดงส่วนการใส่ค่าพารามิเตอร์

ในส่วนของ Input Parameter จะเป็นส่วนที่แสดงข้อมูลที่ผ่านการเตรียมข้อมูลมาแล้ว และข้อมูลจะอยู่ในรูปที่เป็นตัวเลขเท่านั้น ซึ่งจะแสดงอยู่ในส่วนของ Transformed Data

สำหรับในส่วนของ Input Parameter ผู้ใช้จะต้องทำการกำหนดค่าพารามิเตอร์(parameter) ที่จำเป็นต่อการจัดกลุ่มข้อมูลด้วยอัลกอริทึม CLARANS ซึ่งมีอยู่ 3 ค่าคือ จำนวนคลัสเตอร์ที่ต้องการแบ่งกลุ่ม(Number of Cluster) ค่า numlocal และค่า maxneighbor เมื่อใส่ครบทุกค่าให้กดที่ปุ่ม Execute ระบบจะไปยังส่วนของ Output




**รูปที่ 5.13** หน้าจอแสดงผลลัพธ์จากการจัดกลุ่ม

ในส่วนของ Output ระบบจะแสดงคลัสเตอร์ที่ได้จากการจัดกลุ่มข้อมูลตามจำนวนกลุ่มที่ผู้ใช้ระบุมา และแสดงข้อมูลที่อยู่ในแต่ละกลุ่มในส่วน Data in Cluster ว่ามีข้อมูลใดอยู่บ้าง รวมไปถึงแสดงจำนวนข้อมูล และค่าเฉลี่ยในแต่ละกลุ่ม สำหรับในส่วนของ Average Cost จะแสดงค่าเฉลี่ยของ cost ที่อยู่ภายในคลัสเตอร์(In cluster) และระหว่างคลัสเตอร์(Between Cluster)

เมื่อผู้ใช้ต้องการจบการทำงานแล้ว ให้กดที่ปุ่ม Exit เพื่อจบการทำงาน หากผู้ใช้ต้องการจัดกลุ่มข้อมูลใหม่สามารถที่จะกลับไปยังส่วนของ Input Parameter เพื่อใส่ค่าพารามิเตอร์ใหม่ได้

# บทที่ 6
# สรุปผล และข้อเสนอแนะ

## 5.1 สรุปผลงานวิจัย

การพัฒนาระบบงานดาต้าไมน์นิ่งเพื่อการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม CLARANS ได้จัดทำขึ้นเพื่อให้เห็นถึงประโยชน์ของการนำข้อมูลดิบมาวิเคราะห์ เพื่อนำไปใช้ให้เกิดประโยชน์ได้สูงสุด สามารถนำไปประยุกต์กับการทำงาน หรืองานทางด้านต่างๆ ได้ ไม่ว่าจะเป็นงานทางด้านดูแลลูกค้า หรือทางธุรกิจการตลาด

สำหรับอัลกอริทึม CLARANS (Clustering Large Applications based on Randomized Search) ที่ใช้ในระบบงานนี้เป็นอัลกอริทึมที่จัดอยู่ในประเภทของการแบ่งกลุ่มฐานข้อมูล(Database Segmentation) ซึ่งเป็นอัลกอริทึมใช้วิธีการค้นหาแบบสุ่มในการสุ่มหาคลัสเตอร์ที่เหมาะสมออกมาจากข้อมูลที่มีอยู่ ทำให้เหมาะที่จะนำมาใช้จัดกลุ่มกับข้อมูลที่มีปริมาณมากๆ เพราะไม่ต้องใช้วิธีในการค้นหาที่ละเอียดมากนัก

## 5.2 ข้อเสนอแนะ

1) การทำงานของโปรแกรมนั้นควรทำงานงานตามลำดับขั้นตอนที่ได้แสดงไว้แล้ว ถ้าไม่ทำตามลำดับขั้นตอนอาจเกิดข้อผิดพลาดขึ้นกับโปรแกรมได้

2) เพื่อให้ข้อมูลที่วิเคราะห์ออกมาได้ตรงตามความต้องการควรจะต้องทำการกำหนดวัตถุประสงค์ของงานก่อน เพื่อให้สามารถเตรียมข้อมูลที่จะนำมาวิเคราะห์ได้ถูกต้อง ทำให้ผลลัพธ์ที่ออกมาตรงกับความต้องการมากยิ่งขึ้น

3) สำหรับอัลกอริทึม CLARANS ที่ใช้ในระบบงานนี้ สามารถวิเคราะห์ข้อมูลที่เป็นตัวเลขได้อย่างเดียวเท่านั้น ดังนั้นต้องแปลงข้อมูลที่เป็นข้อความหรือตัวอักษรให้อยู่ในรูปของตัวเลขก่อนทำการวิเคราะห์ข้อมูลด้วย

ข้อเสนอแนะข้างต้น หวังว่าจะเป็นประโยชน์ต่อผู้ที่สนใจและต้องการจะนำโปรแกรมไปพัฒนาเพื่อให้โปรแกรมมีความสมบูรณ์ และมีความถูกต้องมากยิ่งขึ้นต่อไปในอนาคต

# บรรณานุกรม

สุรสิทธิ์ คิวประสพศักดิ์ และนันทนี แขวงโสภา. 2546. **อินไซท์ Visual Basic .NET ฉบับสมบูรณ์.** กรุงเทพฯ : โปรวิชัน.

อาทิตยา เชื้อจันอัด. 2549. "การพัฒนาระบบการเตรียมข้อมูลและการสำรวจ สำหรับการทำดาต้าไมนิ่ง". โครงการพัฒนาระบบงานวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ, คณะเทคโนโลยีสารสนเทศ, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง.

A.K. Jain, M.N. Murty and P.J. Flynn. 1999. **Data Clustering : A Review.** ACM Computing Surveys Vol.31. [Online]. Available: http://citeseer.ist.psu.edu/context/1441210/525574

Hinrich Schütze. 2005. **Single-Link, Complete-Link & Average-Link Clustering** . [Online] Available: http://www-csli.stanford.edu/~schuetze/

Maria Halkidi and Yannis Batistakis ET. AL. 2001. **On Clustering Validation Techniques.** Journal of Intelligent Information Systems. 17:2/3. 107-145. [Online] Available: http://www.db-net.aueb.gr/mhalk/papers/validity_survey.pdf

N. Bolshakova and F. Azuaje. 2002. **Improving expression data mining through cluster validation.** Department of Computer Science, Trinity College Dublin, Ireland. [Online] Available: https://www.cs.tcd.ie/publications/tech-reports/reports .02 / TCD-CS-2002-34.pdf

Oracle Corporation. 2003. **Oracle Data Mining Concepts.** Oracle Corporation. 500 Oracle Parkway. Redwood Shores. CA 94065. [Online]. Available: http://www.oracle.com/technology/products/bi/odm/odminer.html

Oracle Corporation. 2004. **Oracle Data Mining Know More, Do More, Spend Less.** Oracle Corporation. 500 Oracle Parkway. Redwood Shores. CA 94065. [Online]. Available: http://download-west.oracle.com/oowsf2004/1439_wp.pdf

Raymond T. Ng and Jiawei Han. 2002. **CLARANS : A Method for Spatial Data Mining.** IEEE Transactions on knowledge and data engineering, Vol.14, NO.5.

Songrit Maneewongvatana. 2546. **Data Mining.** CPE 751. KMUTT. [Online]. Available: http://www.cpe.kmutt.ac.th/~songrit/

# ประวัติผู้เขียน

| | |
|---|---|
| **ชื่อผู้เขียน** | นายสิริภูมิ สุดดีพงศ์ |
| **วันเกิด** | 9 กันยายน พ.ศ. 2525 |
| **สถานที่เกิด** | กรุงเทพมหานคร |
| **วุฒิการศึกษาระดับปริญญาตรี** | วิทยาศาสตรบัณฑิต |
| **สถานที่สำเร็จการศึกษา** | คณะวิทยาศาสตร์<br>สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง |
| **ปีที่สำเร็จการศึกษา** | 2547 |