# การใช้ Data Mining สำหรับการขยายบริการในธุรกิจธนาคารพาณิชย์

# Data Mining for Cross Selling in Banking Business

โดย

**นางสาวปัทมเกสร์ อมาตยกุล**

รหัส 45061737

**อาจารย์ที่ปรึกษา**

**รศ. ดร.อาริต ธรรมโน**





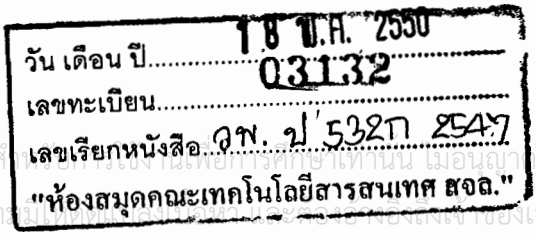
รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการศึกษากรณีพิเศษ

หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

ภาคเรียนที่ 1 ปีการศึกษา 2547

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง



| | |
|---|---|
| **ชื่อหัวข้อ** | การใช้ Data Mining สำหรับการขยายบริการในธุรกิจธนาคารพาณิชย์ |
| **นักศึกษา** | นางสาวปัทมเกสร์ อมาตยกุล |
| **อาจารย์ที่ปรึกษา** | รศ.ดร. อาริต ธรรมโน |
| **ระดับการศึกษา** | วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ |
| **แขนงวิชา** | การจัดการเทคโนโลยีสารสนเทศ |
| **ปีการศึกษา** | 2547 |

# บทคัดย่อ

ในปัจจุบันมีการแข่งขันกันมากขึ้นในธุรกิจต่างๆ รวมทั้งธุรกิจธนาคารพาณิชย์ โดยธนาคารต่างๆ ต่างก็ต้องการที่จะรักษาฐานลูกค้าเดิมของตนและเพิ่มฐานลูกค้าออกไปให้ได้มากที่สุด ดังนั้นจึงมีความจำเป็นอย่างยิ่งที่จะต้องศึกษาพฤติกรรมของลูกค้าเพื่อให้สามารถเข้าถึงลักษณะการใช้บริการและพฤติกรรมการใช้สินค้าประเภทต่างๆ ให้ได้มากที่สุด การนำเอา Data Mining มาใช้ในการสืบค้น และวิเคราะห์ข้อมูล ที่มีอยู่เป็นจำนวนมากในธุรกิจธนาคารซึ่งยังไม่ได้ถูกนำมาใช้ให้เกิดประโยชน์อย่างเต็มที่ จึงเท่ากับเป็นการช่วยธุรกิจให้สามารถตัดสินใจเลือกกลยุทธ์ทางการตลาดที่เหมาะสมได้ การนำเทคนิค Data Mining มาใช้ในการวิเคราะห์ในครั้งนี้จะมุ่งประเด็นไปที่การเพิ่มฐานลูกค้าประเภทบุคคลสำหรับบริการหลากหลายประเภทของธนาคาร รวมทั้งศึกษาลักษณะการขายสินค้าประเภทต่างๆ ว่าควรจะขายหรือใช้กลยุทธ์ใดในช่วงเวลาใด

| | |
|---|---|
| **Title** | Data Mining for Cross Selling in Banking Business |
| **Student** | Ms.Patamakes Amatyakul |
| **Advisor** | Assoc.Prof. Dr. Arit Thammano |
| **Level of Study** | Master of Science in Information Technology |
| **Major** | Information Technology Management |
| **Academic Year** | 2004 |

# ABSTRACT

Nowadays, there are many competitions in several business including the banking business. Banks need to maintain their old customer bases and extend their customer bases as much as possible. Therefore, they need to study customers' behaviors in order to reach the characteristics of services needed and behaviors of products used. The use of Data Mining for searching and analyzing data in banking business, which is used widely but not in full capacity, can guide business to choose the appropriate market strategies. This seminar aims to increase individual customer base for several kinds of bank services including the characteristics of product selling in which the best strategies could be used.

# สารบัญ

หน้า

# สารบัญ (ต่อ)

หน้า

# สารบัญตาราง

หน้า

ตารางที่

# สารบัญภาพ

หน้า

ภาพที่

# สารบัญภาพ (ต่อ)

หน้า

ภาพที่

# สารบัญภาพ (ต่อ)

หน้า

ภาพที่

# บทที่ 1
# บทนำ

## 1.1 ความเป็นมาของปัญหา

ในปัจจุบันมีการแข่งขันกันสูงขึ้นในการดำเนินธุรกิจประเภทต่างๆ ไม่ว่าจะเป็นในด้านสินค้าและการให้บริการ ทำให้นักการตลาดและผู้บริหารของแต่ละธุรกิจต้องนำความรู้ความสามารถที่มีอยู่ทั้งหมดมาใช้ในการกำหนดนโยบายและแผนกลยุทธ์ของตนให้มีประสิทธิภาพมากที่สุด โดยอาศัยทรัพยากรสำคัญที่มีอยู่ คือ ข้อมูลลูกค้า

ในแต่ละธุรกิจมีข้อมูลอยู่เป็นจำนวนมาก โดยผู้ที่สามารถนำข้อมูลที่มีอยู่มาใช้ประโยชน์ในการวิเคราะห์ เพื่อวางแผนกลยุทธ์ และการตลาด เพื่อสร้างสรรค์สินค้าและการบริการให้สามารถสนองตอบต่อความต้องการของลูกค้าได้มากที่สุด ตรงกลุ่มเป้าหมายที่สุด และอยู่ในช่วงเวลาและโอกาสที่เหมาะสมที่สุด จะเป็นผู้ครองส่วนแบ่งตลาดในธุรกิจ ยิ่งไปกว่านั้น การใช้ข้อมูลที่มีอยู่ให้เกิดประโยชน์สูงสุด ยังเท่ากับเป็นการลดต้นทุนทางการแข่งขันและสามารถสร้างผลกำไรให้กับธุรกิจได้อีกด้วย กล่าวคือ การรู้จักนำเอาทรัพยากรที่มีอยู่มาใช้ให้เกิดประโยชน์จะช่วยให้ธุรกิจสามารถเพิ่มขีดจำกัดทางการแข่งขันกับคู่แข่งในธุรกิจได้อย่างแท้จริง

ธุรกิจธนาคารพาณิชย์เป็นธุรกิจอีกประเภทหนึ่งในประเทศไทยที่มีการแข่งขันสูง โดยธนาคารแต่ละแห่งต่างแข่งขันกันสร้างสรรค์สินค้าและบริการให้มีความแตกต่าง สามารถตอบสนองความต้องการของลูกค้าของตนให้ได้มากที่สุด ดังนั้นการนำข้อมูลที่มีอยู่เป็นปริมาณ และมีคุณค่ามากในธุรกิจมาใช้ให้เกิดประโยชน์ จึงเท่ากับเป็นการเพิ่มโอกาสในการดำเนินธุรกิจด้วย

การศึกษาในครั้งนี้จึงได้นำเสนอแนวทางการวิเคราะห์ข้อมูลในเชิงธุรกิจ โดยใช้วิธีการสืบค้นข้อมูลจากฐานข้อมูล (Knowledge Discovery in Database: KDD) หรือที่เรียกว่าการขุดค้นข้อมูล (Data Mining) เพื่อที่จะวิเคราะห์พฤติกรรมการใช้บริการของลูกค้าของธนาคารพาณิชย์ในประเทศไทย สำหรับเป็นต้นแบบสำหรับการศึกษาและวิเคราะห์พฤติกรรมของลูกค้าในธุรกิจธนาคารพาณิชย์ โดยจะโดยจะมุ่งประเด็นไปที่การเพิ่มฐานลูกค้าประเภทบุคคล (ลูกค้ารายย่อย) สำหรับบริการหลากหลายประเภทของธนาคาร รวมทั้งศึกษาลักษณะการใช้สินค้าประเภทต่างๆ เพื่อนำมาเป็นแนวทางปรับปรุงกลยุทธ์ให้เหมาะสม

## 1.2 วัตถุประสงค์

การศึกษาโครงการนี้มีวัตถุประสงค์เพื่อ

1. ศึกษาแนวทางการวิเคราะห์ข้อมูลในเชิงธุรกิจ โดยใช้วิธีการสืบค้นข้อมูลจากฐานข้อมูล (Knowledge Discovery in Database: KDD) หรือที่เรียกว่าการขุดค้นข้อมูล (Data Mining) โดยเน้นการศึกษาการทำงานของระบบโครงข่ายประสาทเทียม (Artificial Neural Networks)
2. ศึกษาผลที่ได้จากการทำงานของระบบโครงข่ายประสาทเทียม (Artificial Neural Network) ใน Kohonen's Self-Organizing Maps (SOM) Algorithm
3. ศึกษาแนวทางและความเป็นไปได้ในการนำเอาผลที่ได้จากการวิเคราะห์โดยใช้ระบบโครงข่ายประสาทเทียม (Artificial Neural Network) มาประยุกต์ใช้ในการแบ่งกลุ่มลูกค้าและทำนายพฤติกรรมการใช้บริการของลูกค้าธนาคารพาณิชย์ในประเทศไทย สำหรับนำมาเป็นแนวทางในการขยายบริการและปรับปรุงกลยุทธ์ให้เหมาะสม

## 1.3 ขอบเขตของการศึกษา

ทำการศึกษา และวิเคราะห์โดยใช้ฐานข้อมูลลูกค้าธนาคารพาณิชย์แห่งหนึ่งจำนวน 1000 ราย เพื่อมาทำการเปรียบเทียบกลุ่มลูกค้าตามลักษณะและพฤติกรรมการใช้บริการที่แตกต่างกัน โดยทำการแบ่งส่วนฐานข้อมูล (Database Segmentation) โดยใช้ Artificial Neural Network คือ Kohonen's Self-Organizing Maps (SOM) Algorithm ในโปรแกรม IBM Intelligent Miner

## 1.4 ประโยชน์ที่คาดว่าจะได้รับ

1. สามารถเข้าใจในหลักการและกระบวนการทำงานของการทำการขุดค้นข้อมูล (Data Mining)
2. สามารถเข้าใจในลักษณะการทำงานและประมวลผลของระบบโครงข่ายประสาทเทียม (Artificial Neural Network)
3. สามารถเข้าใจในระบบการทำงานของ Neural Network Algorithm แบบ Kohonen's Self-Organizing Maps (SOM) Algorithm
4. สามารถนำผลที่จากการศึกษาได้ไปประยุกต์ใช้ในการจัดการ ด้าน Customer Relationship Management (CRM) ในธุรกิจธนาคารพาณิชย์ได้

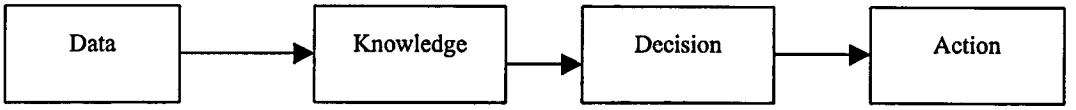
# บทที่ 2
# ทฤษฎีที่เกี่ยวข้อง

## 2.1 ความรู้เบื้องต้นเกี่ยวกับ Data Mining

Data Mining คือ กระบวนการค้นหาสารสนเทศที่เป็นประโยชน์ โดยสารสนเทศที่ได้นี้จะได้มาจากการหาความสัมพันธ์และรูปแบบทั่วไปของข้อมูลที่มีอยู่จำนวนมาก ซึ่งในบางครั้งจะเรียก Data Mining ว่าเป็นการค้นหาความรู้ใหม่จากฐานข้อมูล หรือ Knowledge Discovery in Database (KDD) โดยจุดประสงค์หลักของการค้นหาข้อมูลทั้งหมด ก็เพื่อสร้างรูปแบบที่สามารถเข้าใจได้ง่ายและสะดวกในการที่จะตีความพื้นฐานของข้อมูลนั้นๆ

Data Mining เป็นกระบวนการซึ่งผสมผสานระหว่างความสามารถของมนุษย์ร่วมกับคอมพิวเตอร์ โดยมนุษย์เป็นผู้ออกแบบฐานข้อมูล อธิบายปัญหา และกำหนดจุดมุ่งหมายต่างๆ ส่วนคอมพิวเตอร์ทำหน้าที่กลั่นกรองข้อมูลที่ผ่านเข้ามาและทำการค้นหาแบบแผนที่ตรงตามจุดมุ่งหมายที่ได้กำหนดไว้ ซึ่งเทคนิคของ Data Mining ก็คือพยายามที่จะค้นหากระบวนการ กฎเกณฑ์ที่แน่นอนและมีแบบแผนอัตโนมัติที่จะนำมาใช้ในการดึงข้อมูลที่ถูกจัดเก็บเอาไว้ในฐานข้อมูลที่มีจำนวนมากๆ นำมาใช้ให้เกิดประโยชน์ ซึ่งกระบวนการค้นหาสารสนเทศจากคลังข้อมูลนี้ต้องผ่านกระบวนการจัดเตรียมข้อมูล (Data Preparation) การค้นหาและจัดรูปแบบ (Search for Pattern) จนกระทั่งได้ข้อมูลตามต้องการก่อน เพราะแม้ว่าข้อมูลจะมีการจัดเก็บมาแล้วอย่างเป็นระบบก็ตาม แต่ถ้าขาดกระบวนการในการจัดการสารสนเทศอย่างมีประสิทธิภาพและถูกวิธีแล้ว ข้อมูลต่างๆ ที่เก็บไว้ก็จะไม่มีประโยชน์ ผลลัพธ์ที่ได้จากการทำ Data Mining จะได้รับการแปลงรูปและรวบรวมเข้าด้วยกัน สำหรับใช้เป็นข้อมูลในการช่วยตัดสินใจ ซึ่งกระบวนการในการนำข้อมูลเหล่านี้มาใช้สำหรับช่วยตัดสินใจหลายวิธี ตั้งแต่รูปแบบที่ทำด้วยมือมนุษย์ไปจนถึงรูปแบบที่ต้องการหลักการทางวิทยาศาสตร์เช่น Linear Programming เข้ามาช่วย

ความสามารถในการเป็นระบบสนับสนุนการตัดสินใจ (Decision Support System) นี้เอง Data Mining จึงช่วยให้ข้อมูลที่เรามีอยู่กลายเป็นความรู้อันมีค่า และสร้างคำตอบของอนาคตได้




**รูปที่ 2.1 ข้อมูลสู่การตัดสินใจและปฏิบัติ**

## 2.2 ข้อมูลที่ใช้สำหรับการทำ Data Mining

### 2.2.1 ประเภทของข้อมูลที่สามารถทำ Data Mining

1. Relational Database เป็นฐานข้อมูลที่จัดเก็บอยู่ในรูปแบบของตาราง โดยในแต่ละตารางจะประกอบไปด้วยแถวและคอลัมน์ ความสัมพันธ์ของข้อมูลทั้งหมดสามารถแสดงได้โดย entity relationship (ER) model

2. Data Warehouses เป็นการเก็บรวบรวมข้อมูลจากหลายแหล่งมาเก็บไว้ในรูปแบบเดียวกันและรวบรวมไว้ในที่ๆ เดียวกัน

3. Transactional Database ประกอบด้วยข้อมูลที่แต่ละทรานเซกชั่นแทนด้วยเหตุการณ์ในขณะใดขณะหนึ่ง เช่น ใบเสร็จรับเงิน จะเก็บข้อมูลในรูป ชื่อลูกค้าและรายการสินค้าที่ลูกค้ารายนั้นซื้อ เป็นต้น

4. Advanced Database เป็นฐานข้อมูลที่จัดเก็บในรูปแบบอื่นๆ เช่น ข้อมูลแบบ object-oriented, ข้อมูลที่เป็น text file, ข้อมูลมัลติมีเดีย, ข้อมูลในรูปของ web

### 2.2.2 ลักษณะเฉพาะของข้อมูลที่สามารถทำ Data Mining

1. ข้อมูลขนาดใหญ่ เกินกว่าจะพิจารณาความสัมพันธ์ที่ซ่อนอยู่ภายในข้อมูลได้ด้วยตาเปล่าหรือโดยการใช้ Database Management System ( DBMS ) ในการจัดการฐานข้อมูล

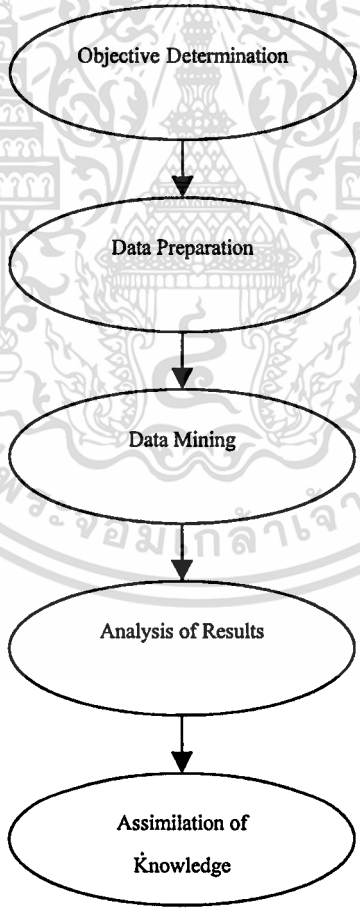
2. ข้อมูลที่มาจากหลายแหล่ง โดยอาจรวบรวมมาจากหลายระบบปฏิบัติการหรือหลาย DBMS เช่น Oracle , DB2 , MS SQL , MS Access เป็นต้น

3. ข้อมูลที่ไม่มีการเปลี่ยนแปลงตลอดช่วงเวลาที่ทำการ Mining หากข้อมูลที่มีอยู่นั้นเป็นข้อมูลที่เปลี่ยนแปลงตลอดเวลาจะต้องแก้ปัญหานี้ก่อน โดยบันทึกฐานข้อมูลนั้นไว้ และนำฐานข้อมูลที่บันทึกไว้มาทำ Mining แต่เนื่องจากข้อมูลนั้นมีการเปลี่ยนแปลงอยู่ตลอดเวลา จึงทำให้ผลลัพธ์ที่ได้จากการทำ Mining สมเหตุสมผลในช่วงเวลาหนึ่งเท่านั้น ดังนั้นเพื่อให้ได้ผลลัพธ์ที่มีความถูกต้องเหมาะสมอยู่ตลอดเวลาจึงต้องทำ Mining ใหม่ทุกครั้งในช่วงเวลาที่เหมาะสม ข้อมูลที่มีโครงสร้างซับซ้อน เช่น ข้อมูลรูปภาพ ข้อมูลมัลติมีเดีย ข้อมูลเหล่านี้สามารถนำมาทำ Mining ได้เช่นกัน แต่ต้องใช้เทคนิคการทำ Data Mining ขั้นสูง

**2.3 ขั้นตอนการทำงานของ Data Mining**

ขั้นตอนการทำงานของ Data Mining เป็นกระบวนการของการสร้างแบบจำลอง (Model) โดยสร้างแบบจำลองของกลุ่มข้อมูลเพื่อสร้างความเข้าใจในแนวโน้ม รูปแบบ และความสัมพันธ์กันของกลุ่มข้อมูลเพื่อใช้ในการทำนายข้อมูลเหล่านั้น ซึ่งกระบวนการของ Data Mining ประกอบด้วย 5 ขั้นตอน คือ

1. กำหนดวัตถุประสงค์ในการทำ Data Mining (Objective Determination)
2. เตรียมข้อมูล (Data Preparation)
3. ทำ Data Mining
4. ทำการวิเคราะห์ผลลัพธ์ที่ได้ (Analysis of Results)
5. นำสารสนเทศที่ได้ไปใช้ประโยชน์ (Assimilation of knowledge)




**รูปที่ 2.2** ขั้นตอนการทำงานของ Data Mining

**2.3.1 ขั้นตอนการกำหนดวัตถุประสงค์ในการทำ Data Mining**

การกำหนดวัตถุประสงค์และปัญหาที่ชัดเจนจะเป็นตัวกำหนดทิศทางการทำ Data Mining ดังนั้นในการกำหนดวัตถุประสงค์ของงานจะต้องเข้าใจถึงปัญหาและความต้องการของงานนั้นๆ รวมทั้งต้องดูถึงความเป็นไปได้ด้วยว่าวิธีการ Data Mining เหมาะกับการหาคำตอบของปัญหานั้น ๆ หรือไม่ การกำหนดถึงความต้องการนั้นจะมุ่งประเด็นถึงคำตอบที่ได้เพื่อจะนำไปใช้ให้เกิดประโยชน์ แต่จะไม่ใช่เกิดจากการตั้งสมมติฐาน และนอกจากนั้นยังเป็นการกำหนดถึงแหล่งที่มาของข้อมูลที่จะทำการ Mining อีกด้วย

**2.3.2 ขั้นตอนการเตรียมข้อมูล**

ขั้นตอนการเตรียมข้อมูลสำหรับการทำ Data Mining นั้น ถือเป็นขั้นตอนที่สำคัญ และเป็นช่วงที่ใช้เวลามากที่สุด โดยปกติแล้วจะใช้เวลาประมาณ 60 เปอร์เซ็นต์ ของเวลาทั้งหมดในการเตรียมข้อมูล เนื่องมาจากอาจต้องมีการนำข้อมูลมาจากหลายแหล่ง และนำมารวมกัน เพื่อดูความสัมพันธ์ของข้อมูล ซึ่งข้อมูลที่ได้จากขั้นตอนนี้จะต้องมีความชัดเจน และถูกต้อง ดังนั้น จุดที่ต้องให้ความสำคัญคือการ Clean ข้อมูลและประเด็นของข้อมูล โดยขั้นตอนการเตรียมข้อมูลนี้จะแบ่งออกเป็น 3 ขั้นตอนย่อย ดังนี้

2.3.2.1 การคัดเลือกข้อมูล (Data Selection)

เป็นการระบุลักษณะและคัดเลือกข้อมูลที่ต้องการ และนำข้อมูลที่ไม่ต้องการออกไป ซึ่งการเลือกข้อมูลนี้ก็ขึ้นอยู่กับวัตถุประสงค์ที่ได้กำหนดไว้ตั้งแต่ต้น และการเลือกข้อมูลนี้จำเป็นจะต้องเข้าใจความหมาย ทราบประเภทของข้อมูล และค่าที่สามารถเป็นไปได้ ซึ่งตัวแปรข้อมูลแบ่งได้ 2 ลักษณะ คือ

*1. แบบ Categorical*

- Nominal : คือตัวแปรที่ลำดับของข้อมูลไม่มีความสำคัญ (ลำดับไม่มีผลกับค่า) เช่น สถานภาพ (Single, Married, Divorced)
- Ordinal : คือตัวแปรที่ลำดับของข้อมูลมีความสำคัญ (ลำดับมีผลกับค่า) เช่น อัตราการใช้บัตรเครดิตของลูกค้า (good, regular, poor) ซึ่งแต่ลักษณะของข้อมูลจะสามารถบอกถึงจำนวนหรือความถี่มากน้อยได้

*2 แบบ Quantitative*

- Continuous : จะเก็บค่าตัวเลขที่เป็นจำนวนจริง (Real number) เช่น ค่าใช้จ่ายของบริษัทเฉลี่ยต่อเดือน
- Discrete : จะเก็บค่าตัวเลขที่เป็นจำนวนเต็ม (Integer) เช่น จำนวนพนักงานในบริษัท

2.3.2.2 การกรองข้อมูล (Data Preprocessing)

ในกระบวนการนี้จะมีปริมาณข้อมูลส่วนหนึ่งที่ถูกเลือกเข้ามาจากกระบวนการ Data Selection ซึ่งข้อมูลเหล่านี้จะต้องเป็นข้อมูลที่ถูกต้องพร้อมสำหรับการทำ Mining แต่บางครั้งข้อมูลที่ได้นี้อาจยังมีข้อมูลที่ไม่ถูกต้อง จึงต้องทำการตรวจสอบก่อนโดยใช้หลักการทางสถิติ เช่น ข้อมูลที่เป็น Categorical การวัดการกระจายของข้อมูลจะทำให้เข้าใจข้อมูลที่มีอยู่ได้ดียิ่งขึ้น วิธีการที่ง่ายที่สุดคือการนำข้อมูลนั้นไปสร้างกราฟ ซึ่งจะช่วยให้เห็นความโน้มเอียงของข้อมูลและข้อมูลที่ผิดปกติได้ ส่วนข้อมูลที่เป็นตัวเลข การวิเคราะห์ข้อมูลทำได้โดย การหาค่าสูงสุด (Max) ค่าต่ำสุด (Min) ค่าเฉลี่ย (Mean) ค่าฐานนิยม (Mode) ค่ามัธยฐาน (Median) ซึ่งเราจะเห็นข้อมูลที่ผิดปกติในขั้นตอนนี้คือ

*1. Noisy Data* เป็นข้อมูลที่มีลักษณะแตกต่างจากข้อมูลที่คาดการณ์ไว้ หรือที่ควรจะเป็น ซึ่งอาจจะเกิดจากการป้อนข้อมูลผิด เช่น บันทึกค่าเงินเดือนพนักงานติดลบ หรือบันทึกส่วนสูงเป็น 560 ซ.ม. เป็นต้น ซึ่งค่าเหล่านี้ควรถูกแก้ไข หรือไม่นำมาวิเคราะห์ ดังนั้นจึงควรมีขั้นตอนของการตรวจสอบข้อมูลก่อนนำไปใช้

*2. Missing Value* ข้อมูลที่ไม่ได้ถูกเลือกมาจากขั้นตอน Data Selection คือมีข้อมูลบางส่วนหายไป อาจเกิดจากความผิดพลาดของคนหรือไม่มีข้อมูลส่วนนี้ในขณะที่รับข้อมูล ถ้าข้อมูลที่ขาดมีจำนวนน้อย อาจแก้ไขโดยการตัดข้อมูลนั้นทิ้งทั้งรายการ แต่ถ้าข้อมูลที่ขาดไปมีมากอาจต้องบันทึกส่วนที่หายไปด้วยค่าเฉลี่ย (สำหรับข้อมูลที่เป็น Categorical อาจบันทึกด้วยค่าฐานนิยมแทน หรือบันทึกเป็น "Unknown")

2.3.2.3 การแปลงข้อมูล (Data Transformation)

เป็นการสร้างข้อมูลชุดใหม่ที่มาจากข้อมูลชุดเดิม ซึ่งทำเพื่อแปลงข้อมูลให้สอดคล้องกับ model ที่ใช้ ได้ เช่นการ map ค่ามาเพื่อคำนวณ เป็นต้น การทำการแปลงข้อมูลนี้ จะทำให้การทำ Data Mining มีคุณภาพดีขึ้น เนื่องจากข้อมูลเมื่อถูกแปลงแล้วจะได้ผลลัพธ์ที่ดีขึ้น นอกจากนี้การทำ transformation ยังรวมไปถึงการสร้างข้อมูลใหม่จากข้อมูลชุดเดิมเพื่อให้ได้ข้อมูลที่มีคุณภาพดีขึ้น การแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมที่พร้อมจะนำไปวิเคราะห์ตาม Algorithm ของ Data Mining ที่จะใช้นั้น จะมีลักษณะเฉพาะแตกต่างกันไป เช่น การแปลงข้อมูลให้เป็นช่วงเพื่อใช้กับ Decision Tree หรือการปรับอัตราส่วนตัวเลขให้อยู่ในช่วง 0-1 เพื่อใช้กับบาง Algorithm ใน Neural Network

**2.3.2.1 ตัวอย่างการเตรียมข้อมูลเพื่อประยุกต์ใช้กับข้อมูลจริง**

หากต้องการนำเทคนิค Data Mining ไปประยุกต์ใช้กับงานด้านการศึกษา เนื่องจากเล็งเห็นว่าในปัจจุบันตามสถาบันการศึกษาส่วนใหญ่มีข้อมูลต่าง ๆ นิสิตที่ได้ถูกจัดเก็บไว้เป็นเวลานาน แต่

ข้อมูลส่วนใหญ่จะได้นำมาใช้ประโยชน์ตอนที่นิสิตศึกษาอยู่เท่านั้น เมื่อนิสิตจบการศึกษาไปแล้ว ข้อมูลก็จะได้รับการจัดเก็บไว้เป็นอย่างดี โดยที่ไม่ได้นำมาใช้ให้เกิดประโยชน์เท่าที่ควร โดยถ้าเราต้องการนำเทคนิค Data Mining เพื่อนำมาช่วยนิสิตในการเลือกสาขาวิชา เช่น สำหรับที่นิสิตคณะวิศวกรรมศาสตร์ จะเห็นได้ว่ามีสาขาวิชาต่าง ๆ มากมายกว่า 10 สาขาวิชา ซึ่งจะเห็นได้ว่า นิสิตส่วนใหญ่เมื่อเข้ามาศึกษาในคณะวิศวกรรมศาสตร์แล้ว พอถึงเวลาที่ต้องเลือกสาขาวิชา นิสิตจะไม่ทราบว่าความสามารถตนเองควรจะเข้าเรียนในสาขาวิชาใดจึงจะมีโอกาสประสบความสำเร็จมากที่สุด ดังนั้น เราจึงเห็นว่าสมควรอย่างยิ่งที่จะนำเทคนิค Data Mining มาประยุกต์ใช้กับฐานข้อมูลนิสิต โดยความรู้ที่ได้จากการทำ Data Mining สามารถนำมาใช้ในการช่วยนิสิตเลือกสาขาวิชาได้เมื่อเราได้เป้าหมายในการทำ Data Mining หลังจากนั้นก็ถึงกระบวนการหาข้อมูลนิสิต

สมมุติว่าเราได้ข้อมูลนิสิตย้อนหลังทั้งหมด 10 ปี มีทั้งหมด 2 ส่วน คือ ข้อมูลประวัติส่วนตัวนิสิตดังตารางที่ 2.1 และข้อมูลการลงทะเบียนเรียนในแต่ละรายวิชาของนิสิตดังตารางที่ 2.2

**ตารางที่ 2.1** แสดงข้อมูลประวัติส่วนตัวนิสิต

| ID | Sex | ชื่อ | Address | SchoolGPA | ... | Major | GPA |
|---|---|---|---|---|---|---|---|
| 1 | นาย | วิโรจน์ พัฒนากุล | 86/9 หมู่ 2..... | 2.5 | .... | ไฟฟ้า | 2.3 |
| 2 | นส. | ดวงพร เอี่ยมสุข | 542 หมู่ 7 .... | 3.4 | .... | โยธา | 3.2 |

จากตารางที่ 2.1 เป็นตัวอย่างข้อมูลประวัติส่วนตัวต่าง ๆ ของนิสิต เช่น รหัสประจำตัวนิสิต ชื่อ เพศสัญชาติ ที่อยู่ วันเกิด สถานภาพทางครอบครัว คะแนนสอบเข้า ผลการเรียนระดับมัธยม สาขาวิชาที่นิสิตศึกษาอยู่ เกรดเฉลี่ยสะสมจนถึงปีปัจจุบัน

**ตารางที่ 2.2** ตัวอย่างข้อมูลการลงทะเบียนเรียนของนิสิต

| ID | Subject | Section | Term | Year | Grade |
|---|---|---|---|---|---|
| 1 | 001 | 1 | 1 | 2537 | C+ |
| 1 | 002 | 1 | 1 | 2537 | D |
| 1 | 005 | 1 | 1 | 2537 | B+ |

จากตารางที่ 2.2 เป็นตารางข้อมูลการลงทะเบียนของนิสิตในแต่ละรายวิชา ในแต่ละภาคการศึกษา พร้อมทั้งหมู่ที่เรียน และผลการเรียนในรายวิชานั้น ๆ ของนิสิตแต่ละคนเมื่อเราได้ข้อมูลทั้งหมดแล้ว ขั้นต่อมาก็คือ การเตรียมข้อมูลเพื่อให้พร้อมที่จะนำไปทำ Data Mining ซึ่งแบ่งเป็นขั้นต่าง ๆ ได้ดังนี้

*1. การทำข้อมูลให้สมบูรณ์ (Data Cleaning)*

ข้อมูลที่ได้มานั้น เป็นข้อมูลที่ยังไม่สมบูรณ์ที่จะสามารถนำไปใช้ผ่านกระบวนการ Data Mining ได้ จึงต้องมีการจัดการข้อมูล การเตรียมข้อมูลเบื้องต้นมีวิธีการดังนี้

- เลือกเฉพาะคอลัมน์สำคัญที่คาดว่าจะสามารถนำมาใช้ประโยชน์ได้ และเป็นคอลัมน์ที่มีข้อมูลค่อนข้างครบถ้วนเมื่อเทียบกับจำนวนนิสิต เช่น จากในตารางที่ 2.1 คอลัมน์สำคัญที่มีข้อมูลค่อนข้างมาก ได้แก่ ข้อมูลรหัสนิสิต ที่อยู่ อายุ เพศ ประวัติครอบครัวโรงเรียน เกรดเฉลี่ยที่จบการศึกษาในมหาวิทยาลัย เป็นต้น ส่วนในบางคอลัมน์ที่มีความสำคัญ แต่มีข้อมูลน้อยมากนั้นจะไม่นำมาพิจารณา เช่น ข้อมูลคะแนนสอบเอ็นทรานซ์ในแต่ละวิชา เหตุผลในการสอบเข้า เป็นต้น

- สำหรับคอลัมน์ที่มีค่าสำหรับทุกแถวเป็นค่าเดียวกัน เช่น “สัญชาติไทย” จะเป็นข้อมูลที่ไม่สามารถแยกความแตกต่างของแต่ละแถวได้เลย ดังนั้นในการทำ Data Mining จะไม่สามารถใช้ประโยชน์จากคอลัมน์นี้ ดังนั้น จึงไม่นำคอลัมน์นี้มาพิจารณา

- คอลัมน์ที่มีค่าที่ไม่ซ้ำกันเลย จากตารางที่ 2.1 ได้แก่ ชื่อผู้ปกครอง หมายเลขโทรศัพท์ เป็นต้น ข้อมูลเหล่านี้ไม่สามารถหาแถวที่มีข้อมูลสัมพันธ์กันได้เลย การทำ Data Mining จึงไม่สามารถนำข้อมูลเหล่านี้มาใช้ประโยชน์ได้ ดังนั้นในการทำ Data Mining ควรกำจัดคอลัมน์ที่มีข้อมูลไม่ซ้ำกันเลยออก

- แก้ไขข้อมูลให้ถูกต้องสมบูรณ์ ได้แก่ การแก้ไขค่าว่างของข้อมูล ซึ่งสามารถแก้ไขได้หลายวิธี เช่น แก้ไขโดยกำจัดข้อมูลที่ในแถวเป็นค่าว่าง (NULL) ยกตัวอย่างเช่น จากในตารางที่ 2 .2 ข้อมูลบางแถวค่าในคอลัมน์ Grade หายไป ซึ่งจะเห็นได้ว่าถ้ามีแต่รหัสนิสิตและวิชาที่ลงทะเบียนโดยที่ไม่มีข้อมูลเกรดแล้ว เราก็ไม่สามารถจะนำแถวนั้นพิจารณาเพื่อหาความสัมพันธ์ที่น่าสนใจได้

- ปรับเปลี่ยนข้อมูลให้มีค่าเหมาะสมในการตัดสินใจ เช่น จากตารางที่ 2.1 ข้อมูลที่เป็นที่อยู่นั้นไม่สามารถที่จะนำมาใช้โดยตรงได้ เพราะจะเป็นปัญหา คือ ข้อมูลที่อยู่ของนิสิตแต่ละคนไม่ซ้ำกันเลย ดังนั้นจึงต้องปรับเปลี่ยนข้อมูลให้อยู่ในรูปแบบที่จะสามารถนำไปใช้ได้ ในกรณีนี้จะปรับข้อมูลในคอลัมน์ที่อยู่ของนิสิตให้เป็น Bangkok และ Non-Bangkok อย่างใดอย่างหนึ่ง เป็นต้น

- การจัดกลุ่มข้อมูลเพื่อลดการกระจาย (Binning Data) ทั้งนี้เนื่องมาจากข้อมูลของนิสิตมีจำนวนไม่มาก แต่เกรดในแต่ละวิชาที่สามารถมีได้นั้นมีจำนวนมากถึง 10 ตัวด้วยกันคือ {A, B+, B, C+, C, D+ ,D, F, W, I} ดังนั้นเพื่อลดการกระจายของข้อมูลเกรดของนิสิตที่มีมากเมื่อเทียบกับ

จำนวนนิสิต จึงได้จัดกลุ่มเกรดของนิสิตเป็น 3 กลุ่ม ดังนี้ คือ เกรด {A, B+, B} เป็น High, เกรด {C+, C} เป็น Medium และ เกรด {D+, D, F, W, I} เป็น Low

ดังนั้น จากตารางที่ 2.1 ที่เป็นข้อมูลประวัตินิสิต จึงได้นำมาปรับเปลี่ยนข้อมูลบางส่วนเพื่อให้สมบูรณ์ขึ้น ได้แก่

- การตัดคอลัมน์ที่ไม่จำเป็นในการทำ Data Mining ออก เช่น คอลัมน์ชื่อนิสิต เพราะชื่อนิสิตแต่ละคนไม่สามารถนำมาทำ Data Mining ได้
- คัดเลือกเฉพาะคอลัมน์ที่คาดว่าจะสามารถนำมาทำ Data Mining ได้ เช่น คัดเลือกคอลัมน์โรงเรียน แต่เนื่องจากชื่อโรงเรียนของนิสิตแต่ละคนมีมากมาย เราจึงต้องปรับข้อมูลโรงเรียนให้เป็นกลุ่มอย่างสมดุลเพื่อที่จะได้สามารถนำไปใช้ในการทำ Data Mining ได้ เช่น แบ่งข้อมูลโรงเรียนเป็น 2 กลุ่ม คือ สอบเทียบ และจบจากมัธยมศึกษาปีที่ 6 โดยกำหนดว่า School = 0 คือจบการศึกษาจากมัธยมศึกษาปีที่ 6 และ School = 1 คือสอบเทียบ เป็นต้น
- ปรับเปลี่ยนข้อมูลในบางคอลัมน์เพื่อให้สามารถนำไป miningได้ เช่น คอลัมน์ที่อยู่ ปรับข้อมูลให้เป็นกลุ่มว่านิสิตอยู่ในกรุงเทพฯหรือไม่ เป็นต้น

ผลที่ได้จากการทำข้อมูลจากตารางที่ 2.1 ให้สมบูรณ์แสดงดังตารางที่ 2.3

**ตารางที่ 2.3 ตัวอย่างข้อมูลประวัตินิสิตที่ทำให้สมบูรณ์**

| ID | Sex | Address | School | ... | Major | GPA |
|---|---|---|---|---|---|---|
| 1 | Female | Bangkok | 1 | .... | ELEC | 2.3 |
| 2 | Male | Non-Bangkok | 0 | .... | CIVIL | 3.2 |

จากตารางที่ 2.3 ที่เป็นตารางข้อมูลการลงทะเบียนเรียนของนิสิต ได้ทำการปรับข้อมูลบางส่วนให้สมบูรณ์ขึ้น ได้แก่

- การตัดบางคอลัมน์ที่ไม่น่าสนใจที่จะนำมาทำ Data Mining ออก เช่น คอลัมน์หมู่การเรียน
- จับกลุ่มข้อมูลในคอลัมน์เกรดเพื่อลดการกระจายของข้อมูล เป็นต้น

ผลที่ได้จากการทำข้อมูลในตารางที่ 2.2 ให้สมบูรณ์แสดงดังตารางที่ 2.4

**ตารางที่ 2.4 ตัวอย่างข้อมูลการลงทะเบียนเรียนของนิสิตที่ทำให้สมบูรณ์**

| ID | Subject | Term | Year | Grade |
|---|---|---|---|---|
| 1 | 001 | 1 | 2537 | Medium |
| 1 | 002 | 1 | 2537 | Low |
| 1 | 005 | 1 | 2537 | High |

*2. การคัดเลือกข้อมูล (Data Selection)*

เราจำเป็นต้องคัดเลือกเฉพาะข้อมูลนิสิตที่สามารถนำมาใช้ประโยชน์ได้ เช่น

- คัดเลือกข้อมูลนิสิตเฉพาะนิสิตคณะวิศวกรรมศาสตร์ และรายวิชาที่นิสิตเรียนทั้งหมดเป็นรายวิชาเดียวกัน เนื่องมาจากถ้าข้อมูลที่เราได้มานั้นย้อนหลังไปถึง 10 ปี ข้อมูลรายวิชาในอดีตอาจเป็นคนละตัวกับรายวิชาในปัจจุบัน เนื่องมาจากความแตกต่างของหลักสูตรการศึกษาในแต่ละปี ดังนั้นเราต้องคัดเลือกเฉพาะข้อมูลนิสิตในปีที่มีรายวิชาแบบเดียวกันเท่านั้น
- คัดเลือกข้อมูลนิสิตในภาควิชาที่สามารถนำมาทำ Data Mining ได้ เช่น คัดเลือกมา 6 สาขาวิชาหลัก ได้แก่ สาขาวิชาวิศวกรรมเคมี สาขาวิชาวิศวกรรมโยธา สาขาวิชาวิศวกรรมคอมพิวเตอร์ สาขาวิชาวิศวกรรมไฟฟ้า สาขาวิชาวิศวกรรมอุตสาหการ และสาขาวิชาวิศวกรรมเครื่องกล สาเหตุที่เลือก 6 สาขาวิชาดังเนื่องมาจากทั้ง 6 สาขาวิชาเป็นสาขาวิชาหลักที่มีทั้งนิสิตและข้อมูลต่าง ๆ อยู่มากพอสมควรที่จะสามารถนำมาวิเคราะห์ได้ สำหรับสาขาวิชาอื่น ๆ ที่ไม่ได้คัดเลือกมานั้นอาจเป็นสาขาวิชาที่เพิ่งก่อตั้งมาได้ไม่นานนัก ทำให้ข้อมูลไม่เพียงพอในการนำมาวิเคราะห์ อาจทำให้มีข้อผิดพลาดได้ในการทดสอบได้

หลังจากที่ทำตามขั้นตอนข้างต้นทั้งหมดแล้ว จะได้ข้อมูลที่มีความสมบูรณ์มากขึ้น

*3. การปรับเปลี่ยนรูปแบบข้อมูล (Data Transformation)*

จากตารางที่ 2.4 จะเห็นได้ว่าข้อมูลอยู่ในระดับรายวิชา เพื่อให้ได้ตรงตามเป้าหมายที่ต้องการจะศึกษาพฤติกรรมและลักษณะของนิสิตแต่ละคน เราจะต้องแปลงข้อมูลให้อยู่ในระดับของนิสิต โดยแบ่งกลุ่มของวิชาต่างๆ ที่ลงทะเบียนตามรหัสนิสิต และคอลัมน์แทนรายชื่อวิชาต่างๆ จากนั้นจะนำตารางที่ 2.3 และ 2.4 มารวมกัน ทำให้ได้เป็นตารางข้อมูลนิสิตขั้นต้นที่แต่ละแถวของตารางแสดงทั้งประวัติส่วนตัวของนิสิตและผลการเรียนของนิสิตในแต่ละรายวิชา เพื่อที่เราจะได้สามารถนำตารางนี้ไปปรับเปลี่ยนเพื่อให้เหมาะสมกับเทคนิคต่าง ๆ ของ Data Mining ต่อไป ผลลัพธ์ที่ได้ทั้งหมดแสดงได้ดังตารางที่ 2.5

**ตารางที่ 2.5 ตัวอย่างตารางข้อมูลนิสิตที่ขั้นต้น**

| ID | Sex | Address | 001 | 002 | ... | Major | GPA |
|---|---|---|---|---|---|---|---|
| 1 | Male | Bangkok | Medium | Low | ... | ELEC | 2.3 |
| 2 | Female | Non-Bangkok | High | High | ... | CIVIL | 3.2 |

ข้อมูลในตารางที่ 2.5 นี้ถือได้ว่าเป็นข้อมูลเบื้องต้นในรูปแบบสมบูรณ์ที่พร้อมจะนำไปทำ Data Mining แล้ว ทั้งนี้จะต้องปรับเปลี่ยนรูปแบบของข้อมูลเพื่อให้เหมาะสมกับแต่ละเทคนิคของ Data Mining ที่เราจะเลือกใช้ด้วย

**2.3.3 ขั้นตอนการทำ Data Mining**

ขั้นตอนนี้ถือว่าเป็นส่วนสำคัญที่สุดของการทำ Data Mining เพราะการเลือกเอาวิธีการและ Algorithm ในการทำ Mining ที่เหมาะสม ก็จะทำให้การ Mining ได้ผลอย่างรวดเร็วและถูกต้องตามจุดประสงค์ที่ต้องการ ในขั้นตอนนี้จะเป็นการประมวลผลข้อมูลตาม Algorithm ที่ได้กำหนดไว้ ซึ่งจะมีความสัมพันธ์กับการวิเคราะห์ข้อมูลและขั้นตอนที่ผ่านมา โดยเมื่อทำในส่วนของ Data Mining แล้วอาจต้องย้อนกลับไปทำในขั้นตอนของการเตรียมข้อมูลใหม่ ในการพัฒนา Data Mining นั้นจะเกี่ยวข้องกับการใช้ Algorithm หลายๆ แบบ ซึ่งแต่ละแบบก็มีข้อดีและข้อเสียแตกต่างกันไป

**2.3.4 ขั้นตอนการวิเคราะห์ผลลัพธ์ที่ได้**

เป็นการวิเคราะห์ผลของการประมวลผล ซึ่งจะทำการแปลความหมายผลลัพธ์ที่ได้จากขั้นตอนการทำ Mining ว่าสามารถนำมาใช้ได้ตามวัตถุประสงค์ที่ต้องการหรือไม่ รวมทั้งเป็นการประเมินถึงความถูกต้องของผลลัพธ์ที่ได้จากการทำ ซึ่งก็เป็นส่วนสำคัญเช่นกัน เนื่องจากบางครั้งผลที่ได้อาจจะยังมีข้อผิดพลาดอยู่บ้าง โดยจะต้องทำการนำแบบจำลองที่ได้ไปทำการทดสอบกับข้อมูลชุดอื่นว่าได้ผลลัพธ์ที่ถูกต้องเช่นเดียวกันหรือไม่ ซึ่งการทำงานในส่วนนี้จำเป็นต้องใช้ทักษะในการวิเคราะห์ข้อมูลและการวิเคราะห์ทางธุรกิจเข้ามาช่วยด้วย

**2.3.5 ขั้นตอนการนำสารสนเทศที่ได้ไปใช้ประโยชน์**

เป็นขั้นตอนสุดท้ายของกระบวนการทั้งหมด ซึ่งเป็นการรวบรวมความเข้าใจในแบบจำลองที่เป็นผลมาจากขั้นตอนการวิเคราะห์ผลลัพธ์ที่ได้ มารวมเข้ากับส่วนความรู้ทางธุรกิจเพื่อที่จะนำเสนอถึงวิธีการที่จะนำผลที่ได้นี้ไปใช้ให้เกิดประโยชน์

**2.4 เทคนิคในการทำ Data Mining**

การทำ Data Mining ประกอบด้วย 4 model หลัก คือ

1. การสร้างแบบจำลองพยากรณ์ (Predictive Modeling)
2. การแบ่งส่วนฐานข้อมูล (Database Segmentation)
3. การวิเคราะห์ความสัมพันธ์ (Link Analysis)
4. การตรวจสอบค่าเบี่ยงเบน (Deviation Detection)

**2.4.1 การสร้างแบบจำลองพยากรณ์ (Predictive Modeling)**

เป็นการทำนายถึงความเป็นไปได้ โดยใช้การสังเกตจากรูปแบบของข้อมูลที่มีอยู่ คือเราจะใช้ Model นี้ในการวิเคราะห์ฐานข้อมูลที่มีอยู่เพื่อตัดสินใจเลือกลักษณะข้อมูลที่ต้องการ โดยมีลักษณะเป็นการเรียนรู้จากกลุ่มข้อมูลที่ได้กำหนดไว้ แล้วจึงนำไปวิเคราะห์กลุ่มข้อมูลที่ต้องการ ซึ่งวิธีนี้เรียกว่า Supervised Learning ดังนั้นข้อมูลที่มีอยู่ต้องสมบูรณ์ จึงจะทำให้ผลลัพธ์ออกมาถูกต้อง เพราะเราต้องนำข้อมูลในอดีตมาสร้างแบบจำลอง การทำงานจะแบ่งออกเป็น 2 ขั้นตอน คือ

1.) Training Phase คือขั้นตอนการสร้างแบบจำลองขึ้นมาใหม่โดยใช้ข้อมูลในอดีต ซึ่งจะใช้ข้อมูลประมาณ 80% ของข้อมูลทั้งหมด
2.) Testing Phase คือขั้นตอนที่ใช้ทำการทดสอบแบบจำลองที่สร้างว่ามีความเหมาะสมหรือไม่ โดยจะนำข้อมูลส่วนที่เหลือ 20% จากช่วง Training Phase มาใช้ทดสอบแบบจำลองที่สร้างขึ้น

Predictive Modeling ยังสามารถแบ่งย่อยได้อีก เป็น 2 เทคนิคคือ

1. Classification : เป็นการทำนายว่าสิ่งนั้นควรอยู่ในกลุ่มไหน ซึ่งเป็นการแบ่งกลุ่มของข้อมูลตามชนิดของกลุ่มข้อมูลที่ควรจะเป็น และสามารถแบ่งกลุ่มข้อมูลได้อย่างชัดเจน เช่น การจัดกลุ่มของลูกค้าเพื่อพิจารณาว่าควรจะให้วงเงินสินเชื่อเพิ่มขึ้นหรือไม่ เป็นต้น ซึ่งวิธีที่นิยมใช้คือ Tree Induction และ Neural Induction

2. Value prediction : เป็นการทำนายถึง ค่าความต่อเนื่องของข้อมูล เป็นการทำนายค่าที่เป็นตัวเลข เช่น การทำนายราคาหุ้น เป็นต้น โดยมีวิธีที่ใช้คือ Linear Regression และ Nonlinear Regression

**2.4.2 การแบ่งส่วนฐานข้อมูล (Database Clustering หรือ Segmentation)**

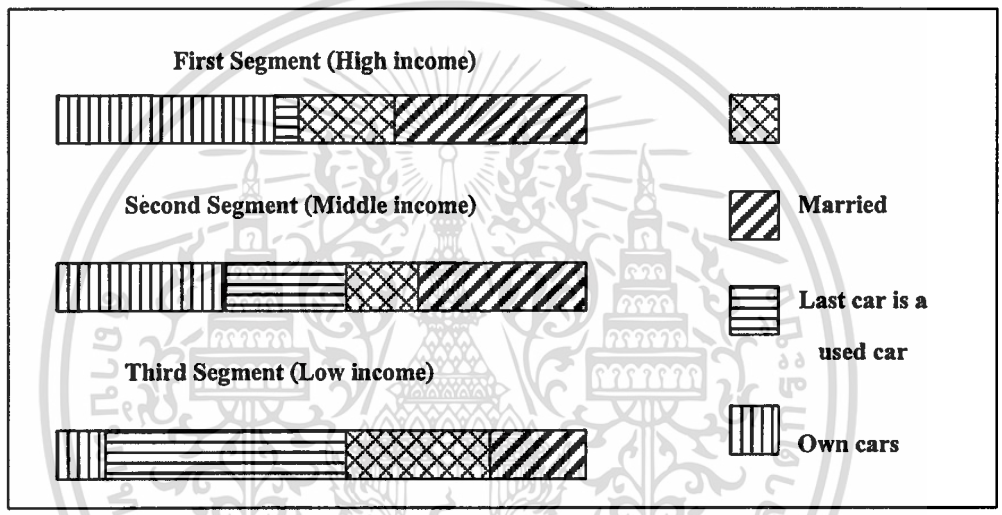
เป็นการแบ่งหรือจัดกลุ่มของข้อมูลที่มีลักษณะคล้ายกันหรือมีคุณสมบัติใกล้เคียงกันในหลายๆ ด้านให้เป็นข้อมูลกลุ่มเดียวกัน ซึ่งแต่ละกลุ่มจะถูกเรียกว่า Segments หรือ Clusters การแบ่งกลุ่มข้อมูลนี้เราจะไม่สามารถกำหนดได้ว่าข้อมูลควรจะอยู่กลุ่มใด แต่จะเป็นการกำหนดกลุ่มของข้อมูลจากธรรมชาติของข้อมูลเองไม่ได้ใช้ความรู้สึกหรือประสบการณ์ในการตัดสินใจแบ่งกลุ่มข้อ

มูลและข้อมูลจะถูกจัดการโดย algorithm ที่เหมาะสม จึงเรียกว่าเป็นรูปแบบของ Unsupervised Learning ซึ่งสามารถแบ่งย่อยตามวิธีที่ใช้ คือ Demographic Clustering และ Neural Clustering

ตัวอย่างเช่น บริษัทจำหน่ายรถยนต์ได้แยกกลุ่มลูกค้าออกเป็น 3 กลุ่ม คือ

a. กลุ่มผู้มีรายได้สูง (>$80,000)
b. กลุ่มผู้มีรายได้ปานกลาง ($25,000 to $ 80,000)
c. กลุ่มผู้มีรายได้ต่ำ (less than $25,000)

และภายในแต่ละกลุ่มยังแยกออกเป็น Have Children, Married, Last car is a used car, Own cars




**รูปที่ 2.3** ตัวอย่าง Clustering

จากข้อมูลข้างต้นทำให้ทางบริษัทรู้ว่าเมื่อมีลูกค้าเข้ามาที่บริษัทควรจะเสนอขายรถประเภทใด เช่น ถ้าเป็นกลุ่มผู้มีรายได้สูงควรจะเสนอรถใหม่ เป็นรถครอบครัวขนาดใหญ่พอสมควร แต่ถ้าเป็นผู้มีรายได้ค่อนข้างต่ำควรเสนอรถมือสอง ขนาดค่อนข้างเล็ก

**2.4.3 การวิเคราะห์ความสัมพันธ์ (Link Analysis)**

เป็นการศึกษาวิเคราะห์ความสัมพันธ์ของข้อมูลหรือกลุ่มของข้อมูล ว่ามีความสัมพันธ์กันหรือไม่ อย่างไร และถ้ามีความสัมพันธ์กันจะสัมพันธ์กันในรูปแบบลักษณะใด โดยเรียกความสัมพันธ์นี้ว่าเป็น "Association" เป็นแบบจำลองที่นิยมกันมากในการวิเคราะห์เพื่อหาความสัมพันธ์ระหว่าง ลูกค้ากับ สินค้าหรือบริการ สามารถแบ่งย่อยได้เป็น 3 ลักษณะ คือ

1. Association Discovery : เป็นการวิเคราะห์ข้อมูลที่เกิดขึ้นพร้อมกันภายในกลุ่มข้อมูลเดียวกัน เป็นเทคนิคหนึ่งที่ได้รับความนิยมมาก ซึ่งมักใช้ในการวิเคราะห์ถึงพฤติกรรมการซื้อของผู้บริโภค จึงมีชื่อเรียกอีกอย่างว่า Market basket analysis

2. Sequential Pattern Discovery : เป็นการศึกษาความสัมพันธ์ระหว่างข้อมูล โดยเทียบข้อมูลกับเวลา ซึ่งเป็นการศึกษาพฤติกรรมในระยะยาว (Long Term Behavior)

3. Similar Time Sequence Discovery : เป็นการศึกษาพฤติกรรมของข้อมูลที่เกิดขึ้นทั้งหมดหรือเกิดขึ้นในช่วงเวลาเดียวกัน เพื่อหาความสัมพันธ์ระหว่างกลุ่มของข้อมูลเหล่านี้

**2.4.4 การตรวจสอบค่าเบี่ยงเบน (Deviation Detection)**

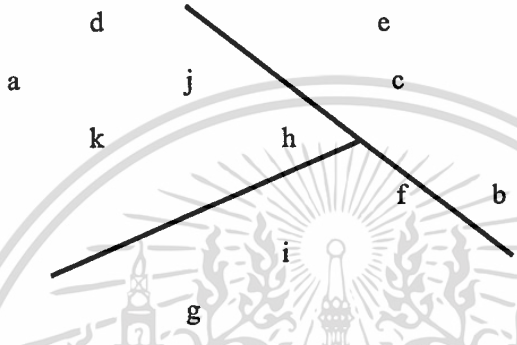
เป็นเทคนิคที่ใช้ทำการหาค่าที่มีความแตกต่างไปจากค่ามาตรฐาน ว่ามีค่ามากน้อยเพียงใด เป็นแบบจำลองที่ใช้เทคนิคทางสถิติ (Statistics) เพื่อใช้วัดความน่าเชื่อถือของข้อมูล และการแสดงให้เห็นภาพ (Visualization) ซึ่งเป็นการสรุปข้อมูลให้แสดงผลออกมาในรูปแบบ Graphic เช่น Histograms Scatter Plots หรือ กราฟวงกลม เป็นต้นเพื่อให้สามารถเข้าใจได้ง่าย นอกจากนี้ Visualization ยังสามารถนำไปใช้ร่วมกับเทคนิคอื่นๆ โดยใช้ในการแสดงผลที่ได้ในรูปแบบของกราฟฟิก ทำให้เข้าใจได้ง่ายขึ้นอีกด้วย

นอกจากนี้แต่ละเทคนิคของ Data Mining ก็ยังมี Algorithm ต่างๆ ของแต่ละวิธีด้วย ซึ่งการที่จะเลือกใช้ Algorithm ใดนั้น ก็ขึ้นอยู่กับปัจจัยหลายๆ อย่างอีก เช่น ข้อจำกัดในการทำ ลักษณะของข้อมูล ชนิดของข้อมูล และจำนวนข้อมูลที่มีอยู่ ซึ่งบางครั้งก็อาจต้องมีการเปลี่ยนแปลง หากเทคนิคนั้นไม่เหมาะสม สิ่งที่สำคัญของกระบวนการนำมาใช้อยู่ที่การกำหนดกลุ่มของข้อมูลที่จะนำมาทำและการสร้าง Model ซึ่งหากทำการกำหนดและเลือกใช้อย่างเหมาะสมแล้ว ก็จะทำให้ผลของการทำ Data mining เป็นไปอย่างถูกต้องและรวดเร็ว โดยจากที่กล่าวมาจะเห็นได้ว่า Data Mining มีเทคนิคและวิธีการที่สามารถนำมาใช้งานอยู่หลายวิธี ซึ่งเราจะต้องเลือกใช้ให้เหมาะสมกับงานประเภทต่างๆ และขึ้นอยู่กับรูปแบบ Application ที่ต้องการนำมาใช้งานด้วย

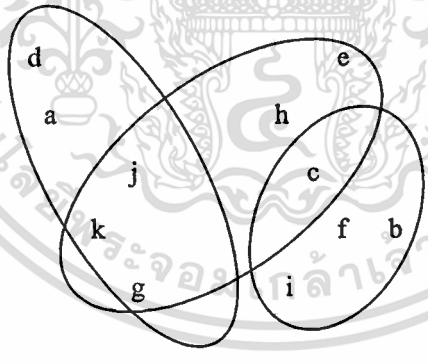
**2.5 ทฤษฎีการแบ่งส่วนฐานข้อมูล (Database Clustering หรือ Segmentation)**

เทคนิคการแบ่งส่วนฐานข้อมูล หรือการแบ่งกลุ่มนี้จะเป็นการทำ Segments หรือ Clustering ในรูปแบบของ Data Mining ซึ่งจะไม่เหมือนกับการทำ Clustering ของทางสถิติ เนื่องจาก Clustering ทางสถิติจะหมายถึงการทำ Sampling ข้อมูล เช่น ถ้าเรามีข้อมูลอยู่ 10 กลุ่มซึ่งมีลักษณะที่ไม่แตกต่างกันมาก ก็จะแบ่งกลุ่มโดยสุ่มใช้ข้อมูลเพียงบางกลุ่ม แต่ถ้าเป็นแบบ Clustering ทางด้าน Mining นั้น จะหมายถึงการแบ่งกลุ่มและต้องใช้ข้อมูลของทุกกลุ่มและหาจุดเด่นของแต่ละกลุ่มออกมาให้เห็นอย่างชัดเจน

Clustering ของทาง Data Mining จะมีลักษณะการแสดงของ Cluster ที่แตกต่างกันออกไป เช่น ภาพการแบ่งกลุ่มในรูปการแบ่งส่วนเพื่อแสดงออกเป็นรูปแบบ Cluster ภาพแสดงให้เห็นว่า ใน 1 ตัวอย่าง (Instance) จะสามารถอยู่ได้ในหลาย Cluster ภาพแสดงความน่าจะเป็น หรือ Degree ของสมาชิกแต่ละ Cluster และ ภาพแสดงการแบ่งกลุ่มเป็นรูปแบบลำดับชั้นจากบนลงล่าง (Hierachical)



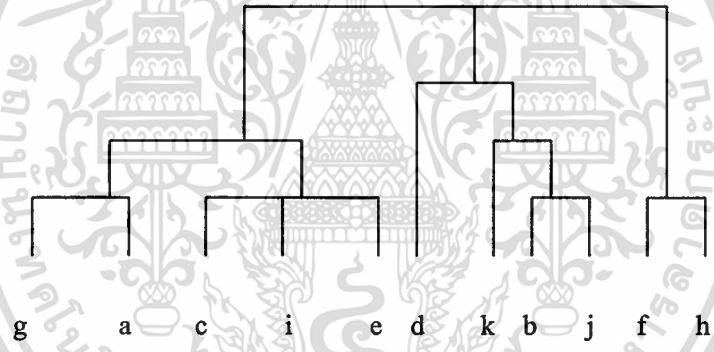

**รูปที่ 2.4** การแบ่งกลุ่มในรูปการแบ่งส่วนเพื่อแสดงออกเป็นรูปแบบ 3 Cluster




**รูปที่ 2.5** แสดงกลุ่มข้อมูลในตัวอย่างซึ่งสามารถอยู่ได้ในหลาย Cluster

| สมาชิก | Cluster | | |
|---|---|---|---|
| A | 0.4 | 0.1 | 0.5 |
| B | 0.1 | 0.8 | 0.1 |
| C | 0.3 | 0.3 | 0.4 |
| D | 0.1 | 0.1 | 0.8 |
| E | 0.4 | 0.2 | 0.4 |
| F | 0.1 | 0.4 | 0.5 |
| G | 0.7 | 0.2 | 0.1 |
| H | 0.5 | 0.4 | 0.1 |

**รูปที่ 2.6** แสดงความน่าจะเป็น หรือ Degree ของสมาชิกแต่ละ Cluster




**รูปที่ 2.7** แสดงการแบ่งกลุ่มเป็นรูปแบบลำดับชั้นจากบนลงล่าง (Hierachical)

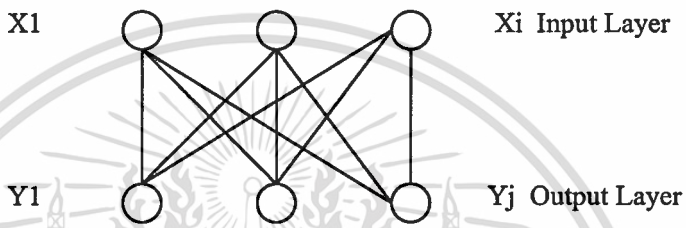
### 2.5.1 Neural Network Algorithm

เป็น algorithm ที่มีที่มาจากงานวิจัยด้านปัญญาประดิษฐ์ (Artificial Intelligence : AI) เพื่อใช้ในการคำนวณค่าฟังก์ชันจากกลุ่มข้อมูล วิธีการของ นิวรอลเน็ต (แท้จริงต้องเรียกให้เต็มว่า Artificial Neural Networks หรือ ANN) เป็นวิธีการที่ให้เครื่องเรียนรู้จากตัวอย่างต้นแบบ แล้วฝึก (train) ให้ระบบได้รู้จักที่จะคิดแก้ปัญหาที่กว้างขึ้นได้ ซึ่งมีแนวความคิดในการเรียนรู้ที่คล้ายคลึงกับระบบสมองของมนุษย์ ขั้นตอนของการนำโครงข่ายประสาทเทียมมาใช้สำหรับการพยากรณ์ก็มีลักษณะเช่นเดียวกับวิธีการพยากรณ์อื่นๆ ซึ่งจะต้องอาศัยข้อมูลป้อนเข้าเพื่อสร้างแบบจำลองในการพยากรณ์ข้อมูลในอนาคต ปรับปรุงให้เหมาะสมกับเงื่อนไขของตลาดที่มีการเปลี่ยนแปลง และมีความสามารถในการรวมการวิเคราะห์พื้นฐานและเทคนิคเพื่อสร้างแบบจำลอง โดยที่โครงข่าย

ประสาทเทียมจะพยายามลดจำนวนของการทำนายที่ผิดพลาดให้ต่ำที่สุด ซึ่งเป็นเหตุผลหลักที่มีการนำมาใช้ในการทำนายข้อมูลทางธุรกิจ

**5.5.1.1 Kohonen's Self-Organizing Maps (SOM)**

เป็น Algorithm ที่นิยมใช้ในการแบ่งกลุ่ม (Cluster) ซึ่งเป็นการใช้งานแบบ Neural Network Algorithm ที่ไม่มี Hidden Layer เพราะฉะนั้นจึงมีเพียงแค่ 2 Layer คือ Input Layer และ Output Layer ดังภาพต่อไปนี้




i = 1, 2,..., n และ j = 1, 2,..., n

**รูปที่ 2.8** Kohonen's Self-Organizing Maps Neural Network

วิธีการคิดของ Kohonen's Self-Organizing Maps Neural Network คือ การทำซ้ำของในแต่ละข้อมูลเพื่อที่จะได้หาค่าของน้ำหนักของข้อมูลที่มีอยู่ทั้งหมดตามจำนวนกลุ่มที่ต้องการแบ่ง มีวิธีการคิดดังนี้

1. กำหนดน้ำหนัก (W) ให้กับข้อมูล และกำหนดค่า Learning Rate ($\alpha$)
2. นำค่าของข้อมูลมาคำนวณด้วยสูตร Distance คือ

$$D(j) = \sum_{i=1}^{N} (W_{ij} - X_i)^2$$

โดยที่ I = ข้อมูลที่ใช้ในการแบ่งกลุ่มข้อมูล (Input)

J = จำนวนกลุ่มข้อมูลที่ต้องการค้นหา (Output)

D = ระยะทางระหว่างจุดข้อมูลกับจุดกึ่งกลางของกลุ่ม (Distance)

X = ค่าของข้อมูล (Attribute)

3. เปรียบเทียบกับ D(j) ของแต่ละกลุ่มข้อมูลว่าค่าของกลุ่มไหนมีค่าน้อยที่สุด

4. นำข้อมูลของกลุ่มที่มีค่าน้อยที่สุดมาคำนวณหาน้ำหนักใหม่ โดยใช้สูตรดังนี้

$$W_{ij}\ (new) = W_{ij}\ (old) + \alpha\ (X_i - W_{ij}\ (old))$$

5. ทำการลดค่า $\alpha$ ลง

6. ทำซ้ำตั้งแต่ข้อ 2 ใหม่จนกว่าจะไม่มีการเปลี่ยนแปลงค่า D(j) ในแต่ละกลุ่ม

**5.5.1.2 ตัวอย่างการคำนวณ Kohonen's Self-Organizing Maps Neural Network**

มีข้อมูลอยู่ 4 รายการ โดยแต่ละชุดข้อมูลมี 4 คอลัมน์ ดังนี้ (1, 1, 0, 0); (0, 0, 0, 1); (1, 0, 0, 0); (0, 0, 1, 1) โดยให้มีการแบ่งกลุ่มข้อมูลออกเป็น 2 กลุ่ม และให้ค่า Learning rate = 0.6

สมมุติกำหนดให้น้ำหนักแก่ค่าโครงข่ายประสาทเทียม ต่าง ๆ ที่ต้องการทำการแบ่งกลุ่มจะได้ดังต่อไปนี้

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \\ W_{31} & W_{32} \\ W_{41} & W_{42} \end{bmatrix} = \begin{bmatrix} 0.2 & 0.8 \\ 0.6 & 0.4 \\ 0.5 & 0.7 \\ 0.9 & 0.3 \end{bmatrix}$$

นำข้อมูลชุดแรก คือ 1100 มาทำการคำนวณ โดยการหาค่าด้วย สูตร Distance

จากสูตร การแทนค่าจะได้ดังต่อไปนี้

$$D(1) = (0.2-1)^2 + (0.6-1)^2 + (0.5-0)^2 + (0.9-0)^2$$
$$= 1.86$$
$$D(2) = (0.8-1)^2 + (0.4-1)^2 + (0.7-0)^2 + (0.3-0)^2$$
$$= 0.98$$

หลังจากได้ค่า D(1), D(2) แล้วให้เลือกตัวที่มีค่าน้อยที่สุดเพื่อมาคำนวณหาน้ำหนักใหม่ จากตัวอย่างที่คำนวณจะต้องคำนวณน้ำหนักใหม่ของ D(2) (เพราะว่า j = 2) ซึ่งจะต้องใช้สูตรดังต่อไปนี้

$$W_{ij}\ (new) = W_{ij}\ (old) + \alpha\ (X_i - W_{ij}\ (old))$$

จากสูตร การแทนค่าจะได้ดังต่อไปนี้

$W_{12} = 0.8 + 0.6 ( 1 - 0.8 )$

$= 0.92$

$W_{22} = 0.4 + 0.6 ( 1 - 0.4 )$

$= 0.76$

$W_{32} = 0.7 + 0.6 ( 1 - 0.7 )$

$= 0.28$

$W_{42} = 0.3 + 0.6 ( 1 - 0.3 )$

$= 0.12$

ซึ่งจะเขียนในรูป Matrix ได้ดังต่อไปนี้

$$\begin{bmatrix} W11 & W_{12} \\ W_{21} & W_{22} \\ W_{31} & W_{32} \\ W_{41} & W_{42} \end{bmatrix} = \begin{bmatrix} 0.2 & 0.92 \\ 0.6 & 0.76 \\ 0.5 & 0.28 \\ 0.9 & 0.12 \end{bmatrix} .$$

นำข้อมูลชุดที่ 2 คือ 0001 มาคำนวณซ้ำ โดยใช้วิธีเดียวกับข้อมูลในแถวแรก ซึ่งจะได้ดังผลต่อไปนี้

สำหรับเวคเตอร์ 0001

$D(1) = 0.66$

$D(2) = 2.2768$

เลือก D(1) เพราะว่ามีค่าน้อยกว่า D(2) หลังจากนั้นทำการคำนวณหาค่าน้ำหนักใหม่ซึ่งจะได้ดังต่อไปนี้

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \\ W_{31} & W_{32} \\ W_{41} & W_{42} \end{bmatrix} = \begin{bmatrix} 0.08 & 0.92 \\ 0.24 & 0.76 \\ 0.20 & 0.28 \\ 0.96 & 0.12 \end{bmatrix}$$

นำค่าจากข้อมูลชุดที่ 3 คือ 1000 มาทำการคำนวณซ้ำ โดยวิธีเดียวกับข้อมูลในแถวแรก ซึ่งจะได้ดังผลต่อไปนี้

สำหรับเวคเตอร์ 1000

D(1) = 1.8656

D(2) = 0.6768

เลือก D(2) เพราะว่ามีค่าน้อยกว่า D(1) หลังจากนั้นทำการคำนวณหาค่าน้ำหนักใหม่ซึ่งจะได้ดังต่อไปนี้

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \\ W_{31} & W_{32} \\ W_{41} & W_{42} \end{bmatrix} - \begin{bmatrix} 0.08 & 0.968 \\ 0.24 & 0.304 \\ 0.20 & 0.112 \\ 0.96 & 0.048 \end{bmatrix}$$

นำค่าจากข้อมูลในแถวต่อไปคือ 0011 มาทำการคำนวณซ้ำในเหมือนกับข้อมูลในแถวแรก ซึ่งจะได้ดังผลต่อไปนี้

สำหรับเวคเตอร์ 0011

D(1) = 0.7056

D(2) = 2.724

เลือก D(1) เพราะว่ามีค่าน้อยกว่า D(2) หลังจากนั้นทำการคำนวณหาค่าน้ำหนักใหม่ซึ่งจะได้ดังต่อไปนี้

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \\ W_{31} & W_{32} \\ W_{41} & W_{42} \end{bmatrix} - \begin{bmatrix} 0.032 & 0.968 \\ 0.096 & 0.304 \\ 0.680 & 0.112 \\ 0.984 & 0.048 \end{bmatrix}$$

ให้ลดค่า $\alpha$ ลงครึ่งหนึ่งดังต่อไปนี้

$$\alpha(1) = \alpha(0) / 2 = 0.6 / 2 = 0.3$$

คำนวณซ้ำใหม่จนกว่าไม่มีการเปลี่ยนแปลงค่า ซึ่งหลังจากคำนวณ 100 รอบได้ผลลัพธ์ดังต่อไปนี้

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \\ W_{31} & W_{32} \\ W_{41} & W_{42} \end{bmatrix} = \begin{bmatrix} 6.7x0^{-16} & 1 \\ 2x10^{-16} & 0.49 \\ 0.51 & 2.3x10^{-16} \\ 1 & 1x10^{-16} \end{bmatrix}$$

เมื่อปรับค่าในเมทริกซ์ให้อยู่ในรูปแบบที่ง่ายขึ้น ได้ผลดังนี้

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \\ W_{31} & W_{32} \\ W_{41} & W_{42} \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ 0 & 0.5 \\ 0.5 & 0 \\ 1 & 0 \end{bmatrix}$$

จากเมทริกซ์นี้ สามารถแบ่งข้อมูลได้เป็น 2 กลุ่มอย่างชัดเจน

# บทที่ 3
# วิธีการดำเนินการศึกษา

## 3.1 การเตรียมข้อมูล (Data Preparation)

การศึกษาในครั้งนี้หลังจากที่ได้มีการกำหนดวัตถุประสงค์ไว้แล้วในบทที่ 1 ก็ได้ทำการเตรียมข้อมูล โดยขั้นตอนการเตรียมข้อมูลนี้จะแบ่งออกเป็น 3 ขั้นตอนย่อย ดังนี้

### 3.1.1 การคัดเลือกข้อมูล (Data Selection)

การเลือกข้อมูล หรือการได้มาของข้อมูล ในการศึกษาครั้งนี้ข้อมูลที่เลือกมานั้นอยู่ในรูปแบบของ excel file เป็นข้อมูลซึ่งสุ่มมาจากฐานข้อมูลลูกค้าธนาคารพาณิชย์แห่งหนึ่งในประเทศไทย เพื่อนำมาใช้เป็นข้อมูลสำหรับการศึกษาพฤติกรรมของลูกค้าธนาคาร โดยข้อมูลที่จะนำมาเพื่อใช้ในการศึกษา แบ่งออกเป็น 4 ลักษณะดังนี้

1. Customer demographic data : เป็นข้อมูลส่วนบุคคลของลูกค้าแต่ละคน ที่ไม่เกี่ยวกับลักษณะการเป็นลูกค้า เช่น อายุ, เพศ, สถานภาพสมรส เป็นต้น

2. Relationship data : เป็นข้อมูลซึ่งบอกถึงปริมาณและคุณภาพการเป็นลูกค้าของลูกค้าแต่ละราย ซึ่งรวมถึงข้อมูลเกี่ยวกับ ทรัพย์สินที่ถือครอง, สินค้าที่ใช้บริการ, ยอดคงเหลือ, วันเริ่มใช้บริการ เป็นต้น

3. Transactional data : เป็นข้อมูลเกี่ยวกับจำนวน และปริมาณการทำธุรกรรมของลูกค้าแต่ละคนที่ใช้บริการกับธนาคาร

4. Additional data : ข้อมูลประกอบอื่นๆ ที่จะทำให้สามารถเข้าใจแก่นแท้ของลูกค้ามากขึ้น

โดยในเบื้องแรกนั้นได้กำหนดตัวแปรที่ต้องการใช้ในการศึกษาไว้ 27 ตัวแปรด้วยกันดังนี้

**ตารางที่ 3.1** แสดง ตัวแปรที่ต้องการในเบื้องแรก 27 ตัวแปร และคำจำกัดความ

| ลำดับที่ | Variable Name | Description |
|---|---|---|
| 1 | Age | อายุ |
| 2 | Sex | เพศ |
| 3 | Address | ที่อยู่ |

| ลำดับที่ | Variable Name | Description |
|---|---|---|
| 4 | Income | รายได้ |
| 5 | Car owner | การเป็นเจ้าของรถยนต์ |
| 6 | Home Owner | การเป็นเจ้าของบ้าน |
| 7 | Marital Status | สถานภาพสมรส |
| 8 | Has Children | จำนวนบุตร |
| 9 | Time as customers | จำนวนครั้งที่เป็นลูกค้า |
| 10 | Total deposit | จำนวนเงินฝาก |
| 11 | Total liabilities | จำนวนหนี้สิน |
| 12 | Has defaulted | เป็นลูกค้าที่มีหนี้เสีย |
| 13 | Salary deposited in bank | การมีเงินเดือนจ่ายเข้าผ่านธนาคาร |
| 14 | Web bank usage | ปริมาณการใช้ website ของธนาคาร |
| 15 | ATM usage | ปริมาณการใช้ ATM |
| 16 | Call center usage | ปริมาณการใช้บริการ Call center |
| 17 | Teller usage | ปริมาณการใช้บริการหน้าเคาท์เตอร์ |
| 18 | No. of auto. Pay. | จำนวนครั้งที่ใช้บริการจ่ายเงินผ่านเครื่องอัตโนมัติ |
| 19 | No. of C/A trans. | จำนวนครั้งที่ทำธุรกรรมในบัญชีกระแสรายวัน |
| 20 | Current account | จำนวนบัญชีกระแสรายวัน |
| 21 | Saving products | จำนวนผลิตภัณฑ์ออมทรัพย์ |
| 22 | Loan products | จำนวนผลิตภัณฑ์สินเชื่อ |
| 23 | Investment products | จำนวนผลิตภัณฑ์การลงทุน |
| 24 | Credit card | การมีผลิตภัณฑ์บัตรเครดิตกับธนาคาร |
| 25 | Average value of transactions | มูลค่าเฉลี่ยของการทำธุรกรรมแต่ละครั้ง |
| 26 | Average value of credit transactions | มูลค่าเฉลี่ยของการใช้สินเชื่อแต่ละครั้ง |
| 27 | Average value of debit transactions | มูลค่าเฉลี่ยของการฝากเงินแต่ละครั้ง |

แต่เนื่องจากฐานข้อมูลของธนาคารอยู่กระจัดกระจายหลายหน่วยงาน ประกอบกับข้อมูลที่ต้องการบางข้อมูลมีการจัดเก็บไม่ครบถ้วนสมบูรณ์ หรือมิได้จัดเก็บ ดังนั้นจากการสุ่มข้อมูลลูกค้าจำนวนทั้งสิ้นจำนวน 1000 ชุดข้อมูล จึงทำให้ได้ Variable เบื้องต้นทั้งสิ้นจำนวน 14 variables ดังต่อไปนี้

**ตารางที่ 3.2** แสดง Variable และคำจำกัดความ

| ลำดับที่ | Variable Name | Description |
|---|---|---|
| 1 | Age | อายุ |
| 2 | Sex | เพศ |
| 3 | Income | รายได้ |
| 4 | Marital Status | สถานภาพสมรส |
| 5 | Has Children | จำนวนบุตร |
| 6 | Web bank usage | ปริมาณการใช้ website ของธนาคาร |
| 7 | ATM usage | ปริมาณการใช้ ATM |
| 8 | Call center usage | ปริมาณการใช้บริการ Call center |
| 9 | Teller usage | ปริมาณการใช้บริการหน้าเคาท์เตอร์ |
| 10 | No. of auto. Pay. | จำนวนครั้งที่ใช้บริการจ่ายเงินผ่านเครื่องอัตโนมัติ |
| 11 | No. of C/A trans. | จำนวนครั้งที่ทำธุรกรรมในบัญชีกระแสรายวัน |
| 12 | Current account | จำนวนบัญชีกระแสรายวัน |
| 13 | Saving products | จำนวนผลิตภัณฑ์ออมทรัพย์ |
| 14 | Loan products | จำนวนผลิตภัณฑ์สินเชื่อ |

### 3.1.2 การกรองข้อมูล (Data Preprocessing )

เป็นขั้นตอนของการทำข้อมูลให้มีคุณภาพดี โดยในที่นี้เนื่องจากข้อมูลที่ได้มาเป็นข้อมูลที่ได้จากการสุ่มตัวอย่างโดยมีรายละเอียดของชุดข้อมูลแตกต่างกันไป มีบางชุดข้อมูลที่ข้อมูลบางส่วนขาดหายไปอันอาจเนื่องมาจากการกรอกข้อมูลไม่ครบของลูกค้าในเบื้องต้น, เกิดความผิดพลาดในระบบฐานข้อมูลของธนาคารเอง หรือความผิดพลาดระหว่างการดึงข้อมูลมาใช้ จึงได้มีการแก้ไขข้อมูลที่เป็นประเภท ข้อมูลที่หายไป (missing value), ข้อมูลล้าสมัย, ข้อมูลที่ format ไม่สอดคล้องกัน เรียบร้อยแล้ว ในที่นี้ได้มีการตัดตัวแปรลำดับที่ 8 Call center usage ออกเนื่องจาก

บริการนี้เป็นการให้บริการใหม่ของธนาคาร ซึ่งเปิดใช้บริการไม่นานนักจึงทำให้ไม่มีข้อมูลในฐานข้อมูล การนำข้อมูลไปใช้จึงได้มีการเลือกเฉพาะตัวแปรซึ่งคาดคิดว่าจะมีผลต่อการสรุปผลตามวัตถุประสงค์ที่ได้ตั้งไว้มาดำเนินการทำ mining เท่านั้น

**3.1.3 การแปลงข้อมูล (Data Transformation)**

เป็นการสร้างข้อมูลชุดใหม่ที่มาจากข้อมูลชุดเดิม โดยในที่นี้ได้มีการแปลงข้อมูลของตัวแปรบางตัวจากตัวแปรทั้งหมด 13 ตัว เพื่อให้สอดคล้องกับ Data Mining model ที่ใช้ และ ให้ผลลัพธ์ของการทำ mining มีประสิทธิภาพมากขึ้น ดังมีรายละเอียดนี้

1. อายุ (Age) : เป็นข้อมูลตัวเลขแสดงอายุของลูกค้าที่มาใช้บริการซึ่งในที่นี้ได้ทำการแบ่งช่วงของข้อมูลดังนี้

- ข้อมูลที่มากกว่า 0 จนถึง 30 แทนค่าด้วย 1
- ข้อมูลที่มากกว่า 30 จนถึง 40 แทนค่าด้วย 2
- ข้อมูลที่มากกว่า 40 จนถึง 50 แทนค่าด้วย 3
- ข้อมูลที่มากกว่า 50 จนถึง 60 แทนค่าด้วย 4
- ข้อมูลที่มากกว่า 60 ขึ้นไป 5

2. เพศ (Sex) เป็นรายละเอียดแสดงเพศของลูกค้า แบ่งได้เป็น 2 รูปแบบ

M = เพศชาย

F = เพศหญิง

ในที่นี้ได้ใช้ ค่า 0 แทนข้อมูลที่เป็น เพศชาย และ 1 แทนข้อมูลที่เป็น เพศหญิง

3. รายได้ (Income) เป็นข้อมูลตัวเลขแบบต่อเนื่องแสดงรายได้ของลูกค้าที่มาใช้บริการ ซึ่งในที่นี้ได้ทำการแบ่งช่วงของข้อมูลดังนี้

- ข้อมูลที่มากกว่า 0 จนถึง 5,000 บาท แทนค่าด้วย 01
- ข้อมูลที่มากกว่า 5,000 บาทจนถึง 10,000 บาทแทนค่าด้วย 02
- ข้อมูลที่มากกว่า 10,000 บาท จนถึง 20,000 บาท แทนค่าด้วย 03
- ข้อมูลที่มากกว่า 20,000 บาท จนถึง 30,000 บาท แทนค่าด้วย 04
- ข้อมูลที่มากกว่า 30,000 บาท จนถึง 45,000 บาท แทนค่าด้วย 05
- ข้อมูลที่มากกว่า 45,000 บาท จนถึง 70,000 บาท แทนค่าด้วย 06
- ข้อมูลที่มากกว่า 70,000 บาท จนถึง 100,000 บาท แทนค่าด้วย 07

4. สถานภาพสมรส (Marital Status) ข้อมูล แบ่งได้เป็น 4 ลักษณะ คือ

M = สมรส

S = โสด

D = หย่า

U =ไม่ทราบสถานะ

ในที่นี้ได้ใช้ ค่า 1 แทนข้อมูลที่เป็น สมรส , 2 แทนข้อมูลที่เป็น โสด, 3 แทนข้อมูลที่เป็น หย่า และ 0 แทนข้อมูลที่ไม่ทราบสถานะ

5. จำนวนบุตร (Has Children) เป็นข้อมูลตัวเลขแบบต่อเนื่องแสดงจำนวนบุตรของลูกค้าซึ่งในที่นี้ใช้รูปแบบของข้อมูลเดิมในการ run

6. ปริมาณการใช้ website ของธนาคาร (Web bank usage) เป็นข้อมูลตัวเลขแบบต่อเนื่องแสดงจำนวนปริมาณการใช้ website ของธนาคาร ซึ่งในที่นี้ใช้รูปแบบของข้อมูลเดิมในการ run

7. ปริมาณการใช้ ATM (ATM usage) เป็นข้อมูลตัวเลขแบบต่อเนื่องแสดงจำนวนปริมาณการใช้ ATM ของลูกค้าซึ่งในที่นี้ใช้รูปแบบของข้อมูลเดิมในการ run

8. ปริมาณการใช้บริการหน้าเคาท์เตอร์ (Teller usage) เป็นข้อมูลตัวเลขแบบต่อเนื่องแสดงจำนวนครั้งที่ลูกค้าคนดังกล่าวมาใช้บริการหน้าเคาท์เตอร์ ซึ่งในที่นี้ได้ใช้รูปแบบของข้อมูลเดิมในการ run

9. จำนวนครั้งที่ใช้บริการจ่ายเงินผ่านเครื่องอัตโนมัติ (No. of auto. Pay.) เป็นข้อมูลตัวเลขแบบต่อเนื่องแสดงจำนวนครั้งที่ใช้บริการจ่ายเงินผ่านเครื่องอัตโนมัติ ซึ่งในที่นี้ใช้รูปแบบของข้อมูลเดิมในการ run

10. จำนวนครั้งที่ทำธุรกรรมในบัญชีกระแสรายวัน (No. of C/A trans.) เป็นข้อมูลตัวเลขแบบต่อเนื่องแสดงจำนวนครั้งที่ลูกค้าทำธุรกรรมในบัญชีกระแสรายวัน ซึ่งในที่นี้ใช้รูปแบบของข้อมูลเดิมในการ run

11. จำนวนบัญชีกระแสรายวัน (Current account) เป็นข้อมูลตัวเลขแบบต่อเนื่องแสดงจำนวนบัญชีกระแสรายวัน ซึ่งลูกค้ามีกับธนาคารซึ่งในที่นี้ใช้รูปแบบของข้อมูลเดิมในการ run

12. จำนวนผลิตภัณฑ์ออมทรัพย์ (Saving products) เป็นข้อมูลตัวเลขแบบต่อเนื่องแสดงจำนวนผลิตภัณฑ์ออมทรัพย์ ซึ่งลูกค้ามีกับธนาคารในที่นี้ใช้รูปแบบของข้อมูลเดิมในการ run

13. จำนวนผลิตภัณฑ์สินเชื่อ (Loan products) เป็นข้อมูลตัวเลขแบบต่อเนื่องแสดงจำนวนผลิตภัณฑ์สินเชื่อซึ่งลูกค้าใช้บริการกับธนาคาร ซึ่งในที่นี้ใช้รูปแบบของข้อมูลเดิมในการ run

จากนั้นจึงก็นำข้อมูลมาดัดแปลงในขั้นต่อไป โดยการทำ Normalization โดยใช้ min-max normalization เพื่อปรับอัตราส่วนตัวเลขให้ข้อมูลอยู่ในช่วงระหว่าง 0-1 เพื่อให้เหมาะสมกับ Algorithm ใน Neural Network โดยแทนค่าที่เป็นตัวเลขตามสมการดังต่อไปนี้

$$V' = V - \min_A / \text{max-min}_A$$

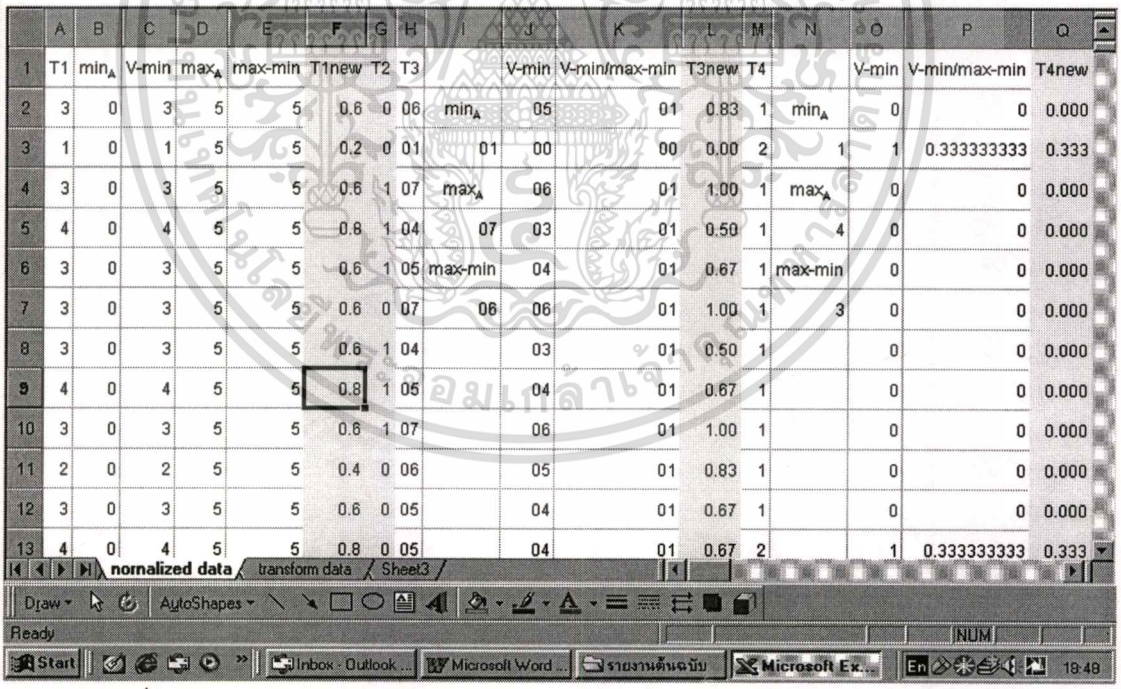
โดยที่ $V'$ = ค่าที่แปลงได้

$V$ = ค่าที่ต้องการจะแปลง (ตัวแปรดั้งเดิม)

$\min_A$ = ค่าน้อยที่สุดของตัวแปรดั้งเดิม

$\max_A$ = ค่ามากที่สุดของตัวแปรดั้งเดิม

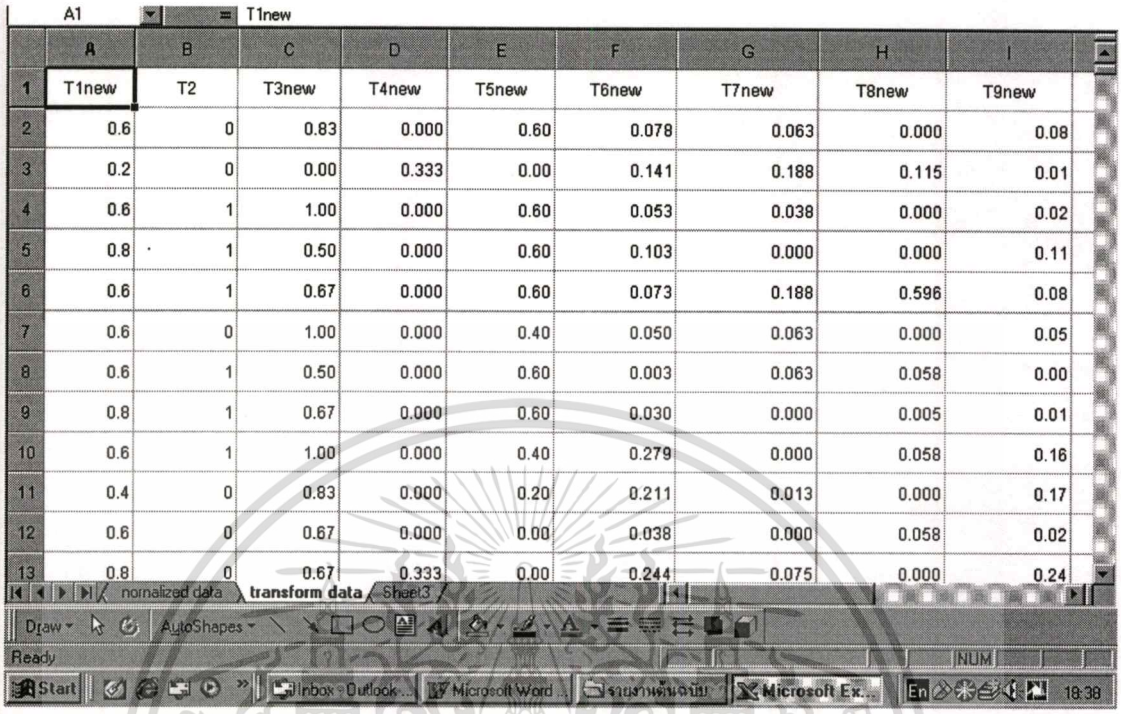
ทั้งนี้ได้ทำกับตัวแปรทั้งหมด 13 ตัวยกเว้นตัวแปรที่ 2 เนื่องจากมีค่าอยู่ที่ 0 และ 1 อยู่แล้ว ซึ่งค่า Normalization แสดงได้ดังรูป



| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | T1 | $\min_A$ | V-min | $\max_A$ | max-min | T1new | T2 | T3 | | V-min | V-min/max-min | T3new | T4 | | V-min | V-min/max-min | T4new |
| 2 | 3 | 0 | 3 | 5 | 5 | 0.6 | 0 | 06 | $\min_A$ | 05 | 01 | 0.83 | 1 | $\min_A$ | 0 | 0 | 0.000 |
| 3 | 1 | 0 | 1 | 5 | 5 | 0.2 | 0 | 01 | 01 | 00 | 00 | 0.00 | 2 | 1 | 1 | 0.333333333 | 0.333 |
| 4 | 3 | 0 | 3 | 5 | 5 | 0.6 | 1 | 07 | $\max_A$ | 06 | 01 | 1.00 | 1 | $\max_A$ | 0 | 0 | 0.000 |
| 5 | 4 | 0 | 4 | 5 | 5 | 0.8 | 1 | 04 | 07 | 03 | 01 | 0.50 | 1 | 4 | 0 | 0 | 0.000 |
| 6 | 3 | 0 | 3 | 5 | 5 | 0.6 | 1 | 05 | max-min | 04 | 01 | 0.67 | 1 | max-min | 0 | 0 | 0.000 |
| 7 | 3 | 0 | 3 | 5 | 5 | 0.6 | 0 | 07 | 06 | 06 | 01 | 1.00 | 1 | 3 | 0 | 0 | 0.000 |
| 8 | 3 | 0 | 3 | 5 | 5 | 0.6 | 1 | 04 | | 03 | 01 | 0.50 | 1 | | 0 | 0 | 0.000 |
| 9 | 4 | 0 | 4 | 5 | 5 | 0.8 | 1 | 05 | | 04 | 01 | 0.67 | 1 | | 0 | 0 | 0.000 |
| 10 | 3 | 0 | 3 | 5 | 5 | 0.6 | 1 | 07 | | 06 | 01 | 1.00 | 1 | | 0 | 0 | 0.000 |
| 11 | 2 | 0 | 2 | 5 | 5 | 0.4 | 0 | 06 | | 05 | 01 | 0.83 | 1 | | 0 | 0 | 0.000 |
| 12 | 3 | 0 | 3 | 5 | 5 | 0.6 | 0 | 05 | | 04 | 01 | 0.67 | 1 | | 0 | 0 | 0.000 |
| 13 | 4 | 0 | 4 | 5 | 5 | 0.8 | 0 | 05 | | 04 | 01 | 0.67 | 2 | | 1 | 0.333333333 | 0.333 |



**รูปที่ 3.1** แสดงการทำ Normalization กับตัวแปรที่ 1-4



A1 = T1new

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | T1new | T2 | T3new | T4new | T5new | T6new | T7new | T8new | T9new |
| 2 | 0.6 | 0 | 0.83 | 0.000 | 0.60 | 0.078 | 0.063 | 0.000 | 0.08 |
| 3 | 0.2 | 0 | 0.00 | 0.333 | 0.00 | 0.141 | 0.188 | 0.115 | 0.01 |
| 4 | 0.6 | 1 | 1.00 | 0.000 | 0.60 | 0.053 | 0.038 | 0.000 | 0.02 |
| 5 | 0.8 | 1 | 0.50 | 0.000 | 0.60 | 0.103 | 0.000 | 0.000 | 0.11 |
| 6 | 0.6 | 1 | 0.67 | 0.000 | 0.60 | 0.073 | 0.188 | 0.596 | 0.08 |
| 7 | 0.6 | 0 | 1.00 | 0.000 | 0.40 | 0.050 | 0.063 | 0.000 | 0.05 |
| 8 | 0.6 | 1 | 0.50 | 0.000 | 0.60 | 0.003 | 0.063 | 0.058 | 0.00 |
| 9 | 0.8 | 1 | 0.67 | 0.000 | 0.60 | 0.030 | 0.000 | 0.005 | 0.01 |
| 10 | 0.6 | 1 | 1.00 | 0.000 | 0.40 | 0.279 | 0.000 | 0.058 | 0.16 |
| 11 | 0.4 | 0 | 0.83 | 0.000 | 0.20 | 0.211 | 0.013 | 0.000 | 0.17 |
| 12 | 0.6 | 0 | 0.67 | 0.000 | 0.00 | 0.038 | 0.000 | 0.058 | 0.02 |
| 13 | 0.8 | 0 | 0.67 | 0.333 | 0.00 | 0.244 | 0.075 | 0.000 | 0.24 |



**รูปที่ 3.2** แสดงข้อมูลที่ทำการแปลงเรียบร้อยแล้ว

เมื่อได้ข้อมูลซึ่งมีการเตรียมเรียบร้อยแล้วจึงเข้าสู่กระบวนการทำ Data Mining ต่อไป

### 3.2 การทำ Data Mining

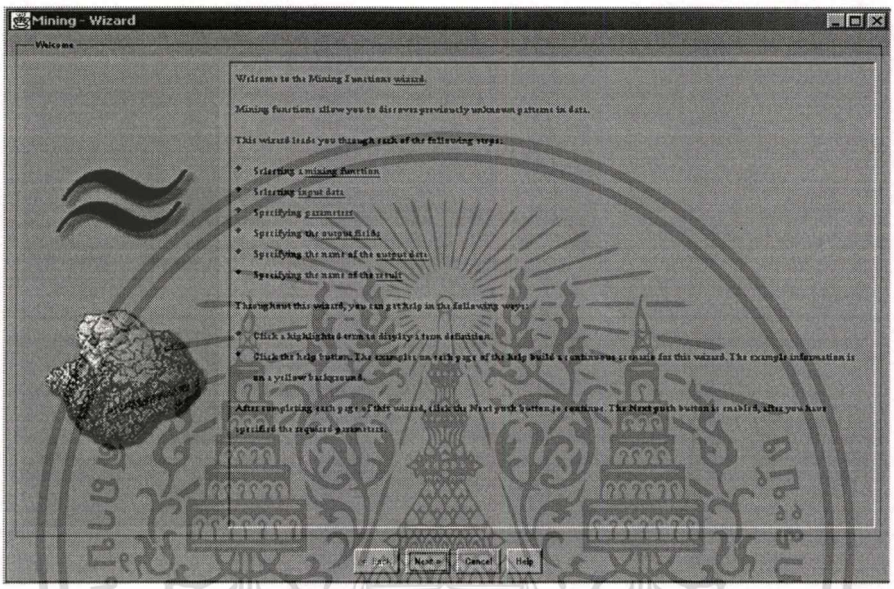
การทำ Data Mining ในที่นี้ได้ทำการศึกษา Algorithm แบบ Neural Network คือ Kohonen's Self-Organizing Maps (SOM) โดยใช้โปรแกรม Intelligent Miner เพื่อวิเคราะห์หาข้อมูลที่เป็นประโยชน์ตามวัตถุประสงค์ที่เราได้ตั้งไว้ในบทที่ 1 คือ การแบ่งส่วนฐานข้อมูล (Database Segmentation) โดยจะทำการแบ่งหรือจัดกลุ่มของข้อมูลที่มีลักษณะคล้ายกัน หรือมีคุณสมบัติใกล้เคียงกันในหลายๆ ด้าน ให้เป็นข้อมูลกลุ่มเดียวกัน ซึ่งแต่ละกลุ่มจะถูกเรียกว่า Segments หรือ Clusters การแบ่งกลุ่มข้อมูลนี้เราจะไม่สามารถกำหนดได้ว่าข้อมูลควรจะอยู่กลุ่มใด แต่จะเป็นการกำหนดกลุ่มของข้อมูลจากธรรมชาติของข้อมูลเอง ไม่ได้ใช้ความรู้สึกหรือประสบการณ์ในการตัดสินใจแบ่งกลุ่มข้อมูล และข้อมูลจะถูกจัดการโดย Algorithm ที่เหมาะสม ทั้งนี้ผลจากการทำ Mining จะกล่าวไว้ในบทถัดไป

จากข้อมูล 1000 ชุดข้อมูลซึ่งแต่ละชุดข้อมูลมีตัวแปร 13 ตัวแปรนั้น ได้นำมาทำการ mining – Clustering โดยใช้ โปรแกรม Intelligent Miner โดยโปรแกรมนี้จะสามารถแบ่งข้อมูลออก

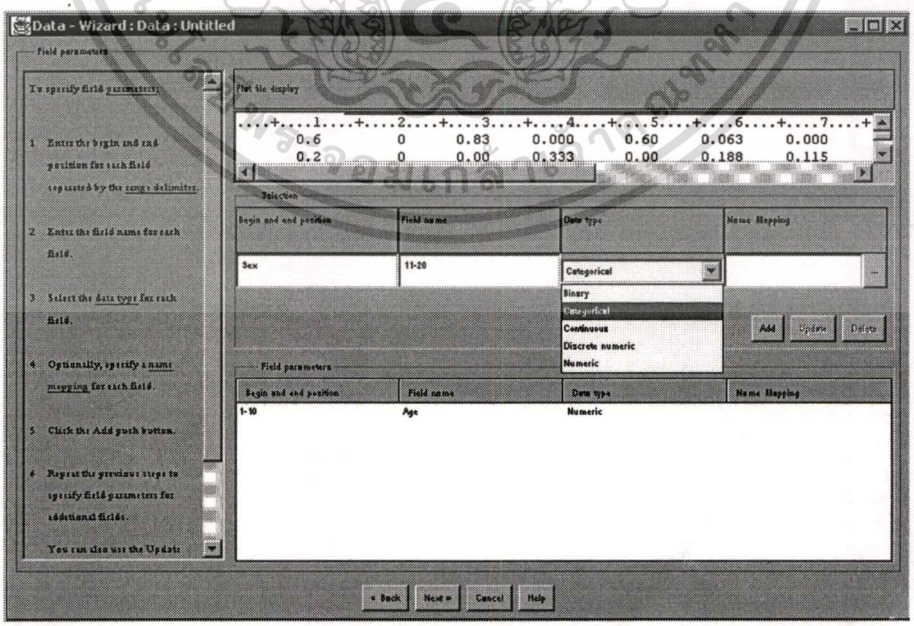
เป็นจำนวน cluster ต่างๆ ซึ่งเราสามารถนำมาวิเคราะห์ประมวลผลเลือกจำนวนกลุ่มกลุ่มที่คิดว่าดีที่สุดได้ รวมถึงบอกลักษณะของแต่ละกลุ่มได้

### 3.2.1 Database Segmentation โดยใช้โปรแกรม Intelligent Miner

การแบ่งกลุ่มข้อมูลโดยวิธี Clustering-Neural โดยใช้ IBM Intelligent Miner เริ่มแรกต้องทำการ Create data ก่อน โดยโปรแกรมจะมีเครื่องมือที่เป็น wizard ในการช่วยทำดังแสดงได้ดังรูป



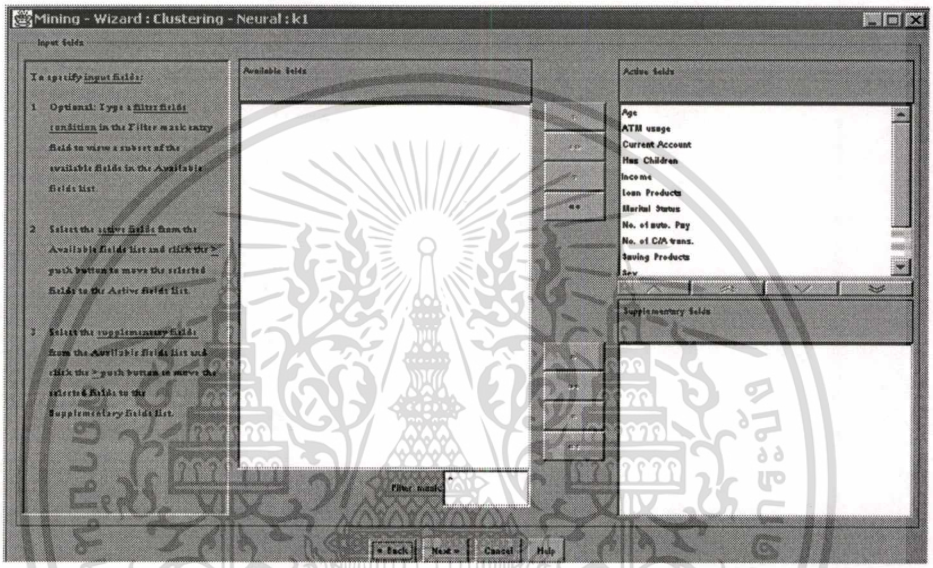

**รูปที่ 3.3** แสดงขั้นตอนการเริ่มต้น Create data



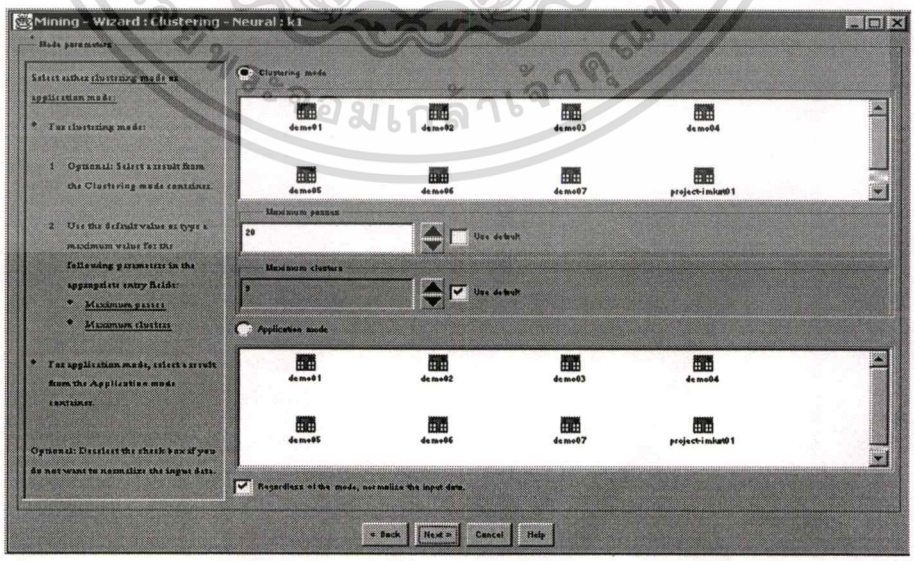

**รูปที่ 3.4** แสดงการกำหนดค่าต่างๆของข้อมูล

ทั้งนี้เราจะต้องนำข้อมูลที่แปลงไว้เรียบร้อยแล้วนำมาเซฟให้อยู่ในรูปแบบของนามสกุล .prn ก่อน หลังจากนั้นจึงจะสามารถนำข้อมูลเข้าไปในระบบได้ โดยจะต้องกำหนดค่าความกว้างของตำแหน่ง ชื่อฟิลด์ และชนิดของข้อมูล

ขั้นตอนต่อไปก็จะทำการ create mining โดยเลือกรูปแบบวิธีการ mining โดยการทำ clustering –neural แล้วเลือกตัวแปรที่จะนำมาเป็นปัจจัยในการแบ่งกลุ่ม (ทั้งนี้สามารถเลือกเพียงบางส่วนหรือทั้งหมดเลยก็ได้)




**รูปที่ 3.5** แสดงการเลือกตัวแปรที่จะนำมาทำ Clustering




**รูปที่ 3.6** แสดงการกำหนดค่าต่างๆ ในการทำ clustering–neural

ในที่นี้ได้เลือกศึกษาโดยใช้ตัวแปรทั้งหมดก่อน และได้ศึกษาเพิ่มเติมโดยเลือกเฉพาะตัวแปรที่เป็น demographic คือ อายุ, เพศ, รายได้, สถานภาพสมรส และจำนวนบุตร มาศึกษา

หลังจากนั้นจะต้องทำการกำหนดค่าต่างๆ สำหรับการทำ clustering–neural โดยสามารถกำหนดค่าดังต่อไปนี้

1. maximum passes : จำนวนครั้งที่จะให้โปรแกรม run ข้อมูล โดยในที่นี้โปรแกรมจะกำหนดค่า default ให้เท่ากับ 5 ครั้ง แต่สามารถเลือกให้มี maximum passes คือ 20 ครั้ง (ค่าที่สามารถเลือกได้สูงสุด) นอกจากนี้พบว่าสามารถที่จะเลือกที่จะกำหนดค่า maximum passes เองได้โดยการคีย์ค่าเข้าไปซึ่งที่นี้พบว่าจะกำหนดเป็นค่าเท่าใดก็ได้ ซึ่งจากการศึกษานี้ได้ใช้จำนวน 20 ครั้งเป็นหลัก และศึกษาโดยเพิ่มเป็น 100, 200 และ 1000 ครั้ง ตามลำดับ ซึ่งพบว่าเมื่อจำนวนครั้งมากขึ้นค่า deviation จะลดลงตามลำดับ
2. จำนวน Cluster : หมายถึงจำนวน กลุ่มที่ต้องการแบ่ง ในที่นี้จะสามารถเลือกได้ตั้งแต่จำนวน 1,4, 9, 16, 25, 36, 49, 64, 81, และ 100 กลุ่มตามลำดับ ที่เป็นค่านี้เนื่องจาก algorithm จะแบ่งกลุ่มโดยจัดในรูปแบบเมตริกซ์

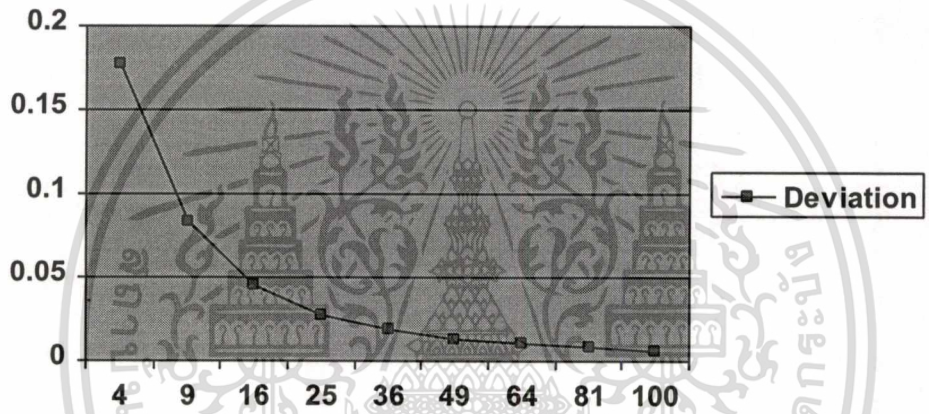
จากนั้นจึงโดยนำค่า Deviation ซึ่งเป็นผลลัพธ์ที่ได้จากการแบ่งกลุ่มแต่ละครั้งมาเปรียบเทียบกันซึ่งพบว่าเมื่อทำการแบ่งกลุ่มเป็นจำนวนมากขึ้นความแตกต่างของ Deviation จะค่อยๆ ลดลงตามลำดับดังแสดงในตารางที่ 3.3

**ตารางที่ 3.3** แสดงค่า Deviation ของแต่ละ Cluster เมื่อใช้ตัวแปรทั้ง 13 ตัว

| Maximum Number of Clusters | Number of Clusters | Deviation | Percentage Diff. |
|---|---|---|---|
| 4 | 4 | 0.1780 | 17.80% |
| 9 | 9 | 0.0836 | -53.03% |
| 16 | 15 | 0.0451 | -46.05% |
| 25 | 24 | 0.0274 | -39.25% |
| 36 | 33 | 0.0195 | -28.83% |
| 49 | 45 | 0.0137 | -29.74% |
| 64 | 57 | 0.0106 | -22.63% |

| Maximum Number of Clusters | Number of Clusters | Deviation | Percentage Diff. |
|---|---|---|---|
| 81 | 75 | 0.0082 | -22.64% |
| 100 | 90 | 0.0065 | -20.73% |

โดยเมื่อนำมาวาดเป็นรูปกราฟจะเห็นชัดเจนขึ้นว่าความชันของกราฟจะค่อยๆ ลดลงเมื่อจำนวน clusters มีมากขึ้น

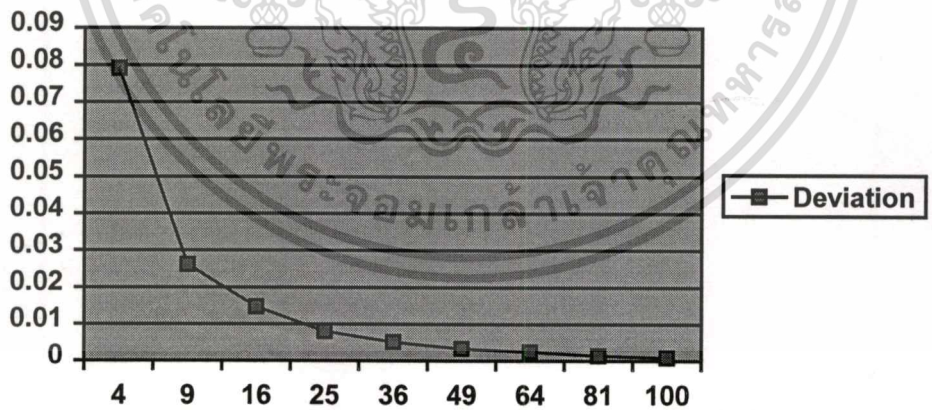



**รูปที่ 3.7** กราฟแสดงการเปรียบเทียบ deviation ของแต่ละครั้งของการแบ่งกลุ่ม

โดยเมื่อพิจารณาจากค่า deviation แล้วพบว่าจำนวนกลุ่มที่เหมาะสมที่สุด คือ ที่ Maximum Number of Clusters 25 กลุ่ม โดยแบ่งกลุ่มออกมาทั้งหมดได้ 24 กลุ่ม เนื่องจากเป็นจุดความชันจะเปลี่ยนแปลงน้อยลงหลังจากแบ่งมากกว่า 24 กลุ่ม แต่เนื่องจากเมื่อพิจารณาจากลักษณะและวัตถุประสงค์ของข้อมูลแล้วเห็นว่าการแบ่งข้อมูลออกเป็นกลุ่มมากกว่า 9 กลุ่มนั้น เป็นจำนวนที่มากเกินความจำเป็น, ยากที่จะทำการวิเคราะห์หาความสัมพันธ์ของข้อมูล และไม่เหมาะสำหรับการนำไปใช้ จึงได้ทดลองแยกส่วนข้อมูลโดยแบ่งเฉพาะตัวแปรที่เป็น demographic data เป็น attribute สำหรับการทำ mining เพื่อที่สามารถจะบอกถึงลักษณะของลูกค้าแต่ละกลุ่มและจำนวนกลุ่มที่เหมาะสมได้

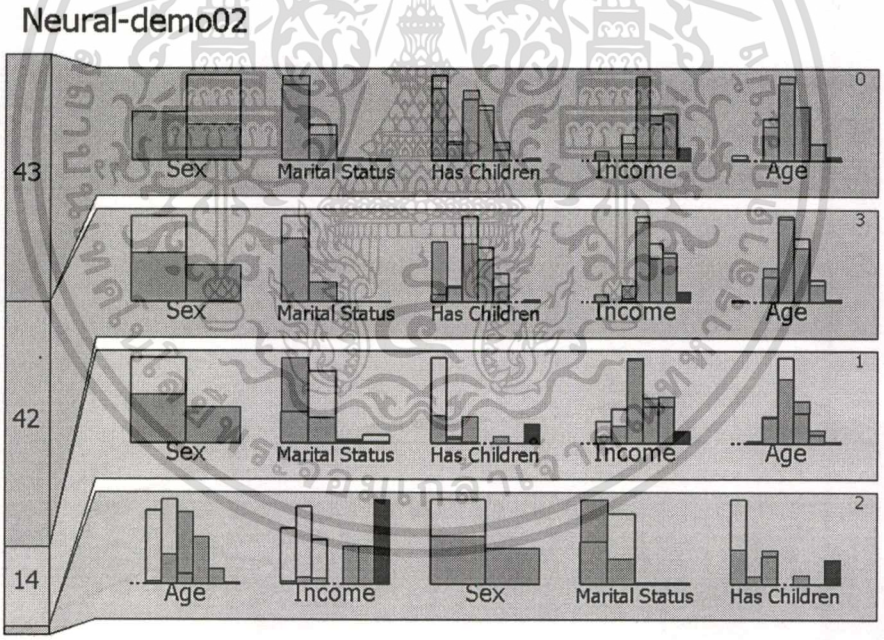
**ตารางที่ 3.4** แสดงค่า Deviation ของแต่ละ Cluster เมื่อใช้ตัวแปรเฉพาะที่เป็น demographic

| Maximum Number of Clusters | Number of Clusters | Deviation | Percentage Diff. |
|---|---|---|---|
| 4 | 4 | 0.0791 | 7.91% |
| 9 | 9 | 0.0261 | -67.00% |
| 16 | 15 | 0.0147 | -43.68% |
| 25 | 24 | 0.0080 | -45.58% |
| 36 | 35 | 0.0051 | -36.25% |
| 49 | 46 | 0.0033 | -35.29% |
| 64 | 60 | 0.0024 | -27.27% |
| 81 | 80 | 0.0014 | -41.67% |
| 100 | 90 | 0.0009 | -35.71% |



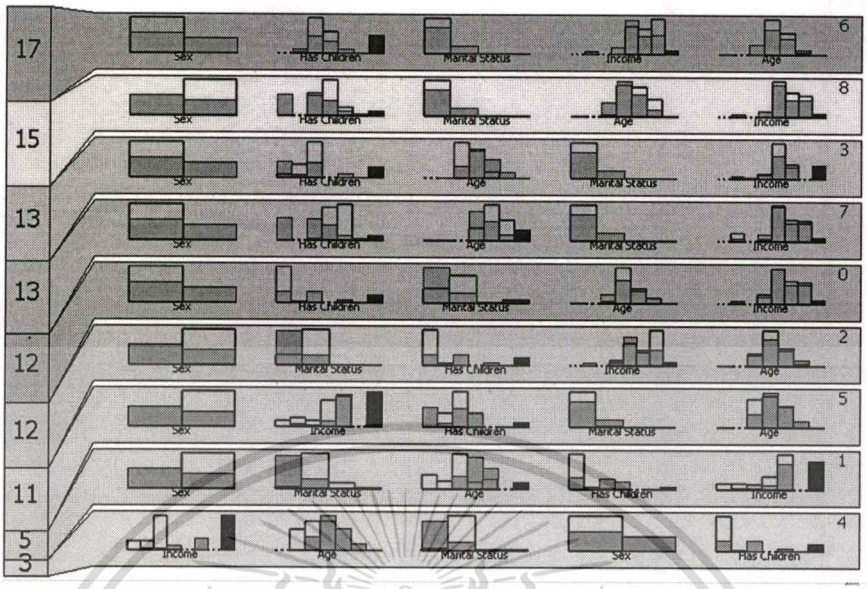

**รูปที่ 3.8** กราฟแสดงการเปรียบเทียบ deviation ของแต่ละครั้งของการแบ่งกลุ่มเมื่อใช้ตัวแปรเฉพาะที่เป็น demographic data

จากตารางที่ 3.4 และรูปที่ 3.8 จะเห็นได้ว่าจำนวนกลุ่มที่เหมาะสมคือ ที่ Maximum Number of Clusters 25 กลุ่ม โดยแบ่งกลุ่มออกมาทั้งหมดได้ 24 กลุ่ม เช่นเดียวกัน แต่เนื่องจากจำนวนกลุ่มที่เหมาะสมนั้นนอกจากจะดูจากค่าความเบียงเบนแล้วยังต้องพิจารณาถึงความเหมาะสมของธุรกิจและวัตถุประสงค์ของการทำ mining ด้วย จึงได้นำรายละเอียดของแต่ละกลุ่มของการแบ่งมาพิจารณาถึงตัวแปรที่ใช้ซึ่งพบว่า จากการแบ่งออกเป็น 4 Cluster ตามรูปที่ 3.9 ข้อมูลซึ่งถูกแบ่งออกเป็น 4 กลุ่ม โดยที่แถวบนแสดงกลุ่มที่มีจำนวนข้อมูลมากที่สุด คือ 43% ซึ่งกลุ่มนี้มีเลข id ของกลุ่มคือ 0 (แสดงทางมุมบนขวามือของแต่ละแถว) ปัจจัยที่มีอิทธิพลต่อการแบ่งกลุ่มมากที่สุดจะแสดงอยู่ในรูปทางซ้ายมือ ส่วนปัจจัยที่มีอิทธิพลน้อยที่สุดจะอยู่ทางขวามือ ซึ่งจากรูปจะเห็นว่าปัจจัยที่มีผลต่อการแบ่งกลุ่มมากที่สุดคือ "เพศ" ซึ่งจะมีอิทธิพลต่อกลุ่มส่วนใหญ่ยกเว้นกลุ่มสุดท้ายซึ่งมีจำนวนข้อมูล 1% ปัจจัยที่มีผลต่อการแบ่งกลุ่มมากที่สุดคือ "อายุ" ส่วนปัจจัยที่มีอิทธิพลน้อยที่สุดสำหรับกลุ่มแรก (id 0) คือ "อายุ" ซึ่งจะมีอิทธิพลน้อยต่อกลุ่มส่วนใหญ่



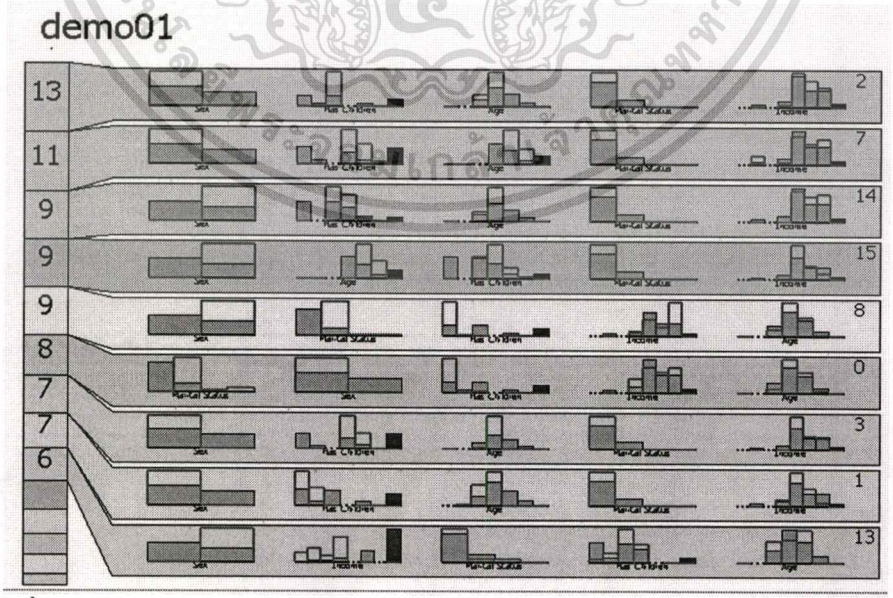

**รูปที่ 3.9** กราฟแสดงการแบ่ง Cluster ออกเป็น 4 กลุ่ม เมื่อใช้ตัวแปรเฉพาะที่เป็น demographic

และเมื่อพิจารณาเพิ่มเติมในรายละเอียดการแบ่งข้อมูลออกเป็น 9 กลุ่ม พบว่าปัจจัยที่มีอิทธิพลหรือบ่งบอกลักษณะของแต่ละกลุ่มอย่างเด่นชัดที่สุดคือ "เพศ" ซึ่งเป็นปัจจัยที่มีอิทธิพลต่อกลุ่มส่วนใหญ่ดังแสดงได้ในรูปที่ 3.10

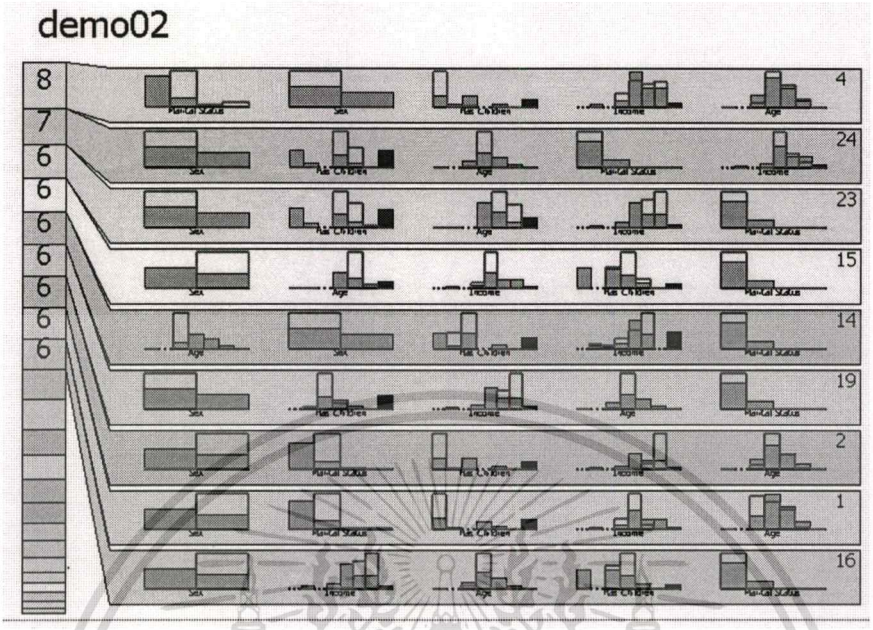


**รูปที่ 3.10** กราฟแสดงการแบ่ง Cluster ออกเป็น 9 กลุ่ม เมื่อใช้ตัวแปรเฉพาะที่เป็น demographic

ต่อจากนั้นจึงทำการพิจารณารายละเอียดของ 16 กลุ่ม และ 25 กลุ่ม ตามลำดับเพื่อดูปัจจัยสำคัญในการแบ่งกลุ่ม ก็พบว่าปัจจัยที่เป็น เพศ เป็นตัวสำคัญในการแบ่งกลุ่มเช่นเดียวกัน ดังนั้นเมื่อนำลักษณะลูกค้ามารวมกลุ่มกันพบว่าจะสามารถแบ่งลักษณะใหญ่ๆ ของลูกค้าได้ โดยมีเพศเป็นตัวหลักในการแบ่ง




**รูปที่ 3.11** กราฟแสดงการแบ่ง Cluster ออกเป็น 16 กลุ่ม เมื่อใช้ตัวแปรเฉพาะที่เป็น demographic



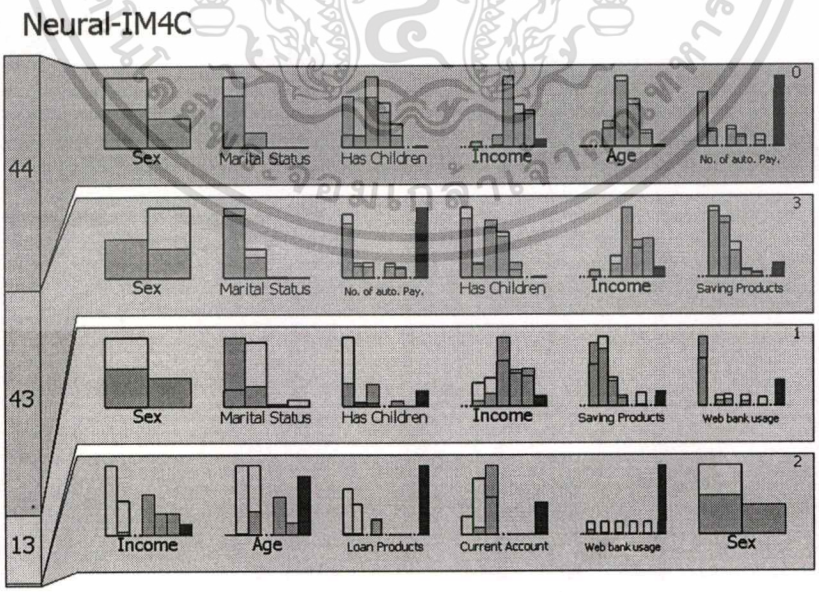

**รูปที่ 3.12** กราฟแสดงการแบ่ง Cluster ออกเป็น 25 กลุ่ม เมื่อใช้ตัวแปรเฉพาะที่เป็น demographic

ทั้งนี้เมื่อได้ปัจจัยหลักในการแบ่งแล้วจึงนำมาพิจารณาร่วมกับการแบ่งกลุ่มโดยใช้ปัจจัยทั้งหมดอีกครั้งหนึ่ง เพื่อที่จะหาผลสรุปที่ได้จากการศึกษาต่อไป

# บทที่ 4
## ผลการศึกษา

### 4.1 ผลที่ได้จาก Kohonen's Self-Organizing Maps โดยใช้โปรแกรม Intelligent Miner

เมื่อทำการศึกษากลุ่มข้อมูลตัวอย่างของลูกค้าธนาคารพาณิชย์แห่งหนึ่งโดยนำมาทำการประมวลผลแบบ Neural Network โดยใช้โปรแกรม IBM Intelligent Miner ในการวิเคราะห์ผล พบว่าเมื่อพิจารณาจากรายละเอียดของการแบ่งกลุ่มหลายรูปแบบตามวิธีดำเนินการศึกษาในบทที่ 3 นั้น สรุปได้ว่าเมื่อนำลักษณะลูกค้ามารวมกลุ่มกันจะสามารถแบ่งลักษณะใหญ่ๆ ของลูกค้าได้ตาม เพศ และสถานภาพสมรส ดังนั้นในการเลือกจำนวนกลุ่มเพื่อความเหมาะสมแก่การวางกลยุทธ์สำหรับธุรกิจธนาคารพาณิชย์ จึงสรุปได้ว่าจำนวนกลุ่มที่เหมาะสมในการทำ Clustering เท่ากับ 4 กลุ่ม เนื่องจากเมื่อทำการพิจารณาจากรายละเอียดของการแบ่งกลุ่มออกเป็น 4 กลุ่ม, 9 กลุ่ม, 16 กลุ่ม และ 25 กลุ่ม ตามลำดับก็พบว่า ปัจจัยหลักที่มีผลในการแบ่งกลุ่ม และลักษณะเด่นของแต่ละกลุ่มนั้นมีลักษณะใกล้เคียงกัน โดยสามารถสรุปผลที่ได้ดังแสดงเป็นกราฟและรายละเอียดโดยรวม พร้อมทั้งแสดงผลเป็นกราฟและรายละเอียดในแต่ละ cluster ได้ดังนี้




**รูปที่ 4.1** แสดง การแบ่งกลุ่มโดยใช้ 13 ปัจจัยโดยแบ่งออกเป็น 4 clusters

จากรูปที่ 4.1 จะเห็นได้ว่าข้อมูลถูกแบ่งออกเป็น 4 กลุ่ม โดยที่แถวบนแสดงกลุ่มที่มีจำนวนข้อมูลมากที่สุด คือ 44% ซึ่งกลุ่มนี้มีเลข id ของกลุ่มคือ 0 ปัจจัยที่มีอิทธิพลต่อการแบ่งกลุ่มมากที่สุด "เพศ" ซึ่งจะมีอิทธิพลต่อกลุ่มส่วนใหญ่ รองลงมาคือ "สถานะภาพสมรส" และ จำนวนบุตร ตามลำดับส่วนปัจจัยที่มีอิทธิพลน้อยที่สุดสำหรับกลุ่มนี้ คือ "จำนวนครั้งที่ทำธุรกรรมในบัญชีกระแสรายวัน" กลุ่มที่ 2 ซึ่งเมื่อแบ่งออกมาแล้วมีจำนวนข้อมูลใกล้เคียงกันคือกลุ่ม id 3 คือมีข้อมูล 43% โดยปัจจัยที่มีอิทธิพลต่อการแบ่งกลุ่มมากที่สุด "เพศ" รองลงมาคือ "สถานะภาพสมรส" เช่นเดียวกันกับกลุ่มแรก ส่วนปัจจัยที่มีอิทธิพลน้อยที่สุดสำหรับกลุ่มนี้ คือ "จำนวนบัญชีกระแสรายวัน"

สำหรับผลในรูปรายละเอียดนั้น โปรแกรมนี้จะแสดงออกมาในรูปของ Result Created ซึ่งจะบอกข้อมูลดังนี้

User Specified Parameters

Maximum Number of Passes : 20

Maximum Number of Cluster : 4

Mining Run Outputs

Number of Passes Performed : 20

Number of Clusters : 4

Deviation : 0.177975

ข้อมูลข้างต้นจะบอกรายละเอียดของจำนวนครั้งที่ทำการ run ข้อมูลซึ่งเท่ากับ 20 ครั้ง จำนวน cluster ที่แบ่ง คือ 4 cluster และค่าความเบี่ยงเบนของข้อมูลซึ่งเท่ากับ 0.177975

ตารางที่ 4.1 แสดง Cluster Characteristics 4 Clusters

| Id | Cluster Size | |
|---|---|---|
| | Absolute | Relative (%) |
| 0 | 440 | 44.00 |
| 1 | 129 | 12.90 |
| 2 | 6 | 0.06 |
| 3 | 425 | 42.50 |

ตาราง Cluster Characteristics จะแสดงปริมาณข้อมูลในแต่ละ cluster ในที่นี้ข้อมูลจะถูกแบ่งออกเป็น 4 cluster โดยที่

- cluster ที่ 0 จะมีจำนวนข้อมูลเท่ากับ 440 ข้อมูล ซึ่งคิดเป็น 44.00%ของข้อมูลทั้งหมด
- cluster ที่ 1 จะมีจำนวนข้อมูลเท่ากับ 129 ข้อมูล ซึ่งคิดเป็น 12.90%ของข้อมูลทั้งหมด
- cluster ที่ 2 จะมีจำนวนข้อมูลเท่ากับ 6 ข้อมูล ซึ่งคิดเป็น 0.60% ของข้อมูลทั้งหมด
- cluster ที่ 3 จะมีจำนวนข้อมูลเท่ากับ 425 ข้อมูล ซึ่งคิดเป็น 42.50% ของข้อมูลทั้งหมด

ซึ่งจะเห็นได้ว่า cluster 2 เป็น cluster ที่มีจำนวนข้อมูลมากที่สุด คือ 440 ข้อมูล

**ตารางที่ 4.2 แสดง** Reference Field Characteristics (For All Field Types)

| Id | Name | Type | Modal Value | Modal Frequency (%) | No.of Possible Values/Buckets |
|---|---|---|---|---|---|
| 1 | Age | DN | 0.6 | 43.00 | 6 |
| 2 | Current account | DN | 0.22 | 61.30 | 7 |
| 3 | Has Children | DN | 0 | 31.40 | 6 |
| 4 | Income | DN | 0.67 | 39.60 | 7 |
| 5 | Loan products | DN | 0.22 | 45.90 | 9 |
| 6 | Marital Status | DN | 0 | 74.70 | 4 |
| 7 | No. of auto. Pay. | DN | 0 | 31.50 | 27 |
| 8 | No. of C/A trans. | DN | 0 | 27.00 | 33 |
| 9 | Saving products | DN | 0.08 | 36.50 | 11 |
| 10 | Sex | DN | 0 | 57.50 | 2 |
| 11 | Teller usage | DN | 0.04 | 13.70 | 35 |
| 12 | Web bank usage | DN | 0 | 59.70 | 27 |
| 13 | ATM usage | CO | 0.0125 | 53.90 | 11 |

ตาราง Reference Field Characteristics (For All Field Types) แสดงรายละเอียดทั่วไปของแต่ละ ปัจจัยโดยจะบอกถึงชนิด, Modal Value, Modal Frequency และ No.of Possible Values/Buckets โดยจะสังเกตได้ว่าข้อมูลนี้จะปรากฏอยู่ในผลสรุปของทุกๆ การแบ่งกลุ่มในข้อมูล

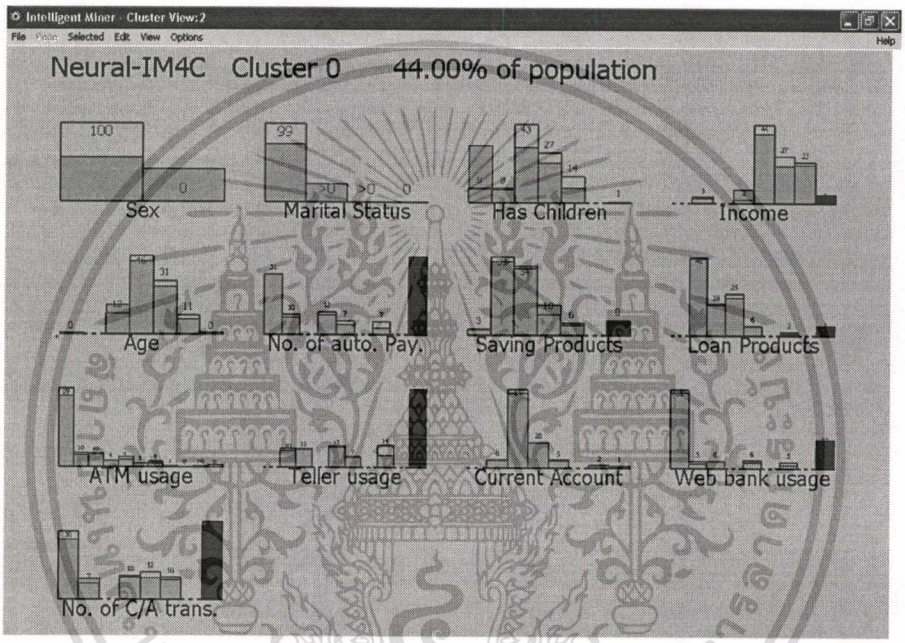
ชุดเดียวกันไม่ว่าจะแบ่งโดยใช้กี่ปัจจัย หรือ แบ่งเป็นกี่ cluster โดยที่ CO = Continous Numeric และ DN = Discrete Numeric

**ตารางที่ 4.3** แสดง Reference Field Characteristics (For Numeric Fields Only)

| Id | Name | Minimum Value | Maximum Value | Mean | Standard Deviation |
|---|---|---|---|---|---|
| 1 | Age | 0 | 1 | 0.6456 | 0.1912 |
| 2 | Current account | 0 | 1 | 0.2635 | 0.1141 |
| 3 | Has Children | 0 | 1 | 0.3400 | 0.2713 |
| 4 | Income | 0 | 1 | 0.7188 | 0.2266 |
| 5 | Loan products | 0 | 1 | 0.3183 | 0.1493 |
| 6 | Marital Status | 0 | 1 | 0.0963 | 0.1830 |
| 7 | No. of auto. Pay. | 0 | 1 | 0.1442 | 0.1962 |
| 8 | No. of C/A trans. | 0 | 1 | 0.0507 | 0.0716 |
| 9 | Saving products | 0 | 1 | 0.1693 | 0.1428 |
| 10 | Sex | 0 | 1 | 0.4250 | 0.4946 |
| 11 | Teller usage | 0 | 1 | 0.0699 | 0.0820 |
| 12 | Web bank usage | 0 | 1 | 0.0360 | 0.0808 |
| 13 | ATM usage | 0 | 1 | 0.0579 | 0.0970 |

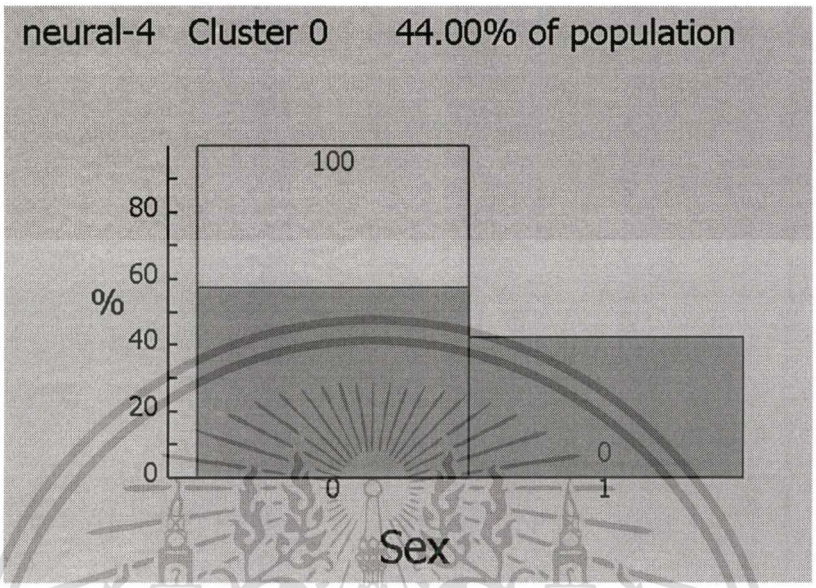
ตาราง Reference Field Characteristics (For Numeric Fields Only) จะแสดงรายละเอียดทั่วไปของแต่ละปัจจัยเช่นเดียวกับตาราง Reference Field Characteristics (For All Field Types) โดยจะบอกถึงค่าของข้อมูลที่มีค่าต่ำสุดและสูงสุด, ค่าเฉลี่ยของข้อมูล, และ ค่าความเบี่ยงเบนมาตรฐานของข้อมูล ซึ่งในที่นี้ทุกๆ ปัจจัยจะมีค่าของข้อมูลต่ำสุดอยู่ที่ 0 และสูงสุดที่ 1 เนื่องจากได้มีการทำ normalization ข้อมูลก่อนที่จะทำการ mining โดยที่ "อายุ" จะเป็นปัจจัยที่มีค่าเฉลี่ยของข้อมูลสูงสุดเท่ากับ 0.65 ส่วนปัจจัยที่มีค่าความเบี่ยงเบนมาตรฐานของข้อมูลสูงสุด คือ "เพศ" คือมีค่าเท่ากับ 0.49 ซึ่งค่าของข้อมูลในตารางนี้ก็จะปรากฏอยู่ในผลสรุปของทุกๆ การแบ่งกลุ่มในข้อมูลชุดเดียวกันไม่ว่าจะแบ่งโดยใช้กี่ปัจจัย หรือ แบ่งเป็นกี่ cluster เช่นเดียวกับตาราง Reference Field Characteristics (For All Field Types)

นอกจากนี้วิธีการแบบ Neural โดยใช้ IBM Intelligent Miner นี้สามารถแสดงรายละเอียดของแต่ละ cluster ได้เป็นกราฟและรายละเอียดของข้อมูล ดังรูปที่ 4.2 ซึ่งแสดงถึงข้อมูลในกลุ่มที่ 0 ซึ่งมีจำนวนข้อมูลมากที่สุดคือคิดเป็น 44.00% ของข้อมูลทั้งหมด โดยที่กราฟสีทึบจะแสดงถึง % ของแต่ละปัจจัยในแต่ละ cluster ส่วน กราฟใสจะแสดงถึงจำนวนข้อมูลทั้งหมด ในที่นี้ถ้าพิจารณาทีละปัจจัยแล้วจะพบว่า ในกลุ่มนี้จะมีประชากรซึ่งเป็นเพศชายทั้งหมด และส่วนใหญ่จะเป็นคนที่สมรสแล้ว

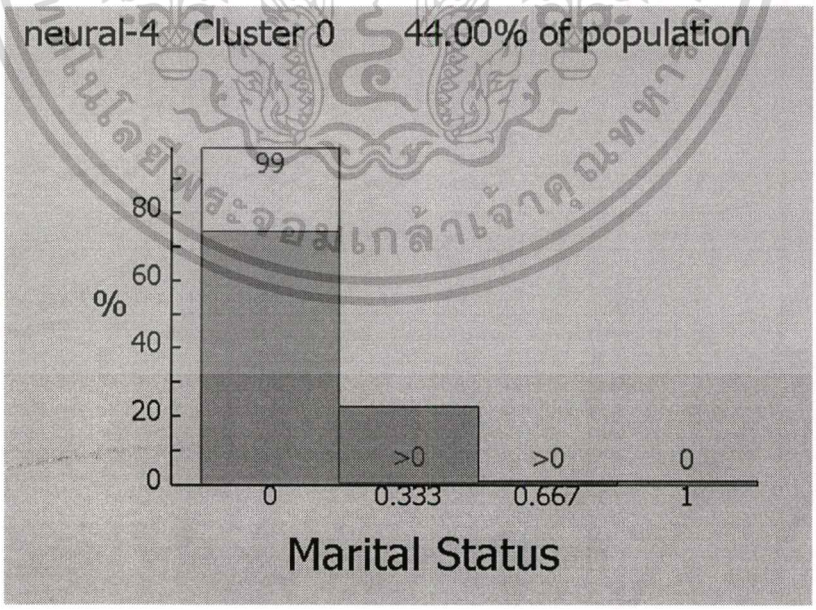



**รูปที่ 4.2** แสดงการแบ่งกลุ่มของ cluster ที่ 0 ซึ่งมีจำนวนข้อมูลมากที่สุดในการแบ่งกลุ่มจำนวน 4 Clusters

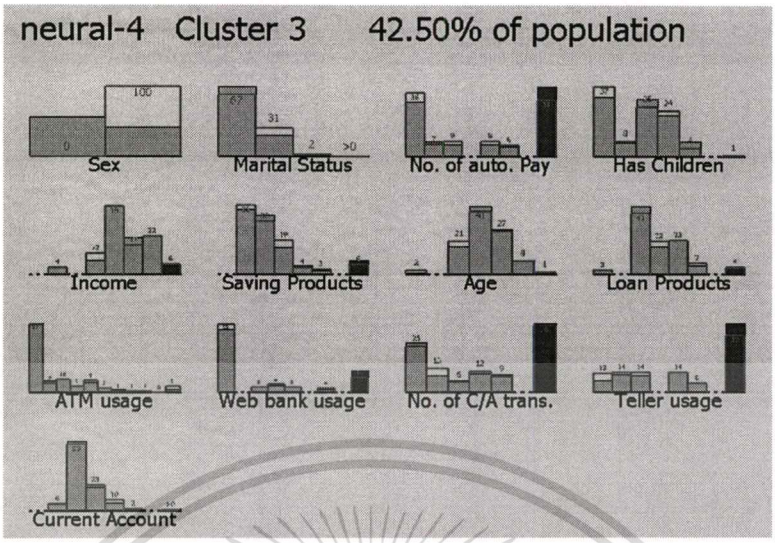
และเมื่อมาพิจารณาแยกเฉพาะปัจจัยที่เป็น “เพศ” และ “สถานภาพสมรส” ดังรูปที่ 4.3 และ รูปที่ 4.4 แล้วพบว่าใน cluster นี้มีประชากรเพศชายถึง 58% ของข้อมูลประชากรที่เป็นเพศชายทั้งหมด และมีประชากรที่เป็นคนที่สมรสแล้วถึง 75% ของข้อมูลของคนที่เป็นคนที่สมรสแล้วทั้งหมด




**รูปที่ 4.3** แสดงเฉพาะปัจจัย "เพศ" ของ cluster ที่ 0 จากการแบ่งกลุ่มจำนวน 4 Clusters

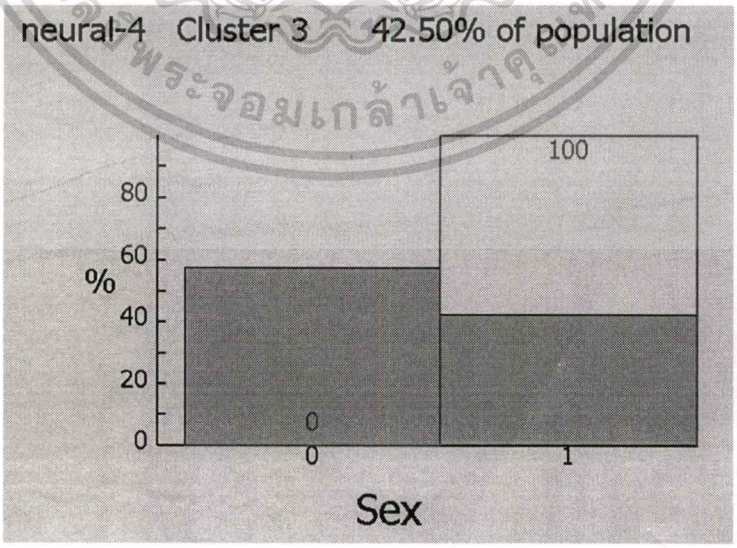



**รูปที่ 4.4** แสดงเฉพาะปัจจัย "สถานภาพสมรส" ของ cluster ที่ 0 จากการแบ่งกลุ่ม จำนวน 4 Clusters



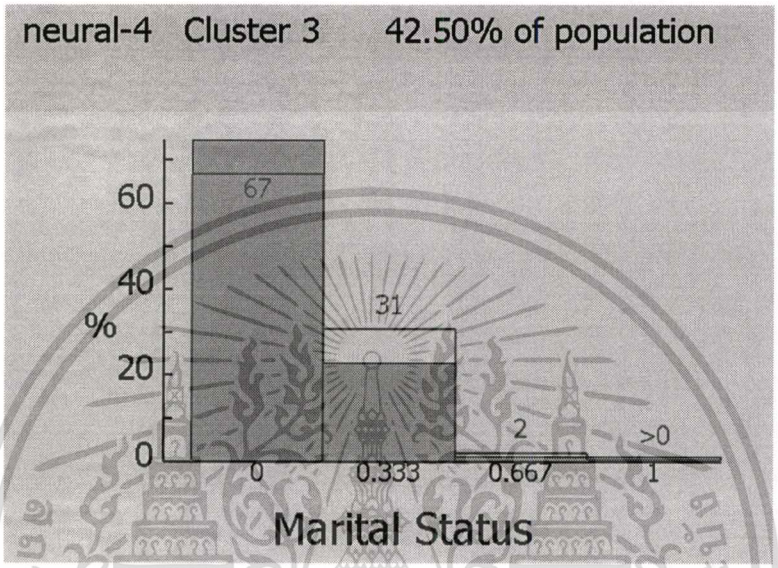

**รูปที่ 4.5** แสดงการแบ่งกลุ่มของ cluster ที่ 3 ซึ่งมีจำนวนข้อมูลมากเป็นอันดับ 2 ในการแบ่งกลุ่มจำนวน 4 Clusters

รูปที่ 4.5 แสดงถึงข้อมูลในกลุ่มที่ 3 ซึ่งมีจำนวนคิดเป็น 42.50% ของข้อมูลทั้งหมด กลุ่มนี้มีปัจจัยที่เป็น "เพศ" เป็นปัจจัยหลักในการแบ่งกลุ่ม โดยกลุ่มนี้จะมีสมาชิกของกลุ่มเป็นเพศหญิงทั้งหมด และส่วนใหญ่จะเป็นคนที่มีสถานภาพสมรสแล้ว ซึ่งถ้าพิจารณาทีละปัจจัยดังรูปที่ 4.6 และรูปที่ 4.7 แล้วจะพบว่า ในกลุ่มนี้จะมีประชากรที่เป็นเพศหญิงอยู่ประมาณ 40% ของข้อมูลประชากรที่เป็นเพศหญิงทั้งหมด

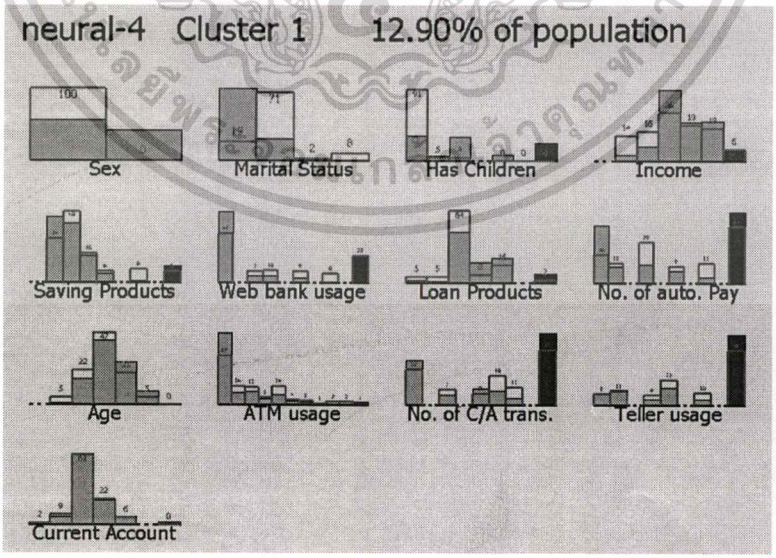



**รูปที่ 4.6** แสดงเฉพาะปัจจัย "เพศ" ของ cluster ที่ 3 จากการแบ่งกลุ่มจำนวน 4 Clusters

สำหรับส่วนของ "สถานภาพสมรส" นั้นเมื่อเทียบจำนวนข้อมูลของคนที่สมรสแล้วทั้งหมด สมาชิกของกลุ่มนี้นั้นจะเป็นคนที่เป็นคนซึ่งสมรสแล้วมากกว่าข้อมูลอื่นๆ เช่นเดียวกับกลุ่มแรก ทั้งนี้ในกลุ่มนี้ปัจจัยที่มีอิทธิพลในการแบ่งกลุ่มน้อยที่สุด คือ จำนวนบัญชี Current

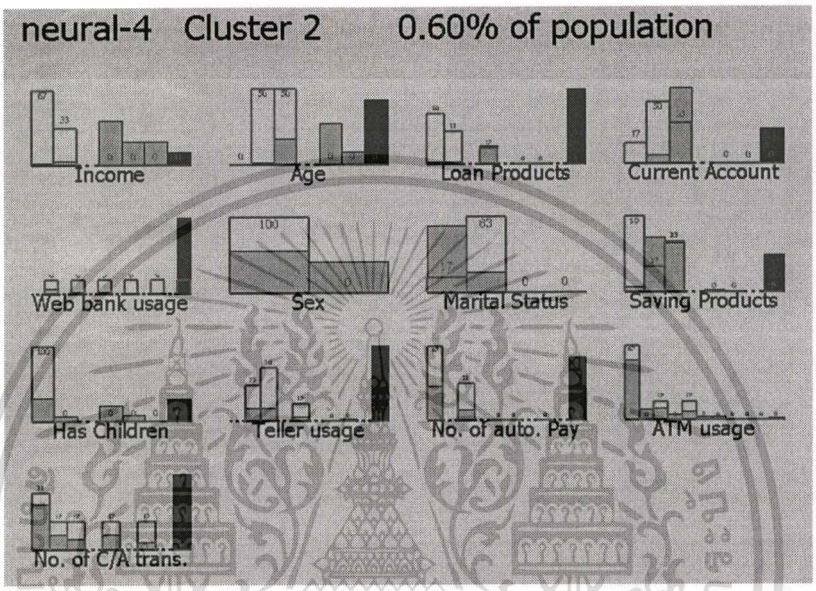



**รูปที่ 4.7** แสดงเฉพาะปัจจัย "สถานภาพสมรส" ของ cluster ที่ 3 จากการแบ่งกลุ่ม จำนวน 4 Clusters



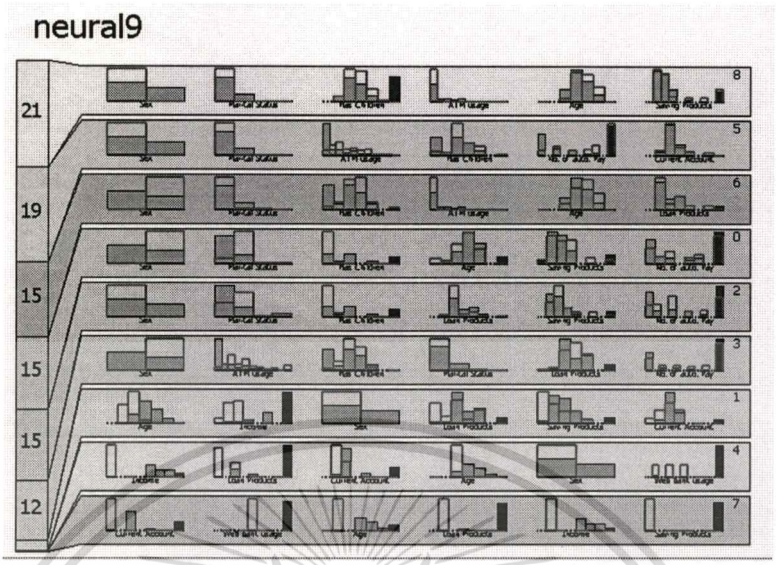

**รูปที่ 4.8** แสดงการแบ่งกลุ่มของ cluster ที่ 1 ซึ่งมีจำนวนข้อมูลมากเป็นอันดับ 3 ในการแบ่งกลุ่มจำนวน 4 Clusters

รูปที่ 4.8 แสดงถึงข้อมูลในกลุ่มที่ 1 ซึ่งมีจำนวนคิดเป็น 12.90% ของข้อมูลทั้งหมด โดยถ้าพิจารณาทีละปัจจัยแล้วจะพบว่ายังคงเป็นเพศ และสถานภาพสมรส ที่เป็นปัจจัยหลักในการแบ่งกลุ่ม ซึ่งในกลุ่มนี้จะมีประชากรที่เป็นเพศชาย และ เป็นคนโสด ทั้งนี้ในกลุ่มนี้ปัจจัยที่มีอิทธิพลในการแบ่งกลุ่มน้อยที่สุด คือ จำนวนบัญชี Current




**รูปที่ 4.9** แสดงการแบ่งกลุ่มของ cluster ที่ 2 ซึ่งมีจำนวนข้อมูลน้อยที่สุด ในการแบ่งกลุ่มจำนวน 4 Clusters

รูปที่ 4.9 แสดงถึงข้อมูลในกลุ่มที่ 2 ซึ่งมีจำนวนข้อมูลน้อยที่สุดในการแบ่งกลุ่มจำนวน 4 Clusters โดยมีข้อมูลเพียง 0.60% ของข้อมูลทั้งหมด โดยถ้าพิจารณาแล้วพบ “รายได้” และ “อายุ” เป็นปัจจัยหลักในการแบ่งกลุ่มนี้ ทั้งนี้ในปัจจัยที่มีอิทธิพลในการแบ่งกลุ่มน้อยที่สุด คือ จำนวนครั้งที่ทำธุรกรรมในบัญชีกระแสรายวัน และปริมาณการใช้ ATM ตามลำดับ




**รูปที่ 4.10** แสดง การแบ่งกลุ่มโดยใช้ 13 ปัจจัยโดยแบ่งออกเป็น 9 clusters

จากรูปที่ 4.10 จะเห็นได้ว่าข้อมูลถูกแบ่งออกเป็น 9 กลุ่ม โดยที่แถวบนแสดงกลุ่มที่มีจำนวนข้อมูลมากที่สุด คือ 21.40% ซึ่งกลุ่มนี้มีเลข id ของกลุ่มคือ 8 กลุ่มที่ 2 ซึ่งเมื่อแบ่งออกมาแล้วมีจำนวนข้อมูลใกล้เคียงกันคือกลุ่ม id 5 คือมีข้อมูล 19.20% ส่วนกลุ่มอื่นๆ จะมีจำนวนข้อมูล 15.20%, 15.10%, 14.50%, 12.20%, 2%, 0.30% และ 0.10% ตามลำดับ

สำหรับผลในรูปรายละเอียดนั้น โปรแกรมนี้จะแสดงออกมาในรูปของ Result Created ซึ่งจะบอกข้อมูลดังนี้

neural9

Result created: 10/20/04 22:40:12

| | | |
|---|---|---|
| Result File | : | C:\DOCUME~1\SCB_ST~1\LOCALS~1\Temp\VDMMFY4P.C2T |
| Mode | : | Training |
| User Specified Parameters | | |
| Maximum Number of Passes | : | 20 |
| Maximum Number of Clusters | : | 9 |
| Mining Run Outputs | | |
| Number of Passes Performed | : | 20 |
| Number of Clusters | : | 9 |
| Deviation | : | 0.0835927 |

**Cluster Characteristics :**

| Id | Cluster Size | | Id | Cluster Size | |
|---|---|---|---|---|---|
| | Absolute | Relative(%) | | Absolute | Relative(%) |
| 0 | 151 | 15.10 | 5 | 192 | 19.20 |
| 1 | 20 | 2.00 | 6 | 152 | 15.20 |
| 2 | 145 | 14.50 | 7 | 1 | 0.10 |
| 3 | 122 | 12.20 | 8 | 214 | 21.40 |
| 4 | 3 | 0.30 | | | |

**รูปที่ 4.11** แสดง Result Created การแบ่งกลุ่มโดยใช้ 13 ปัจจัยโดยแบ่งออกเป็น 9 clusters

User Specified Parameters

Maximum Number of Passes : 20

Maximum Number of Cluster : 9

Mining Run Outputs

Number of Passes Performed : 20

Number of Clusters : 9

Deviation : 0.0835927

ข้อมูลข้างต้นจะบอกรายละเอียดของจำนวนครั้งที่ทำการ run ข้อมูลซึ่งเท่ากับ 20 ครั้ง จำนวน cluster ที่แบ่ง คือ 4 cluster และค่าความเบี่ยงเบนของข้อมูลซึ่งเท่ากับ 0.0835927

ตารางที่ 4.4 แสดง Cluster Characteristics 9 Clusters

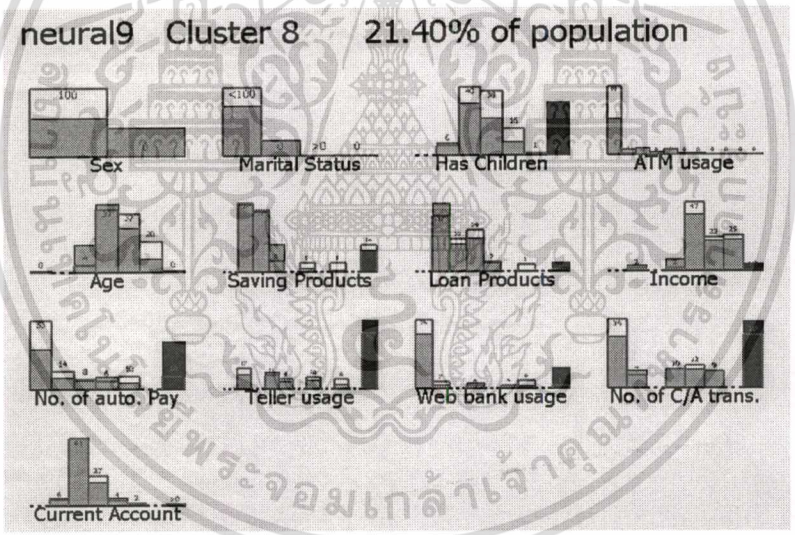
| Id | Cluster Size | |
|---|---|---|
| | Absolute | Relative (%) |
| 0 | 151 | 15.10 |
| 1 | 20 | 2.00 |
| 2 | 145 | 14.50 |
| 3 | 122 | 12.20 |
| 4 | 3 | 0.30 |
| 5 | 192 | 19.20 |
| 6 | 152 | 15.20 |
| 7 | 1 | 0.10 |
| 8 | 214 | 21.40 |

ตาราง Cluster Characteristics จะแสดงปริมาณข้อมูลในแต่ละ cluster ในที่นี้ข้อมูลจะถูกแบ่งออกเป็น 4 cluster โดยที่

- cluster ที่ 0 จะมีจำนวนข้อมูลเท่ากับ 151 ข้อมูล ซึ่งคิดเป็น 15.10%ของข้อมูลทั้งหมด
- cluster ที่ 1 จะมีจำนวนข้อมูลเท่ากับ 20 ข้อมูล ซึ่งคิดเป็น 2.00%ของข้อมูลทั้งหมด
- cluster ที่ 2 จะมีจำนวนข้อมูลเท่ากับ 145 ข้อมูล ซึ่งคิดเป็น 14.50% ของข้อมูลทั้งหมด

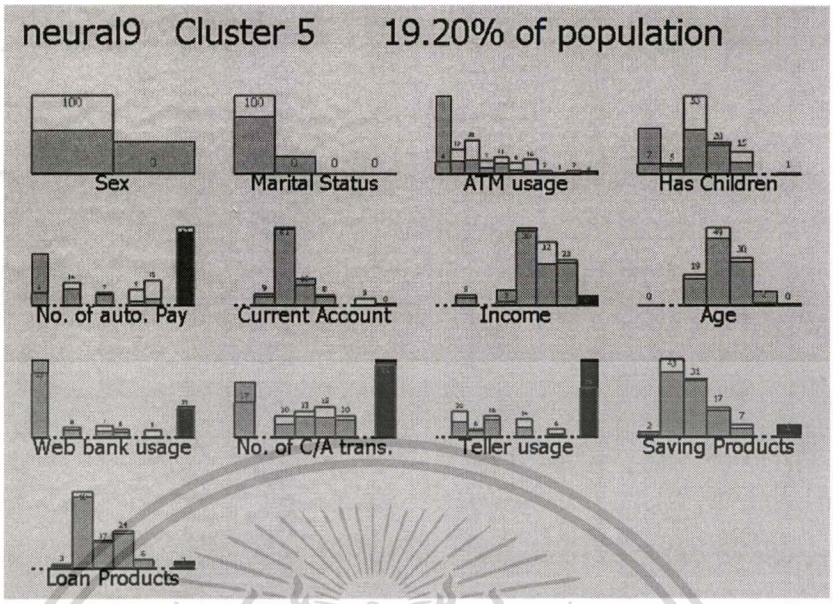
- cluster ที่ 3 จะมีจำนวนข้อมูลเท่ากับ 122 ข้อมูล ซึ่งคิดเป็น 12.20% ของข้อมูลทั้งหมด
- cluster ที่ 4 จะมีจำนวนข้อมูลเท่ากับ 3 ข้อมูล ซึ่งคิดเป็น 0.30% ของข้อมูลทั้งหมด
- cluster ที่ 5 จะมีจำนวนข้อมูลเท่ากับ 192 ข้อมูล ซึ่งคิดเป็น 19.20% ของข้อมูลทั้งหมด
- cluster ที่ 6 จะมีจำนวนข้อมูลเท่ากับ 152 ข้อมูล ซึ่งคิดเป็น 15.20% ของข้อมูลทั้งหมด
- cluster ที่ 7 จะมีจำนวนข้อมูลเท่ากับ 1 ข้อมูล ซึ่งคิดเป็น 0.10% ของข้อมูลทั้งหมด
- cluster ที่ 8 จะมีจำนวนข้อมูลเท่ากับ 214 ข้อมูล ซึ่งคิดเป็น 21.40% ของข้อมูลทั้งหมด

ถ้าพิจารณาทีละกลุ่มข้อมูลแล้วจะพบว่า cluster 8 นั้นเป็น cluster ที่มีจำนวนข้อมูลมากที่สุด คือ 214 ข้อมูล หรือมีข้อมูล 21.40% โดยเพศ และสถานภาพสมรสเป็นปัจจัยที่มีอิทธิพลในการแบ่งกลุ่ม ส่วนปัจจัยที่มีอิทธิพลในการแบ่งกลุ่มน้อยที่สุด คือ จำนวนบัญชีกระแสรายวัน และจำนวนครั้งที่ทำธุรกรรมในบัญชีกระแสรายวัน ตามลำดับ

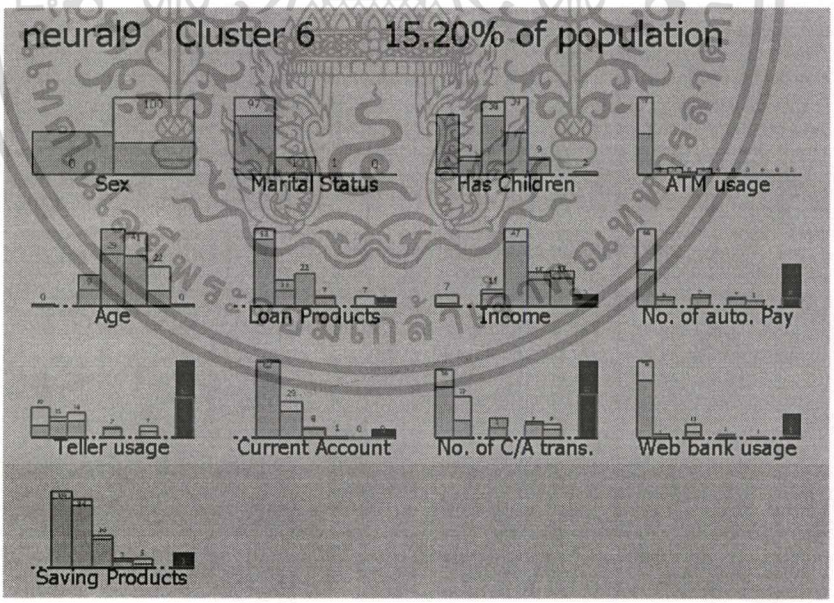



**รูปที่ 4.12** แสดงการแบ่งกลุ่มของ cluster ที่ 8 ซึ่งมีจำนวนข้อมูลมากที่สุดในการแบ่งกลุ่มจำนวน 9 Clusters

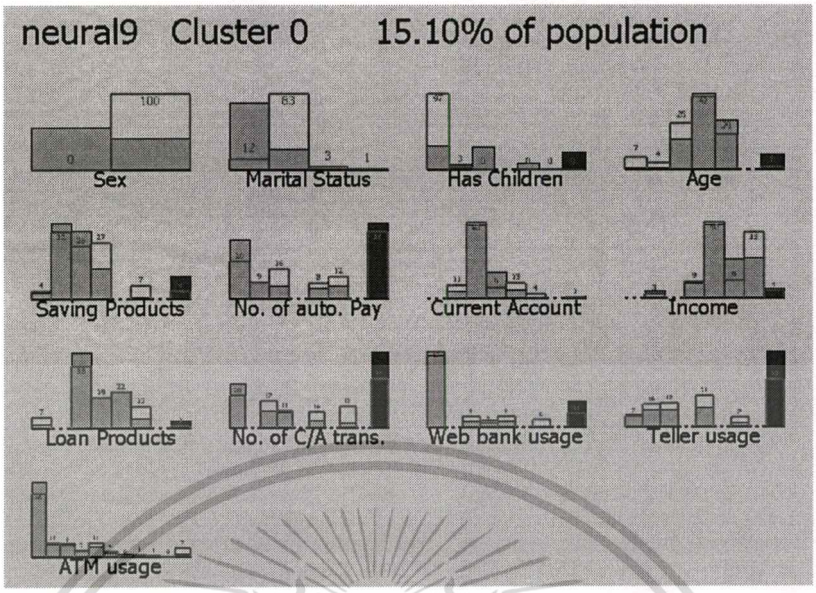
ทั้งนี้เมื่อพิจารณา Cluster อื่นๆ ดังแสดงในรูปที่ 4.13 -4.20



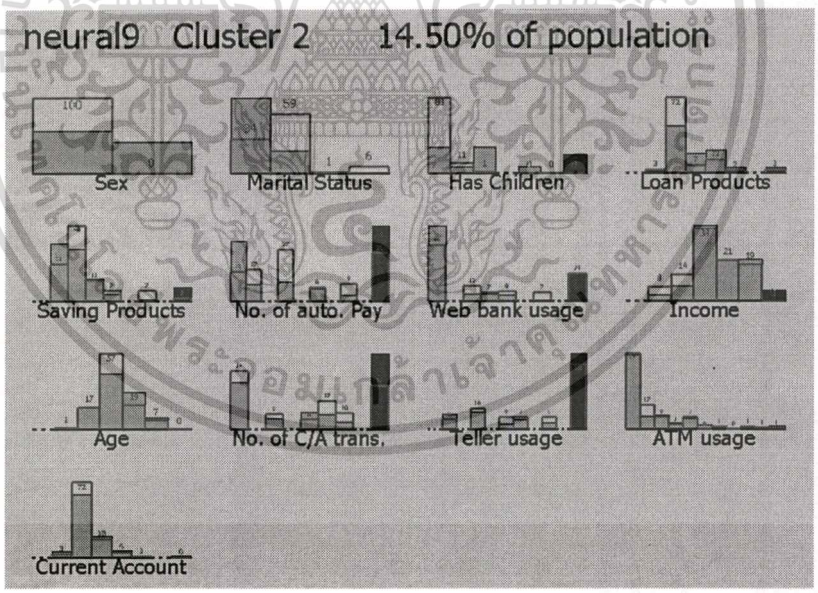

**รูปที่ 4.13** แสดงการแบ่งกลุ่มของ cluster ที่ 5 ซึ่งมีจำนวนข้อมูล 19.20% ในการแบ่งกลุ่มจำนวน 9 Clusters



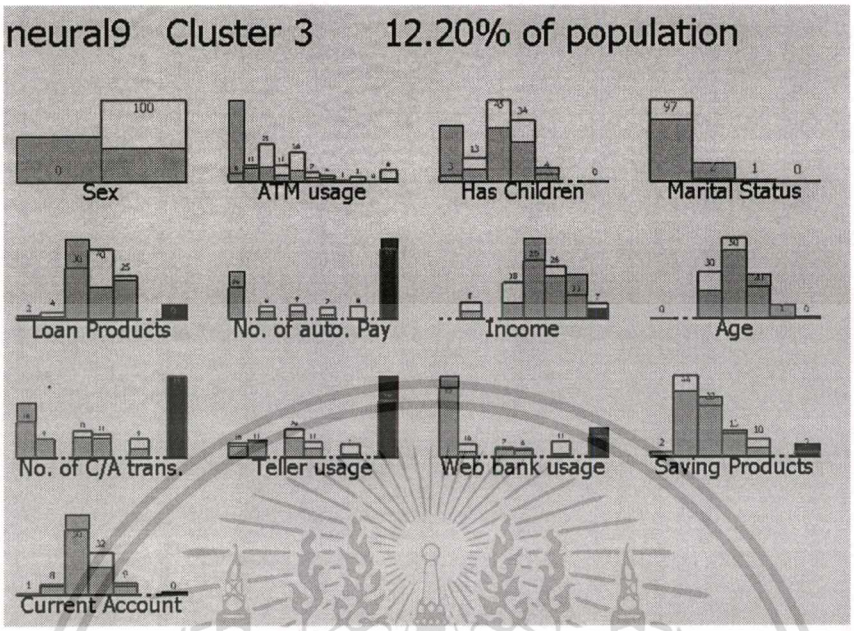

**รูปที่ 4.14** แสดงการแบ่งกลุ่มของ cluster ที่ 6 ซึ่งมีจำนวนข้อมูล 15.20% ในการแบ่งกลุ่มจำนวน 9 Clusters



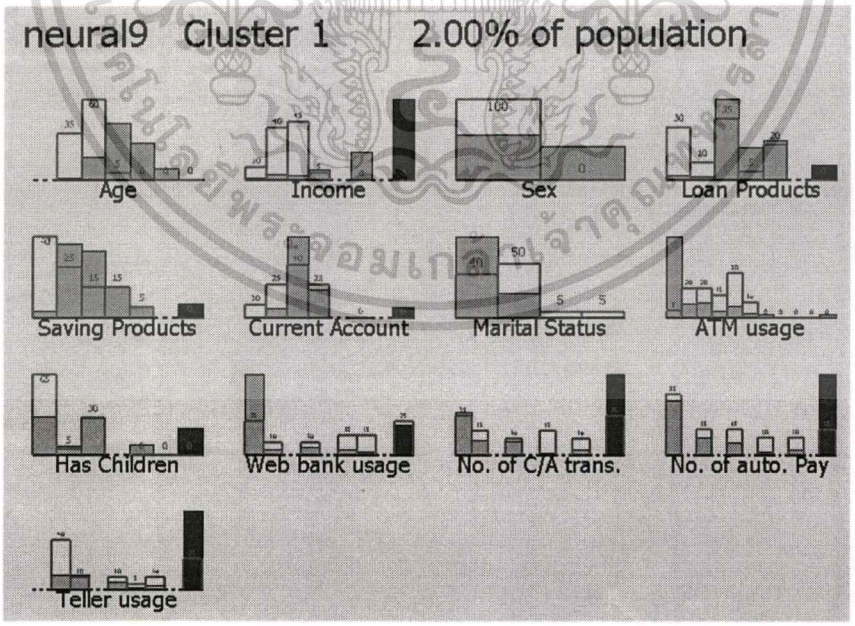

**รูปที่ 4.15** แสดงการแบ่งกลุ่มของ cluster ที่ 0 ซึ่งมีจำนวนข้อมูล 15.10% ในการแบ่งกลุ่มจำนวน 9 Clusters



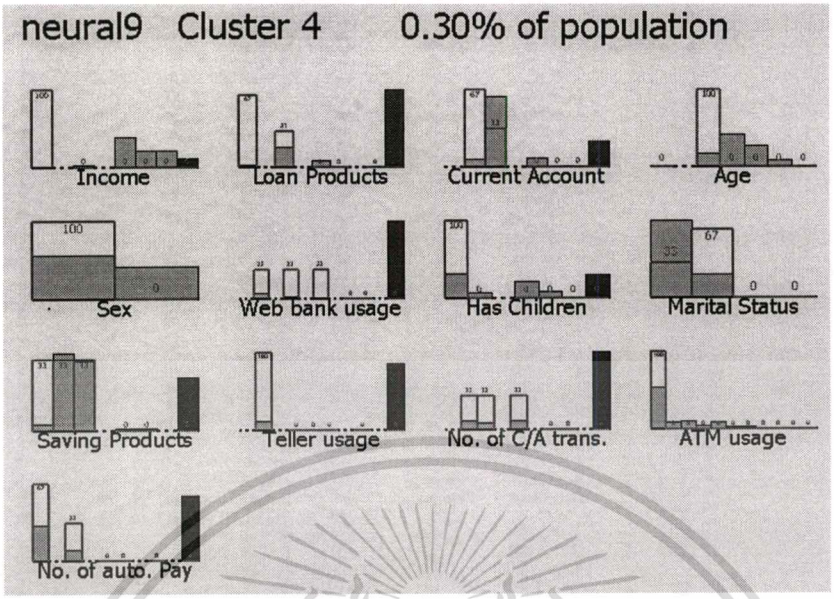

**รูปที่ 4.16** แสดงการแบ่งกลุ่มของ cluster ที่ 2 ซึ่งมีจำนวนข้อมูล 14.50% ในการแบ่งกลุ่มจำนวน 9 Clusters



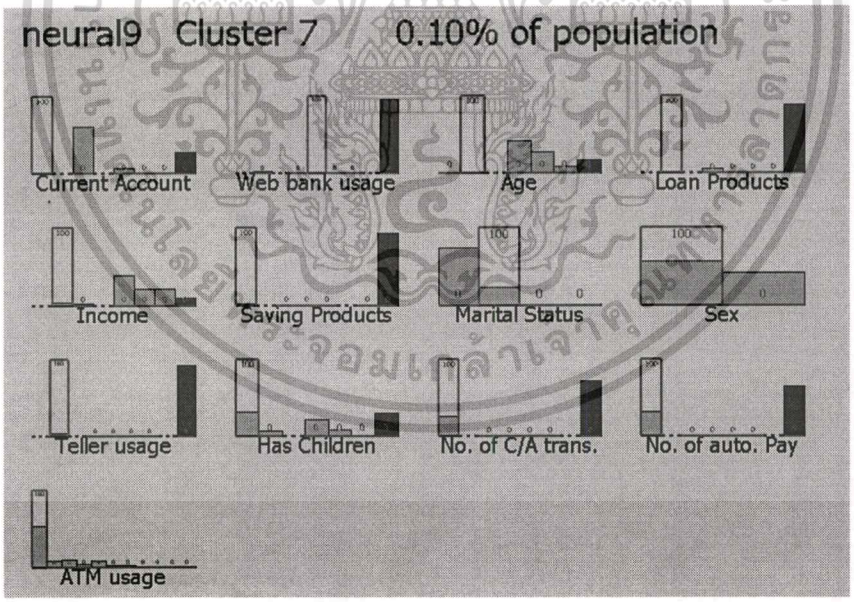

**รูปที่ 4.17** แสดงการแบ่งกลุ่มของ cluster ที่ 3 ซึ่งมีจำนวนข้อมูล 12.20% ในการแบ่งกลุ่มจำนวน 9 Clusters




**รูปที่ 4.18** แสดงการแบ่งกลุ่มของ cluster ที่ 1 ซึ่งมีจำนวนข้อมูล 2.00% ในการแบ่งกลุ่มจำนวน 9 Clusters




**รูปที่ 4.19** แสดงการแบ่งกลุ่มของ cluster ที่ 4 ซึ่งมีจำนวนข้อมูล 0.30% ในการแบ่งกลุ่มจำนวน 9 Clusters

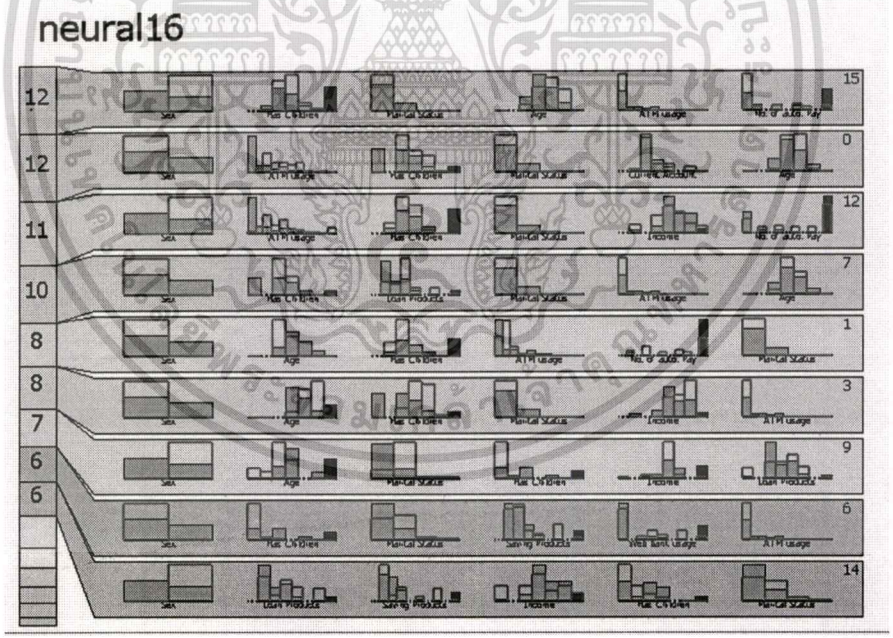



**รูปที่ 4.20** แสดงการแบ่งกลุ่มของ cluster ที่ 7 ซึ่งมีจำนวนข้อมูล 0.10% ในการแบ่งกลุ่มจำนวน 9 Clusters

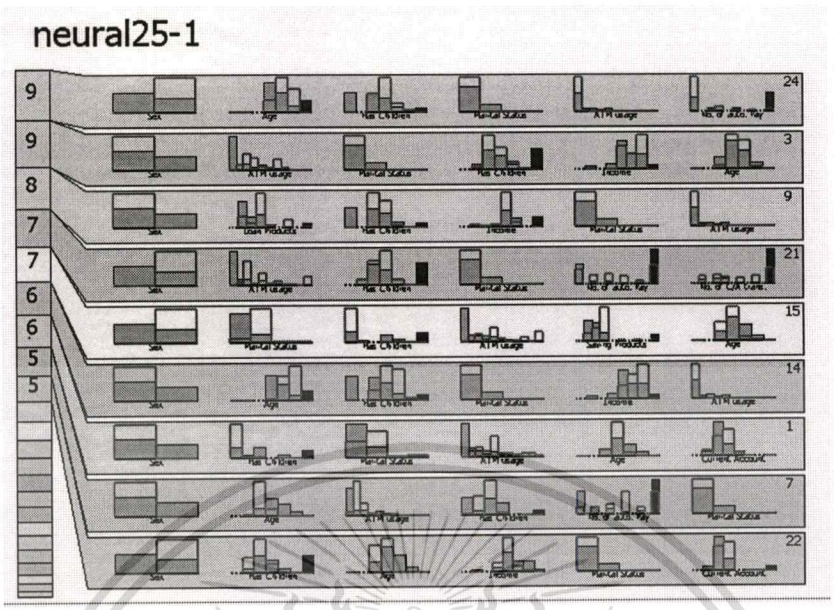
จะเห็นได้ว่ากลุ่มโดยส่วนใหญ่ถูกแบ่งโดยใช้ปัจจัยที่เป็น เพศ และสถานภาพสมรสเป็นปัจจัยหลักในการแบ่ง มีเพียง 2.4% ของข้อมูลทั้งหมดที่มี อายุ และรายได้ เป็นปัจจัยหลักในการแบ่ง ซึ่งถ้านำมายุบกลุ่มรวมกันเข้าโดยมีปัจจัยคือเพศและสถานภาพสมรสเป็นปัจจัยหลักก็จะสามารถแบ่งกลุ่มลูกค้าออกเป็น 4 ประเภท ดังนี้

1. ลูกค้าที่เป็นเพศชาย สถานภาพสมรสแล้ว
2. ลูกค้าที่เป็นเพศหญิง สถานภาพสมรสแล้ว
3. ลูกค้าที่เพศชาย/หญิง ที่มีสถานภาพโสด
4. ลูกค้าที่มีอายุและรายได้น้อย

ทั้งนี้เมื่อพิจารณาจากข้อมูลของการแบ่งกลุ่มออกเป็น 16 กลุ่ม และ 25 กลุ่มก็พบว่าเป็นเช่นเดียวกัน ดังแสดงได้ในรูปที่ 4.21 และ รูปที่ 4.22 โดยปริมาณข้อมูลในกลุ่มอื่นๆ นั้นมีปริมาณน้อยมากจนไม่มีนัยสำคัญในการนำมาทำกลยุทธ์ทางการตลาด




**รูปที่ 4.21** แสดง การแบ่งกลุ่มโดยใช้ 13 ปัจจัยโดยแบ่งออกเป็น 16 clusters




**รูปที่ 4.21** แสดง การแบ่งกลุ่มโดยใช้ 13 ปัจจัยโดยแบ่งออกเป็น 25 clusters

จึงสรุปว่าในเบื้องต้นนั้นในปริมาณข้อมูลขนาด 1000 ชุดข้อมูลควรจะแบ่งกลุ่มเพื่อทำการตลาดเพียง 4 กลุ่มเท่านั้น นอกจากนี้เมื่อทำการแบ่งกลุ่มแล้วควรทำการสรุปลักษณะเด่นของแต่ละกลุ่มเพื่อหาลักษณะและพฤติกรรมของลูกค้าแต่ละกลุ่มซึ่งแตกต่างกัน ทั้งนี้เพื่อช่วยในการสร้างกลยุทธ์ให้เหมาะสมสำหรับลูกค้าแต่ละกลุ่มตามวัตถุประสงค์ที่ได้วางไว้ตั้งแต่ต้น นอกจากนี้หากสามารถทำได้ควรนำลักษณะการใช้บริการมาเปรียบเทียบกับรายรับซึ่งได้จากลูกค้ากลุ่มดังกล่าวด้วย จะทำให้เราสามารถทราบได้ว่าว่าคนกลุ่มไหนจะสร้างหารายได้ได้มากที่สุด และเป็นทางหนึ่งที่จะช่วยให้เจาะกลุ่มลูกค้าได้ตรงกับความต้องการมากที่สุด

# บทที่ 5
# สรุปผล และข้อเสนอแนะ

## 5.1 สรุปผลที่ได้จากการศึกษา

การใช้เทคนิค Data Mining โดย ใช้ Algorithm คือ Neural Network Clustering เพื่อใช้ในการแบ่งส่วนฐานข้อมูล (Database Segmentation) จะทำการแบ่งหรือจัดกลุ่มของข้อมูลที่มีลักษณะคล้ายกัน หรือมีคุณสมบัติใกล้เคียงกันในหลายๆ ด้าน ให้เป็นข้อมูลกลุ่มเดียวกัน ซึ่งแต่ละกลุ่มจะถูกเรียกว่า Segments หรือ Clusters การแบ่งกลุ่มข้อมูลนี้เราจะไม่สามารถกำหนดได้ว่าข้อมูลควรจะอยู่กลุ่มใด แต่จะเป็นการกำหนดกลุ่มของข้อมูลจากธรรมชาติของข้อมูลเอง ข้อมูลที่ได้จากการศึกษาและวิเคราะห์ข้อมูลลูกค้าธนาคารพาณิชย์ โดยการใช้เทคนิค Data Mining เพื่อแบ่งกลุ่มลูกค้าตามลักษณะพฤติกรรมที่แตกต่างกันนั้นในครั้งนี้ ธนาคารพาณิชย์สามารถใช้เป็นแนวทางเพื่อทราบได้ว่าลูกค้ากลุ่มหลักของธนาคารมีลักษณะและพฤติกรรมเป็นแบบใด สามารถนำมาใช้เป็นฐานข้อมูลในการทำ CRM (Customer Relationship Management) และช่วยในการตัดสินใจสำหรับผู้บริหารในการวางแผนกลยุทธ์ของธนาคาร ในการขยายบริการธุรกิจอื่นๆ ของธนาคารให้กับลูกค้ากลุ่มดังกล่าวต่อไป โดยนอกจากเทคนิคที่ได้ทำการศึกษาดังกล่าวข้างต้น สามารถนำเทคนิค Data Mining มาพัฒนาการศึกษาวิเคราะห์ข้อมูลต่อไปในด้านอื่นๆ เช่นการหาความสัมพันธ์ของข้อมูลในแต่ละกลุ่มข้อมูล (Association Rules), สร้างแบบจำลองพยากรณ์ (Predictive Modeling) เป็นต้น

## 5.2 ข้อเสนอแนะ

1. การศึกษาในครั้งนี้ใช้ข้อมูลของธนาคารพาณิชย์แห่งหนึ่งในประเทศไทย ซึ่งอาจจะยังไม่ได้ครอบคลุมในบางกรณีที่อาจจะเกิดขึ้นได้ในการแบ่งกลุ่ม และข้อจำกัดในเรื่องของข้อมูลที่นำมาใช้ในการทดสอบไม่สามารถเปิดเผยได้ทั้งหมด ดังนั้นหน่วยงานที่ต้องการใช้งานและมีผู้ที่สามารถหาข้อมูลจำนวนมากได้หลากหลายและครบถ้วนจะทำให้ได้ผลลัพธ์ที่น่าสนใจและนำไปใช้ได้ดีขึ้น
2. ข้อมูลของลูกค้าเป็นข้อมูลที่เปลี่ยนแปลงตลอดเวลา ดังนั้นควรมีการ up date ข้อมูลรวมทั้งตัวแปรอยู่เสมอเพื่อให้สามารถแข่งขันกับคู่แข่งได้

3. เนื่องจากมีปัญหาในการใช้โปรแกรม โดยไม่สามารถนำโปรแกรมที่มี license มาใช้ได้ โปรแกรมที่เป็น freeware นั้นไม่อำนวยความสะดวกในการประมวลผล และแสดงผลลัพธ์บางอย่างในข้อมูลได้ดังนั้น ถ้ามีโปรแกรมซึ่งมีความสามารถและหาใช้ได้ไม่ยาก จะทำให้สามารถศึกษาได้กว้างขึ้น
4. ควรจะมีการพัฒนาต่อไปในด้านอื่น ๆ ต่อเช่นการหาความสัมพันธ์ของข้อมูลในแต่ละกลุ่มข้อมูล (Association Rules) เป็นต้น
5. จากข้อมูลที่สรุปผลนั้นเราสามารถที่จะนำไปเป็นแนวทางในการศึกษา วางแผน และวิเคราะห์ ข้อมูลในธุรกิจธนาคารพาณิชย์บริษัทอื่นโทรคมนาคมในประเทศไทยเพื่อนำผลสรุปที่ได้ไปทำการวางแผนกลยุทธ์ทางการตลาดเพื่อตอบสนองความต้องการของลูกค้ากลุ่มต่างๆได้
6. จากข้อมูลที่สรุปผลได้ 4 กลุ่มนั้นเราสามารถที่จะนำเอาข้อมูลของคนในกลุ่มที่ 1 ไปทำการวางแผนกลยุทธ์ทางการตลาด เพื่อดึงความสนใจของคนในกลุ่มนี้ให้ใช้บริการกับธนาคารมากขึ้น ซึ่งจะทำให้สร้างรายได้ให้กับธนาคารได้มากขึ้น

# บรรณานุกรม


กฤษณะ ไวยมัย และคณะ. 2546. Data Mining : การเตรียมข้อมูลสำหรับดาต้าไมน์นิง. **ไมโครคอมพิวเตอร์.** [Online]. Available: Http://www.micro.se-ed.com

ชัยรัตน์ แสงทอง. 2546, กรกฎาคม. "สมรภูมิเอ็นพีแอลแบงก์ปีนี้แข่งกันเดือด." **Business.com.** หน้า 21.

Backer, E. 1988. **Reasoning in Cluster Analysis**. London: Prentice Hall.

Baragoin, C. et al. 2001. **Mining Your Own Business in Banking**. [Online]. Available: Http://www.ibm.com/redbooks

Berson, A. and Smith, S. J. 2001. **Data Warehousing, Data Mining and OLAP.** New York: McGraw – Hill .

Carpenter, G. A. et al. 1991. **Fuzzy ART: An adaptive resonance algorithm for rapid, stable classification of analog patterns**. Center for Adaptive Systems and Graduate Program in Cognitive and Neural Systems, Boston University.

Chongstitvatana, P. 2002. **Data clustering**. [Online]. Available: Http://www.cp.eng.chula.ac.th

Han, J. and Kamber, M. 2003. **Data Mining : Concepts and Techniques**. [Online]. Available: Http://www.cs.cfu.ca

Haruechaiyasak, C. 2003. **Self-Organization Maps**. [Online]. Available: Http://www.nectec.or.th

Jan, A. K. and Dubes, R. C. 1988. **Algorithms for Clustering Data**. London: Prentice Hall.

Nukoolkit, C. 2004. CS_LectureNote : **Clustering**. [Online]. Available: Http://www.cpe.kmutt.ac.th

# ประวัติผู้เขียน

| | |
|---|---|
| **ชื่อ** | นางสาวปัทมเกสร์ อมาตยกุล |
| **วัน/เดือน/ปี เกิด** | 6 พฤศจิกายน |
| **สถานที่เกิด** | กรุงเทพมหานคร |
| **ประวัติการศึกษา** | จบการศึกษาระดับปริญญาตรี คณะเศรษฐศาสตร์ มหาวิทยาลัยธรรมศาสตร์ |
| **ประวัติการทำงาน** | ปัจจุบันเป็นเจ้าหน้าที่บริหารงานบริการกองทุนสำรองเลี้ยงชีพ สังกัด ผลิตภัณฑ์การเงิน ธนาคารไทยพาณิชย์ จำกัด (มหาชน) |