ระบบวิเคราะห์จัดแบ่งกลุ่มลูกค้าด้วยวิธี Fuzzy C-Means

Customer Segmentation using Fuzzy C-Means Algorithm

โดย

ต่อพงษ์ โลหะรังสิกุล

รหัสประจำตัว 46066840

อาจารย์ที่ปรึกษา

ผศ.ดร.วรพจน์ กรีสุระเดช

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน

หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

ภาคเรียนที่ 2 ปีการศึกษา 2548

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ชื่อหัวข้อ	ระบบวิเคราะห์จัดแบ่งกลุ่มลูกค้าด้วยวิธี Fuzzy C-Means
นักศึกษา	นายต่อพงษ์ โลหะรังสิกุล
อาจารย์ที่ปรึกษา	ผศ.ดร. วรพจน์ กรีสุระเดช
ระดับการศึกษา	วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2548

บทคัดย่อ

การเติบโต และการแข่งขันกันในธุรกิจปัจจุบันจำเป็นที่จะต้องอาศัยข้อมูลที่ถูกต้อง และแม่นยำเข้ามามีส่วนช่วยในการบริหารจัดการตลอดจนการตัดสินใจในเรื่องต่างๆเพิ่มมากขึ้น โดยเฉพาะข้อมูลของลูกค้าหรือกลุ่มเป้าหมายต่างๆที่เราสนใจ เราจำเป็นที่จะต้องศึกษา เรียนรู้ลักษณะ และพฤติกรรมของลูกค้าตลอดจนทำการวิเคราะห์ข้อมูลของลูกค้าต่างๆเพื่อที่จะได้สามารถจัดแบ่งลูกค้าเป้าหมายจำนวนมากที่มีความแตกต่างกันออกเป็นกลุ่มที่มีความชัดเจนมากยิ่งขึ้นได้

Customer Segmentation เป็นกระบวนการหนึ่งในการทำ Data mining ซึ่งจะสามารถช่วยให้เราทำการศึกษา และวิเคราะห์จัดแบ่งกลุ่มลูกค้าได้ง่ายดายยิ่งขึ้น โดยในโครงงานพัฒนาระบบนี้จะเน้นในเรื่องของการวิเคราะห์แบบ Cluster Analysis และประยุกต์ใช้ด้วยวิธีการของ Fuzzy C-Means Algorithm เพื่อค้นหา และแยกแยะความแตกต่างของกลุ่มลูกค้าเป้าหมายให้มีความชัดเจนมากขึ้น ตลอดจนสามารถนำผลลัพธ์ของการจัดแบ่งกลุ่มลูกค้าที่ได้นี้ไปประเมินผลเพื่อตอบสนองต่อกลยุทธ์ทางการตลาดของธุรกิจต่อไปได้

Title	Customer Segmentation using Fuzzy C-Means Algorithm
Student	Mr. Torpong Loharungsikul
Advisor	Asst. Prof. Dr. Worapoj Kreesuradej
Level of Study	Master of Science in Information Technology
Major	Information Science
Academic Year	2005

ABSTRACT

Nowadays, the high competition and growth rapidly in The Business World drives our needs to have precise data for managing and making a decision. Especially the details of targeted prospect, such as, characteristics, profiles and several behaviors of customers so that we can analyze and classify the distinctions accurately.

Customer Segmentation is a process of Data mining which can assist to classify customers easily. So in this System Project, I will use especially the Cluster Analysis Methodology and apply with Fuzzy C-Means Algorithm in order to search and clarify the clusters of customers and use the result to assess our business strategies successfully.

กิตติกรรมประกาศ

โครงการพัฒนาระบบงานฉบับนี้จัดทำขึ้นโดยการรวบรวม ประมวลผลทั้งความรู้ทางภาคทฤษฎี และปฏิบัติเกี่ยวกับทางด้านคาค้าไมนิ่งเพื่อทำการวิเคราะห์จัดแบ่งกลุ่มลูกค้า ซึ่งตลอดระยะเวลาในการศึกษา และพัฒนาระบบงานนั้นเกิดอุปสรรคต่างๆในการทำงานมากมายไม่ว่าจะเป็นความรู้พื้นฐานในภาคทฤษฎี การเตรียมข้อมูล วิธีการในการวิเคราะห์ข้อมูล ตลอดจนการพัฒนาระบบด้วยเครื่องมือสำหรับเขียนโปรแกรม แต่อย่างไรก็ตามโครงการพัฒนาระบบงานฉบับนี้จะไม่สามารถสำเร็จลุล่วงไปได้ด้วยดีเลย ถ้าหากขาดความช่วยเหลือ ข้อเสนอแนะ และการให้คำปรึกษาจากอาจารย์ วรพจน์ กรีสุระเดช ซึ่งเป็นที่ปรึกษาให้กับโครงการพัฒนาระบบงานชิ้นนี้

ขอกราบขอบพระคุณอาจารย์ วรพจน์ กรีสุระเดช เป็นอย่างสูงที่ได้คอยให้ความช่วยเหลือตลอดจนชี้แนะแนวทางการศึกษาในเรื่องต่างๆไม่ว่าจะเป็นทั้งความรู้ในภาคทฤษฎี และภาคปฏิบัติเป็นอย่างดีเสมอมา

ขอกราบขอบพระคุณบิดา และมารดาอันเป็นที่รักยิ่ง ผู้ที่ให้กำเนิดชีวิต และร่างกายที่สมบูรณ์ นอกจากจะคอยให้กำลังใจตลอดเวลาแล้ว ยังคอยช่วยส่งเสริมเรื่องการศึกษาด้วยดีตลอดมา และยังปลูกฝังคุณความดี และสติปัญญาแก่ข้าพเจ้า เพื่อที่จะนำไปใช้ในชีวิตและการงานให้มีความสำเร็จต่อไปในอนาคต

สุดท้ายนี้ต้องขอขอบคุณบุคคลรอบข้างทั้งพี่ชาย และพี่สาว ตลอดจนเพื่อนๆ IS 16.2 ที่คอยเติมกำลังใจให้กันเสมอมา ข้าพเจ้าหวังว่าโครงการพัฒนาระบบงานฉบับนี้จะเป็นประโยชน์ต่อผู้ที่ได้อ่าน หรือทำการศึกษาไม่มากก็น้อย

ด้วยความเคารพอย่างสูง

นายต่อพงษ์ โลหะรังสิกุล

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

หน้า

ตารางที่

สารบัญรูป

หน้า

รูปที่

สารบัญรูป (ต่อ)

หน้า

รูปที่

บทที่ 1
บทนำ

1.1 ความเป็นมาของโครงการ

เนื่องด้วยสภาพเศรษฐกิจ และสภาวะความเป็นไปทางด้านเศรษฐศาสตร์ในปัจจุบัน ส่งผลให้การเจริญเติบโต และการขยายตัวของตลาดธุรกิจประเภทต่างๆเพิ่มสูงขึ้นตามไปด้วย รวมไปถึงช่องทางใหม่ๆที่เกิดขึ้นในการทำธุรกรรม อาทิเช่น ช่องทางการใช้จ่ายเงินผ่านบัตรเครดิต ช่องทางการโฆษณา การขายสินค้าบนอินเทอร์เน็ต เป็นต้น ช่องทางต่างๆเหล่านี้มีส่วนช่วยทำให้การซื้อและการบริโภคสินค้าประเภทต่างๆเป็นไปอย่างง่ายดายมากยิ่งขึ้น จากความสะดวกสบายในการประกอบธุรกรรมผ่านช่องทางอันหลากหลายเหล่านี้มีผลให้บริษัทต่างๆทั่วโลกต่างก็ให้ความสนใจและความสำคัญกับสิ่งเหล่านี้ เพื่อที่จะได้รองรับความต้องการของลูกค้าที่เพิ่มมากขึ้นได้ พร้อมทั้งอำนวยความสะดวกให้กับลูกค้าเก่าของตนเองที่มีอยู่ นอกจากนี้ยังช่วยดึงดูดลูกค้าหน้าใหม่ที่จะเข้ามาใช้บริการได้อีกด้วย ด้วยเหตุนี้จึงส่งผลให้ในตลาดธุรกิจมีการแข่งขันกันสูงมาก คู่แข่งของแต่ละบริษัทก็พยายามที่จะรักษาลูกค้าของตนเองไว้ รวมไปถึงความพยายามในการที่จะเพิ่มลูกค้าให้มากขึ้นกว่าเดิมด้วย ดังนั้นลักษณะ และพฤติกรรม รวมไปถึงข้อมูลที่เกี่ยวข้องกับลูกค้าจึงมีความหมายอย่างมากในการบริหารจัดการเพื่อที่แต่ละบริษัทจะได้สามารถศึกษา และวางแผนกลยุทธ์ทางการตลาดให้ตรงกับความต้องการของกลุ่มลูกค้าเป้าหมายมากที่สุด

สำหรับการศึกษา และการวิเคราะห์ข้อมูลของลูกค้าจำเป็นที่จะต้องอาศัยความรู้ และเทคนิคเฉพาะด้านของดาต้าไมนิ่ง (Data mining) เข้ามามีส่วนเกี่ยวข้อง แต่เนื่องจากข้อมูลของลูกค้านั้นมีอยู่เป็นจำนวนมาก ซึ่งเกินกว่าความสามารถของมนุษย์เราที่จะทำการวิเคราะห์ได้ทั้งหมด ดังนั้นผู้จัดทำโครงงานพัฒนาระบบนี้จึงได้เลือกที่จะออกแบบ และพัฒนาเครื่องมือที่ใช้ช่วยในการวิเคราะห์ ศึกษาถึงลักษณะ และพฤติกรรมของลูกค้าที่หลากหลาย โดยใช้ความรู้ และเทคนิคทางด้านดาต้าไมนิ่ง (Data mining) มาประยุกต์ใช้กับทฤษฎีฟัซซี่เซต (Fuzzy Set Theory) เพื่อที่จะได้สามารถทำการแบ่งแยกประเภทข้อมูลที่มีความคลุมเครือ หรือความไม่ชัดเจนให้ออกเป็นกลุ่มๆได้อย่างถูกต้อง และเพื่อให้ได้กลุ่มลูกค้าเป้าหมายที่ตรงตามวัตถุประสงค์ที่เรากำหนดได้อย่างแม่นยำ นอกจากนี้ยังคาดหวังว่าจะสามารถนำผลลัพธ์ หรือกลุ่มลูกค้าเป้าหมายที่ได้ไปทำการวิเคราะห์ ประกอบการตัดสินใจในการวางแผนกลยุทธ์ทางการตลาดต่อไปได้ รวมไปถึงผลประโยชน์ในเชิงธุรกิจในอนาคตอีกด้วย

1.2 วัตถุประสงค์ของโครงงานพัฒนาระบบ

1. เพื่อศึกษาถึงกระบวนการในการทำงานของดาต้าไมนิ่ง (Data Mining) และกระบวนการที่ใช้ในการจัดแบ่งกลุ่มลูกค้า (Customer Segmentation) ว่ามีขั้นตอนอะไรบ้าง และแต่ละขั้นตอนมีการทำงานอย่างไร
2. เพื่อให้เข้าใจถึงเทคนิคในการวิเคราะห์ข้อมูลแบบ Cluster Analysis และการประยุกต์ใช้งานกับทฤษฎีฟัซซี่เซต (Fuzzy Set Theory) พร้อมทั้งสามารถเลือกใช้อัลกอริทึมที่เหมาะสมได้
3. เพื่อศึกษาถึงขั้นตอน และกลไกการทำงานของอัลกอริทึม Fuzzy C-Means ซึ่งเป็นอัลกอริทึมหนึ่งที่ใช้ในการวิเคราะห์ จัดกลุ่มข้อมูลว่าเป็นอย่างไร และมีประโยชน์อย่างไรบ้าง รวมไปถึงความสามารถในการนำไปประยุกต์ใช้กับงานประเภทต่างๆ
4. สามารถนำผลลัพธ์ที่ได้ไปวางแผน ทำนายผล หรือสนับสนุนในการทำงานบางอย่างได้ เช่น การวิเคราะห์เกี่ยวกับการใช้งานโทรศัพท์มือถือ เป็นต้น

1.3 ขอบเขตของโครงงานพัฒนาระบบ

สำหรับโครงงานพัฒนาระบบนี้จะเป็นการศึกษาถึงกระบวนการในการทำ Customer Segmentation โดยจะนำความรู้ และกระบวนการทางด้านดาต้าไมนิ่ง (Data mining) รวมไปถึงทฤษฎีฟัซซี่เซต (Fuzzy Set Theory) มาประยุกต์ใช้เข้าด้วยกัน สำหรับเทคนิคที่จะใช้ในการจัดแบ่งกลุ่มลูกค้านั้นจะเป็นแบบ Cluster Analysis ซึ่งเป็นการจัดแบ่งลูกค้าจำนวนมากให้ออกเป็นกลุ่มย่อยๆ ได้ และอัลกอริทึมที่เลือกนำมาใช้ในการวิเคราะห์ข้อมูลของลูกค้ามีชื่อว่า Fuzzy C-Means ซึ่งเป็นอัลกอริทึมหนึ่งที่สามารถทำงาน และจัดการข้อมูลได้ดีสำหรับข้อมูลที่มีความคลุมเครือ และความไม่ชัดเจนอยู่มาก ทำให้เห็นลักษณะของกลุ่มลูกค้าที่มีความแตกต่างกันได้อย่างถูกต้องมากขึ้น

1.4 ขั้นตอนการดำเนินงานของโครงงานพัฒนาระบบ

1. กำหนดวัตถุประสงค์ หรือขอบเขตในการจัดแบ่งกลุ่มลูกค้าให้เป็นที่แน่นอน พร้อมทั้งเก็บรวบรวมข้อมูลที่เกี่ยวข้องกับลูกค้า เช่น ประวัติลูกค้า ลักษณะ และพฤติกรรมต่างๆ เป็นต้น เพื่อที่จะได้กลุ่มลูกค้าที่เป็นผลลัพธ์ได้อย่างถูกต้อง แม่นยำ
2. ศึกษากระบวนการเกี่ยวกับการจัดแบ่งกลุ่มลูกค้า (Customer Segmentation) โดยใช้หลักการ และขั้นตอนการทำงานของดาต้าไมนิ่ง (Data mining)

3. ทำความเข้าใจเกี่ยวกับเทคนิคที่จะใช้ในการจัดแบ่งลูกค้าออกเป็นกลุ่มย่อยๆหรือที่เรียกว่าเทคนิคการทำ Cluster Analysis พร้อมทั้งศึกษาทฤษฎีที่เกี่ยวข้องอันได้แก่ ทฤษฎีฟัซซี่เซต (Fuzzy Set Theory)
4. เลือกใช้อัลกอริทึม Fuzzy C-Means และทำความเข้าใจเกี่ยวกับขั้นตอนการทำงานของอัลกอริทึมในแต่ละขั้นว่าเป็นอย่างไร รวมไปถึงข้อมูลที่จะนำเข้าไปใช้ในอัลกอริทึม และข้อมูลที่จะได้จากอัลกอริทึมด้วย
5. กำหนดแหล่งข้อมูลที่จะนำมาใช้ พร้อมทั้งศึกษาประเภท หรือชนิดข้อมูลนั้นๆว่าเป็นอย่างไร เพื่อจะได้ทำการวิเคราะห์ได้อย่างถูกต้อง สำหรับกรณีศึกษาในโครงงานพัฒนาระบบนี้จะเป็นข้อมูลเกี่ยวกับการใช้งานโทรศัพท์พื้นฐานแบบพกพาของบริษัทโทรคมนาคมแห่งหนึ่ง
6. ออกแบบ และพัฒนาระบบโดยใช้อัลกอริทึม Fuzzy C-Means ตามที่ได้ศึกษามา
7. ตรวจสอบ และประเมินผลที่ได้จากการทำงานของระบบ พร้อมทั้งสรุปผลที่ได้จากการศึกษาว่าเป็นไปตามวัตถุประสงค์ตามที่เรากำหนดไว้ตั้งแต่แรกหรือไม่

1.5 ประโยชน์ที่คาดว่าจะได้รับ

หลังจากที่ได้ทำการศึกษา และพัฒนาโครงงานจนเสร็จสิ้นแล้ว คาดว่าจะได้รับประโยชน์ดังต่อไปนี้

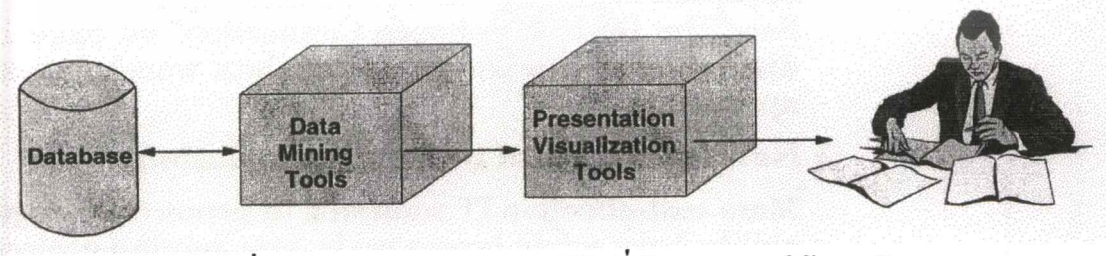
1. ทำให้เข้าใจถึงกระบวนการในการจัดแบ่งกลุ่มลูกค้า (Customer Segmentation) ด้วยหลักการ และขั้นตอนของการทำดาต้าไมนิ่ง (Data mining) อย่างถ่องแท้
2. ด้วยกระบวนการจัดกลุ่มแบบ Cluster Analysis และวิธีการของอัลกอริทึม Fuzzy C-Means จะช่วยให้เราสามารถแบ่งกลุ่มในลูกค้าประเภทต่างๆได้อย่างชัดเจน และถูกต้อง แม่นยำมากขึ้น นอกจากนี้ยังสามารถนำไปประยุกต์ใช้กับธุรกิจประเภทอื่นได้ เช่น ลูกค้าที่เกี่ยวกับธุรกิจบัตรเครดิต หรือลูกค้าที่มาซื้อสินค้ากับห้างสรรพสินค้าต่างๆ เป็นต้น
3. ผลลัพธ์ หรือกลุ่มลูกค้าเป้าหมายที่ได้จากการจัดกลุ่มสามารถนำไปสร้างเป็นข้อสารสนเทศ เพื่อใช้ประกอบการตัดสินใจในการวางแนวทาง การวางแผนกลยุทธ์ทางการตลาดที่บริษัทจะตอบสนองต่อกลุ่มลูกค้าเหล่านั้นได้ นับว่าเป็นประโยชน์อย่างมากทีเดียว นอกจากนี้ก็ยังอาจจะพบข้อสารสนเทศใหม่ๆ หรือความสัมพันธ์ภายในตัวของข้อมูล โดยเฉพาะลักษณะ และพฤติกรรมของลูกค้าที่บริษัทอื่นอาจจะไม่เคยพบมาก่อนก็เป็นได้

บทที่ 2
ทฤษฎีทางด้านดาต้าไมนิ่ง

2.1 ความรู้ทางด้านดาต้าไมนิ่ง (Data mining)

คำจำกัดความหรือความหมายของดาต้าไมนิ่ง (Data mining) ซึ่งเป็นที่ยอมรับโดยทั่วไปนั้น มีความหมายว่า เป็นกระบวนการในการค้นหา และวิเคราะห์จากกลุ่มข้อมูลที่มีขนาดใหญ่ หรือที่มีอยู่เป็นจำนวนมาก เพื่อหาความสัมพันธ์ภายในข้อมูลนั้น หรือเพื่อให้ได้ข้อสารสนเทศบางประการที่เราสนใจออกมา โดยลักษณะของข้อสารสนเทศที่ได้นั้นจะต้องไม่ทราบมาก่อน และจะต้องตรงกับความเป็นจริง หรือสามารถเชื่อถือได้นั่นเอง นอกจากนี้ข้อสารสนเทศที่ได้จะต้องสามารถนำไปใช้ช่วยประกอบแนวทางการตัดสินใจที่ก่อให้เกิดผลดีในการทำธุรกิจต่อไปได้อีกด้วย จะเห็นได้ว่าการทำดาต้าไมนิ่งนั้นจะมีลักษณะที่คล้ายคลึงกับวิธีการวิเคราะห์ข้อมูลทางสถิติในแบบอื่นๆ อยู่ แต่จะมีข้อแตกต่างกันอยู่บ้างยกตัวอย่างเช่น ดาต้าไมนิ่งจะสามารถทำการวิเคราะห์จำนวนข้อมูลที่มีขนาดใหญ่มากๆ ได้ พร้อมทั้งสามารถเรียนรู้ และตรวจสอบข้อมูลที่ผิดพลาดได้ นอกจากนี้ดาต้าไมนิ่งยังรองรับในเรื่องของชนิดข้อมูลได้หลากหลายกว่าอีกด้วย ในขณะที่การวิเคราะห์ทางสถิติทั่วๆ ไปนั้นสามารถรองรับชนิดข้อมูลได้เพียงข้อมูลที่เป็นตัวเลขเท่านั้น และไม่สามารถตรวจจับหรือตรวจสอบข้อมูลที่มีความผิดพลาดได้นั่นเอง (Han Jiawei and Kamber Micheline. 2001)

จากความหมายของดาต้าไมนิ่งทำให้เราสามารถสรุปภาพรวม หรือตำแหน่งของดาต้าไมนิ่งในระบบการใช้งานจริงได้ดังรูป



รูปที่ 2.1 แสดงภาพรวมของดาต้าไมนิ่งในระบบการใช้งานจริง

จากรูปนั่นก็หมายความว่าเราสามารถนำดาต้าไมนิ่งไปเป็นเครื่องมือในการวิเคราะห์ข้อมูลและประยุกต์ใช้ในทางธุรกิจเพื่อประกอบการตัดสินใจในรูปแบบต่างๆ ได้อย่างมากมาย ตัวอย่างธุรกิจที่มีการนำดาต้าไมนิ่งไปใช้จริงแล้วได้แก่ (Cabena Peter. et al. 1997)

❖ **ธุรกิจการบริหารจัดการด้านการตลาด (Market Management Applications)** สำหรับธุรกิจทางด้านนี้มักจะมีการนำดาต้าไมนิ่งไปใช้กันอย่างแพร่หลายเป็นอย่างมาก และรู้จักกันเป็นอย่างดี ตัวอย่างเช่น

- การศึกษา และการวิเคราะห์ชีวิตความเป็นอยู่ของคน หรือที่เราเรียกว่า "Lifestyles Studies" เรื่องที่จะศึกษานั้นจะเริ่มตั้งแต่อายุ เพศ รายได้ ความสนใจ งานอดิเรกต่างๆ เป็นต้นเพื่อที่จะได้สามารถจัดโครงการ หรือโปรโมชั่นได้ตรงกับความต้องการของบุคคลนั้นๆออกมา
- การศึกษาลักษณะ หรือรูปแบบการซื้อสินค้าของลูกค้าในแต่ละวัย หรือสถานะการแต่งงานของคน ตัวอย่างเช่น ถ้าแต่งงานแล้วอาจจะต้องการกู้เงินซื้อบ้าน หรือต้องการซื้อประกันชีวิต เป็นต้น
- การทำ Cross-selling นั่นคือการดึงดูดลูกค้าให้มาสนใจในผลิตภัณฑ์ชนิดหนึ่งเพื่อที่จะได้สนใจสินค้าอีกชนิดหนึ่งซึ่งมีความเกี่ยวข้องกันตามไปด้วย ตัวอย่างเช่น ลูกค้าที่มีความต้องการในการฝากเงินกับธนาคารแห่งหนึ่ง ธนาคารก็มักจะยื่นข้อเสนอในการทำบัตรเครดิตให้กับลูกค้าด้วยเสมอ
- การสำรวจรายการสั่งซื้อทางโทรศัพท์ เพื่อปรับปรุงให้ตรงกับความต้องการของลูกค้า ยกตัวอย่างเช่น การโทรศัพท์สั่งพิซซ่า ถ้ามีการบันทึกว่าส่วนใหญ่ลูกค้าชอบสั่งพิซซ่าหน้าอะไรไว้ คราวต่อไปเมื่อลูกค้าโทรศัพท์เข้ามาอีกก็สามารถที่จะเสนอรายการ หรือโปรโมชั่นที่เหมาะกับลูกค้าท่านนั้นได้
- การออกบัตรประเภทต่างๆให้กับลูกค้า แล้วกำหนดสิทธิให้กับบัตรแต่ละประเภทไม่เหมือนกัน ยกตัวอย่างเช่น บัตรเครดิตที่ใช้เป็นส่วนลดในการซื้อสินค้าตามห้างสรรพสินค้า บัตรเครดิตที่ใช้เป็นส่วนลดตามร้านอาหาร เป็นต้น ดังนั้นจะต้องมีการศึกษาข้อมูลของลูกค้าเสียก่อนว่าลูกค้าแต่ละท่านนั้นชอบเกี่ยวกับอะไรเป็นพิเศษ
- ด้านเกี่ยวกับอัตราค่าเบี้ยประกันภัย เพราะลูกค้าแต่ละระดับมีเงินเดือนไม่เท่ากัน ดังนั้นจึงมีความสามารถในการจ่ายค่าเบี้ยประกันภัยได้ไม่เท่ากันด้วย จึงเป็นเหตุให้ต้องทำการวิเคราะห์ข้อมูลของลูกค้าต่างๆก่อนจะยื่นข้อเสนอในการทำประกันภัย ในบางครั้งอาจจะต้องรู้ด้วยว่าควรจะเลือกประกันภัยประเภทใดให้แก่ลูกค้าท่านนั้น

❖ **ธุรกิจการบริหารจัดการด้านความเสี่ยง (Risk Management Applications)** ความเสี่ยงที่ว่านี้ไม่ได้เกี่ยวข้องกับการประกันภัย หรือการลงทุนแต่อย่างใด แต่จะหมายความถึง

ความเสี่ยงที่เกิดขึ้นอันเนื่องมาจากการขยายตัวของคู่แข่งทางธุรกิจ ความเสี่ยงของคุณภาพสินค้าที่ลดต่ำลง และความเสี่ยงที่เกิดจากการที่จำนวนลูกค้าลดลง ดังนั้นธุรกิจเหล่านี้จึงจำเป็นที่จะต้องใช้ดาต้าไมนิ่ง (Data mining) เข้ามีส่วนช่วยในการวิเคราะห์และทำนาย (Forecasting) ว่าแนวโน้มในอนาคตควรจะเป็นอย่างไร เพื่อจะได้สามารถตัดสินใจในการบริหารจัดการได้อย่างถูกต้อง ตัวอย่างธุรกิจเหล่านี้ได้แก่

- การทำนาย หรือคาดเดาเกี่ยวกับกับการเงิน(Financial) ยกตัวอย่างเช่น การกำหนดราคา (Pricing) ของสินค้าชนิดต่างๆว่าควรที่จะมีแนวโน้มในการปรับราคาเป็นอย่างไรต่อไปในอนาคต ขึ้นราคาดี หรือว่าจะลดราคาสินค้าลงเพื่อให้เข้ากับสภาวะตลาด รวมไปถึงผลกำไรขาดทุนของสินค้าชนิดนั้นๆด้วย

❖ **ธุรกิจการบริหารจัดการตรวจสอบความผิดพลาด (Fraud Management Applications)** ตัวอย่างธุรกิจที่มีการนำดาต้าไมนิ่งไปใช้กับธุรกิจทางด้านนี้ได้แก่

- ระบบการตรวจสอบเกี่ยวกับการรักษาคนไข้ โดยจะนำประวัติการรักษาของคนไข้แต่ละคนที่เก็บไว้มาทำการวิเคราะห์ว่าเคยรักษากับแพทย์ทางด้านไหนมาบ้าง และที่ผ่านมานั้นเคยรักษาด้วยวิธีการอย่างไร รวมไปถึงคนไข้นั้นเคยใช้ตัวยาอะไรบ้างในการรักษา และมีอาการแพ้ยาเหล่านั้นหรือไม่ ข้อมูลเหล่านี้มีความสำคัญเป็นอย่างมากเพื่อที่ว่าครั้งต่อไปเมื่อคนไข้เข้ามารับการรักษากับทางโรงพยาบาล แพทย์จะได้สามารถวินิจฉัยโรคได้อย่างรวดเร็วแม่นยำ และเหมาะสม นอกจากนี้ระบบนี้ยังมีส่วนช่วยลดต้นทุน และค่าใช้จ่ายอื่นๆภายในโรงพยาบาลอีกด้วย

❖ **ด้านอื่นๆ และที่กำลังจะเกิดขึ้นในอนาคต (Emerging and Future Application Areas)** ปัจจุบันดาต้าไมนิ่งถูกนำไปใช้เกี่ยวกับข้อมูลทางการค้ามากขึ้น และมีแนวโน้มว่าต่อไปจะเพิ่มขึ้นเรื่อยๆอีกด้วย สำหรับเรื่องที่เห็นได้ชัดในปัจจุบันได้แก่

- **Text mining** เป็นการวิเคราะห์ และเก็บรวบรวมข้อมูลต่างๆที่เป็น text ซึ่งเก็บอยู่ในฐานข้อมูลขนาดใหญ่ เพื่อที่จะได้สามารถอ้างอิง หรืออ้างถึงเอกสารฉบับนั้นได้อย่างรวดเร็ว ด้วยหลักการที่ว่านี้จึงได้มีผู้คิดค้นนำไประยุกต์ใช้กับการตอบจดหมายอิเล็กทรอนิกส์ (E-mail) ซึ่งถูกส่งเข้ามาที่ศูนย์บริการลูกค้าสัมพันธ์ (Call Center) โดยที่ระบบจะทำการตอบจดหมายกลับไปหายังลูกค้าโดยอัตโนมัติเมื่อได้รับจดหมายอิเล็กทรอนิกส์จากลูกค้าโดยปราศจากการพึ่งพาให้คนเข้ามาเปิดอ่าน เป็นต้น แต่อย่างไรก็ตามการวิเคราะห์ในลักษณะนี้

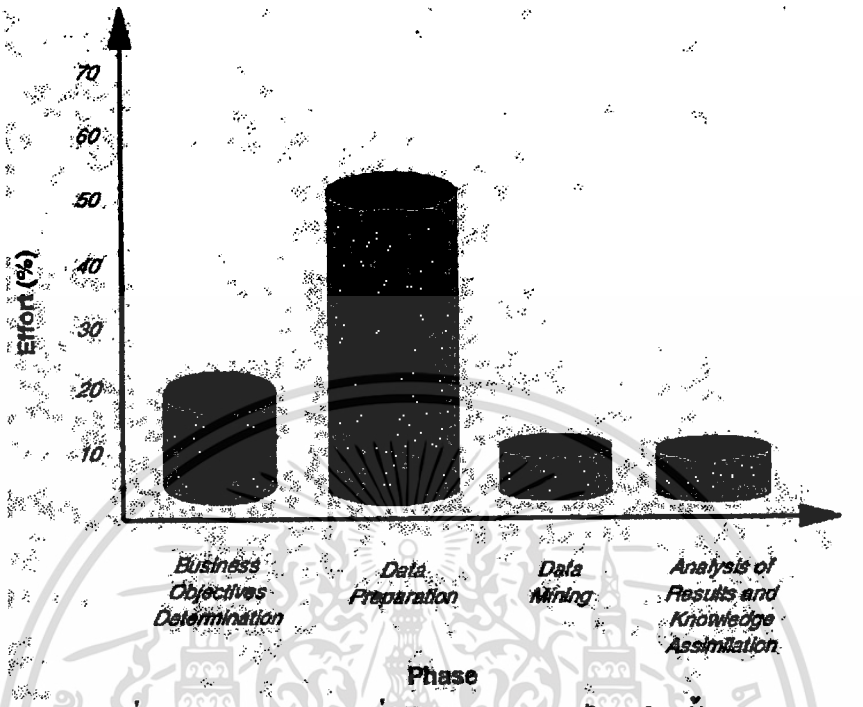
ก็ยังมีปัญหาอยู่บ้างกับการใช้งานในบางเรื่อง แต่ก็มีโอกาสที่จะนำมาใช้มากขึ้นในอนาคตอันใกล้นี้

- **Web analytics** เป็นการนำดาต้าไมนิ่งไปประยุกต์ใช้กับข้อมูลบนอินเทอร์เน็ตนั่นเอง โดยการสังเกตการณ์จากรูปแบบการใช้งานของผู้ใช้ว่ามีลักษณะการใช้งานบนเว็บไซต์ไหนบ้าง เพื่อที่จะได้ทำการวิเคราะห์ และคาดเดาการใช้งานของผู้ใช้ในครั้งต่อๆไปได้ และยังช่วยในการดึงเอาเว็บไซต์ที่เกี่ยวข้องกันกับสิ่งที่ผู้ใช้ต้องการหานั้นออกมาแสดงผลพร้อมๆกันอีกด้วย ซึ่งจะมีผลต่อการตลาด การประชาสัมพันธ์ และการโฆษณาต่างบนเว็บไซต์ต่อไป

2.1.1 กระบวนการของการทำดาต้าไมนิ่ง (The Data Mining Process)

คนส่วนใหญ่มักจะคิดว่ากระบวนการในการทำดาต้าไมนิ่งนั้นจะเริ่มจากการค้นหาสิ่งที่เราสนใจจากกลุ่มข้อมูลที่มีขนาดใหญ่เลย แต่ในความเป็นจริงกลับไม่เป็นอย่างนั้น ขั้นตอนที่เป็นการสืบค้นข้อมูลเพื่อให้ได้ข้อสารสนเทศ (mine data) จริงๆนั้นกลับเป็นเพียงส่วนหนึ่งของกระบวนการทั้งหมด สำหรับกระบวนการในการทำดาต้าไมนิ่งแล้วสิ่งที่สำคัญที่สุดก็คือเพื่อตอบสนองต่อวัตถุประสงค์ในการทำธุรกิจนั่นเอง ดังนั้นเราจึงสามารถทำการวัด และตรวจสอบการทำดาต้าไมนิ่งได้ว่ามีความถูกต้องมากน้อยเพียงใดด้วยการเปรียบเทียบผลของข้อสารสนเทศที่ได้ว่ามีคุณค่ากับวัตถุประสงค์ของธุรกิจนั้นๆหรือไม่ และมากน้อยเพียงใด

กระบวนการในการทำดาต้าไมนิ่งนั้นสามารถแบ่งออกได้เป็น 5 ขั้นตอนด้วยกัน (Cabena Peter. et al. 1997) ซึ่งในแต่ละขั้นตอนนั้นต่างก็มีความสำคัญที่แตกต่างกันออกไป และใช้ระยะเวลาในการวิเคราะห์ข้อมูลไม่เท่ากัน ซึ่งสามารถแสดงได้ดังภาพ



รูปที่ 2.2 แสดงระยะเวลาที่ใช้ในการวิเคราะห์ในแต่ละขั้นตอน

จากรูปจะเห็นได้ว่าขั้นตอนที่มีระยะเวลาในการทำงานยาวนานมากที่สุดในกระบวนการของดาต้าไมนิ่งนั่นก็คือ การเตรียมข้อมูล (Data Preparation) นั่นเอง ซึ่งเราจะอธิบายต่อไปว่าเหตุใดขั้นตอนนี้จึงกินระยะเวลายาวนานมากกว่าขั้นตอนอื่นๆ ในการทำดาต้าไมนิ่งจะประกอบด้วยกัน 5 ขั้นตอนดังต่อไปนี้

1. การกำหนดวัตถุประสงค์ทางธุรกิจ (Business Objectives Determination)

ก่อนที่จะลงมือทำการวิเคราะห์ข้อมูลได้นั้น เราจำเป็นที่จะต้องทำความเข้าใจถึงปัญหาและความต้องการของธุรกิจนั้นๆเสียก่อนเพื่อที่จะได้สามารถกำหนด หรือระบุปัญหาในธุรกิจที่พบเจอได้ นอกจากนี้ยังช่วยเป็นตัวกำหนดแนวทาง และทิศทางในการทำดาต้าไมนิ่งได้อย่างถูกต้องต่อไป ถ้าเรากำหนดวัตถุประสงค์ที่ไม่ชัดเจนแล้ว อาจจะส่งผลให้การวิเคราะห์ข้อมูลในขั้นตอนต่อไปเกิดความคลุมเครือ ซึ่งจะนำไปสู่ผลลัพธ์ที่คลาดเคลื่อนไม่ตรงกับวัตถุประสงค์ และไม่สามารถนำไปใช้ช่วยในการตัดสินใจได้ นอกจากนี้ยังอาจทำให้เสียเวลาไปโดยเปล่าประโยชน์ โดยจะต้องกลับไปเริ่มต้นใหม่อีกครั้งหนึ่ง

2. ขั้นตอนการเตรียมข้อมูล (Data Preparation)

ในกระบวนการทำดาต้าไมนิ่งทั้งหมด ขั้นตอนการเตรียมข้อมูลจะเป็นขั้นตอนซึ่งกินระยะเวลามากที่สุดประมาณ 60% ของกระบวนการทั้งหมด สำหรับขั้นตอนการเตรียมข้อมูลนี้จะแบ่งออกเป็นขั้นตอนย่อยๆ 3 ขั้นตอน (Cabena Peter. et al. 1997) ด้วยกันประกอบด้วย

1) **การเลือกข้อมูล (Data Selection)** เป็นการกำหนด และคัดเลือกแหล่งของข้อมูลที่เราสนใจ หรือที่เราจำเป็นต้องใช้ในการศึกษา เพื่อที่จะใช้วิเคราะห์ในขั้นตอนของการทำไมนิ่งต่อไป การคัดเลือกข้อมูลที่จะนำมาใช้จากแหล่งข้อมูลแต่ละแห่งนั้น เราจะต้องทำความเข้าใจกับเนื้อหา หรือรายละเอียดของข้อมูล ชนิดของข้อมูล และค่าที่เป็นไปได้ของข้อมูล รวมไปถึงลักษณะ และรูปแบบของข้อมูลอื่นๆอีกด้วย ซึ่งสามารถแบ่งประเภท หรือลักษณะของข้อมูลที่พบเจอได้ดังต่อไปนี้

- **ลักษณะของข้อมูลที่สามารถจัดประเภทได้ (Categorical)** จะแบ่งออกได้เป็น 2 ประเภทด้วยกันได้แก่
    a) **Nominal Variables** คือข้อมูลที่สามารถแบ่งประเภทได้ แต่ไม่สามารถเรียงลำดับ หรือมีลำดับอยู่ในตนเอง ยกตัวอย่างเช่น สถานะภาพแต่งงาน ค่าที่เป็นไปได้คือ โสด (Single), แต่งงานแล้ว (Married), หย่าร้าง (Divorced) หรือสถานะของเพศ ค่าที่เป็นไปได้คือ เพศชาย (Male), เพศหญิง (Female) เป็นต้น
    b) **Ordinal Variables** จะแตกต่างกับ Nominal Variables ตรงที่ข้อมูลประเภทนี้สามารถเรียงลำดับค่าในตัวของมันเองได้ ยกตัวอย่างเช่น ระดับเครดิตของลูกค้า (Customer Credit Rating) ค่าที่เป็นไปได้คือ ดี (Good), ทั่วไป (Regular), แย่ (Poor) เป็นต้น
- **ลักษณะข้อมูลที่สามารถวัดและบอกปริมาณได้ (Quantitative)** ข้อมูลประเภทนี้สามารถแบ่งออกได้เป็น 2 ลักษณะด้วยกันได้แก่
    a) **Continuous** เป็นข้อมูลที่มีค่าเป็นจำนวนจริง ดังนั้นค่าที่เป็นไปได้ในของมันจึงมีความต่อเนื่อง ยกตัวอย่างเช่น รายได้ (income) เป็นต้น
    b) **Discrete** เป็นข้อมูลที่มีลักษณะของค่าในตัวแบบไม่ต่อเนื่อง ค่าที่เป็นไปได้จึงเป็นจำนวนเต็ม ยกตัวอย่างเช่น จำนวนของพนักงาน (Number of Employees) เป็นต้น

การคัดเลือกข้อมูลที่ดี และถูกต้องนั้นจะต้องเลือกให้เหมาะสม และตรงกับปัญหา หรือความต้องการที่เราได้กำหนดไว้ในขั้นตอนแรกด้วย เพราะมิเช่นนั้นแล้วอาจทำให้การไมนิ่งข้อมูลได้ผลลัพธ์ที่คลาดเคลื่อน และไม่สามารถที่จะทำการประเมินผลได้อย่างถูกต้องตรงกับสิ่งที่เราสนใจได้นั่นเอง นอกจากนี้แล้วเพื่อให้ได้ข้อมูลที่มีความทันสมัยอยู่เสมอก็จำเป็นที่จะต้องมีการ

ตรวจสอบค่าที่เป็นไปได้ของข้อมูลประเภทนั้นอีกด้วย เพราะข้อมูลบางประเภทจะมีการเปลี่ยนแปลงไปตามกาลเวลา เช่น ความชอบของลูกค้าแต่ละคน เป็นต้น

ข้อมูลที่ถูกคัดเลือกออกมานั้นจะถูกเรียกว่า "Active Variables" (Cabena Peter. et al. 1997) และข้อมูลเหล่านี้จะต้องสามารถที่จะช่วยให้เราแยกความแตกต่าง จัดแบ่งกลุ่มได้ หรือแม้กระทั่งทำการคาดเดาเรื่องที่จะเกิดขึ้นในอนาคตได้อีกด้วย ข้อมูลบางส่วนที่ถูกคัดเลือกมาใช้นั้นบางครั้งอาจไม่จำเป็นต้องใช้ทั้งหมด มีเพียงแค่บางส่วนที่จะถูกนำไปใช้ในขั้นตอนของการไมนิ่งเท่านั้น แต่ถึงแม้ข้อมูลเหล่านั้นจะไม่ได้ถูกนำไปใช้ก็ไม่ได้หมายความว่าข้อมูลนั้นจะไม่มีประโยชน์ ข้อมูลที่เหลืออาจจะสามารถใช้ในการช่วยอธิบายผลลัพธ์ที่ได้จากการไมนิ่งในขั้นตอนของการวิเคราะห์และประเมินผลด้วยก็เป็นได้

2) **การเตรียมข้อมูลก่อนการนำไปประมวลผล (Data Preprocessing)** ในขั้นตอนนี้จะเป็นการวัด และตรวจสอบดูว่าข้อมูลที่ผ่านการคัดเลือกมานั้นมีคุณภาพมากน้อยเพียงใดสามารถที่จะนำไปทำการไมนิ่งแล้วจะได้ผลลัพธ์ที่ควรจะเป็นหรือไม่ ซึ่งทำให้เป็นปัญหาอย่างมากในขั้นตอนของการเตรียมข้อมูล เนื่องจากข้อมูลที่ได้มาจากแหล่งต่างๆนั้นไม่ได้ถูกออกแบบมาให้ใช้กับการทำดาต้าไมนิ่งตั้งแต่แรก

สำหรับเทคนิคที่ใช้ในการวัดคุณภาพของข้อมูลที่คัดเลือกมานั้น โดยปกติจะใช้วิธีการคำนวณทางสถิติเข้ามาช่วย และอาจนำมาแสดงผลในรูปแบบของกราฟ หรือแผนภาพเพื่อทำการวิเคราะห์อีกครั้งหนึ่ง การสุ่มตัวอย่างข้อมูลขึ้นมาก็เป็นวิธีหนึ่งที่มักจะนิยมใช้กันเพื่อที่จะวัดคุณภาพ เทคนิคที่นิยมใช้ในการวัดและตรวจสอบข้อมูลได้แก่

- ในกรณีที่ข้อมูลที่คัดเลือกมามีลักษณะที่เป็นแบบสามารถจัดประเภทได้ (Categorical Variables) การวัดคุณภาพของข้อมูลก็จะพิจารณาจากขอบเขตของข้อมูลชนิดนั้นๆ โดยการนำค่าที่เป็นไปได้ของข้อมูลมาแจกแจง และแบ่งเป็นประเภทต่างๆว่า แต่ละประเภทนั้นมีจำนวนข้อมูลเป็นเท่าไหร่ แล้วจึงนำไปแสดงผลเป็นแผนภาพในรูปแบบต่างๆ เช่น แผนภาพวงกลม หรือกราฟแท่งเปรียบเทียบกัน เป็นต้น เพื่อที่จะได้สามารถตรวจสอบความผิดพลาดของข้อมูล หรือค่าที่หายไปได้ (Missing Values)
- สำหรับลักษณะข้อมูลที่สามารถวัดและบอกปริมาณได้นั้น (Quantitative Variables) ส่วนใหญ่ก็จะใช้วิธีการวัดจากค่าสูงสุด (Maxima) และค่าต่ำสุด (Minima) หรืออาจจะดูจากค่าเฉลี่ย (Average) ความถี่ในการเกิดของข้อมูล ค่ากึ่งกลาง และค่าเบี่ยงเบนมาตรฐานต่างๆของข้อมูล เพื่อที่จะได้สามารถทำการวิเคราะห์หาแนวโน้ม และความผิดปกติของชุดข้อมูลนั้นๆได้

สำหรับความไม่สมบูรณ์ หรือความผิดปกติของข้อมูลส่วนใหญ่ที่มักจะพบเป็นประจำ สามารถแบ่งออกได้เป็น 3 ลักษณะดังต่อไปนี้ (Han Jiawei and Kamber Micheline. 2001)

- **Incomplete Data/Missing Values** หมายถึงการที่ค่าของข้อมูล หรือกลุ่มข้อมูลที่เราคัดเลือกมาใช้นั้นบางค่านั้นขาดหายไป หรือบางทีอาจจะไม่มีข้อมูลก็เป็นได้ การเกิดความผิดปกติกับข้อมูลในรูปแบบนี้นั้นอาจจะเกิดจากการที่ ณ ขณะนั้นไม่มีค่าของข้อมูลจริงๆ หรืออาจจะเกิดจากการที่นำเอาแหล่งข้อมูลจากที่อื่นมาใช้ในระบบการเก็บข้อมูลของเราแล้วเกิดการไม่เข้ากันของข้อมูลส่งผลให้ค่าของข้อมูลบางกลุ่มนั้นก็จะหายไป ยกตัวอย่างเช่น การอิมพอร์ต (Import) ข้อมูลจากแหล่งข้อมูลหนึ่งมาใส่ไว้ที่อีกแหล่งข้อมูลหนึ่ง เป็นต้น
- **Noisy Data** หมายถึง ค่าของข้อมูลซึ่งมีความผิดปกติไปจากที่ควรจะเป็น ยกตัวอย่างเช่น ข้อมูลที่ควรจะเก็บเป็นตัวเลข(Number) แต่ค่าที่เก็บอยู่จริงอาจจะเป็นตัวอักษร (Character) เป็นต้น ข้อมูลที่มีความผิดปกติเช่นนี้เราจะเรียกว่า "Outlier" แต่ความผิดปกติที่พบนี้ก็อาจจะเป็นเรื่องดีก็เป็นได้ เช่น ถ้าข้อมูลที่เราคัดเลือกมาใช้เป็นรหัสของสินค้า และมีการเพิ่มรหัสของสินค้าตัวใหม่เข้าไปซึ่งเป็นรหัสที่เราไม่เคยรู้จักมาก่อน ความผิดปกตินี้ก็จะไม่มีผลกระทบต่อการทำไมนิ่ง เราก็เพียงแต่ปรับปรุงข้อมูลที่เราเคยรู้จักให้เข้าใจว่ามีรหัสสินค้าตัวใหม่เพิ่มเข้ามาแล้ว แต่ถ้าเป็นความผิดปกติในแง่ที่ไม่ดีอย่างเช่น ข้อมูลที่เป็นอายุของคน ถ้าเราป้อนเข้าไปเก็บเป็น 650 หรือตัวเลขที่ติดลบก็จะทำให้ค่าของข้อมูลที่ได้ผิดไปจากความเป็นจริง ซึ่งจะนำไปสู่การวิเคราะห์ที่ไม่ถูกต้องอีกด้วย ดังนั้นข้อมูลเหล่านี้จึงจำเป็นที่เราจะต้องกำจัดออกไปเสียก่อนที่จะไปถึงการทำไมนิ่งข้อมูล
- **Inconsistent Data** หมายถึง ข้อมูลที่ถูกอ้างถึงนั้นอาจจะหายไป หรือถูกลบออกไปจากฐานข้อมูลแล้วทำให้เราไม่สามารถนำข้อมูลที่จำเป็นต้องใช้เหล่านั้นมาใช้งานได้

กระบวนการในการกำจัดความผิดปกติ หรือความไม่สมบูรณ์ของข้อมูลนี้นั้นเราจะเรียกว่า Data Cleaning ซึ่งเมื่อข้อมูลผ่านกระบวนการนี้แล้วก็จะช่วยให้ข้อมูลที่ได้นั้นสามารถทำการวิเคราะห์ได้อย่างถูกต้อง และมีความแม่นยำมากขึ้น

3) **การแปลงข้อมูล (Data Transformation)** เป็นขั้นตอนที่จะมีการปรับเปลี่ยน หรืออาจจะมีการสร้างโครงสร้างของข้อมูลเสียใหม่ให้อยู่ในรูปแบบที่เหมาะสมกับแต่ละเทคนิค หรืออัลกอริทึมที่ใช้ในกระบวนการของการทำดาต้าไมนิ่ง เพื่อให้สามารถทำการวิเคราะห์ต่อไปได้ง่ายดายยิ่งขึ้นนั่นเอง วิธีการปรับเปลี่ยนเหล่านี้มีอยู่ด้วยกันหลายวิธีด้วยกัน ซึ่งอาจจะเป็นแค่เพียงการแปลงข้อมูลอย่างง่าย (Simple

Data Conversion) หรืออาจจะเป็นกระบวนการทางสถิติที่ใช้ในการปรับเปลี่ยน และมีความซับซ้อนมากขึ้นก็เป็นไปได้ ยกตัวอย่าง การแปลงข้อมูลอย่างง่ายๆเช่น การคำนวณอายุของคนจากวันเดือนปีเกิด การคิดค่าเฉลี่ยจากฟิลด์ข้อมูลต่างๆ เป็นต้น

วิธีการที่ช่วยในการปรับเปลี่ยนข้อมูล เพื่อที่จะช่วยเพิ่มประสิทธิภาพในการวิเคราะห์ข้อมูล และได้รับความนิยมมีดังต่อไปนี้

- **Data Reduction and Integration** จากสถิติการใช้งานนับว่าวิธีนี้เป็นวิธีการที่ได้รับความนิยมมากที่สุดก็ว่าได้ จุดประสงค์หลักของวิธีการนี้ก็คือการรวบรวม และสรุปข้อมูลให้เหลือตัวแปร หรือจำนวนของข้อมูลที่น้อยลงเพื่อให้เหมาะสมกับการทำงานของอัลกอริทึมในการทำไมนิ่ง ซึ่งอาจจะเป็นการช่วยลดความซ้ำซ้อนของข้อมูลได้อีกทางหนึ่งด้วย ยกตัวอย่างการเก็บรายละเอียดข้อมูลของลูกค้าเกี่ยวกับการจำแนกแยกแยะระดับความสำคัญของลูกค้า อาทิเช่น ลูกค้าใหม่ ลูกค้าที่มีระดับสิทธิพิเศษ (Premium Level) เป็นต้น ส่วนใหญ่เรามักจะพิจารณาจากรายได้ ระดับการศึกษา หรือที่อยู่ของลูกค้าแล้วสรุปไปเป็นระดับความสำคัญของลูกค้าได้ อย่างไรก็ตามวิธีการทำ Data Reduction and Integration นั้นก็ยังมีจุดอ่อน หรือข้อด้อยอยู่นั่นเองก็คือ การตีความหรือการที่จะตัดสินใจว่าจะทำการพิจารณารวมข้อมูลอะไรบ้าง แล้วสรุปให้เหลือเพียงข้อมูลที่เป็นตัวแปรเพียงตัวเดียวนั้นทำได้ยาก เพราะในบางครั้งการรวมหรือยุบกลุ่มของข้อมูลให้เหลือเพียงข้อมูลเพียงตัวเดียวในการพิจารณาอาจจะทำให้เกิดความคลาดเคลื่อน หรือการสูญเสียข้อมูลที่สำคัญที่เราต้องการไปก็เป็นได้
- **Data remodeling and refining** เป็นวิธีการปรับเปลี่ยนโครงสร้างข้อมูล เพื่อทำให้ขั้นตอนของการทำ Data Transformation นั้นเกิดความรวดเร็ว และมีประสิทธิภาพเพิ่มมากขึ้นนั่นเอง
- **Discretization** เป็นเทคนิคหนึ่งในการปรับเปลี่ยนประเภทข้อมูลที่มีลักษณะเชิงปริมาณ (Quantitative variables) ให้กลายมาเป็นลักษณะข้อมูลที่สามารถทำการจำแนกให้เป็นประเภทได้ (Categorical variables) โดยการแบ่งช่วงของข้อมูลออกเป็นช่วงๆ แล้วแทนด้วยค่าคงที่ค่าหนึ่งให้กับช่วงข้อมูลนั้น ยกตัวอย่างการแบ่งอัตราเงินเดือนออกเป็นช่วงต่างๆกัน อาทิเช่น 0 – 5,000 บาทให้มีค่าเท่ากับ 1 ช่วง 5,001 – 10,000 บาทให้มีค่าเท่ากับ 2 เป็นต้น

นอกจากวิธีที่กล่าวมาข้างต้นแล้วนั้นก็ยังมีวิธีการที่ได้รับการยอมรับในแบบอื่นๆอีก เช่น วิธีการแปลงข้อมูลประเภทที่สามารถจำแนกได้ให้อยู่ในรูปของการแทนด้วยตัวเลข ซึ่งส่วนใหญ่วิธี

นี้มักจะใช้เฉพาะกับการแก้ปัญหาของดาต้าไมนิ่งที่เป็นแบบนิวรอลเน็ตเวิร์ค (Neural Network) เพราะการแก้ปัญหาแบบนี้นั้นจะใช้ข้อมูลนำเข้าที่อยู่ในลักษณะที่เป็นตัวเลขมากกว่าตัวอักษรทั่วไป ดังนั้นจึงต้องมีการแปลงข้อมูลเหล่านั้นให้อยู่ในรูปของตัวเลขเสียก่อนจึงจะสามารถทำการวิเคราะห์ในส่วนอัลกอริทึมของนิวรอลเน็ตเวิร์ค (Neural Network) ต่อไปได้ ตัวอย่างการแปลงข้อมูลเหล่านี้อาทิเช่น ประเภท หรือยี่ห้อของรถต่างๆ ได้แก่ ยี่ห้อฟอร์ด (Ford), ยี่ห้อโตโยต้า (Toyota), ยี่ห้อนิสสัน (Nissan) ซึ่งยี่ห้อรถประเภทต่างๆต่อไปนี้นั้นเราจะแทนด้วยค่าของตัวเลขที่มีค่าต่างกันไปดังนี้ ยี่ห้อฟอร์ดให้มีค่าเท่ากับ 100, ยี่ห้อโตโยต้าให้มีค่าเท่ากับ 010, ยี่ห้อนิสสันให้มีค่าเท่ากับ 001 เหล่านี้เป็นต้น

3. การทำดาต้าไมนิ่ง (Data Mining)

เป็นขั้นตอนของการไมนิ่งข้อมูลนั่นเอง ซึ่งในความเป็นจริงแล้วก็คือการนำข้อมูลที่ได้จากการเตรียมข้อมูลเหล่านั้นมาประยุกต์ใช้ และส่งเข้าไปทำงานภายในอัลกอริทึมของดาต้าไมนิ่งที่เรากำหนดไว้ เพื่อที่จะได้สามารถนำผลลัพธ์ที่ได้ไปสู่ขั้นตอนของการวิเคราะห์ และการตีความ (Analysis of Results) และสามารถนำไปใช้งานให้เกิดประโยชน์มากที่สุดต่อไปได้

การที่เราต้องแยกขั้นตอนการทำไมนิ่งข้อมูลออกจากขั้นตอนของการทำการวิเคราะห์ผลลัพธ์(Analysis of Results) และการเตรียมข้อมูล (Data Preprocessing) นั่นก็เพราะว่าในบางครั้งที่เราทำการไมนิ่งข้อมูล หรือประมวลผลข้อมูลด้วยอัลกอริทึมไปแล้วนั้น เมื่อนำผลลัพธ์ที่ได้ทำการวิเคราะห์เราอาจจะยังไม่ได้สิ่งที่เราสนใจ หรืออาจจะยังไม่สามารถตีความผลลัพธ์ได้ตรงตามวัตถุประสงค์ที่เราได้กำหนดไว้ตั้งแต่ตอนแรก ดังนั้นเราจึงอาจจะต้องย้อนกลับไปทำในส่วนของการทำไมนิ่งใหม่อีกครั้งหนึ่ง หรืออาจจะต้องย้อนกลับไปเริ่มทำใหม่ตั้งแต่ขั้นตอนการเตรียมข้อมูลก็เป็นไปได้ ส่งผลให้ในกระบวนการของการทำดาต้าไมนิ่งอาจมีการทำงานซ้ำไปซ้ำมาในแต่ละขั้นตอนหลายครั้ง ทั้งนี้ทั้งนั้นก็เพื่อที่จะทำให้ได้ข้อสารสนเทศที่เป็นประโยชน์ต่อการนำไปใช้งานหรือสิ่งที่เราสนใจกับวัตถุประสงค์ที่เรากำหนดมากที่สุด

สำหรับอัลกอริทึมที่เราเลือกใช้ในการทำไมนิ่งข้อมูลนั้นในบางครั้งเราอาจจะจำเป็นต้องใช้จำนวนของอัลกอริทึมที่ช่วยในการวิเคราะห์มากกว่าหนึ่งอัลกอริทึมก็เป็นไปได้ ทั้งนี้ก็ขึ้นอยู่กับรูปแบบ หรือลักษณะของแอปพลิเคชั่นที่เราทำงานด้วยกับอัลกอริทึมนั้นๆ ยกตัวอย่างเช่น แอปพลิเคชั่นที่ทำงานเกี่ยวกับการจัดการ การจำแนกแบ่งกลุ่มข้อมูล (Database segmentation) ก็อาจจะมีการใช้งานอัลกอริทึมที่มากกว่าหนึ่งอัลกอริทึมก็ได้ หรือแอปพลิเคชั่นที่ทำงานเกี่ยวกับการทำนาย การคาดเดาต่างๆ (Predictive Modeling) ซึ่งจะมีลักษณะการทำงานเป็นวงจร นั่นคือมีการทำงานเกี่ยวกับข้อมูลที่ถูกสุ่มออกมาซ้ำหลายครั้ง (Training phase) ก่อนที่จะไปทำงานกับข้อมูลจริงๆ ตรงส่วนการทำงานนี้นี่เองที่ทำให้เราต้องใช้อัลกอริทึมมากกว่าหนึ่งอัลกอริทึมขึ้นไปเข้ามา

ช่วยจัดการในเรื่องของการทำงาน การวิเคราะห์ข้อมูลต่างๆ เป็นต้น แต่อย่างไรก็ตามการเลือกใช้อัลกอริทึมแต่ละตัวให้มีความเหมาะสมกับรูปแบบ หรือลักษณะของแอปพลิเคชั่นต่างๆ เราจำเป็นที่จะต้องพิจารณาปัจจัยต่างๆประกอบกันด้วย อาทิเช่น ความสามารถของอัลกอริทึมในการจัดการเกี่ยวกับชนิดข้อมูลประเภทต่างๆ, การแจกแจง หรือความสามารถในการอธิบายผลลัพธ์ที่ได้จากการวิเคราะห์ของอัลกอริทึมเอง รวมไปถึงความยืดหยุ่น และความฉลาดของอัลกอริทึมอีกด้วย ซึ่งแต่ละอัลกอริทึมเองนั้นต่างก็มีข้อดี และข้อด้อยในตัวของมันที่แตกต่างกันออกไป จึงทำให้มีความเหมาะสมกับการใช้งานในรูปแบบที่ไม่เหมือนกันไปด้วยนั่นเอง สำหรับเทคนิค หรืออัลกอริทึมที่นำไปประยุกต์ใช้กับลักษณะงานในประเภทต่างๆนั้นจะได้ทำการศึกษาเกี่ยวกับคุณสมบัติ และรายละเอียดต่างๆในหัวข้อถัดไป เพื่อจะได้สามารถเลือกใช้ได้ตรงกับความต้องการของแอปพลิเคชั่นที่เราใช้งานได้อย่างถูกต้อง และเหมาะสมที่สุด

4. การวิเคราะห์ และการตีความผลลัพธ์ที่ได้ (Analysis of Results)

หลังจากที่เราได้นำข้อมูลจากแหล่งต่างๆมาผ่านกระบวนการจัดเตรียมการ เพื่อที่จะได้สามารถนำเข้าไปผ่านกระบวนการ หรืออัลกอริทึมที่ใช้ในการไมนิ่งข้อมูลเป็นที่เรียบร้อยแล้ว ในที่สุดเราก็จะได้ผลลัพธ์ซึ่งสามารถที่จะใช้มาทำการวิเคราะห์ และพิจารณาในแง่มุมต่างๆได้ง่ายขึ้น เพราะผลลัพธ์ที่ได้จะเป็นข้อมูลที่ช่วยให้เรามองเห็นรูปแบบของข้อมูลที่มีอยู่ว่าเป็นอย่างไร การตีความจากผลลัพธ์ที่ได้ของการทำดาต้าไมนิ่งนั้นเราจะไม่ถือว่าเป็นการตีความที่ถูก หรือผิดแต่เราจะมองว่าผลลัพธ์ที่ได้เมื่อนำไปวัดกับวัตถุประสงค์ที่ได้ทำการกำหนดไว้ตั้งแต่ตอนแรกนั้นว่าตรงกับสิ่งที่เราสนใจ หรือที่เราต้องการหรือไม่ ถ้าเรายังไม่สามารถสรุปผลออกมาได้ก็ต้องมีการย้อนกลับไปพิจารณาถึงข้อมูลที่เราได้ทำการคัดเลือกมาใช้ว่ามีประโยชน์ และเป็นสิ่งที่เราสนใจหรือไม่นั่นเอง ซึ่งในบางครั้งเราอาจจะต้องทำคัดเลือกข้อมูลที่จะนำมาใช้ในการไมนิ่งเสียใหม่ก็เป็นได้ทั้งนี้ทั้งนั้นก็ขึ้นอยู่กับรูปแบบ หรือลักษณะของแอปพลิเคชั่นที่เราทำงานด้วย ยกตัวอย่างเช่นในเรื่องการจัดแบ่งกลุ่มของข้อมูล(Database segmentation) เราต้องกำจัดตัวแปรข้อมูลที่ทำให้เกิดความเหมือนกันในแต่ละกลุ่มออกไป แล้วเลือกตัวแปรข้อมูลที่มีลักษณะเฉพาะที่ทำให้เกิดความแตกต่างกันในแต่ละกลุ่มเข้ามาแทนที่กันไป เป็นต้น

นอกจากนี้ในการตรวจสอบ และการประเมินผลที่ได้จากการไมนิ่งข้อมูลนั้นเรายังอาจจะใช้เครื่องมือในรูปแบบที่มีการแสดงผลเป็นแบบกราฟิคเข้ามามีส่วนช่วยในการตัดสินใจประกอบการพิจารณา หรือสรุปผลลัพธ์ได้อย่างแม่นยำอีกด้วย

5. การนำความรู้มาใช้งาน (Assimilation of Knowledge)

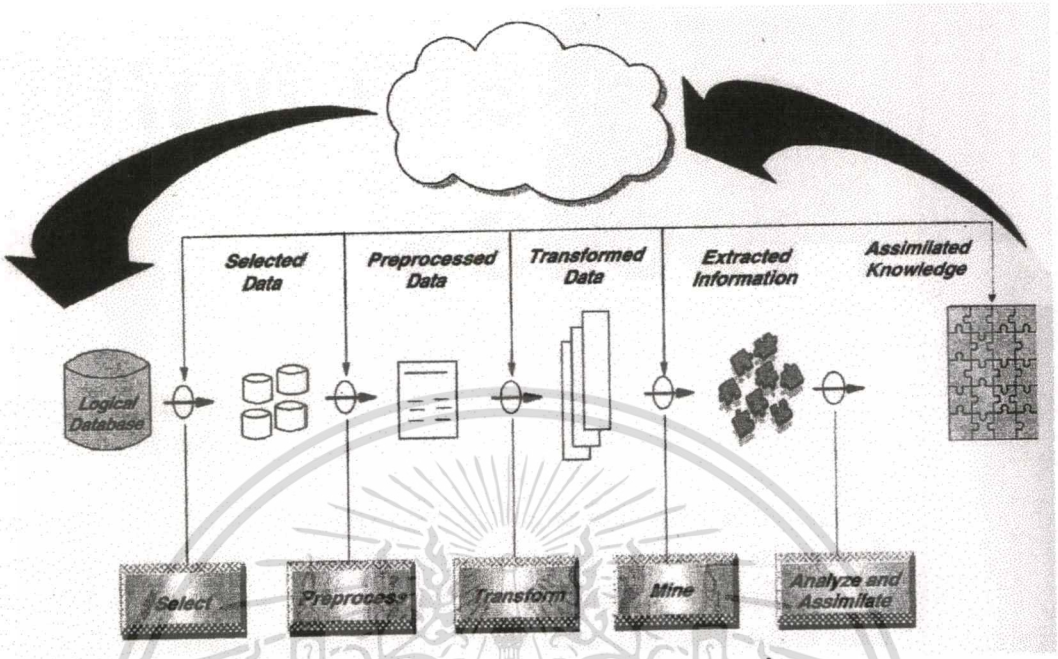
เมื่อเราได้ข้อสรุปที่แน่นอนจากการวิเคราะห์ และประเมินผลแล้วว่าผลลัพธ์ที่ได้นั้นเป็นประโยชน์ และมีคุณค่ากับสิ่งที่เราสนใจโดยผลลัพธ์ที่ได้มานั้นอาจจะเป็นการค้นพบข้อสารสนเทศ

ใหม่ๆซึ่งยังไม่เคยพบมาก่อน หรืออาจจะพบความสัมพันธ์ของข้อมูลในรูปแบบต่างๆที่อาจยังไม่ทราบมาก่อนก็อาจเป็นได้ ซึ่งจะก่อให้เกิดเป็นความรู้ หรือแนวความคิดใหม่ เพื่อให้สามารถนำไปแก้ไขปัญหาได้ตรงตามวัตถุประสงค์ที่เราได้กำหนดไว้ในขั้นตอนแรกก็ได้ การนำความรู้ หรือแนวความคิดใหม่ที่ว่านี้ไปประยุกต์ใช้ในเชิงธุรกิจจะมีหลักที่สำคัญอยู่ 2 ประการดังต่อไปนี้

1) การนำเสนอความรู้ หรือแนวความคิดใหม่ที่ค้นพบออกไปในเชิงธุรกิจ
2) การหาแนวทางที่เหมาะสม กับการประยุกต์ใช้ความรู้ หรือแนวความคิดใหม่ในการทำงานต่างๆ เพื่อให้เกิดประโยชน์สูงสุด

จากความรู้ หรือแนวความคิดใหม่ที่ได้อาจจะส่งผลให้เกิดการเปลี่ยนแปลงในเชิงธุรกิจหลายอย่างได้ อาทิเช่น เกิดแอปพลิเคชั่นทางด้านดาต้าไมนิ่งใหม่ๆ หรืออาจจะเป็นการเปลี่ยนแปลงแก้ไขรูปแบบเดิมที่มีอยู่แล้วให้ดีขึ้นกว่าเดิม อาทิเช่น การนำโปรโมชั่นใหม่ๆไปใช้กับกลุ่มลูกค้าประเภทใหม่ หรือการมอบส่วนลดของสินค้าให้กับลูกค้าที่มีความชอบในแต่ละประเภทที่แตกต่างกันไป หรือการส่งเสริมการขายผลิตภัณฑ์ที่มีอยู่ รวมไปถึงการจัดพื้นที่ในการวางขายสินค้าเพื่อให้เหมาะสมกับผู้ซื้อที่เข้ามาใช้บริการ นอกจากนี้ก็ยังทำให้เกิดการพัฒนาเกี่ยวกับคุณภาพของข้อมูล การจัดการเอกสารข้อมูลในประเภทต่างๆของระบบ และอาจสามารถป้องกันการเกิดข้อผิดพลาดที่อาจจะเกิดขึ้นได้ภายในระบบ เป็นต้น สิ่งเหล่านี้จะเป็นประโยชน์อย่างมากในการทำธุรกิจต่างๆในปัจจุบัน เนื่องจากคู่แข่ง และจำนวนลูกค้าที่เพิ่มขึ้นอย่างรวดเร็ว จึงทำให้ศักยภาพในการแข่งขันสูงขึ้นตามไปด้วย ดังนั้นการให้ความสำคัญกับการเก็บข้อมูลจึงนับได้ว่าเป็นสิ่งที่ช่วยให้กระบวนการทำดาต้าไมนิ่งนั้นสามารถดำเนินได้อย่างมีประสิทธิภาพนั่นเอง

เมื่อเราทำความเข้าใจถึงลักษณะความสำคัญ และกระบวนการในการทำดาต้าไมนิ่งในแต่ละขั้นตอนเป็นที่เรียบร้อยแล้ว เพื่อให้มองเห็นภาพรวมเกี่ยวกับความต่อเนื่องของกระบวนการดาต้าไมนิ่งทั้งหมดจะสามารถแสดงได้ดังรูปต่อไปนี้



รูปที่ 2.3 ภาพรวมของกระบวนการดาต้าไมนิ่ง

2.1.2 รูปแบบ และเทคนิคที่ใช้ในการทำดาต้าไมนิ่ง

ในหัวข้อที่ผ่านมาเราได้ทำการศึกษา และทำความเข้าใจเกี่ยวกับกระบวนการต่างๆในการทำดาต้าไมนิ่งไปแล้วว่าประกอบด้วยขั้นตอนการทำงานอะไรบ้าง และแต่ละขั้นตอนนั้นมีความสำคัญอย่างไร สำหรับในหัวข้อนี้เราจะมาทำการศึกษาต่อไปว่ารูปแบบที่ใช้ในการทำดาต้าไมนิ่งกับลักษณะ ของแอปพลิเคชั่นต่างๆนั้นเป็นอย่างไร ซึ่งจะช่วยให้เราได้ทราบ และมองเห็นภาพการทำงานในกระบวนการของการทำดาต้าไมนิ่งมากยิ่งขึ้น

แอปพลิเคชั่นของดาต้าไมนิ่งนั้นมีอยู่มากมายหลายประเภทด้วยกัน ซึ่งขึ้นอยู่กับลักษณะของงาน และข้อมูลที่เราจะนำไปใช้ทำการวิเคราะห์ หรือค้นหาเพื่อให้ได้ข้อสารสนเทศที่เป็นประโยชน์ สำหรับส่วนของรูปแบบการดำเนินการภายในแอปพลิเคชั่นของดาต้าไมนิ่ง หรือที่เราเรียกกันว่า Data Mining Operations นั้นจะมีอยู่ 4 รูปแบบด้วยกันได้แก่ (Cabena Peter. et al. 1997)

1) **Predictive Modeling** เป็นรูปแบบหนึ่งในการทำดาต้าไมนิ่ง ซึ่งมีลักษณะการทำงานเลียนแบบ หรือคล้ายคลึงกับการเรียนรู้ประสบการณ์ของมนุษย์ โดยอาศัยหลักการจากการสังเกตุลักษณะของสิ่งต่างๆที่เราสนใจ จากนั้นจึงรวบรวมข้อมูล และจดจำเอาไว้เพื่อใช้ในการคาดเดา และจำแนกประเภท แยกแยะสิ่งที่ค้นพบขึ้นมาใหม่ได้ เราสามารถนำรูปแบบของ Predictive Modeling ไปประยุกต์ใช้ในการทำดาต้าไมนิ่งได้โดยการตรวจสอบ และสังเกตุลักษณะที่สำคัญของข้อมูลที่มีอยู่ เพื่อให้เกิดรูปแบบหรือประเภทที่ชัดเจนของชุดข้อมูลนั้นๆ และสามารถที่จะนำลักษณะที่สำคัญเหล่านั้น

ไปคาดเดาชุดข้อมูลอื่นๆได้อย่างแม่นยำว่าเป็นข้อมูลประเภทไหน ยกตัวอย่างเช่น การคาดเดาเกี่ยวกับอัตราการเพิ่ม และลดลงของจำนวนลูกค้าในบริษัทประกันภัยแห่งหนึ่ง ซึ่งสามารถจำแนกประเภทของลูกค้าได้ออกเป็นสองประเภทด้วยกัน คือ กลุ่มลูกค้าประเภทที่ยังคงทำประกันกับบริษัทต่อไป (Stay) กับกลุ่มลูกค้าประเภทที่ยกเลิกการทำประกันกับบริษัท (Leave) โดยเราจะทำการวิเคราะห์ลักษณะข้อมูลของลูกค้าอันได้แก่ ระยะเวลาที่ลูกค้าเคยทำประกันกับบริษัทมา (Tenure) และจำนวนของประกันภัยประเภทต่างๆที่ลูกค้าทำกับบริษัท (Services) เพื่อคาดเดาว่าหากลูกค้ามีระยะเวลาที่ทำประกันภัยกับบริษัทมาน้อยกว่า 2 ปีครึ่ง และทำประกันภัยในประเภทต่างๆน้อยกว่า 3 ประเภท แสดงว่าลูกค้าคนนั้นน่าจะถูกจัดอยู่ในประเภทที่จะยกเลิกการทำประกันภัยกับบริษัทนั้นๆ อย่างแน่นอน ส่วนลูกค้าในกรณีเงื่อนไขอื่นๆ ก็จะถูกจัดอยู่ในประเภทหรือกลุ่มลูกค้าที่จะทำประกันกับบริษัทต่อไปนั่นเอง ซึ่งสามารถแสดงได้ดังภาพ

รูปที่ 2.4 แสดงการทำนาย เพื่อแยกประเภทลูกค้าที่ทำประกันกับบริษัท

จากตัวอย่างการทำนาย และการคาดเดาลูกค้าแต่ละคนว่าถูกจัดอยู่ในประเภทไหน โดยพิจารณาจากข้อมูลของลูกค้านั้นจะช่วยส่งผลให้เราสามารทำการประเมินผลอัตราการเพิ่ม และการลดจำนวนลงของลูกค้าที่ทำประกันกับบริษัทได้เป็นอย่างดี

สำหรับการตรวจสอบ และการสังเกตุลักษณะที่สำคัญของข้อมูลนั้น เราสามารถแบ่งออกได้เป็นสองขั้นตอนใหญ่ๆอันได้แก่

1. **Training Phases** เป็นขั้นตอนที่ใช้ในการค้นหารูปแบบ หรือลักษณะสำคัญที่ซ่อนอยู่ภายในชุดข้อมูลนั้นๆ ข้อมูลที่ใช้ในขั้นตอนนี้จะมีจำนวนมาก หรือมีขนาดใหญ่นั่นเอง

2. **Testing Phases** เป็นขั้นตอนการประยุกต์ใช้รูปแบบที่ถูกค้นพบจากขั้นตอนของ Training Phases กับชุดข้อมูลจริงๆ หรือชุดข้อมูลที่ไม่เคยพบมาก่อนเพื่อให้สามารถคาดเดาการจำแนก แยกแยะข้อมูลในแต่ละประเภทได้อย่างแม่นยำ สำหรับข้อมูลที่ใช้ในขั้นตอนนี้จะมีขนาดเล็ก หรือจำนวนน้อย

เทคนิคที่ใช้กับรูปแบบของ Predictive Modeling สามารถแบ่งออกได้เป็นดังนี้

- **Classification** เป็นเทคนิคที่ใช้ในการคาดเดาข้อมูลแต่ละชุดว่าจัดอยู่ในประเภทใด โดยอาศัยหลักการ และเงื่อนไขในการคาดเดาข้อมูลจากการพิจารณาลักษณะความสำคัญของประเภทข้อมูลนั้นๆว่าประกอบด้วยอะไรบ้าง และมีคุณสมบัติตรงตามประเภทที่เราจะจำแนกหรือไม่ยกตัวอย่างเช่น การลดลงของจำนวนลูกค้าหรือการที่ลูกค้าจะยกเลิกการทำประกันกับบริษัทในอนาคตนั้นจะสามารถเกิดขึ้นได้เฉพาะในกรณีที่ลูกค้าเคยทำประกันกับบริษัทมาแล้วน้อยกว่า 2 ปีครึ่ง และจะต้องทำประกันภัยในประเภทต่างๆน้อยกว่า 3 ประเภทอีกด้วย เป็นต้น ดังนั้นถ้าเราพบชุดข้อมูลของลูกค้าท่านใดที่มีคุณสมบัติตรงตามที่กล่าวไว้ เราก็จะสามารถทำนายได้ว่าลูกค้าคนนั้นจะยังคงทำประกันกับบริษัทของเราต่อไปหรือไม่นั่นเอง
- **Value Prediction** เป็นเทคนิคที่ใช้ในการคาดเดาข้อมูลอีกวิธีหนึ่ง แต่จะต่างกับเทคนิคแรกตรงที่การคาดเดาด้วยวิธีนี้จะได้ออกมาเป็นค่าบางอย่างของข้อมูลในแต่ละชุด ยกตัวอย่างเช่น การทำนายช่วงเวลาในการซื้อ หรือเปลี่ยนรถคันใหม่ของลูกค้า การคาดเดาช่วงระยะเวลาเหล่านี้อาจทำได้ด้วยการพิจารณาจากประวัติ หรือข้อมูลการซื้อรถของลูกค้าท่านนั้นๆ เป็นต้น นอกจากนี้ก็ยังสามารถพิจารณาจากปัจจัยอื่นๆได้อีก เช่น ความมั่นคงทางการเงิน รายได้ จำนวนบุคคลในครอบครัว ฐานะทางสังคม ระดับการศึกษา และระยะเวลาที่เป็นลูกค้า สิ่งเหล่านี้จะช่วยให้เราสามารถทำนาย หรือคาดเดาช่วงระยะเวลาที่ลูกค้าจะเปลี่ยน หรือซื้อรถคันใหม่ได้ง่ายดายยิ่งขึ้น

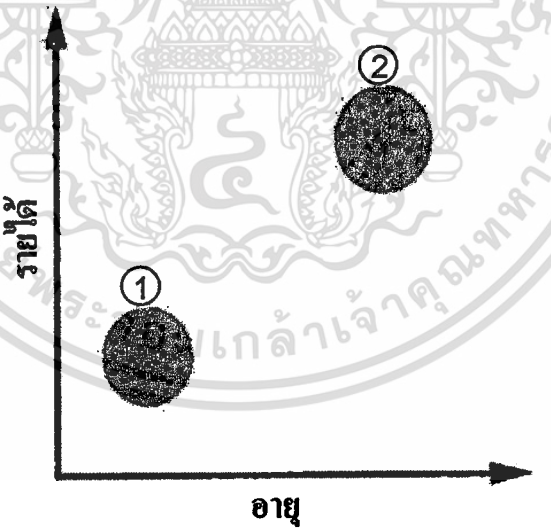
แอปพลิเคชั่นที่นำรูปแบบของ Predictive Modeling ไปประยุกต์ใช้อย่างเหมาะสมได้แก่ ธุรกิจเกี่ยวกับการจัดการ และการรักษาลูกค้าเอาไว้ (Customer Retention Management), ธุรกิจการทำเครดิตต่างๆ (Credit Approval), การขายสินค้าแบบ Cross Selling และการเจาะกลุ่มเป้าหมายในตลาด (Target Marketing) เป็นต้น

2) **Database Segmentation** เป็นอีกรูปแบบหนึ่งในการทำดาต้าไมนิ่งมีจุดประสงค์หลักในการทำงานคือ การจัดแบ่งข้อมูลออกเป็นกลุ่มๆ โดยที่ข้อมูลในแต่ละกลุ่มนั้นๆจะต้องมีคุณสมบัติที่มีความคล้ายคลึงกัน และแตกต่างจากข้อมูลในกลุ่มอื่นๆโดย

สิ้นเชิง ซึ่งแต่ละกลุ่มของข้อมูลที่มีความคล้ายคลึงกัน หรือเหมือนกันนั้นเราจะเรียกว่า Segments หรือ Clusters ส่วนวิธีการในการจัดแบ่งข้อมูลออกเป็นกลุ่มๆเหล่านี้เราจะเรียกว่า Segmentation หรือ Clustering นั่นเอง

วิธีการจัดแบ่งข้อมูลออกเป็นกลุ่มๆมีความสำคัญเนื่องจากว่าข้อมูลที่มีอยู่ในระบบฐานข้อมูลนั้นมีอยู่เป็นจำนวนมาก การหาข้อสารสนเทศ หรือข้อสรุปให้ได้ตรงตามที่เราต้องการจึงเป็นไปได้ยาก ดังนั้นจึงมีความจำเป็นที่จะต้องใช้วิธีการในการจัดแบ่งข้อมูลที่มีความเกี่ยวข้องกันหรือมีความคล้ายคลึงกันออกเป็นกลุ่มเสียก่อนเพื่อให้สามารถนำไปวิเคราะห์ต่อได้ในรูปแบบอื่นๆ ของการทำดาต้าไมนิ่ง อย่างเช่นรูปแบบ Predictive Modeling เป็นต้น

ตัวอย่างการจัดแบ่งข้อมูลของลูกค้าที่มีคุณสมบัติ หรือลักษณะที่คล้ายคลึงกันออกเป็นกลุ่มโดยพิจารณาจากรายได้ และอายุ ซึ่งสามารถแบ่งออกได้เป็นสองกลุ่มใหญ่ๆ ดังต่อไปนี้ กลุ่มแรกกำหนดให้เป็นกลุ่มของลูกค้าที่มีอายุน้อย และมีการศึกษาดี ส่วนกลุ่มที่สองกำหนดให้เป็นกลุ่มของลูกค้าที่มีอายุมาก และมีการใช้จ่ายเงินสูง ทั้งสองกลุ่มสามารถแสดงได้ดังภาพ



รูปที่ 2.5 แสดงการจัดแบ่งข้อมูลลูกค้าที่มีลักษณะคล้ายคลึงกันออกเป็นกลุ่ม (Segmentation)

จากตัวอย่างจะเห็นว่าเราจะทำการจัดแบ่งข้อมูลออกเป็นกลุ่มเสียก่อน จึงจะสามารถตีความคุณสมบัติ หรือลักษณะของข้อมูลภายในกลุ่มเหล่านั้นได้ สำหรับเทคนิคที่ใช้ในกับรูปแบบของ Database Segmentation แบ่งออกได้ดังนี้

- **Demographic Clustering** เป็นเทคนิคซึ่งเหมาะกับการจัดแบ่งข้อมูลชนิดที่เป็นแบบสามารถแจกแจงประเภทได้ (Categorical Variables) โดยจะอาศัยหลักเกณฑ์ในการวัดเพื่อแยกแยะความคล้ายคลึงกันของข้อมูลแต่ละชุดด้วยการนำเข้าข้อมูลแต่ละชุดมาเปรียบเทียบกัน ถ้าหากว่าค่าของข้อมูลในแต่ละชุดนั้นมีค่าที่ใกล้เคียงกัน เราก็จะเพิ่มคะแนนให้กับข้อมูลชุดนั้นๆ แต่ถ้าหากว่าค่าของข้อมูลที่นำมาเปรียบเทียบกันนั้นมีค่าที่แตกต่างกันมาก เราก็จะทำการลดคะแนนให้กับชุดข้อมูลที่นำมาเปรียบเทียบกันนั่นเอง เมื่อผลคะแนนสรุปออกมาเราก็สามารถที่จะจำแนกข้อมูลแต่ละชุดไปไว้ยังกลุ่มต่างๆได้ โดยที่แต่ละกลุ่มก็จะประกอบด้วยข้อมูลชุดที่มีคะแนนใกล้เคียงกัน ไม่แตกต่างกันมาก หลักการในการให้คะแนนแบบนี้เปรียบเสมือนเป็นการโหวตคะแนนให้กับข้อมูลแต่ละชุดซึ่งเราเรียกว่า Condorset (Cabena Peter. et al. 1997)
- **Neural Clustering** เป็นเทคนิคที่ใช้ความรู้พื้นฐานเกี่ยวกับโครงข่ายประสาทเทียม (Neural networks) ซึ่งจะเหมาะสมกับการจัดแบ่งข้อมูลชนิดที่เป็นตัวเลข หรือข้อมูลในลักษณะเชิงปริมาณเสียมากกว่า แต่ถ้าต้องการนำเข้าข้อมูลชนิดอื่นๆ ก็ทำได้โดยจะต้องทำการแปลงข้อมูลเหล่านั้นให้เป็นตัวเลขเสียก่อนเพื่อความเหมาะสมในการทำงานต่อไป ส่วนหลักเกณฑ์ที่ใช้ในการคัดเลือก เพื่อจัดแบ่งข้อมูลนั้นจะใช้วิธีการที่เรียกว่า Euclidean distance เพื่อหาระยะห่าง หรือค่าความแตกต่างจากจุดศูนย์กลางของข้อมูลในแต่ละกลุ่ม ถ้าหากค่าที่ได้มีค่าน้อยก็แสดงว่าข้อมูลชุดนั้นมีความคล้ายคลึงกับสมาชิกของข้อมูลในกลุ่มนั้นๆนั่นเอง

จากเทคนิคที่กล่าวมาจะเห็นได้ว่าทั้งสองเทคนิคนั้นมีข้อแตกต่างกันอยู่บ้างพอสมควร ซึ่งสามารถสรุปได้ดังนี้

1. ชนิดของข้อมูลที่ยอมให้นำเข้าไปวิเคราะห์ในแต่ละเทคนิค
2. วิธีการคำนวณหาค่าความแตกต่าง เพื่อใช้ในการวัดความเหมือนกัน หรือความคล้ายคลึงกันของข้อมูล
3. วิธีในการจัดการ และการประเมินผลลัพธ์ที่ได้จากการวิเคราะห์การจัดแบ่งข้อมูลออกเป็นกลุ่ม

ในความเป็นจริงแล้วการทำ Database Segmentation นั้นค่อนข้างที่จะมีความแม่นยำ และความถูกต้องน้อยกว่ารูปแบบอื่นๆของการทำดาต้าไมนิ่งอย่างเช่นรูปแบบการดำเนินการของ Predictive Modeling เป็นต้น ผลอันเนื่องจากอัลกอริทึมที่ใช้ในการทำงานนั้นมีความอ่อนไหวต่อข้อมูลที่นำเข้านั่นเอง ซึ่งจะทำให้ผลของการจัดแบ่งข้อมูลนั้นผิดพลาดไปได้ เช่น เกิดความซ้ำซ้อน

หรือเกิดความไม่เกี่ยวเนื่องกันของข้อมูล ส่วนการแก้ไขอาจจะต้องนำการคำนวณทางสถิติเข้ามามีส่วนช่วยในกระบวนการทำงานให้มากขึ้นเพื่อการทำงานที่ถูกต้อง สมบูรณ์ต่อไป

เราสามารถเห็นการประยุกต์ใช้รูปแบบของ Database Segmentation ได้กับแอปพลิเคชั่นจำพวกการจัดการแบ่งส่วนทางการตลาด (Target Marketing) หรือการจัดแบ่งลูกค้าทางการตลาด (Customer Segmentation) การขายสินค้าแบบ Cross Selling และธุรกิจอื่นๆอีกมากมาย

3) **Link Analysis** เป็นรูปแบบของการทำดาต้าไมนิ่งที่แตกต่างจากรูปแบบก่อนๆที่กล่าวมา จุดประสงค์หลักในการทำงานของรูปแบบนี้ก็คือ การค้นหาความสัมพันธ์ซึ่งซ่อนเร้นอยู่ภายในฐานข้อมูล ความเกี่ยวข้องกันระหว่างข้อมูลแต่ละชุด หรือแม้กระทั่งกลุ่มข้อมูลแต่ละกลุ่ม โดยความสัมพันธ์ หรือความเกี่ยวข้องกันเหล่านี้จะต้องมีการเชื่อมโยง และมีความหมายที่สำคัญ (associations) ดังนั้นแอปพลิเคชั่นที่เหมาะสมต่อการนำรูปแบบของ Link Analysis ไปใช้ก็จะเป็นไปในเรื่องของการค้นหาความสัมพันธ์ หรือการเชื่อมโยงกันระหว่างสินค้า หรือบริการซึ่งลูกค้ามักจะมีแนวโน้มในการเลือกซื้อ หรือใช้บริการเหล่านั้นควบคู่กันไปในช่วงระยะเวลาใดเวลาหนึ่ง การประยุกต์ใช้ที่เราเห็นได้อย่างชัดเจนได้แก่ การขายสินค้าเชื่อมโยงกันแบบ cross selling การหาแนวโน้มของราคาของสินค้าในคลังสินค้า เป็นต้น

สำหรับเทคนิคที่ใช้ในการค้นหาความสัมพันธ์ หรือการเชื่อมโยงกันระหว่างข้อมูลนั้นมีอยู่ด้วยกัน 3 แบบ และเพื่อให้มองเห็นภาพการนำไปใช้งาน และความเข้าใจในแต่ละเทคนิคจะขอยกตัวอย่างที่เกี่ยวกับการดำเนินการซื้อขายสินค้าของลูกค้าจากร้านค้าแห่งหนึ่ง ซึ่งสามารถอธิบายได้ดังนี้

- **Association discovery** เป็นการค้นหาความสัมพันธ์ หรือความเกี่ยวข้องกันระหว่างข้อมูลแต่ละตัว หรือสินค้าแต่ละอย่างที่ถูกลูกค้าซื้อภายใต้การดำเนินการหนึ่งๆ (Transaction) เพื่อดูว่าเหตุใดลูกค้าจึงซื้อสินค้าเหล่านี้พร้อมกัน หรือเพราะอะไรลูกค้าจึงเลือกซื้อสินค้าเหล่านี้ควบคู่กันไป วิธีการวิเคราะห์หาความสัมพันธ์แบบนี้ถูกเรียกว่า Market Basket Analysis (MBA) หรือ Product Affinity Analysis นั่นเอง
- **Sequential pattern discovery** เป็นเทคนิคที่ใช้ในการค้นหาความสัมพันธ์ หรือความเกี่ยวข้องกันของลำดับการเลือกซื้อสินค้าระหว่างการดำเนินการการเลือกซื้อสินค้า (ระหว่าง Transactions) ของลูกค้าภายในช่วงระยะเวลาใดเวลาหนึ่ง เพื่อศึกษาพฤติกรรมลูกค้า กับลำดับการเลือกซื้อสินค้าเหล่านั้น เช่น ลูกค้ากลุ่มหนึ่งซื้อเบียร์ และไวน์พร้อมกัน ส่วนลูกค้าอีกกลุ่มหนึ่งจะซื้อเบียร์ และไวน์พร้อมด้วยมัน

ฝรั่งทอดกรอบ เป็นต้น ซึ่งจะช่วยให้ได้รับข้อสารสนเทศรูปแบบใหม่ และนำไปสู่การจัดโปรโมชั่นใหม่ตามช่วงระยะเวลาได้อีกด้วย

- ❖ **Similar time sequence discovery** เป็นเทคนิคที่ใช้ในการศึกษาสินค้าสองกลุ่มในช่วงระยะเวลาเดียวกันโดยอาจจะดูจากยอดขายของสินค้าแต่ละตัวในกลุ่มนั้น หรือประวัติการถูกเลือกซื้อสินค้า รวมไปถึงความเคลื่อนไหวของราคาสินค้าในคลังสินค้า เพื่อหาการเชื่อมโยง และความเกี่ยวข้องกันระหว่างสินค้าสองกลุ่มดังกล่าวว่ามีทิศทาง และแนวโน้ม หรือมีรูปแบบที่เหมือนกันอย่างไร

4) **Deviation Detection** เป็นรูปแบบใหม่ในการทำดาต้าไมนิ่ง ซึ่งมีความสำคัญจุดประสงค์หลักในการทำงานของรูปแบบนี้ก็คือ การตรวจสอบ หรือตรวจจับความเบี่ยงเบน และค่าความคลาดเคลื่อนของข้อมูลในรูปแบบต่างๆ แอปพลิเคชั่นที่เหมาะสมกับรูปแบบนี้ได้แก่ การตรวจจับความผิดพลาด (Fraud Detection) และการปลอมแปลงการใช้งานบัตรเครดิต การตรวจสอบการเรียกร้องประกัน การใช้งานบัตรโทรศัพท์ การควบคุมคุณภาพ และการตรวจจับข้อบกพร่องในเรื่องต่างๆ สำหรับเทคนิคที่มีการนำมาใช้กับรูปแบบของ Deviation detection มีดังต่อไปนี้

- ❖ **Visualization** เป็นเทคนิคซึ่งใช้ในการตรวจจับหาความแปรปรวน และความเบี่ยงเบนของข้อมูลแล้วแสดงออกมาในรูปแบบของกราฟ แผนภาพ หรือแม้กระทั่งแผนภูมิต่างๆเพื่อให้สามารถทำความเข้าใจ หรือตีความได้ง่ายดายยิ่งขึ้น ดังนั้นการแสดงผลลัพธ์ด้วยวิธีการของเทคนิคนี้จึงมีอยู่หลากหลายรูปแบบด้วยกัน แต่จะใช้รูปแบบไหนนั้นก็จะขึ้นอยู่กับความเหมาะสมของจำนวนตัวแปรที่นำมาวิเคราะห์ ยกตัวอย่างเช่น ฮิสโตแกรม (histogram) แผนภูมิวงกลม (pie-charts) กราฟแบบแท่ง เป็นต้น กราฟผลลัพธ์ที่ได้ส่วนใหญ่มักจะถูกประมวลผลจากเครื่องคอมพิวเตอร์ที่มีสมรรถนะสูงเพื่อให้เกิดความรวดเร็วในการทำงานนั่นเอง เทคนิคนี้มักจะถูกนำไปใช้ร่วมกับรูปแบบอื่นๆในการทำดาต้าไมนิ่ง อย่างเช่น รูปแบบของการจัดแบ่งข้อมูลออกเป็นกลุ่ม (Database segmentation) เป็นต้น
- ❖ **Statistic** สำหรับเทคนิคนี้จะใช้ความรู้ทางด้านสถิติเข้ามามีส่วนเกี่ยวข้องกับการทำงานเสียมากกว่า โดยมีจุดประสงค์เพื่อการวัดความสำคัญ และความน่าสนใจของข้อมูล เทคนิคนี้จะจัดการกับความแปรปรวนของข้อมูลทันทีที่พบในชุดข้อมูล เทคนิคนี้มักจะถูกนำไปประยุกต์ใช้ในส่วนของขั้นตอนการเตรียมข้อมูล (Data Preparation) รวมไปถึงขั้นตอนของการวิเคราะห์ และการตีความผลลัพธ์ที่ได้ (Analysis of Results) ซึ่งอยู่ในกระบวนการของการทำดาต้าไมนิ่งในส่วนอื่นๆอีก

ด้วย เกณฑ์ที่ใช้ในการวัดความแปรปรวนของข้อมูลได้แก่ ค่ากลาง ค่าเฉลี่ย ค่าความถี่ต่างๆ รวมไปถึงค่าเบี่ยงเบนมาตรฐาน เป็นต้น

ตามที่ได้กล่าวมารูปแบบทั้ง 4 ต่างก็เหมาะกับการทำงานของแอปพลิเคชั่นดาต้าไมนิ่งที่มีลักษณะแตกต่างกันออกไป ซึ่งในบางครั้งแอปพลิเคชั่นของดาต้าไมนิ่งหนึ่งๆนั้นอาจจะมีความเหมาะสม หรือมีความจำเป็นต้องใช้กับรูปแบบหลายๆอย่างก็เป็นไปได้ยกตัวอย่างเช่น แอปพลิเคชั่นของดาต้าไมนิ่งทางด้านการจัดการด้านการตลาด (Market Management) อันได้แก่ ลูกค้าสัมพันธ์ (Customer Relationship Management), การแบ่งส่วนทางการตลาด (Market Segmentation) เหล่านี้เป็นต้น แอปพลิเคชั่นจำพวกนี้มักจะใช้รูปแบบการดำเนินงานภายในเป็น 2 รูปแบบแรกนั่นคือ รูปแบบการใช้ Predictive Modeling และการทำ Database Segmentation ควบคู่กันไปเพื่อให้กระบวนการในการทำดาต้าไมนิ่งทางด้านการจัดการ หรือการแบ่งส่วนทางการตลาดนั้นมีความสมบูรณ์ ลดข้อผิดพลาด และยังช่วยส่งผลให้ได้ข้อสารสนเทศจากการวิเคราะห์ที่ถูกต้องมากที่สุดอีกด้วย

อย่างไรก็ตามการเลือกใช้รูปแบบการดำเนินการของดาต้าไมนิ่ง (Data Mining Operations) นั้นควรจะพิจารณาจากลักษณะของงาน ประกอบกับประเภทของข้อมูลในส่วนของแอปพลิเคชั่นให้มากที่สุดเพื่อให้ได้ผลลัพธ์จากการทำงานของแอปพลิเคชั่นนั้นๆมีประสิทธิภาพ เราสามารถสรุปการเลือกใช้รูปแบบการดำเนินการของดาต้าไมนิ่ง (Data Mining Operations) กับการใช้งานแอปพลิเคชั่นในด้านต่างๆ รวมไปถึงเทคนิคในการวิเคราะห์อื่นๆได้ดังตารางต่อไปนี้ (Cabena Peter. et al. 1997)

ตารางที่ 2.1 แสดงความสัมพันธ์การใช้งานแอปพลิเคชั่น (Data Mining Applications) กับการเลือกใช้รูปแบบการดำเนินการของดาต้าไมนิ่ง (Data Mining Operations) และเทคนิคที่ใช้ในการวิเคราะห์ต่างๆ

	Market Management	Risk Management	Fraud Management
Applications	✓ Target marketing ✓ Customer relationship management ✓ Market basket analysis ✓ Cross-selling ✓ Market segmentation	✓ Forecasting ✓ Customer retention ✓ Improved underwriting ✓ Quality control ✓ Competitive analysis	✓ Fraud detection

Operations	**Predictive Modeling**	**Database Segmentation**	**Link Analysis**	**Deviation Detection**
Techniques	✓ Classification ✓ Value Prediction	✓ Demographic Clustering ✓ Neural Clustering	✓ Association Discovery ✓ Sequential Pattern Discovery ✓ Similar Time Sequence Discovery	✓ Visualization ✓ Statistics

สำหรับโครงงานพัฒนาระบบฉบับนี้ถูกพัฒนาขึ้นในรูปแบบแอปพลิเคชั่นที่มีลักษณะเกี่ยวกับการแบ่งส่วนทางการตลาด (Market Segmentation) โดยเลือกศึกษา และเน้นในรูปแบบของการทำดาต้าไมนิ่งที่เกี่ยวข้องกับการจัดแบ่งลูกค้าออกเป็นกลุ่มๆ ซึ่งจะทำการวิเคราะห์ข้อมูลของลูกค้าเป็นสำคัญ และจะจัดแบ่งลูกค้าที่มีลักษณะ หรือพฤติกรรมคล้ายคลึงกันเข้าไว้ในกลุ่มเดียวกัน ผ่านการทำงานของอัลกอริทึมที่เลือกนำมาใช้ในการศึกษาอีกด้วย ดังนั้นรูปแบบที่จะนำมาใช้ในการพัฒนาระบบจึงถูกเรียกว่า Customer Segmentation ส่วนรายละเอียดเกี่ยวกับกระบวนการทำงานของรูปแบบ Customer Segmentation นั้นจะได้นำเสนอในบทถัดไปตามลำดับ

บทที่ 3

การจัดลูกค้าเป็นกลุ่ม (Customer Segmentation)

Customer Segmentation เป็นการศึกษา วิเคราะห์ข้อมูลต่างๆที่เกี่ยวข้องกับลูกค้า เช่น พฤติกรรม ข้อมูลส่วนตัวของลูกค้า เป็นต้น เพื่อที่จะสามารถทำการจัดแบ่งกลุ่มลูกค้าหรือกลุ่มเป้าหมายทางธุรกิจได้อย่างชัดเจน แม่นยำ การจะจัดแบ่งข้อมูลลูกค้าออกเป็นกลุ่มๆได้นั้น จำเป็นที่จะต้องอาศัยกระบวนการที่เรียกว่า Cluster Analysis ซึ่งเป็นกระบวนการที่จะช่วยทำการวิเคราะห์ข้อมูลลูกค้าทางด้านการตลาดเหล่านี้ โดยใช้หลักเกณฑ์ในการคัดแยกข้อมูลที่มีความเหมือน หรือคล้ายคลึงกันเอาไว้ในกลุ่มเดียวกัน ส่วนข้อมูลอื่นที่มีความแตกต่างกันก็จะถูกแยกออกจากกันไปอยู่ในคนละกลุ่มข้อมูลกันนั่นเอง

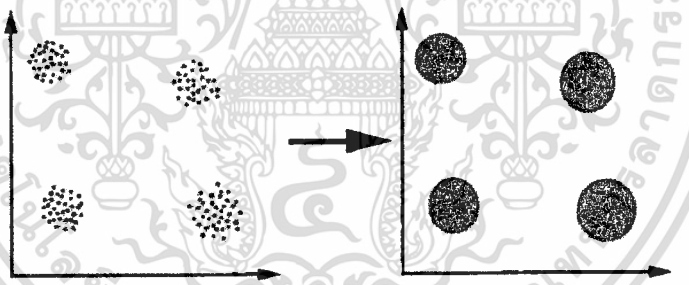
ดังนั้นเนื้อหาภายในบทจึงเป็นการนำเสนอเกี่ยวกับวิธีการ หรืออัลกอริทึมที่เลือกนำมาใช้ในการทำงานกับรูปแบบของการจัดแบ่งลูกค้า (Customer Segmentation) ว่าอัลกอริทึมเหล่านั้นสามารถแก้ปัญหา และแยกความแตกต่างระหว่างข้อมูลของลูกค้าแต่ละคนได้อย่างไร ก่อนที่จะเริ่มทำการศึกษาการทำงานของอัลกอริทึมที่เลือกมาใช้ เราจำเป็นจะต้องมีพื้นฐาน และความเข้าใจในส่วนของทฤษฎีฟัซซี่เป็นอันดับแรกเสียก่อน เนื่องจากทฤษฎีนี้จะช่วยให้เราสามารถเข้าใจถึงส่วนของข้อมูลลูกค้าแต่ละคนที่มีความคลุมเครือได้อย่างชัดเจน

3.1 ทฤษฎีฟัซซี่เซต (Fuzzy Set Theory)

ความสามารถในการจดจำรูปแบบ วิเคราะห์ แยกแยะ และจัดการแบ่งข้อมูลนับว่าเป็นความชาญฉลาดอย่างหนึ่งของสมองของมนุษย์ แต่ในบางครั้งการทำงานเหล่านี้ก็มีข้อจำกัดอยู่บ้าง เพราะการรับรู้ทางประสาทสัมผัสของมนุษย์เราไม่สามารถรับรู้ข้อมูลบางอย่างได้ทั้งหมด ยกตัวอย่าง การแยกระดับของสีในรูปภาพ เช่น สีเทา บางครั้งตาของคนเราก็ไม่สามารถบอกได้ว่าสีที่เราเห็นนั้นเป็นสีเทาจริงแท้ 100 เปอร์เซ็นต์ เพราะสีเทานั้นอาจจะมีความเป็นสีขาว หรือสีดำอยู่ในตัวของมันอยู่ เป็นต้น สิ่งเหล่านี้จึงทำให้มนุษย์ไม่สามารถที่จะตัดสินใจบอกได้ว่าข้อมูลที่ได้มานั้นมีรูปแบบหรือลักษณะที่ชัดเจนเป็นอย่างไร ทำให้เกิดความคลุมเครือ และไม่สามารถนำข้อมูลต่างๆเหล่านี้ไปใช้งานได้อย่างมีประสิทธิภาพ ดังนั้นกระบวนการการวิเคราะห์ แยกแยะ และจัดการแบ่งข้อมูลเพื่อให้ได้ข้อมูลที่มีลักษณะที่ชัดเจน หรือประเภทที่ชัดเจนนั้นจึงเริ่มเข้ามามีบทบาท และเป็นส่วนสำคัญในการทำงานต่างๆเพิ่มมากขึ้น กระบวนการการวิเคราะห์ และการจัดการแบ่งกลุ่มข้อมูลที่ว่า

นี้เราเรียกว่า Cluster Analysis (Shing Jyh. et al. 1997) นอกจากนี้ยังได้มีการนำเอาความรู้ และทฤษฎีทางด้านฟัซซี่เซต (Fuzzy Set) มาประยุกต์ใช้ในการจัดกลุ่มข้อมูลที่มีความความคลุมเครืออีกด้วย เพื่อที่จะได้สามารถแยกประเภทของข้อมูลที่เราต้องการได้อย่างถูกต้อง

ดังนั้นการศึกษาถึงการวิเคราะห์ และการจัดแบ่งกลุ่มของข้อมูล หรือแยกประเภทของข้อมูลจึงใช้ความรู้ทั้งสองด้านรวมเข้าด้วยกัน เราจึงเรียกวิธีการในการวิเคราะห์แบบนี้ว่า Fuzzy Clustering Algorithm ซึ่งมีความหมายว่า เป็นกระบวนการ หรือวิธีการในการพยายามที่จะจัดการแบ่งกลุ่มข้อมูลต่างๆ (input data) ให้มีรูปแบบ หรือโครงสร้างที่ชัดเจน โดยอาศัยทฤษฎี หรือหลักการทางคณิตศาสตร์เข้าช่วย การแบ่งข้อมูลออกเป็นกลุ่มๆด้วยวิธีนี้เราจะทำการแยกข้อมูลที่มีความเหมือน หรือคล้ายคลึงกันจัดเข้าไว้ในกลุ่มเดียวกัน ส่วนข้อมูลที่มีความแตกต่างกันก็จะถูกแยกออกจากกัน และนำไปไว้ในอีกกลุ่มหนึ่ง แต่ละกลุ่มที่ประกอบไปด้วยข้อมูลที่มีความคล้ายคลึงกันนั้น เราจะเรียกกลุ่มนั้นๆว่า Cluster ดังนั้นแต่ละ Cluster ก็จะมีลักษณะของข้อมูลที่แตกต่างกันไปด้วย



รูปที่ 3.1 แสดงถึงลักษณะของการวิเคราะห์ และจัดกลุ่มข้อมูล (Cluster Analysis)

กระบวนการทางการวิเคราะห์ และจัดกลุ่มข้อมูลหรือ Cluster Analysis ดังที่กล่าวมานั้นในยุคสมัยแรกๆถูกนำไปใช้ในงานเฉพาะด้านที่เกี่ยวกับเรื่อง Pattern Recognition แต่เพียงอย่างเดียวเท่านั้น แต่ต่อมาได้มีการนำไปประยุกต์ใช้กับงานทางด้านอื่นๆ อีกอย่างมากมาย ซึ่งจะเน้นงานที่มีความเกี่ยวข้องกับการแบ่งแยกประเภทของข้อมูลออกเป็นกลุ่มๆเสียเป็นส่วนใหญ่ นอกจากนี้ก็ยังนำไปใช้เกี่ยวกับการทำนายผล หรือวางแผนการทำงานบางอย่าง เพื่อให้ได้ข้อสารสนเทศที่เราต้องการหรือที่เราสนใจจริงๆ และสามารถนำข้อสารสนเทศเหล่านี้ไปใช้ประโยชน์ต่อไปได้ ยกตัวอย่างของการนำกระบวนการ Cluster Analysis ไปใช้งานเช่น (Shing Jyh. et al. 1997)

- การวิเคราะห์ด้านการตลาด (Marketing) เราจำเป็นที่จะต้องทราบถึงกลุ่มลูกค้าแต่ละประเภทของบริษัทว่ามีความต้องการอะไร ส่วนใหญ่อยู่ในวัยไหนบ้าง เป็นต้น

เพื่อที่จะได้นำข้อมูลเหล่านี้ไปผ่านกระบวนการจัดการแบ่งกลุ่ม แบ่งประเภทของลูกค้า (Data Clustering) และสามารถที่จะนำ Information ของลูกค้าแต่ละประเภทเหล่านี้ไปวิเคราะห์ และวางแผนออกเป็นแนวทาง หรือนโยบายเพื่อตอบสนองความต้องการของลูกค้าต่อไป

- การวางผังเมือง (City-Planning) โดยการนำข้อมูลที่อยู่อาศัยของประชากร ซึ่งอาจจะเป็นข้อมูลที่เกี่ยวกับที่ตั้ง หรือราคาของที่อยู่อาศัย เป็นต้น มาทำการวิเคราะห์ เพื่อให้ทราบว่าบริเวณไหน หรือพื้นที่ตรงไหนของเมืองที่มีประชากรอาศัยอยู่หนาแน่น และจะได้ทำการขยายผังเมืองต่อไปได้ในอนาคต
- การศึกษาเกี่ยวกับแผ่นดินไหว โดยการวัดค่าการสั่นสะเทือนในพื้นที่บริเวณต่างๆ เพื่อที่จะนำมาทำการวิเคราะห์ และทำนายผลว่าบริเวณใดที่มีการสั่นสะเทือนมากที่สุด ตรงไหนปลอดภัย หรือว่าอันตรายน้อยกว่ากัน เป็นต้น เพื่อประโยชน์ในการเตือนภัยให้กับผู้ที่อาศัยอยู่ในแถบบริเวณเหล่านั้นได้ทราบกัน

จากตัวอย่างที่ยกมากล่าวถึงทั้งหมดนี้เป็นเพียงตัวอย่างบางส่วน ซึ่งแสดงถึงการนำกระบวนการ Cluster Analysis ไปประยุกต์ใช้ในงานด้านต่างๆ ในหัวข้อถัดไปเราจะกล่าวถึงนิยามของฟัซซี่เซต (Fuzzy Set) เสียก่อนซึ่งจะใช้เป็นองค์ความรู้พื้นฐานที่สำคัญ ในการศึกษาถึงวิธีการของ Fuzzy Clustering Algorithm ต่างๆต่อไป

3.1.1 นิยามของฟัซซี่เซต (Fuzzy Set)

ปกติเวลาที่เรากล่าวถึงเซตทั่วๆไปนั้น เรามักจะนิยามเซตที่มีขอบเขตของสมาชิกที่แน่นอน นั่นคือเมื่อพิจารณาแล้วสามารถเข้าใจได้ว่า ข้อมูลตัวไหนบ้างที่เป็นสมาชิก หรือไม่เป็นสมาชิกของเซตนั้นๆ ยกตัวอย่างเช่น

กำหนด Set A เป็นเซตของจำนวนจริงที่มีค่ามากกว่า 6 เราสามารถแสดงในรูปของสัญลักษณ์ทางคณิตศาสตร์ได้ดังนี้

$$A = \{x \mid x > 6\}$$

จากนิพจน์ที่แสดง เราสามารถอธิบายได้ว่า ข้อมูลใดบ้างที่เป็นสมาชิกของเซตนี้ นั่นคือข้อมูลที่มีค่ามากกว่า 6 ก็จะเป็นสมาชิกของ Set A เช่น ค่า "6.001" เป็นต้น แต่ถ้าข้อมูลใดที่มีค่าน้อยกว่าค่า 6 ก็จะถือว่าไม่ได้เป็นสมาชิกของ Set A นั่นเอง การนิยามเซตในลักษณะนี้นั้นถูกนำไปใช้และมีความเหมาะสมกับงานต่างๆมาก แต่บางครั้งก็ไม่สามารถที่จะอธิบายถึงการทำงานบางอย่าง

ได้ เช่น ถ้าเรากำหนด Set A ใหม่ให้เป็น เซตของคนสูง โดยที่สมาชิกที่อยู่ใน Set A นี้มีค่าเป็นส่วนสูงของแต่ละคน จากการนิยามเซตแบบนี้เราจะบอกได้อย่างไรว่าสมาชิกในเซตนี้ประกอบด้วยสมาชิกอะไรบ้าง ถ้าเราให้คนที่มีส่วนสูงมากกว่า 6 ฟุตถือว่าเป็นคนสูง และคนที่มีส่วนสูงน้อยกว่านี้ถือว่าเป็นคนเตี้ย แต่เมื่อเราพิจารณาส่วนสูงของแต่ละคนพบว่ามีคนที่สูง 5.999 ฟุตกับคนที่สูง 6.001 ฟุตอยู่ด้วย จึงทำให้เกิดปัญหาขึ้นเพราะเราไม่สามารถบอกได้ว่าคนที่มีส่วนสูงดังกล่าวนั้นเราจะให้เป็นคนสูง หรือคนเตี้ย นั่นคือ เราไม่สามารถแยกแยะข้อมูลเหล่านี้ได้นั่นเอง นอกจากนี้การนิยามเซตในแบบแรกนั้นก็ยังไม่สามารถที่จะอธิบายปัญหาที่พบนี้ได้ด้วยเช่นกัน เนื่องจากข้อมูลที่มีอยู่มีความคลุมเครือ ไม่ชัดเจน และโดยสัญชาติญาณทางความคิดของมนุษย์ก็จะเกิดความไม่แน่ใจในการตัดสินใจจึงไม่สามารถแยกความแตกต่างเหล่านี้ได้

ดังนั้นการนิยามเซตอีกรูปแบบหนึ่งเพื่อใช้ในการแก้ปัญหาที่พบต่างๆเหล่านี้จึงเกิดขึ้น และมีเรียกชื่อเซตที่ถูกนิยามแบบนี้ว่าฟัซซี่เซต (Fuzzy Set) การนิยามเซตในลักษณะนี้จะสามารถอธิบายความเป็นสมาชิกของเซตนั้นๆได้ โดยจะมีการกำหนดลักษณะความเป็นสมาชิกให้กับข้อมูลแต่ละตัว ซึ่งลักษณะความเป็นสมาชิกที่ว่านี้เราเรียกว่า Characteristic Function หรือ Membership Function (Shing Jyh. et al. 1997) ซึ่งข้อมูลแต่ละตัวก็จะมีลักษณะความเป็นสมาชิกที่แตกต่างกันไป โดยทั่วไปการนิยามฟัซซี่เซต (Fuzzy Set) นั้นจะนิยามสมาชิกของเซตในรูปแบบของคู่ลำดับกล่าวคือจะเป็นคู่ลำดับของค่าของข้อมูล และลักษณะความเป็นสมาชิก ซึ่งสามารถแสดงได้ดังนี้

กำหนดให้ X เป็นเอกภพสัมพัทธ์ และ A เป็นฟัซซี่เซต (Fuzzy Set) และเขียนในรูปของสัญลักษณ์ทางคณิตศาสตร์เป็นได้ดังนี้

$$A = \{(x, \mu_A(x)) \mid x \notin X\}$$

โดยที่

- x เป็นสมาชิกในเอกภพสัมพัทธ์ของ X ซึ่งอาจจะเป็นหรือไม่เป็นสมาชิกในเซต A ก็ได้
- $\mu_A(x)$ หมายถึง ลักษณะความเป็นสมาชิกของ x ในเซต A (Membership Function)

3.1.2 ความเป็นสมาชิก (Membership)

จากการที่เรานิยามการเขียนเซตขึ้นใหม่เป็นแบบ Fuzzy Set นั้นจะเห็นได้ว่ามีการเพิ่มเอาลักษณะความเป็นสมาชิกใส่เข้าไปไว้ด้วยใน Fuzzy Set แต่ละเซต ซึ่งลักษณะความเป็นสมาชิก หรือ

Membership Function นี้จะเป็นค่าที่ใช้ในการบ่งบอกถึงระดับความเป็นสมาชิกของข้อมูลแต่ละตัว ดังนั้นข้อมูลแต่ละตัวก็จะมีค่านี้แตกต่างกันไป สำหรับค่าที่ใช้ในการบอกระดับความเป็นสมาชิกเหล่านี้นั้น เราจะใช้ค่าความน่าจะเป็นในการเป็นสมาชิกของข้อมูลเป็นตัวแสดงนั่นเอง และเราจะเรียกค่าเหล่านี้ว่า Membership Grade ซึ่งสามารถแสดงได้ดังตัวอย่างต่อไปนี้

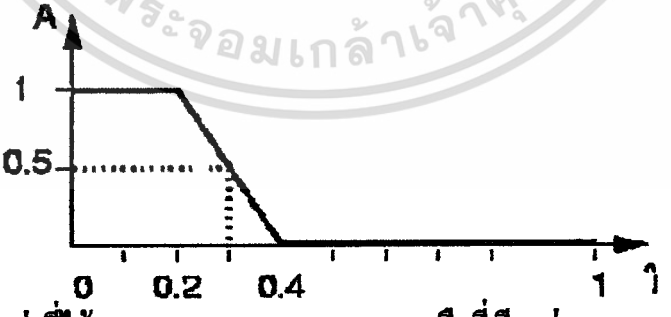
กำหนดให้ X เป็นเอกภพสัมพัทธ์ที่มีเซตเป็น

$$X = \{San\ Francisco, Boston, Los\ Angeles\}$$

และให้ C เป็น Fuzzy Set ซึ่งแสดงเมืองที่เราอยากไปอยู่ เมื่อเขียนในรูปแบบของสัญลักษณ์ทางคณิตศาสตร์แบบ Fuzzy Set จะแสดงได้เป็น

$$C = \{(San\ Francisco, 0.9), (Boston, 0.8), (Los\ Angeles, 0.6)\}$$

จากตัวอย่างเราจะเห็นได้ว่า สมาชิกในเซตจะประกอบไปด้วยคู่ลำดับต่างๆ และภายในแต่ละคู่ลำดับก็จะประกอบด้วยชื่อเมือง ซึ่งเป็นสมาชิกที่อยู่ในเอกภพสัมพัทธ์ X และค่าที่ใช้ในการบ่งบอกระดับความเป็นสมาชิกของแต่ละข้อมูล ซึ่งได้มาจาก Membership Function อีกทีหนึ่ง สำหรับค่าที่ได้มาเหล่านี้เนื่องจากเป็นค่าของความน่าจะเป็น ดังนั้นค่าที่ได้จึงอยู่ในช่วงของ [0,1] ซึ่งสามารถแสดงได้ดังรูปต่อไปนี้



รูปที่ 3.2 แสดงค่าที่ได้จาก Membership Function หรือที่เรียกว่า Membership Grade

3.2 Fuzzy C-Means Clustering Algorithm

หลังจากที่เราได้ศึกษาถึงเรื่องของทฤษฎี และการใช้งานของ Fuzzy Set ซึ่งเป็นการอธิบายถึงข้อมูลที่มีความคลุมเครืออยู่เป็นที่เรียบร้อยแล้ว ต่อไปเราจะนำความรู้ที่ได้มาประยุกต์ใช้กับกระบวนการ Cluster Analysis เพื่อให้เกิดการจัดแบ่งกลุ่มของข้อมูลที่ถูกต้องตามลำดับ ตามที่ได้

กล่าวมาแล้วในข้างต้นว่า วิธีการในการจัดการแบ่งกลุ่มข้อมูลที่มีความคลุมเครืออยู่นั้นเราจะเรียกวิธีการเหล่านี้ว่า Fuzzy Clustering Algorithm ซึ่งมีอยู่ด้วยกันหลายวิธี ยกตัวอย่างเช่น K-Means Clustering Algorithm, Hierarchical Clustering Algorithm หรือ Mixture of Gaussians เป็นต้น แม้ว่าจะมีอยู่หลายวิธีด้วยกันแต่หลักการทำงานของการทำ Fuzzy Clustering จริงๆนั้นก็จะเหมือนๆกัน คือ ต้องการหาโครงสร้างของข้อมูลที่ชัดเจน และทำการจัดกลุ่มของข้อมูลที่มีความคล้ายคลึงกันเข้าไว้ด้วยกันนั่นเอง สำหรับวิธีการที่จะเลือก และนำมาศึกษาในหัวข้อนี้นั้นจะเป็นวิธีการที่มีชื่อว่า Fuzzy C-Means Clustering Algorithm

Fuzzy C-Means Clustering Algorithm หรือที่รู้จักกันในอีกชื่อหนึ่งว่า Fuzzy ISODATA เป็นเทคนิคหนึ่งซึ่งนิยมใช้กันอย่างแพร่หลายมากสำหรับการจำแนกข้อมูลที่มีความคลุมเครือออกเป็นกลุ่มๆ เทคนิคนี้ถูกพัฒนาและปรับปรุงโดยศาสตราจารย์ Jim Bezdek ในปีค.ศ.1981 โดยพื้นฐานวิธีการของ Fuzzy C-Means Clustering Algorithm นั้นจริงๆแล้วถูกดัดแปลงและพัฒนามาจาก Dunn's Algorithm ลักษณะพิเศษของ Fuzzy C-Means Clustering Algorithm ที่แตกต่างจากวิธีการอื่นๆนั้นก็คือ มันสามารถที่จะบ่งบอกถึงระดับความความเป็นสมาชิก(Membership Grade) ของแต่ละกลุ่ม (Clusters) ได้นั่นเอง ซึ่งทำให้วิธีนี้แตกต่างจากวิธีการอื่นๆ ซึ่งไม่มีความสามารถในลักษณะนี้ สำหรับวิธีการอื่นๆ โดยทั่วไปนั้นจะสามารถบอกค่าระดับของความเป็นสมาชิกได้เพียงแค่ค่าเดียวเท่านั้น ซึ่งค่าที่บ่งบอกนี้ก็จะเป็นเฉพาะของกลุ่มเดียวเท่านั้นอีกด้วย นอกจากนี้วิธีการแบบ Fuzzy C-Means Clustering Algorithm ยังมีการใช้ค่าของ Membership Grade ในการถ่วงน้ำหนัก เพื่อให้ค่าของความเป็นสมาชิกมีความเป็น Fuzzy น้อยลงอีกด้วย (Shing Jyh. et al. 1997)

วิธีการวัด หรือประเมินผลของ Fuzzy Clustering เพื่อให้ทราบว่าข้อมูลแต่ละตัวที่ผ่าน Fuzzy Clustering Algorithm นั้นจัดอยู่ใน Cluster กลุ่มใดนั้นมีหลากหลายวิธีด้วยกัน ได้แก่ hierarchical, graph theoretic หรือ decomposing of density function เป็นต้น แต่สำหรับวิธีการวัดและประเมินผลของวิธีการแบบ Fuzzy C-Means Clustering Algorithm นั้นคือการลดค่าที่ได้จากการคำนวณ Criteria Function หรือ Cost Function นั่นเอง สำหรับการคำนวณค่าของ Cost Function นั้นผลลัพธ์ที่ได้จะไม่ได้เป็นแบบ linear แต่จะได้ผลลัพธ์ในรูปแบบของสัญลักษณ์ทางคณิตศาสตร์นั่นคือ Matrix นั่นเอง การคำนวณค่าของ Cost Function สำหรับวิธีการแบบ Fuzzy C-Means Clustering Algorithm นั้นสามารถอธิบายให้อยู่ในรูปของสมการทางคณิตศาสตร์ได้ดังต่อไปนี้

กำหนดให้ X เป็นกลุ่มของข้อมูลเข้า ซึ่งมีลักษณะเป็น Input Vectors และมีขนาดเป็น N ตัว เขียนในรูปแบบของเซตจะได้เป็น

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

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

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

ดังนั้นสูตรการคำนวณค่า Cost Function ของวิธีการแบบ Fuzzy C-Means Clustering Algorithm จะสามารถแสดงเป็นสัญลักษณ์ทางคณิตศาสตร์ได้ดังนี้

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

โดยที่

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

$$U = (u_{ij})_{N\ x\ C}$$

นอกจากนี้เมตริกซ์ U ยังต้องมีคุณสมบัติดังต่อไปนี้ จึงจะเป็นเมตริกซ์ที่สมบูรณ์ และถูกต้องตามสูตรการคำนวณ

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

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

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

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

- ตัวแปร m (Fuzziness Degree) เป็นตัวแปรที่ใช้กำหนดระดับของความเป็นฟัซซี่ ซึ่งจะมีค่าอยู่ในช่วงตั้งแต่ [1,∞) การกำหนดค่าของตัวแปร m จะขึ้นอยู่กับลักษณะของแต่ละแอปพลิเคชั่น ค่าของตัวแปร m ที่สามารถทำงานได้ดีจะอยู่ในช่วงระหว่าง 1.5-2.5 แต่โดยปกติทั่วไปแล้วจะกำหนดค่าไว้ให้เท่ากับ 2 เพราะถ้าหากกำหนดค่ามากเกินไปจะทำให้แต่ละกลุ่มไม่เกิดความแตกต่างกัน ส่งผลให้ไม่สามารถจัดแบ่งกลุ่มได้นั่นเอง
- ส่วนการคำนวณ $\|x_i - v_j\|$ นั้นเป็นการคำนวณเพื่อหาระยะห่าง หรือระยะทางระหว่างจุดสองจุดที่เรียกว่า การหา Distance Norm สำหรับวิธีการหา Distance ในรูปแบบของ Fuzzy C-Means Clustering Algorithm นั้นจะใช้วิธีการหาแบบ Euclidean Distance เพราะเป็นมาตรฐานในการหาระยะห่างระหว่างจุดสองจุด หรือขนาดของเวกเตอร์สองเวกเตอร์ที่เอามาลบกันนั่นคือ ข้อมูลตัวที่ i กับ Cluster Center ตัวที่ j

ในการคำนวณโดยใช้สูตรคำนวณดังกล่าวนี้ จะเป็นการแก้ปัญหาแบบ nonlinear ซึ่งจะใช้กระบวนการในการทำงานซ้ำๆ หรือวนลูปในการทำงานนั่นเอง โดยสามารถที่จะอธิบายเป็นขั้นตอนต่างๆ ได้ดังต่อไปนี้

Step1: กำหนดค่าแสดงความเป็นสมาชิกแต่ละตัว ในเมตริกซ์ U โดยอาจจะทำการสุ่มเอาค่าต่างๆขึ้นมาใช้ได้ แต่ต้องตรงกับคุณสมบัติที่ได้กล่าวไว้แล้วข้างต้น และกำหนดค่าระดับความเป็นฟัซซี่ หรือค่าตัวแปร m (Fuzziness Degree) ที่เหมาะสมเพื่อใช้ในการคำนวณด้วย

Step2: คำนวณหาค่า Cluster Center (V_j) ของทุกๆ Cluster ตามสมการข้างล่างนี้

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

Step3: คำนวณหาค่า Distance Norm ตามที่ได้อธิบายไว้แล้วข้างต้น โดยใช้สมการข้างล่างดังต่อไปนี้

$$d_{ij} = \|x_i - v_j\| = \sqrt{(x - v_1)^2 + (y - v_2)^2}$$

Step4: อัพเดตค่า u_{ij} ในเมตริกซ์ U โดยการตรวจสอบค่าของ Distance Norm ที่คำนวณได้จากขั้นตอนที่ 3 ว่ามีค่ามากกว่า 0 หรือไม่ ถ้าค่า Distance Norm มีค่ามากกว่า 0 จริงนั่นหมายความว่า $x_i \neq v_j$ นั่นเองก็ให้อัพเดตค่าของ u_{ij} ตามสมการข้างล่าง

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

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

Step5: ตรวจสอบค่าของ J(U,V) ว่ามีค่าเปลี่ยนแปลงจากเดิมหรือไม่ ถ้ามีการเปลี่ยนแปลงก็ให้กลับไปเริ่มทำในขั้นตอนที่ 2 อีกครั้งหนึ่งจนกว่าของ J(U,V) จะไม่มีการเปลี่ยนแปลงอีกจึงหยุดทำงาน ค่าของ J(U,V) ในที่นี้นั้นจริงๆแล้วก็คือค่าของเมตริกซ์ U นั่นเอง

หลังจากที่เราได้ทำการคำนวณตามขั้นตอนทั้ง 5 เสร็จเรียบร้อยแล้ว ให้ทำการวัด และประเมินผลค่าต่างๆในเมตริกซ์ U โดยให้ตรวจสอบข้อมูลแต่ละตัวว่ามีค่าของความเป็นสมาชิกใน Cluster กลุ่มใดสูงที่สุด แสดงว่าข้อมูลตัวนั้นน่าจะปรากฏอยู่ใน Cluster หรือกลุ่มนั้นนั่นเอง เพื่อให้เกิดความเข้าใจ และมองเห็นภาพในการทำงานได้ง่ายขึ้น ลำดับต่อไปจะขอแสดงตัวอย่าง ซึ่งจะช่วยอธิบายขั้นตอนการทำงานทั้งหมดที่ผ่านมาอีกครั้งหนึ่งดังต่อไปนี้

<u>ตัวอย่างการทำงานของ Fuzzy C-Means Clustering Algorithm</u>

กำหนดข้อมูลเข้า (input vectors) ต่างๆ ดังนี้

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

ต้องการแบ่งข้อมูลที่กำหนดมาให้ออกเป็น 2 กลุ่มหรือ 2 Cluster เมื่อทำตามขั้นตอนการทำงานตามที่ได้กล่าวมาสามารถแสดงได้ดังนี้

1) กำหนดค่าเริ่มต้นของเมทริกซ์ U

$$U = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

พร้อมทั้งกำหนดค่าบอกระดับความเป็นฟัซซี่ หรือค่าตัวแปร m (Fuzziness Degree) ให้มีค่าเท่ากับ 2

$$m = 2$$

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

$$v_1 = \begin{bmatrix} 2 \\ 3 \end{bmatrix} \qquad v_2 = \begin{bmatrix} 2.5 \\ 3 \end{bmatrix}$$

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

- ระยะห่างระหว่างข้อมูลชุดที่ 1 (x_1) กับ Cluster

$$d_{11} = \|x_1 - v_1\| = 2.24$$
$$d_{12} = \|x_1 - v_2\| = 2.5$$

- ระยะห่างระหว่างข้อมูลชุดที่ 2 (x_2) กับ Cluster

$$d_{21} = \|x_2 - v_1\| = 1.41$$
$$d_{22} = \|x_2 - v_2\| = 1.8$$

➢ ระยะห่างระหว่างข้อมูลชุดที่ 3 (x_3) กับ Cluster

$$d_{31} = \|x_3 - v_1\| = 1.41$$
$$d_{32} = \|x_3 - v_2\| = 1.12$$

➢ ระยะห่างระหว่างข้อมูลชุดที่ 4 (x_4) กับ Cluster

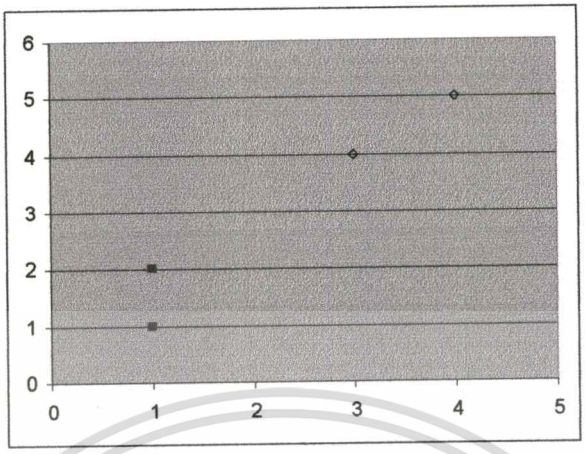
$$d_{41} = \|x_4 - v_1\| = 2.83$$
$$d_{42} = \|x_4 - v_2\| = 2.5$$

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

$$U = \begin{bmatrix} 0.56 & 0.44 \\ 0.62 & 0.38 \\ 0.39 & 0.61 \\ 0.44 & 0.56 \end{bmatrix}$$

5) กลับไปทำในขั้นตอนที่สองอีกครั้งหนึ่งเพราะภายใน Matrix U ยังมีการเปลี่ยนแปลงค่าของสมาชิกอยู่ ดังนั้นจึงต้องกลับไปทำในขั้นตอนที่สองจนกว่า Matrix U จะไม่มีการเปลี่ยนแปลงอีก

หลังจากที่ได้ทำการปรับค่าใน Matrix U จนค่าภายในไม่มีการเปลี่ยนแปลงแล้วก็พอที่จะสรุปได้ว่าข้อมูลตัวที่ 1 และตัวที่ 2 นั้นน่าจะมีโอกาสที่จะอยู่ในกลุ่มเดียวกันคืออยู่ใน Cluster กลุ่มที่ 1 ส่วนข้อมูลตัวที่ 3 และตัวที่ 4 นั้นก็มีโอกาสที่น่าจะอยู่ใน Cluster กลุ่มที่ 2 ซึ่งสามารถแสดงผลได้ดังกราฟข้างล่างนี้



รูปที่ 3.3 แสดงผลตัวอย่างการจัดแบ่งกลุ่มข้อมูลออกเป็น 2 กลุ่ม (2 Clusters)

จากที่กล่าวมาทั้งหมดจะเห็นได้ว่าการทำ Fuzzy Clustering ด้วยการใช้วิธีการแบบ Fuzzy C-Means Clustering Algorithm นั้นสามารถทำการวิเคราะห์ และแยกกลุ่มของข้อมูลได้อย่างมีประสิทธิภาพมากทีเดียว ซึ่งในบทถัดไปเราจะนำวิธีการของอัลกอริทึมเหล่านี้ไปประยุกต์ใช้ในการพัฒนาระบบเพื่อศึกษาถึงการนำไปใช้งานจริงๆว่าเป็นอย่างไร สามารถที่จะจัดแบ่งข้อมูลจริงของลูกค้าได้หรือไม่ และน่าจะเกิดข้อผิดพลาดอะไรบ้าง

บทที่ 4
การออกแบบระบบวิเคราะห์จัดแบ่งกลุ่มลูกค้า

เนื้อหาภายในบทนี้จะเป็นการกล่าวถึงเกี่ยวการออกแบบระบบวิเคราะห์จัดแบ่งกลุ่มลูกค้าทั้งหมดก่อนที่เราจะเริ่มนำความรู้ทางด้านดาต้าไมนิ่งไปประยุกต์ใช้กับระบบต่อไป การออกแบบระบบจะเป็นการอธิบายภาพรวมของการทำงานภายในระบบว่ามีโครงสร้างเป็นอย่างไร และระบบสามารถตอบสนองการทำงานของผู้ใช้ หรือมีส่วนของการทำงานที่ประกอบไปด้วยอะไรบ้าง ซึ่งจะส่งผลให้ในขั้นตอนการพัฒนาเป็นไปอย่างรวดเร็ว และสามารถเข้าใจได้อย่างง่ายดายมากยิ่งขึ้น

4.1 สถาปัตยกรรมระบบ

ในส่วนของการออกแบบสถาปัตยกรรมของระบบนั้นได้เลือกให้โปรแกรม (Software Application) มีรูปแบบ หรือลักษณะเป็นแบบ Stand-Alone Application โดยที่สามารถที่จะนำไปติดตั้งบนเครื่องไคลเอนต์ประเภทใดก็ได้ ส่วนการทำงานของตัวระบบนั้นยังคงยึดหลักการทำงานและสถาปัตยกรรมแบบ Client-Server based ซึ่งจะแบ่งออกเป็น 3 ส่วน (3 Tiers) ด้วยกันดังต่อไปนี้

1) **Data Tier** จะเป็นส่วนของการจัดการ และการเก็บข้อมูลภายในระบบฐานข้อมูล โดยการทำงานของส่วนนี้นั้นจะถูกควบคุมโดยโปรแกรมทางด้านการจัดการระบบฐานข้อมูลโดยเฉพาะ หรือที่เราเรียกกันว่า Database Management System (DBMS) นั่นเอง นอกจากนี้ยังเปิดการเชื่อมต่อเพื่อให้สามารถดึงข้อมูลจากส่วนของระบบฐานข้อมูลนำมาประมวลผลต่อไปได้ผ่านทาง JDBC (Java Database Connectivity) ซึ่งเป็น Middleware ที่ใช้ในการติดต่อสื่อสารระหว่างฝั่งของไคลเอนต์กับเซิร์ฟเวอร์อีกด้วย
2) **Application Tier** เป็นส่วนที่ใช้ในการควบคุมการทำงานหลักของโปรแกรม โดยจะทำหน้าที่ในการดึงข้อมูลจาก Data Tier เพื่อนำมาประมวลผล และการวิเคราะห์ สำหรับการติดต่อสื่อสารเพื่อดึงข้อมูลจากส่วนของ Data Tier นั้นก็จะผ่านทาง JDBC (Java Database Connectivity) ดังกล่าว และเมื่อผ่านการประมวลผลเป็นที่เรียบร้อยแล้วจึงจะส่งผลลัพธ์ที่ได้ไปยังส่วนของ Presentation Tier ต่อไป

3) **Presentation Tier** เป็นส่วนที่ใช้ในการติดต่อสื่อสาร รวมไปถึงการแสดงผลลัพธ์ที่ได้จากการวิเคราะห์ให้กับผู้ใช้งาน โปรแกรมได้รับทราบนั่นเอง ดังนั้นผู้ใช้หรือ Client ก็จะเป็นผู้ที่กำหนดการทำงานในส่วนนี้เอง

4.2 ปัญหา และผลประโยชน์ที่ได้รับจากระบบ (Problems and Benefits)

❖ **ปัญหาของระบบ (Problems)**

เนื่องจากทั้งจำนวน และรายละเอียดของลูกค้ามีเป็นจำนวนมาก เราจึงจำเป็นที่จะต้องสร้างระบบที่ใช้ในการวิเคราะห์ข้อมูลของลูกค้าว่ารายละเอียดไหนบ้างที่มีความสำคัญ และส่งผลกระทบต่อผลกำไรขาดทุนขององค์กร

❖ **ผลประโยชน์ที่ได้รับจากระบบ (Benefits)**

ระบบสามารถรองรับการวิเคราะห์ และประมวลผลจำนวนข้อมูลมากๆ ได้ นอกจากนี้ผู้ใช้ที่เกี่ยวข้องกับระบบยังสามารถทำความเข้าใจกับผลลัพธ์ที่ได้ออกมาอย่างง่ายดาย และเพื่อที่จะได้สามารถตอบสนองความต้องการของลูกค้าให้มากยิ่งขึ้นได้อีกด้วย

4.3 เครื่องมือที่ใช้ในการพัฒนาระบบ

สำหรับระบบวิเคราะห์จัดแบ่งกลุ่มลูกค้านี้ถูกพัฒนาขึ้นโดยใช้องค์ประกอบต่างๆ ดังต่อไปนี้

1) ส่วนของระบบจัดการฐานข้อมูล (DBMS: Database Management System) ที่ใช้ในระบบคือ MySQL ซึ่งเป็นระบบฐานข้อมูลที่มีขนาดเล็ก สามารถจัดการ และใช้งานได้อย่างง่ายดายไม่มีความซับซ้อนมากนัก สะดวกต่อการติดตั้ง นอกจากนี้ยังมีความเหมาะสมในการนำไปประยุกต์ใช้กับงานต่างๆ ได้มากมายหลายประเภท สำหรับเวอร์ชั่นของระบบฐานข้อมูล MySQL ที่ใช้คือ เวอร์ชั่น 5.0
2) ภาษาที่เลือกใช้ในการพัฒนาระบบได้แก่ ภาษาจาวา (Java) เนื่องจากเป็นภาษาโปรแกรมเชิงวัตถุ (Object-Oriented Programming) ดังนั้นจึงมีความยืดหยุ่นในการเขียนโปรแกรมเป็นอย่างมาก นอกจากนี้ยังช่วยให้สามารถทำการพัฒนาระบบได้เร็วขึ้น ด้วยเหตุที่ว่าเราสามารถนำ Code ที่เขียนขึ้นไว้แล้วนั้นนำมากลับมาใช้ได้อีก อย่างเช่น Class และ Method เป็นต้น อีกทั้งภาษาจาวายังสามารถนำไปติดตั้ง และทำงานได้แทบทุก Platform จึงทำให้ไม่เป็นปัญหาสำหรับการทำงานของระบบนี้เลย
3) คอมไพเลอร์ของภาษาจาวา (Java) เป็นองค์ประกอบที่สำคัญในการพัฒนาโปรแกรม และการทำงานของภาษาจาวา สำหรับคอมไพเลอร์ที่ใช้ในการพัฒนาระบบนี้คือ Java

Development Kit 5.0 (J2SE 5.0) ซึ่งเป็นเวอร์ชั่นล่าสุด และมีฟังก์ชั่นที่เอื้ออำนวยต่อการทำงานเป็นอย่างมาก

4) ส่วนของเครื่องมือที่ใช้ช่วยในการเขียนโปรแกรมภาษาจาวา (Java) นั้นคือ โปรแกรมที่ชื่อว่า Eclipse มีลักษณะเป็นโปรแกรมประเภท Freeware ทำให้ไม่ต้องเสียค่าลิขสิทธิ์ในการใช้งานตัวโปรแกรม เวอร์ชั่นที่ใช้คือ เวอร์ชั่น 3.1 ซึ่งเป็นเวอร์ชั่นล่าสุดที่ทางเว็บไซต์เปิดให้ดาวน์โหลดมาใช้งาน นอกจากนี้การติดตั้ง และการใช้งานยังสามารถเข้าใจได้ง่าย สะดวก รวดเร็วในการทำงานนับว่าเป็นเครื่องมือที่ช่วยในการพัฒนาระบบไม่แพ้เครื่องมือตัวอื่นๆได้เลย

4.4 Functional Requirement

ระบบวิเคราะห์จัดแบ่งกลุ่มลูกค้าสามารถแบ่งออกเป็นกระบวนการต่างๆได้ดังต่อไปนี้

Process 1 Specify the Data Source

เป็นการระบุ และเลือกใช้แหล่งข้อมูลที่จำเป็น หรือแหล่งข้อมูลที่เราสนใจในการทำดาต้าไมนิ่ง สำหรับข้อมูลที่ระบบจะนำมาใช้ในการทำงานนั้นจะมาจากระบบฐานข้อมูลนั่นเอง ซึ่งสามารถอธิบายกระบวนการทำงานได้ดังนี้

- **Process 1.1** การเชื่อมต่อกับฐานข้อมูล (Connect Database)

- List of Function: เชื่อมต่อกับระบบฐานข้อมูลที่ต้องการเพื่อที่จะนำข้อมูลจากระบบฐานข้อมูลมาทำการวิเคราะห์

- Class of Input

ตารางที่ 4.1 แสดง Class of Input ของ Process 1.1 Connect Database

Input	Type of Input	Valid	Invalid	Description
DBMS Name	VARCHAR2(1)	เลือกจาก Combo Box List		รายชื่อของระบบฐานข้อมูลที่สามารถเลือกได้แก่ MySQL, DB2 และ Oracle
Database URL	VARCHAR2(50)	ตามรูปแบบที่แสดงไว้ ซึ่งขึ้นอยู่กับระบบฐานข้อมูลที่เลือก	ใส่ข้อมูลไม่ตรงตามรูปแบบที่ระบุไว้	ชื่อของ Database URL ซึ่งจะต้องใส่ค่าให้ตรงตามรูปแบบที่กำหนด

Username	VARCHAR2(15)	Username ที่ใช้สำหรับเข้าถึงระบบฐานข้อมูล	Username ที่ไม่สามารถเข้าถึงระบบฐานข้อมูล	username ที่ใช้ในการติดต่อกับระบบฐานข้อมูล
Password	VARCHAR2(15)	Password ที่ใช้ในการเข้าถึงระบบฐานข้อมูล	Password ที่ไม่สามารถเข้าถึงระบบฐานข้อมูล	Password ที่ใช้ในการติดต่อกับระบบฐานข้อมูล
Table Name	VARCHAR2(50)	ชื่อตาราง (Table) ที่มีอยู่จริงในฐานข้อมูลที่ต้องการเชื่อมต่อ	ชื่อตาราง (Table) ที่ไม่มีอยู่จริงในฐานข้อมูลที่เชื่อมต่อ	ชื่อของตารางในฐานข้อมูลที่จะเชื่อมต่อ ซึ่งอาจจะมีมากกว่า 1 ตารางก็ได้

- Error Message

กรณีป้อน Database URL, Username, Password หรือ Table Name ผิดอย่างใดอย่างหนึ่งไม่ถูกต้องระบบจะฟ้องข้อผิดพลาดที่เกี่ยวข้องกับ SQLException และโปรแกรมจะไม่สามารถทำการเชื่อมต่อกับระบบฐานข้อมูลได้ พร้อมทั้งหยุดการทำงานทันที

Process 2 Data Preparation

เป็นกระบวนการในการจัดเตรียมข้อมูลต่างๆก่อนจะเริ่มทำการไมนิ่งข้อมูล ซึ่งจะประกอบด้วยขั้นตอนย่อยๆ 3 ขั้นตอนด้วยกันอธิบายได้ดังต่อไปนี้

- **Process 2.1** การคัดเลือกข้อมูล (Data Selection)

- List of Function: ทำหน้าที่ในการกลั่นกรอง และคัดเลือกข้อมูลต่างๆที่เราต้องการหรือสนใจ โดยการเลือกเขตข้อมูลที่มีอยู่ในตาราง นอกจากนี้ยังสามารถทำการกรองข้อมูลที่ไม่ต้องการในเบื้องต้นได้อีกด้วย โดยการใส่เงื่อนไข (Condition Filters) ลงไป

- Class of Input

ตารางที่ 4.2 แสดง Class of Input ของ Process 2.1 การคัดเลือกข้อมูล (Data Selection)

Input	Type of Input	Valid	Invalid	Description
Available Fields	VARCHAR2(100)	เลือกเขตข้อมูลตามที่ได้แสดงไว้ใน List	ไม่เลือกเขตข้อมูลตามที่แสดงไว้	เขตข้อมูลของตารางที่สามารถเลือกได้

Condition Filter	TEXT	เงื่อนไขในรูปแบบของภาษาSQL Condition	ใส่เงื่อนไขไม่ถูกต้อง	เป็นเงื่อนไขในรูปแบบของภาษา SQL ซึ่งอาจจะใส่หรือไม่ก็ได้

- Error Message

กรณีไม่ได้เลือกเขตข้อมูล (Field) ที่ต้องการในการคัดเลือกข้อมูลระบบจะฟ้องข้อผิดพลาดที่เกี่ยวข้องกับ SQLException และโปรแกรมจะไม่สามารถดำเนินงานต่อไปได้

กรณีใส่เงื่อนไขที่ต้องการกลั่นกรองไม่อยู่ในรูปเงื่อนไขของภาษา SQL ระบบจะฟ้องข้อผิดพลาดที่เกี่ยวข้องกับ SQLException และโปรแกรมจะไม่สามารถดำเนินงานต่อไปได้

- **Process 2.2** การกำจัดความไม่สมบูรณ์ของข้อมูล (Data Cleaning)

- List of Function: ทำหน้าที่ในการกำจัดความไม่สมบูรณ์ของข้อมูลในรูปแบบต่างๆ ที่อาจจะสามารถเกิดขึ้นได้กับเขตข้อมูลแต่ละชนิดนั้นๆ ยกตัวอย่างเช่น ความไม่สมบูรณ์อันเนื่องมาจากความผิดปกติกับค่าของข้อมูลที่หายไป ได้แก่ ค่าว่าง หรือค่าเป็น Null เป็นต้น ค่าเหล่านี้จะส่งผลให้การวิเคราะห์ และการคำนวณเกิดการผิดพลาดไม่สามารถทำงานได้นั่นเอง ดังนั้นขั้นตอนนี้จึงมีความสำคัญในการช่วยป้องกันไม่ให้เกิดความผิดปกติขึ้นกับการจัดแบ่งข้อมูลลูกค้าออกเป็นกลุ่มได้

- Class of Input

ชุดข้อมูลที่ถูกคัดเลือกไว้จาก Process 2.1

- **Process 2.3** การแปลงข้อมูล (Data Transformation)

ในส่วนของโพรเซสการแปลงข้อมูลนี้จะถูกแบ่งออกเป็น 2 ขั้นตอนย่อยๆดังต่อไปนี้

- **Process 2.3.1** การปรับปรุง และเปลี่ยนแปลงค่าของข้อมูล (Transform Data)

- List of Function: มีหน้าที่ในการปรับปรุง หรือเปลี่ยนแปลงค่าของข้อมูลให้เป็นอีกค่าหนึ่ง เพื่อให้เกิดความเหมาะสมในการวิเคราะห์ และการคำนวณกับอัลกอริทึมที่นำมาใช้ในการจัดแบ่งกลุ่มข้อมูล อันเนื่องจากอัลกอริทึมที่เลือกมาใช้ในการวิเคราะห์ข้อมูลนั้นไม่สามารถที่จะทำงานได้กับชุดข้อมูลบางประเภท ยกตัวอย่างเช่น ข้อมูลประเภทตัวอักษร ซึ่งควรจะเปลี่ยนแปลง และแทนด้วยค่าๆหนึ่ง โดยค่าที่จะมาแทนที่ควรกำหนดขึ้นมาให้มีความเหมาะสมด้วย

- **Class of Input**

ตารางที่ 4.3 แสดง Class of Input ของ Process 2.3.1 การแปลงข้อมูล (Transform Data)

Input	Type of Input	Valid	Invalid	Description
ค่าที่ต้องการแทน	INTEGER(10)	ค่าที่เป็นตัวเลข	ค่าที่ไม่เป็นตัวเลข	ค่าตัวเลขที่ใช้ในการเปลี่ยนแปลงหรือแทนที่ค่าข้อมูลจริงๆ

- **Error Message**

กรณีที่ใส่ค่าที่ไม่เป็นตัวเลขจะทำให้โปรแกรมเกิดการคำนวณที่ผิดพลาด ซึ่งไม่สามารถทำงานต่อได้ เนื่องจากเป็น Runtime Error

○ **Process 2.3.2** การปรับเปลี่ยนช่วงของข้อมูล (Normalize Data)

- **List of Function:** ทำหน้าที่ในการปรับเปลี่ยนค่าของข้อมูลให้อยู่ในช่วงที่เราต้องการ และยังช่วยป้องกันไม่ให้ค่าของข้อมูลเกิดความแตกต่างกันมากจนเกินไป หรือเกิดความลำเอียงระหว่างข้อมูลที่มีค่ามากกับข้อมูลที่มีค่าน้อยๆ ได้ ซึ่งจะส่งผลให้เกิดความถูกต้องในการวิเคราะห์ข้อมูลต่อไป

- **Class Of Input**

ตารางที่ 4.4 แสดง Class of Input ของ Process 2.3.2 การปรับเปลี่ยนช่วงของข้อมูล (Normalize Data)

Input	Type of Input	Valid	Invalid	Description
ค่าสูงสุดที่ใช้ในการปรับค่าของเขตข้อมูลนั้นๆ (Optional)	INTEGER(10)	ค่าที่เป็นตัวเลข	ไม่ใส่ค่าตัวเลข	ค่าที่ใช้ในการปรับค่าของข้อมูลเพื่อป้องกันไม่ให้เกิดความแตกต่างมากจนเกินไป
ค่าต่ำสุดของช่วงใหม่	INTEGER(10)	ค่าที่เป็นตัวเลข	ไม่ใส่ค่าตัวเลข	ค่าต่ำสุดของช่วงที่เราต้องปรับ
ค่าสูงสุดของช่วงใหม่	INTEGER(10)	ค่าที่เป็นตัวเลข	ไม่ใส่ค่าตัวเลข	ค่าสูงสุดของช่วงที่เราต้องการปรับ

- Error Message

ในกรณีที่ไม่ได้ใส่ค่าของช่วงที่เราต้องการปรับไม่ว่าจะเป็นค่าต่ำสุด หรือค่าสูงสุด ระบบจะแสดงข้อผิดพลาดด้วยการเตือนให้ระบุช่วงใหม่ด้วย และในกรณีที่ใส่ข้อมูลที่ไม่ใช่ตัวเลขลงไป ระบบจะไม่สามารถทำงานต่อไปได้เนื่องจากจะเกิด Runtime Error ในการทำงาน

Process 3 Segmentation by Fuzzy C-Means Algorithm

เป็นส่วนของการไมนิ่งข้อมูลด้วยอัลกอริทึมที่ได้เลือกมาทำการศึกษาไว้นั่นก็คืออัลกอริทึมที่มีชื่อว่า Fuzzy C-Means Algorithm

- List of Function: ทำหน้าที่ในการประมวลผล และจัดแบ่งข้อมูลลูกค้าออกเป็นกลุ่มๆ ด้วยกลไกการทำงานของวิธี Fuzzy C-Means Algorithm ข้อมูลที่จะผ่านกระบวนการนี้ได้จะต้องผ่านกระบวนการของการเตรียมข้อมูล (Data Preparation) เป็นที่เรียบร้อยมาก่อน เมื่อถูกจัดแบ่งออกเป็นกลุ่มๆ ก็จะถูกแสดงอยู่ในรูปของ Cost Function มีรูปแบบ และลักษณะเป็นแบบเมทริกซ์ โดยจำนวนแถวจะบ่งบอกถึงจำนวนของข้อมูลที่นำเข้า ส่วนจำนวนคอลัมน์จะบ่งบอกถึงจำนวนกลุ่มของข้อมูลที่เราต้องการ และสำหรับค่าที่อยู่ภายในจะเป็นการแสดงถึงระดับความเป็นสมาชิกของข้อมูลแต่ละชุดในแต่ละกลุ่ม การพิจารณาว่าข้อมูลแต่ละตัวนั้นปรากฏอยู่ในกลุ่มไหน เราจะสังเกตจากค่าระดับความเป็นสมาชิกที่มากที่สุดอยู่ที่กลุ่มใดก็แสดงว่าข้อมูลตัวนั้นถูกจัดอยู่ในกลุ่มนั้นนั่นเอง

- Class of Input

ตารางที่ 4.5 แสดง Class of Input ของ Process 3 Segmentation by Fuzzy C-Means Algorithm

Input	Type of Input	Valid	Invalid	Description
Number of Cluster	INTEGER(2)	ค่าที่เป็นตัวเลข	ค่าที่ไม่ใช่ตัวเลข	จำนวนกลุ่มที่ต้องการในการจัดแบ่งข้อมูล
Fuzziness Degree	INTEGER(5)	ค่าที่เป็นตัวเลขและมีค่ามากกว่า 1	ค่าที่ไม่เป็นตัวเลข หรือค่าที่เป็นตัวเลขแต่มีค่าน้อยกว่าหรือเท่ากับ 1	ระดับความเป็นฟัซซี่ หรือค่าตัวแปร m (Fuzziness Degree)

- Error Message

กรณีที่ใส่ค่าที่ไม่เป็นตัวเลขจะทำให้โปรแกรมเกิดการคำนวณที่ผิดพลาด ซึ่งไม่สามารถทำงานต่อได้ เนื่องจากเป็น Runtime Error

Process 4 Display Results of Cluster

เป็นส่วนของการแสดงผลลัพธ์ที่ได้อันเนื่องมาจากการวิเคราะห์ และการจัดแบ่งกลุ่มข้อมูลของลูกค้า โดยจะแยกส่วนของการแสดงผลออกเป็นสองส่วนด้วยกันอันได้แก่

- **Process 4.1** ส่วนผลลัพธ์ของแต่ละกลุ่ม

- List of Function: ทำหน้าที่ในการแสดงผลลัพธ์ที่ได้จากการวิเคราะห์ในแต่ละกลุ่ม

- Class of Input

Cost Function ในรูปแบบของเมทริกซ์ที่คำนวณได้จาก Process 3 และจุดศูนย์กลางของแต่ละกลุ่มที่คำนวณได้ ซึ่งอยู่ในรูปของเมทริกซ์เช่นกัน

- Output Fields

ตารางที่ 4.6 แสดง Output Fields กับส่วนการแสดงผลลัพธ์ของแต่ละกลุ่ม

Output Fields	Description
Cluster No.	กลุ่มที่
Cluster Center	ค่าจุดศูนย์กลางของแต่ละกลุ่มในแต่ละเซตข้อมูลที่เลือกมา
Data No. of Cluster	จำนวนของข้อมูลในแต่ละกลุ่ม

- **Process 4.2** ส่วนผลลัพธ์ของข้อมูลแต่ละตัว

- List of Function: ทำหน้าที่ในการแสดงผลลัพธ์ที่ได้จากการวิเคราะห์ข้อมูลแต่ละตัวว่ามีค่าความเป็นสมาชิกในแต่ละกลุ่มเป็นอย่างไร และสมควรจะจัดอยู่ในกลุ่มไหน

- Class of Input

Cost Function ในรูปแบบของเมทริกซ์ที่คำนวณได้จาก Process 3

- Output Fields

ตารางที่ 4.7 แสดง Class of Output กับส่วนการแสดงผลลัพธ์ในข้อมูลแต่ละตัว

Output Fields	Description
Data No.	ข้อมูลตัวที่
Membership Value	ค่าบ่งบอกระดับความเป็นสมาชิกในแต่ละกลุ่ม ของแต่ละชุดข้อมูล
Cluster No.	กลุ่มที่ข้อมูลชุดนั้นๆอยู่

Process 5 การแสดงผลลัพธ์ในรูปแบบอื่นๆ

นอกจากการแสดงผลลัพธ์ของวิเคราะห์จัดแบ่งกลุ่มข้อมูลของลูกค้าในรูปแบบของตารางทั้งสองใน Process ที่ 4 แล้วยังมีรูปแบบการแสดงผลลัพธ์ในรูปแบบอื่นๆอีก ซึ่งประกอบด้วย 2 รูปแบบด้วยกันดังต่อไปนี้

- **Process 5.1** การแสดงผลลัพธ์ของค่าจุดศูนย์กลางของแต่ละกลุ่มในแต่ละเขตข้อมูล

- List of Function: เป็นการแสดงผลลัพธ์ค่าจุดศูนย์กลางของแต่ละกลุ่มในแต่ละเขตข้อมูลด้วยรูปแบบกราฟเชิงเส้น

- Class of Input

ค่าจุดศูนย์กลางของแต่ละกลุ่มในแต่ละเขตข้อมูลที่ได้จาก Process 4.1

- Output

กราฟเชิงเส้นซึ่งแกนตั้งจะแสดงค่าจุดศูนย์กลางของแต่ละเขตข้อมูล และแกนนอนจะแสดงชื่อของเขตข้อมูลที่ได้เลือกมาทำการวิเคราะห์จาก Process 2.1 ส่วนจำนวนเส้นของกราฟจะใช้แสดงแทนจำนวนกลุ่มที่จัดแบ่ง

- **Process 5.2** การนำผลลัพธ์ของข้อมูลแต่ละชุดที่ผ่านการวิเคราะห์จัดแบ่งกลุ่มออกเป็นไฟล์

- List of Function: ทำหน้าที่ในการนำผลลัพธ์ของข้อมูลที่ผ่านการวิเคราะห์ และจัดแบ่งกลุ่มเรียบร้อยแล้ว หรือผลลัพธ์ที่ได้จาก Process 4.2 มาทำการส่งออกเป็นไฟล์ประเภทต่างๆ ได้แก่ ไฟล์ประเภท Microsoft Office Excel, CSV File (Comma Delimited), Text File (Tab Delimited) ซึ่งสามารถที่จะกำหนดได้ว่าจะให้เป็นประเภทไหน

- Class of Input

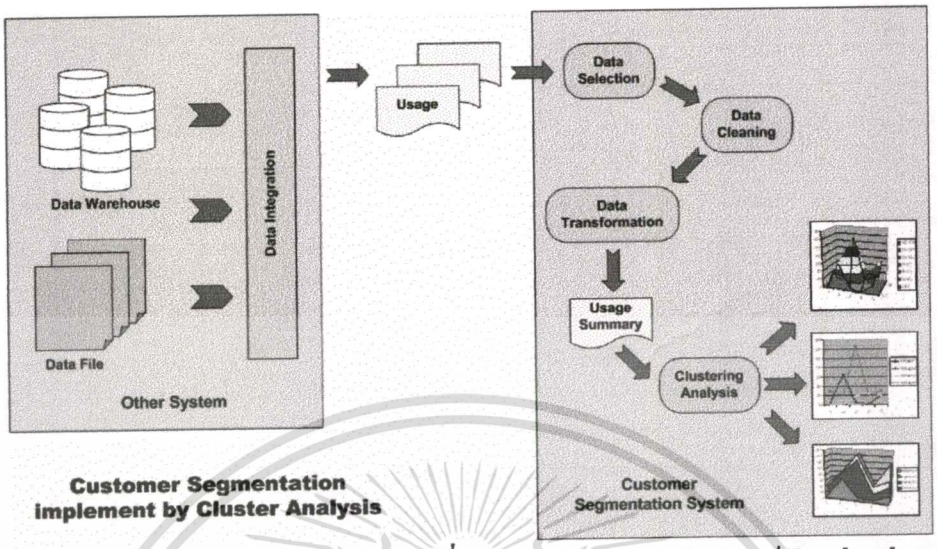
ผลลัพธ์ของข้อมูลแต่ละชุดที่ได้จาก Process 4.2

- Output

ไฟล์ข้อมูลตามที่ได้กำหนดไว้ก่อนจะเริ่มการนำข้อมูลออก

4.5 สรุปขั้นตอนการทำงานของระบบ

Process หรือขั้นตอนที่ได้ออกแบบสำหรับระบบวิเคราะห์จัดแบ่งกลุ่มลูกค้านั้นสามารถที่จะทำงานร่วมกันได้อย่างเป็นระบบ ซึ่งเราสามารถสรุปเป็นแผนภาพได้ดังรูปข้างล่าง



รูปที่ 4.1 แสดงการทำงานร่วมกันของฟังก์ชั่นต่างๆภายในระบบวิเคราะห์จัดแบ่งกลุ่มลูกค้า

จากแผนภาพข้างต้นเราสามารถอธิบายการทำงานที่เป็นลำดับขั้นตอนต่างๆได้ดังต่อไปนี้

1. ขั้นแรกจะเป็นกระบวนการของ Data Integration ซึ่งเป็นการรวบรวมข้อมูลจากที่ต่างๆ จาก Data Warehouse รวมไปถึงข้อมูลที่เป็น Meta Data อย่างเช่น Data Dictionary ซึ่งจะสามารถนำมาใช้อธิบายรายละเอียดต่างๆของข้อมูลที่รวบรวมมาได้ เพื่อให้ได้ข้อมูลที่เพียงพอ และจำเป็นต้องใช้ในการวิเคราะห์ของระบบ สำหรับการรวบรวมข้อมูลต่างๆในขั้นตอนนี้นั้นจะถูกทำงานโดยระบบอื่นซึ่งจะไม่เกี่ยวข้องกับระบบ Customer Segmentation ที่จะพัฒนา เมื่อผ่านกระบวนการ Data Integration จากระบบอื่นแล้วเราก็จะได้ข้อมูลที่ใช้เป็นข้อมูลนำเข้าของระบบ Customer Segmentation เพื่อทำการวิเคราะห์ต่อไป ข้อมูลจากภายนอกระบบจะถูกนำเข้าสู่ระบบผ่านทางการจัดการของระบบฐานข้อมูล (Database Management System)

2. หลังจากที่ได้นำข้อมูลเข้าระบบเป็นที่เรียบร้อยแล้ว กลไกการทำงานที่เป็น Data Selection ของระบบจะอนุญาตให้มีการเลือกข้อมูลที่จำเป็นต้องใช้ โดยเราจะต้องพิจารณาแต่ละเขตข้อมูลเพื่อเลือกให้ได้ใกล้เคียง หรือตรงกับวัตถุประสงค์ที่ได้กำหนดไว้ในตอนแรกมากที่สุด สำหรับวัตถุประสงค์ในบทความนี้จะเป็นการเลือกเขตข้อมูลที่มีความเกี่ยวข้องกับลักษณะ และพฤติกรรมการใช้งานของการใช้โทรศัพท์พื้นฐานเพียงเท่านั้น

3. จากนั้นกลไก Data Cleaning ของระบบก็จะพิจารณาแต่ละเขตข้อมูลอย่างคร่าวๆ ว่าแต่ละเขตข้อมูลนั้นมีความไม่สมบูรณ์อย่างไร และจะแก้ไขปัญหาที่เกิดขึ้นเหล่านั้นอย่างไรบ้าง

4. ขั้นตอนต่อไปจะเป็นกระบวนการของ Data Transformation ซึ่งจะทำการปรับเปลี่ยนข้อมูลทั้งหมดเพื่อให้เหมาะสมต่อการการ Cluster Analysis โดยในบทความนี้จะแปลงข้อมูลต่างๆ ให้อยู่ในรูปของสรุปผลการใช้งานโทรศัพท์พื้นฐานแบบรายวัน

5. ขั้นตอนสุดท้ายที่ระบบจะทำงานก็คือการนำข้อมูลที่ผ่านกระบวนการแปลงแล้วเข้าสู่การทำ Cluster Analysis โดยในที่นี้ได้นำ Fuzzy C-Means Clustering Algorithm นำมาประยุกต์ใช้เพื่อให้ได้ผลลัพธ์ของการจัดแบ่งกลุ่มข้อมูลของลูกค้าต่อไป

6. ผลลัพธ์ที่ได้จะถูกนำมาสรุปและ ตีความเพื่อให้เกิดความสำคัญ ด้วยการตรวจสอบกับวัตถุประสงค์ ถ้าสามารถที่จะตีความได้ตรงกับวัตถุประสงค์ที่กำหนดไว้ก็แสดงว่าการทำ Cluster Analysis นั้นประสบผลสำเร็จตามที่คาดไว้ แต่ถ้าไม่สามารถสรุปได้ก็อาจที่จะต้องย้อนกลับไปทำตั้งแต่กระบวนการแรกใหม่ก็เป็นได้

ในบทถัดไปเราจะนำฟังก์ชั่นการทำงานต่างๆที่ได้ออกแบบไว้ไปประยุกต์ใช้ในการทำไมนิ่งข้อมูลกับระบบวิเคราะห์จัดแบ่งกลุ่มลูกค้าต่อไป

บทที่ 5
การประยุกต์ใช้ดาต้าไมนิ่งกับ
ระบบวิเคราะห์จัดแบ่งกลุ่มลูกค้าที่ใช้โทรศัพท์พื้นฐาน

เทคโนโลยีทางด้านการติดต่อสื่อสารในปัจจุบันนับว่าเป็นเทคโนโลยีที่เติบโตไปอย่างรวดเร็วมาก การให้บริการที่มีความรวดเร็ว ถูกต้อง และแม่นยำ ย่อมเป็นสิ่งจำเป็นที่ผู้ให้บริการแต่ละบริษัทจำเป็นที่จะต้องคำนึงถึง เพื่อสร้างความพึงพอใจให้กับลูกค้ามากที่สุด นอกจากนี้การให้บริการที่มีความสะดวก และทันสมัยย่อมช่วยเพิ่มความเชื่อมั่นทางธุรกิจ และสนองต่อความต้องการของลูกค้าเพิ่มมากขึ้นอีกด้วย สำหรับช่องทางในด้านการสื่อสารโทรคมนาคมในปัจจุบันนั้น มีอยู่ด้วยกันหลายช่องทางด้วยกันไม่ว่าจะเป็นทางด้านสายโทรศัพท์ คลื่นวิทยุ หรือแม้กระทั่งดาวเทียมสื่อสาร เป็นต้น ซึ่งแต่ละช่องทางก็จะมีค่าใช้จ่ายที่แตกต่างกันไปขึ้นอยู่กับอุปกรณ์ที่ใช้ในการติดต่อสื่อสาร แต่ช่องทางที่นิยมใช้กันมากที่สุดก็คือ สายโทรศัพท์เพราะสายสื่อสารประเภทนี้สามารถส่งข้อมูลได้ดี และใช้กับงานต่างๆได้หลายประเภทด้วยกัน ได้แก่ การสื่อสารด้วยเสียง รูปภาพ Fax E-mail Pager โทรศัพท์เคลื่อนที่ หรือแม้กระทั่งการส่งข้อมูลผ่านทางด้านเว็บ และอื่นๆ ดังนั้นจึงเกิดการใช้งานหลากหลายรูปแบบ หลายลักษณะเพิ่มมากขึ้นซึ่งจะขึ้นอยู่กับการใช้งานของผู้ใช้เป็นหลัก ผู้ให้บริการจึงจำเป็นที่จะต้องทราบเกี่ยวรูปแบบ และลักษณะการใช้งานต่างๆเหล่านี้เพื่อที่จะได้จัดให้บริการเพื่อรองรับกับการใช้งานที่เกิดขึ้นได้อย่างถูกต้อง

วัตถุประสงค์ในการพัฒนาระบบเพื่อการวิเคราะห์ข้อมูล คือเพื่อจะได้สามารถจัดแบ่งกลุ่มลูกค้าที่ต้องการได้อย่างถูกต้อง ยกตัวอย่างเช่น จัดแบ่งกลุ่มลูกค้าที่มีการใช้งานในส่วนของโทรศัพท์พื้นฐานดูว่ามีการใช้งานโทรศัพท์หรือโทรออกในช่วงเวลาไหนมากที่สุด หรือจัดแบ่งกลุ่มลูกค้าในส่วนของการใช้งานโทรศัพท์พื้นฐานดูว่ามีการโทรออกไปที่ไหนบ้าง โทรไปต่างประเทศหรือภายในประเทศมากกว่ากัน เป็นต้น สิ่งเหล่านี้นอกจากจะสามารถทำการวิเคราะห์แบ่งกลุ่มลูกค้าได้อย่างชัดเจนแล้ว ยังสามารถนำผลการวิเคราะห์ที่ได้ไปสร้างคุณค่า และประโยชน์ในการที่บริษัทจะได้จัดโปรแกรมเพื่อสนับสนุน หรือโปรโมชั่นสนองต่อความต้องการของลูกค้าให้ได้มากที่สุดอีกด้วย นอกจากนี้ยังช่วยส่งเสริมการแข่งขันกันทางธุรกิจโทรคมนาคมกับองค์กรอื่นๆได้อีกด้วย

จากความรู้ทั้งทางด้านทฤษฎีที่เกี่ยวข้องกับดาต้าไมนิ่ง รวมไปถึงรูปแบบ และขั้นตอนในการจำแนกข้อมูลออกเป็นกลุ่มตามที่ได้กล่าวมาแล้วในบทก่อนหน้านี้ เราสามารถนำความรู้ดังกล่าวมาประยุกต์ใช้อย่างเป็นขั้นเป็นตอนได้ดังรายละเอียดต่อไปนี้

5.1 การกำหนดวัตถุประสงค์ทางธุรกิจ (Business Objectives Determination)

วัตถุประสงค์ของระบบวิเคราะห์ข้อมูลนี้คือ

"ศึกษาการจัดแบ่งกลุ่มลูกค้าในส่วนของการใช้งานโทรศัพท์พื้นฐานไปยังบริการ 1900 เพื่อดูลักษณะ และพฤติกรรมการใช้งานของลูกค้าว่าเป็นอย่างไร"

5.2 แหล่งที่มาของข้อมูล

สำหรับข้อมูลของลูกค้าที่จะนำมาใช้ในการวิเคราะห์กับระบบนี้นั้นได้รับการอำนวยความสะดวกมาจากบริษัทโทรคมนาคมแห่งหนึ่ง ซึ่งรายละเอียดเกี่ยวกับแต่ละเขตข้อมูลนั้นจะมีอยู่เป็นจำนวนมาก แต่ส่วนของข้อมูลที่ได้ทำการรวบรวมมาจากคลังฐานข้อมูลขนาดใหญ่ (Data Warehouse) จะมีลักษณะ และรายละเอียดดังต่อไปนี้

ตารางที่ 5.1 แสดงรายละเอียดของแหล่งข้อมูลที่รวบรวมจากคลังฐานข้อมูล (Data Warehouse)

Field Name	Field Type	Description
CODE	INTEGER(12)	รหัสของลูกค้า
PRODUCT_ID	CHAR(16)	เบอร์โทรศัพท์
PRODUCT_TYPE	CHAR(1)	ชนิดของ Product
CALL_TYPE	CHAR(2)	ชนิดของการโทร
DIALED_TN	VARCHAR2(18)	เบอร์ที่โทรออก
SUB_TYPE	CHAR(6)	ชนิดแยกย่อยของการโทร
C_DURATION	INTEGER(9)	ช่วงเวลาที่โทร

จากตารางเป็นข้อมูลดิบที่ได้ทำการรวบรวมมา โดยจะเป็นข้อมูลที่เกี่ยวกับการใช้งานโทรศัพท์ของลูกค้าในการโทรแต่ละครั้ง ดังนั้นขนาด และจำนวนของข้อมูลจึงใหญ่มาก เราจึงทำการสรุปเป็นยอดการใช้งานโทรศัพท์ในแต่ละวันตามประเภทการโทร และประเภทย่อยของการโทร ไม่รวมประเภทการใช้งานโทรศัพท์ที่เกี่ยวข้องกับทางด้านอินเทอร์เน็ต และอินเทอร์เน็ตความเร็วสูง อีกทั้งยังนับจำนวนนาทีที่ผู้ใช้เริ่มมีการโทรออกแต่ละครั้ง โดยจำนวนข้อมูลที่ถูกรวบรวมมาทั้งหมดนั้นนับเป็นระยะเวลาทั้งสิ้น 3 เดือนด้วยกัน กระบวนการในการรวบรวม และ

สรุปข้อมูลที่จำเป็นต้องใช้ออกมานี้เราเรียกว่าการรวมข้อมูล (Data Integration) และการลดขนาดของข้อมูล (Data Reduction) ซึ่งเป็นกระบวนการหนึ่งในการทำดาต้าไมนิ่งนั่นเอง รายละเอียดของแต่ละเซตข้อมูลที่รวบรวม และสรุปออกมาสามารถแสดงได้ดังตารางต่อไปนี้

ตารางที่ 5.2 แสดงข้อมูลที่ใช้นำเข้าสู่ระบบเพื่อใช้ในการคัดเลือกข้อมูล

Field Name	Field Type	Description
COUNT	CHAR(6)	จำนวนครั้งที่โทร
CALL_TYPE	CHAR(6)	ชนิดของการโทร
SUB_CALL_TYPE	INTEGER(10)	ชนิดแยกย่อยของการโทร
DURATION	FLOAT(10,2)	เวลาที่ใช้ในการโทร (คิดเป็นนาที)

หมายเหตุ เนื่องจากค่าที่ได้ความละเอียดของข้อมูลสูง ยกตัวอย่างเช่น หากมีการโทรอยู่ในช่วง 0.1-0.49 จะปัดเศษไปเป็นครึ่งนาที เป็นต้น

5.3 กระบวนการเตรียมข้อมูล (Data Preparation)

กระบวนการเตรียมข้อมูลจะจำแนกออกเป็นหลายๆขั้นตอนด้วยกันดังต่อไปนี้

1) การเลือกข้อมูล (Data Selection)

สำหรับขั้นตอนในการคัดเลือกเซตข้อมูลที่จะนำมาใช้เป็นปัจจัยในการวิเคราะห์ จัดแบ่งกลุ่มลูกค้านั้น นับว่าเป็นขั้นตอนที่มีความสำคัญ และจะต้องทำก่อนที่จะนำไปวิเคราะห์เสมอ ไม่ว่าจะเริ่มต้นด้วยวิธีการในรูปแบบใดก็ตาม เพราะเนื่องจากการกำหนดวัตถุประสงค์ในการวิเคราะห์และการทำดาต้าไมนิ่งนั้นแตกต่างกันออกไปขึ้นอยู่กับจุดมุ่งหมายของแอปพลิเคชั่นนั้นๆ จึงจำเป็นที่เราจะต้องทำการคัดเลือกเซตข้อมูลที่มีความสำคัญ หรือที่คิดว่าน่าจะเป็นปัจจัยหลักซึ่งส่งผลกระทบโดยตรงต่อสิ่งที่เราต้องการศึกษาในการทำดาต้าไมนิ่งเพื่อที่จะได้สามารถตอบสนองต่อวัตถุประสงค์ รวมไปถึงการประเมินผลลัพธ์ที่ได้ว่ามีความถูกต้อง แม่นยำตรงตามที่เราต้องการหรือไม่นั่นเอง

การคัดเลือกข้อมูล (Data Selection) จำเป็นจะต้องทราบรายละเอียด หรือค่าที่เป็นไปได้ของแต่ละเซตข้อมูลเสียก่อนว่าเป็นอย่างไร รวมไปถึงความหมายของค่าที่จะคัดเลือกด้วย ซึ่งจากข้อมูลที่ได้ทำการรวบรวม และสรุปมาแล้วนั้นเซตข้อมูลหลักที่ควรจะทราบได้แก่ เซตข้อมูล CALL_TYPE ซึ่งสามารถอธิบายถึงรายละเอียดต่างๆได้ดังต่อไปนี้

ตารางที่ 5.3 แสดงค่าที่เป็นไปได้ของเขตข้อมูล CALL_TYPE

Value Field	Description
A0	บริการ1900 ราคา 9 บาท/ครั้ง
A1	บริการ1900 ราคา 3 บาท/ครั้ง
A2	บริการ1900 ราคา 3 บาท/นาที
A3	บริการ1900 ราคา 5 บาท/ครั้ง
A4	บริการ1900สำหรับโหวต ราคา 9 บาท/ครั้ง
A7	บริการ1900 ราคา 6 บาท/ครั้ง
AA	บริการ1900 ราคา 6 บาท/นาที
AB	บริการ1900 ราคา 9 บาท/นาที
AC	บริการ1900 ราคา 15 บาท/นาที
AD	บริการ1900 ราคา 25 บาท/นาที
AE	บริการ1900 ราคา 5 บาท/นาที
AF	บริการ1900 ราคา 10 บาท/นาที
AG	บริการ1900 ราคา 20 บาท/นาที
AJ	บริการ1900 ราคา 13 บาท/นาที
AK	บริการ1900 ราคา 30 บาท/นาที

เมื่อเราทราบแล้วว่าแต่ละเขตข้อมูลมีลักษณะเป็นอย่างไรแล้ว ก็สามารถที่จะคัดเลือกข้อมูลได้อย่างถูกต้องเพื่อทำการเตรียมข้อมูลในขั้นตอนต่อไปได้

2) การเตรียมข้อมูลก่อนการนำไปประมวลผล (Data Preprocessing)

ในส่วนของขั้นตอนการเตรียมข้อมูลให้พร้อมก่อนนำไปประมวลผลนี้นั้น เราจะทำการกำจัดความไม่สมบูรณ์ต่างๆของข้อมูลที่ค้นพบหลังจากที่เราได้ทำการคัดเลือกข้อมูลเป็นที่เรียบร้อยแล้ว ซึ่งกระบวนการที่ว่านี้เราเรียกว่า Data Cleaning ของการทำดาต้าไมนิ่งนั่นเอง ส่วนวิธีการในการกำจัดความไม่สมบูรณ์ของข้อมูลนั้นจะแตกต่างกันออกไปขึ้นอยู่กับว่าสิ่งผิดปกติที่พบในข้อมูลนั้นเกิดขึ้นในรูปแบบใด สำหรับระบบวิเคราะห์จัดแบ่งกลุ่มลูกค้าที่พัฒนาขึ้นนี้จะใช้กรรมวิธีอยู่ 2 แบบด้วยกันคือ

a) **<u>วิธีที่ 1</u>** การกำจัดความไม่สมบูรณ์ของข้อมูลด้วยวิธีตัดชุดข้อมูลที่มีค่าบางค่าขาดหายไป คือค่าว่าง หรือค่าที่เป็น Null โดยเมื่อชุดข้อมูลชุดนั้นถูกตัดออกไปแล้ว ชุด

ข้อมูลชุดนั้นก็จะไม่ถูกนำไปประมวลผล หรือทำการวิเคราะห์ในการจัดแบ่งกลุ่มนั่นเอง

b) **<u>วิธีที่ 2</u>** การกำจัดความไม่สมบูรณ์ของข้อมูลด้วยวิธีการแทนค่าบางค่าลงไป เพื่อทดแทนค่าว่าง หรือค่าที่เป็น Null โดยค่าที่แทนลงไปควรจะมีความสอดคล้องกับค่าที่เป็นไปได้ในเขตข้อมูลนั้นๆ นอกจากนี้ควรจะสามารถนำไปประมวลผลต่อได้ ยกตัวอย่างค่าที่ควรจะนำไปทดแทน ได้แก่ ค่าเฉลี่ย หรือค่ากลาง เป็นต้น

วิธีดังกล่าวเป็นเพียงวิธีการเบื้องต้นในการกำจัดความผิดปกติที่เกิดขึ้นกับข้อมูล ซึ่งยังมีวิธีอื่นๆอีกมากมาย เมื่อได้ข้อมูลของลูกค้าที่มีความสมบูรณ์เป็นที่เรียบร้อยแล้วขั้นตอนต่อไปก็จะเข้าสู่กระบวนการในการปรับปรุง หรือเปลี่ยนแปลงข้อมูลเพื่อให้เกิดความเหมาะสมกับอัลกอริทึมที่เลือกใช้ในการจัดแบ่งกลุ่มข้อมูลนั่นเอง

3) การแปลงข้อมูล (Data Transformation)

สำหรับขั้นตอนในการแปลงข้อมูล (Data Transformation) ภายในระบบจะแบ่งออกเป็น 2 ขั้นตอนย่อยๆด้วยกันดังนี้

a) **<u>การปรับปรุง และเปลี่ยนแปลงค่าของข้อมูล (Transform Data)</u>**

ขั้นตอนนี้มีจุดมุ่งหมายเพื่อที่จะทำการปรับปรุง หรือเปลี่ยนแปลงค่าต่างๆของข้อมูลให้อยู่ในรูปที่เหมาะสมกับการนำเข้าไปประมวลผลกับอัลกอริทึมที่เราได้เลือกใช้ในการ ไมนิ่งข้อมูลนั่นเอง ระบบวิเคราะห์จัดแบ่งกลุ่มข้อมูลของลูกค้าที่พัฒนาก็เช่นเดียวกัน เนื่องมาจากว่าอัลกอริทึมที่ได้เลือกใช้นั้นคือ Fuzzy C-Means Algorithm ซึ่งเป็นอัลกอริทึมที่สามารถรองรับข้อมูลนำเข้าที่มีลักษณะเป็นตัวเลขได้เท่านั้น ดังนั้นจึงต้องมีการปรับปรุง และเปลี่ยนแปลงค่าที่มีลักษณะเป็นตัวอักษร หรือข้อความต่างๆให้อยู่ในรูปของค่าที่เป็นตัวเลข เพื่อที่จะได้สามารถนำไปประมวลผลต่อได้

จากตารางที่ 5.3 ซึ่งแสดงค่าที่เป็นไปได้ของเขตข้อมูล CALL_TYPE เราจึงจำเป็นที่จะต้องทำการปรับเปลี่ยนค่าต่างๆทั้ง 16 ค่าให้อยู่ในรูปของค่าที่เป็นตัวเลข และควรจะปรับไม่ให้เกิดช่วงของตัวเลขที่มีความแตกต่างกันมากมายนัก เพราะผลลัพธ์ที่ได้จากการคำนวณอาจจะเกิดความผิดพลาดขึ้นได้อย่างมาก รวมไปถึงเขตข้อมูล SUB_CALL_TYPE ที่มีชนิดข้อมูลเป็นแบบตัวอักษรด้วย

b) **<u>การปรับเปลี่ยนช่วงของข้อมูล (Normalize Data)</u>**

หลังจากที่ได้กำจัดความผิดปกติของข้อมูล รวมไปถึงปรับปรุง และเปลี่ยนแปลงค่าต่างๆให้เป็นตัวเลขที่เหมาะสมต่อการนำไปใช้งานกับอัลกอริทึม Fuzzy C-Means เป็นที่เรียบร้อยแล้ว ขั้นตอนนี้จะเป็นการปรับเปลี่ยนค่าของข้อมูลให้อยู่ในช่วงที่เราต้องการนั่นเอง โดยจะเริ่ม

ปรับเปลี่ยนไปทีละเขตข้อมูล เนื่องจากบางครั้งในแต่ละเขตข้อมูลที่เราทำการปรับปรุงค่ามาก่อนหน้านั้นอาจเกิดความแตกต่างกันมากระหว่างค่าภายในเขตข้อมูลเดียวกัน ซึ่งจะส่งผลให้เกิดความลำเอียงของข้อมูลระหว่างค่ามากกับค่าที่น้อยๆ ดังนั้นการปรับเปลี่ยนช่วงค่าของข้อมูลจึงมีส่วนสำคัญในการช่วยป้องกันไม่ให้ค่าของข้อมูลเกิดความแตกต่างกันมากจนเกินไป นอกจากนี้ก็ยังจะช่วยส่งผลให้การจัดแบ่งกลุ่มสามารถทำงานได้อย่างถูกต้องเพิ่มมากขึ้นอีกด้วย แต่สำหรับบางเขตข้อมูลที่ค่าภายในไม่มีความแตกต่างกันมากนักอย่างเห็นได้ชัดเจนก็อาจจะไม่จำเป็นต้องผ่านการทำงานของขั้นตอนนี้ก็ได้

ในบทถัดไปเราจะแสดงรายละเอียด รวมไปถึงขั้นตอนการใช้งานต่างๆภายในโปรแกรมของระบบวิเคราะห์จัดแบ่งกลุ่มลูกค้าที่พัฒนาขึ้นว่าเป็นอย่างไร และแต่ละขั้นตอนของกระบวนการดาต้าไมนิ่งที่กล่าวมามีความสอดคล้องกับโปรแกรมมากน้อยเพียงใด

บทที่ 6
ระบบวิเคราะห์จัดแบ่งกลุ่มลูกค้าด้วยวิธี Fuzzy C-Means

บทที่ผ่านมาเป็นการนำเสนอกระบวนการในการจัดเตรียมข้อมูล เพื่อให้ได้ข้อมูลที่มีคุณภาพ ซึ่งหลังจากเตรียมข้อมูลเสร็จเรียบร้อยแล้ว เนื้อหาในส่วนนี้จะเป็นการอธิบายเกี่ยวกับรายละเอียดของการทำงาน และขั้นตอนการใช้งานโปรแกรมของระบบวิเคราะห์จัดแบ่งกลุ่มลูกค้าด้วยอัลกอริทึม Fuzzy C-Means Clustering Algorithm โดยสามารถที่รับข้อมูลลูกค้านำเข้าผ่านทางระบบฐานข้อมูล เพื่อการเตรียมข้อมูล ตลอดจนวิเคราะห์ และประมวลผลผ่านทางกระบวนการไมนิ่งข้อมูลด้วยอัลกอริทึมจนได้ผลลัพธ์ที่สามารถนำไปจัดแบ่งข้อมูลของลูกค้าเกี่ยวกับการใช้งานโทรศัพท์พื้นฐานไปยังบริการ 1900 ออกเป็นกลุ่มได้

6.1 โปรแกรมของระบบวิเคราะห์จัดแบ่งกลุ่มลูกค้า

ระบบวิเคราะห์จัดแบ่งกลุ่มลูกค้าด้วยวิธีของ Fuzzy C-Means Algorithm นี้ถูกพัฒนาขึ้นด้วยภาษาจาวา ซึ่งทำงานอยู่บนระบบปฏิบัติการวินโดวส์ (Windows Operating System) ในความเป็นจริงแล้วโปรแกรมนี้สามารถที่จะนำไปติดตั้งบนระบบปฏิบัติการใดๆก็ได้ที่มี Java Virtual Machine ทำงานอยู่ แต่การที่เลือกพัฒนาบนระบบปฏิบัติการวินโดวส์ก็เพราะวินโดวส์เป็นระบบปฏิบัตการที่นิยมแพร่หลายใช้กันมากในปัจจุบัน อีกทั้งยังช่วยอำนวยความสะดวกในเรื่องของโปรแกรมอื่นๆอีกมากมาย ไม่ว่าจะเป็นเครื่องมือที่ใช้ในการช่วยเขียนโปรแกรมภาษาจาวา หรือระบบฐานข้อมูลต่างๆ เป็นต้น นอกจากนี้ระบบปฏิบัติการวินโดวส์ยังใช้งานง่ายกว่าอีกด้วย

6.2 รายละเอียดและขั้นตอนการใช้งานโปรแกรมของระบบ

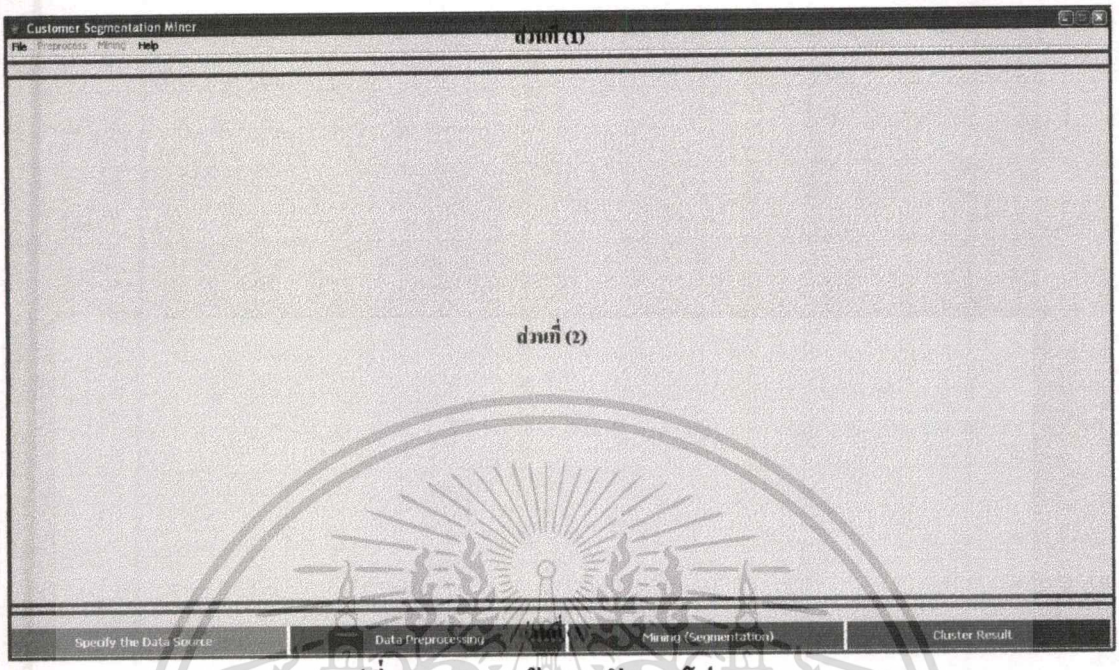
การใช้งานโปรแกรมของระบบวิเคราะห์จัดแบ่งกลุ่มลูกค้า เริ่มจากให้ดับเบิ้ลคลิก (Double Click) ที่ Shortcut ของโปรแกรมเพื่อรันโปรแกรมให้เริ่มทำงานดังรูปข้างล่าง

รูปที่ 6.6 แสดง Shortcut ของโปรแกรมระบบวิเคราะห์จัดแบ่งกลุ่มลูกค้า

หลังจากที่เริ่มรันโปรแกรมระบบ จะเข้าสู่หน้าจอแรกแสดงข้อความต้อนรับการใช้งานของโปรแกรมก่อน ซึ่งแสดงได้ดังภาพ

รูปที่ 6.2 แสดงหน้าจอต้อนรับผู้ใช้ของโปรแกรมระบบ

จากนั้นให้กดที่ปุ่ม OK เพื่อตอบตกลง โปรแกรมระบบวิเคราะห์ก็จะเริ่มทำงานและเข้าสู่หน้าจอหลักของโปรแกรมดังภาพ



รูปที่ 6.3 แสดงหน้าจอหลักของโปรแกรม

จากหน้าจอหลักของโปรแกรมเราสามารถแบ่งแยกออกเป็น 3 ส่วนด้วยกันดังต่อไปนี้

1. **ส่วนแรก** คือ ส่วนที่อยู่ด้านบนสุดจะแสดงรายการเมนูคำสั่ง (Command Menu) ซึ่งจะคอยควบคุมการทำงานของโปรแกรมให้สามารถดำเนินงานได้ โดยแต่ละเมนูก็มีความสำคัญที่แตกต่างกันออกไป สำหรับรายการเมนูคำสั่งที่เปิดให้สามารถใช้งานได้ประกอบด้วยรายการคำสั่งดังต่อไปนี้

   - **File** ประกอบด้วยรายการคำสั่งย่อย 3 รายการด้วยกัน อันได้แก่

     - Load From Database ใช้ในการโหลดเพื่อนำเข้าข้อมูลจากระบบฐานข้อมูล

     - Reset ใช้เมื่อต้องการจะล้างผลลัพธ์ของข้อมูลที่เลือกมาแสดงในส่วนแสดงผลส่วนที่สองของหน้าจอหลัก

     - Exit ใช้เมื่อต้องการออกจากโปรแกรม

   - **Preprocess** ประกอบด้วยรายการคำสั่งย่อย 2 รายการด้วยกัน ได้แก่

     - Clean Data ใช้เพื่อดำเนินการกำจัดความไม่สมบูรณ์ของข้อมูลที่เลือกมาในกระบวนการของดาต้าไมนิ่ง

     - Data Transform ใช้ในการปรับปรุง หรือเปลี่ยนแปลงข้อมูลบางส่วนเพื่อให้เกิดความเหมาะสมต่อการจัดแบ่งกลุ่มข้อมูลลูกค้าด้วยวิธี Fuzzy C-Means Clustering Algorithm ซึ่งสามารถแบ่งออกได้เป็น 2 วิธีด้วยกัน ได้แก่

การแปลงข้อมูลให้เป็นตัวเลข (Transform) และการปรับค่าของข้อมูลให้อยู่ในช่วง (Normalize) ซึ่งจะได้กล่าวถึงต่อไป

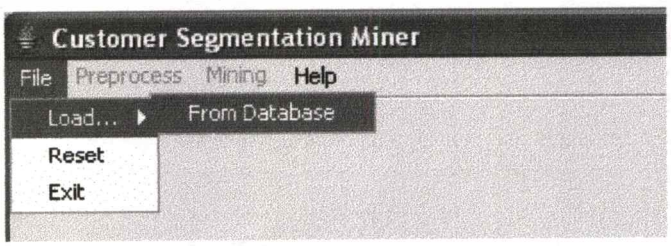
- **Mining** ใช้เพื่อให้โปรแกรมดำเนินการจัดแบ่งกลุ่มข้อมูลที่เลือกมาด้วยวิธี Fuzzy C-Means Clustering Algorithm
- **Help** ใช้เมื่อต้องการทราบข้อมูลทั่วไปเกี่ยวกับโปรแกรม

2. **ส่วนที่สอง** จากหน้าจอหลักของโปรแกรมจะสังเกตเห็นพื้นที่ว่างอยู่ในส่วนตรงกลางของหน้าจอ พื้นที่ส่วนนี้จะใช้ในการแสดงผลลัพธ์ของข้อมูลที่ได้เลือกมาจากระบบฐานข้อมูล รวมไปถึงการแสดงค่าต่างๆที่ได้จากขั้นตอนการเตรียมข้อมูล (Preprocess) อีกด้วย
3. **ส่วนที่สาม** เป็นส่วนสุดท้ายซึ่งจะสังเกตุเห็นเป็นแถบสีที่อยู่ด้านล่างโดยจะมีสีเขียวและสีแดง สำหรับแถบสีนี้จะเปรียบเสมือนเป็นแถบบอกสถานะการทำงานของโปรแกรมว่า ณ ขณะนี้โปรแกรมดำเนินงานถึงขั้นตอนตรงส่วนไหนแล้ว ถ้าแถบสีเขียวไปหยุด ณ ตำแหน่งใดก็หมายความว่าโปรแกรมกำลังดำเนินงานอยู่ในส่วนของขั้นตอนนั้นๆ

สำหรับขั้นตอนการทำงาน และการใช้งานของโปรแกรมนั้นสามารถอธิบายรายละเอียดและวิธีการใช้งานได้ดังต่อไปนี้

❖ **<u>ขั้นตอนที่ 1</u>** ระบุที่มาของข้อมูล (Specify the Data Source)

เราสามารถสังเกตุเห็นแถบสีเขียวที่แสดงสถานะการทำงานของโปรแกรมจะหยุดอยู่ที่ขั้นตอนแรกเพื่อบอกว่าให้เราทำการระบุที่มาของข้อมูล (Specify the Data Source) ดังนั้นจึงให้ทำการเลือกที่เมนู File จากหน้าจอหลักของโปรแกรม จากนั้นเลือก Load แล้วเลือก From Database ดังภาพ



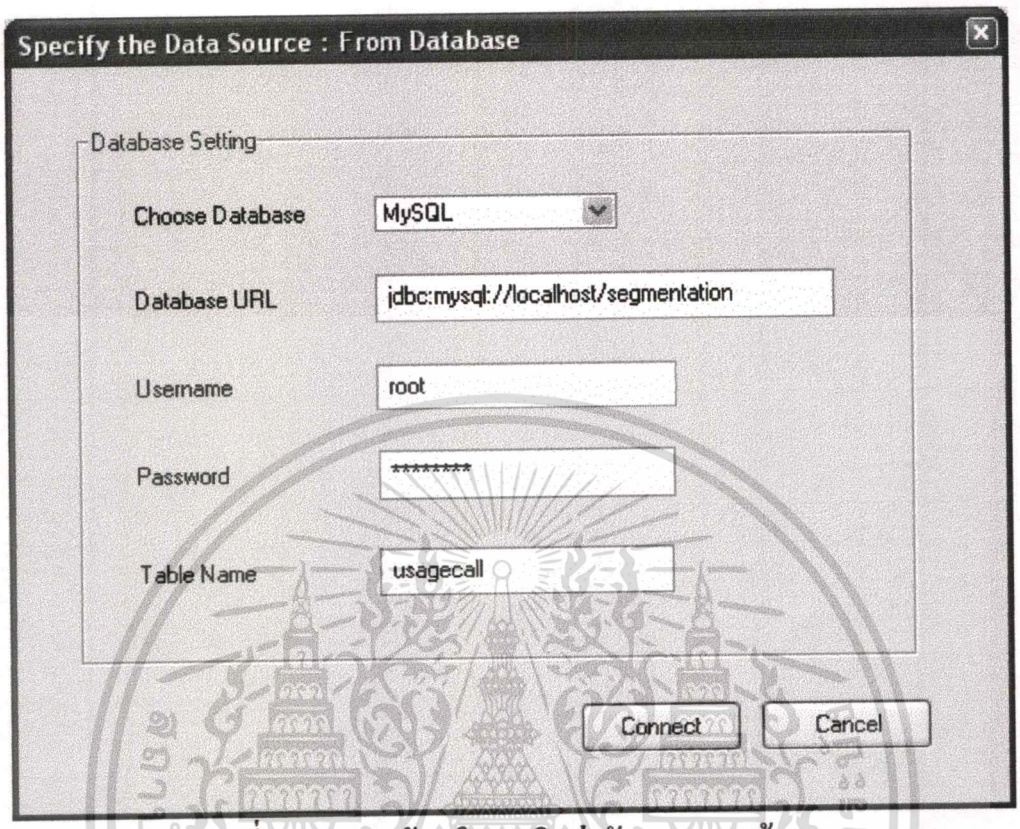
รูปที่ 6.4 แสดงการเลือกแหล่งข้อมูลจากระบบฐานข้อมูล (Database System)

โปรแกรมจะเปิดหน้าจอใหม่ขึ้นมาเพื่อจะทำการติดต่อกับระบบฐานข้อมูลที่ผู้ใช้ต้องการโดยมีรายละเอียดต่างๆที่ต้องป้อนเข้าสู่โปรแกรมระบบดังต่อไปนี้

ตารางที่ 6.1 แสดงข้อมูลต่างๆที่จำเป็นต้องป้อนเข้าสู่ระบบเพื่อทำการติดต่อกับระบบฐานข้อมูล

Field Input	Description
Database	ระบบฐานข้อมูลที่ผู้ใช้ต้องการติดต่อ ซึ่งมีอยู่ 3 แบบด้วยกัน ได้แก่ ■ MySQL ■ DB2 ■ ORACLE สำหรับค่ามาตรฐานที่โปรแกรมระบบนี้ใช้จะเป็น MySQL เนื่องจากการพัฒนาระบบได้เลือกใช้ระบบฐานข้อมูลนี้ในการติดต่อ
Database URL	เป็นรูปแบบที่ใช้ในการติดต่อกับระบบฐานข้อมูล ซึ่งจะมีความแตกต่างกันออกไปตามแต่ละระบบฐานข้อมูล ส่วนการติดต่อนั้นจะติดต่อผ่านทาง JDBC (Java Database Connectivity)
Username	ชื่อที่ใช้ในการติดต่อกับระบบฐานข้อมูล
Password	รหัสผ่านในการติดต่อกับระบบฐานข้อมูล
Table name	ชื่อตารางข้อมูลที่ผู้ใช้ต้องการไปดึงข้อมูล

ส่วนของหน้าจอในการติดต่อกับระบบฐานข้อมูลสามารถแสดงได้ดังภาพข้างล่าง

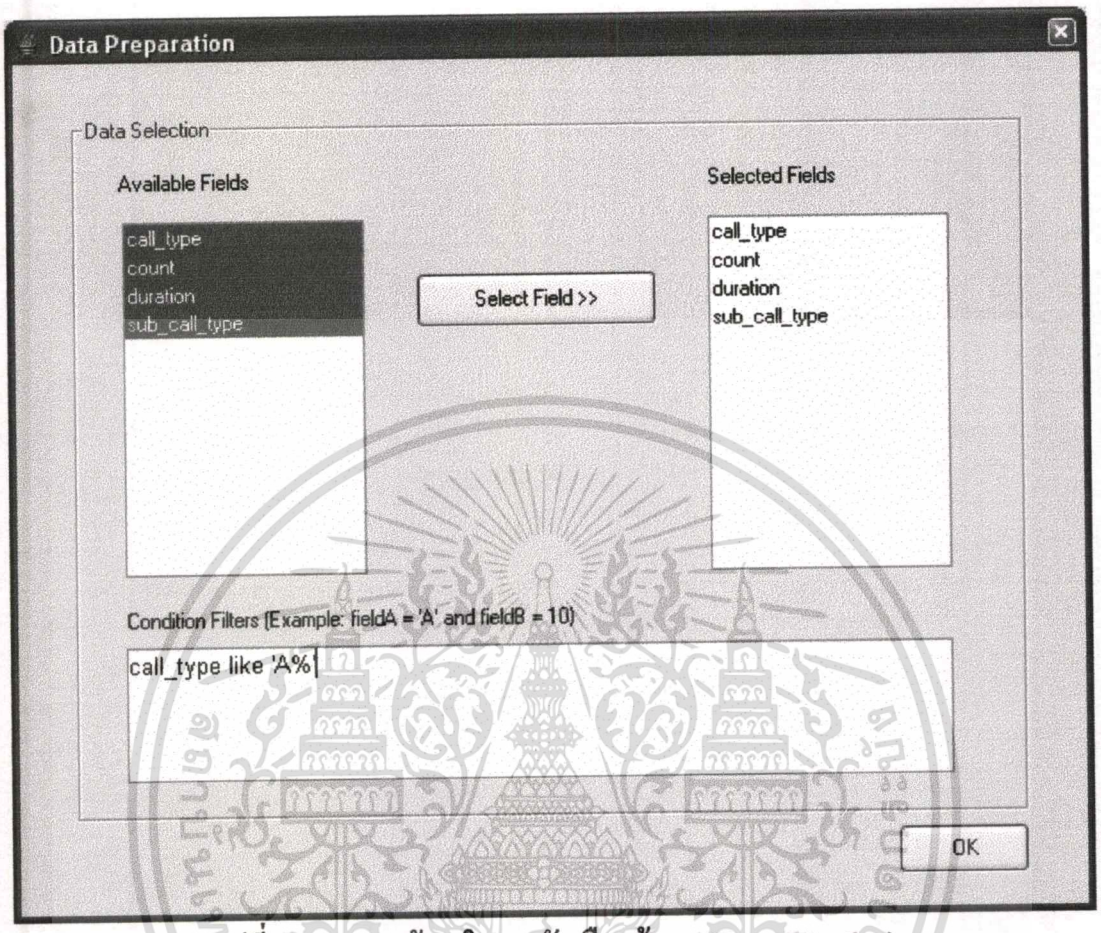


รูปที่ 6.5 แสดงหน้าจอในการติดต่อกับระบบฐานข้อมูล

หลังจากที่กรอกข้อมูลที่จำเป็นเรียบร้อยแล้ว ให้เลือกที่ปุ่ม Connect จากนั้นระบบก็จะทำการเชื่อมต่อกับระบบฐานข้อมูลที่ผู้ใช้ต้องการ และเข้าสู่ขั้นตอนที่ 2 ซึ่งเป็นการคัดเลือกข้อมูล (Data Selection) ต่อไป

❖ **ขั้นตอนที่ 2** การคัดเลือกข้อมูล (Data Selection)

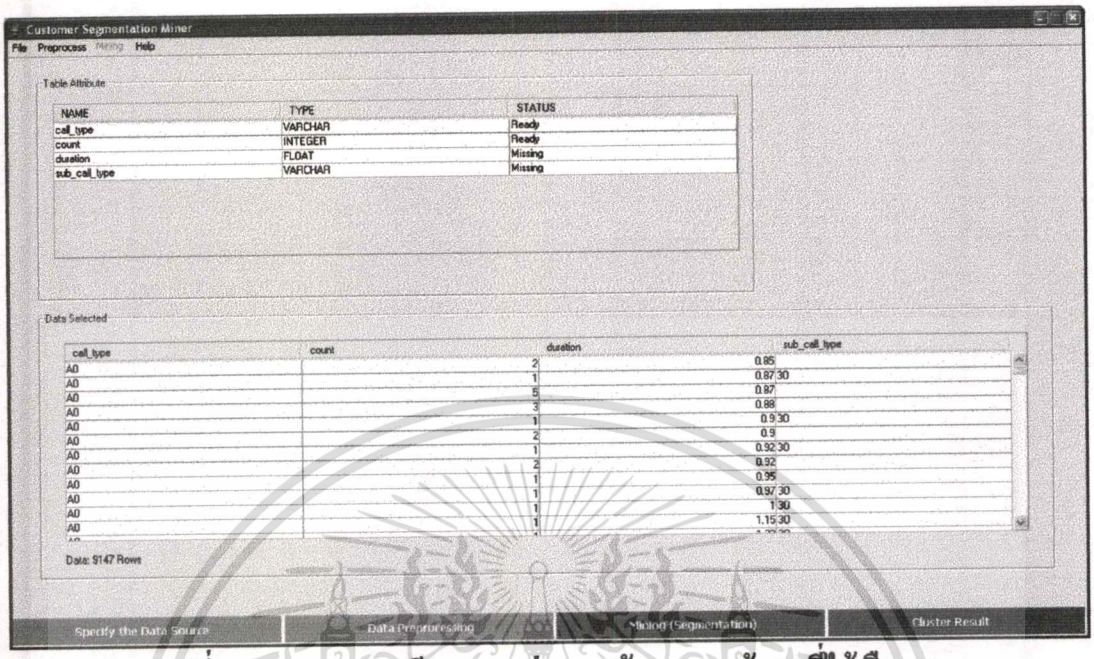
การคัดเลือกข้อมูลเป็นกระบวนการในการเลือกเซตข้อมูล และการเลือกเอาเฉพาะข้อมูลที่เราสนใจออกมา เพื่อที่จะได้สามารถทำการจัดแบ่งกลุ่มข้อมูลลูกค้าได้อย่างถูกต้องตรงตามวัตถุประสงค์ที่ระบุไว้ในการศึกษาตอนแรก ดังนั้นการคัดเลือกข้อมูลควรจะระมัดระวังในการเลือกเซตข้อมูลที่มีผลกระทบต่อสิ่งที่เราต้องการศึกษาจริงๆ หากเลือกไม่ถูกต้องก็จะส่งผลให้ผลลัพธ์ที่ออกมาอาจจะไม่ถูกต้องตามไปด้วยก็เป็นได้ สำหรับหน้าจอของการคัดเลือกข้อมูลสามารถแสดงได้ดังภาพด้านล่าง



รูปที่ 6.6 แสดงหน้าจอในการคัดเลือกข้อมูล (Data Selection)

จากหน้าจอจะแสดงรายการเขตข้อมูลที่สามารถเลือกได้อยู่ทางด้านซ้าย ซึ่งผู้ใช้สามารถเลือกได้หลายๆเขตข้อมูลพร้อมๆกัน ด้วยการใช้คีย์ Shift หรือ Ctrl ในการเลือก เมื่อผู้ใช้เลือกเขตข้อมูลเสร็จเรียบร้อยแล้วก็ให้ทำการกดปุ่ม Select Fields เพื่อให้ระบบรับทราบว่าได้เลือกเขตข้อมูลอะไรไปบ้าง หลังจากนั้นระบบก็จะทำการแสดงรายการของเขตข้อมูลที่ผู้ใช้เลือกไว้ทางด้านขวา

สำหรับผู้ใช้ที่ต้องการกรองข้อมูลเพียงบางส่วนเพื่อนำไปใช้ในการวิเคราะห์ก็สามารถทำได้เช่นกัน โดยการใส่เงื่อนไขเพิ่มเข้าไปในรูปแบบเงื่อนไขของภาษา SQL ในช่องของ Text area ทางด้านล่าง ยกตัวอย่างเช่น call_type like 'A%' เป็นต้น เมื่อเลือกรายการเขตข้อมูล และใส่เงื่อนไขเพื่อกรองเอาข้อมูลที่เราต้องการเสร็จเรียบร้อยแล้วให้กดปุ่ม OK เพื่อทำการยืนยันว่ากระบวนการคัดเลือกข้อมูลนั้นถูกต้อง หลังจากนั้นระบบจะแสดงรายละเอียดเกี่ยวกับแต่ละเขตข้อมูล รวมไปถึงข้อมูลทั้งหมดที่ได้เลือกมาอีกด้วย ซึ่งแสดงให้เห็นบนหน้าจอดังภาพ

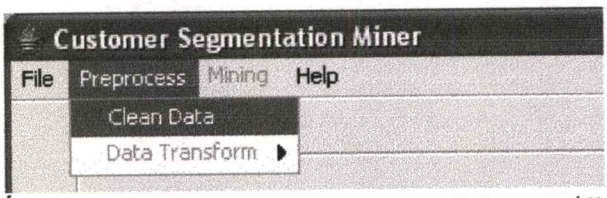


รูปที่ 6.7 แสดงรายละเอียดของแต่ละเซตข้อมูล และข้อมูลที่ได้เลือกมา

❖ **ขั้นตอนที่ 3** การเตรียมข้อมูลก่อนการประมวลผล (Data Preprocessing)

หลังจากที่เราได้ทำการเลือกข้อมูลที่จะนำมาใช้ในการจัดแบ่งกลุ่มเป็นที่เรียบร้อยแล้ว ขั้นตอนต่อไปก็คือ การเตรียมข้อมูลเพื่อให้เหมาะสมก่อนที่จะนำไปประมวลผล หรือการไมนิ่งข้อมูลด้วยอัลกอริทึมที่ได้เลือกไว้ก่อนหน้า เมื่อเข้าสู่ขั้นตอนนี้โปรแกรมจะแสดงแถบสถานะสีเขียวไปอยู่ที่ส่วนของการเตรียมข้อมูล (Data Preprocessing) การทำงานในส่วนนี้จะแบ่งออกเป็น 2 ขั้นตอนย่อยๆด้วยกันดังต่อไปนี้

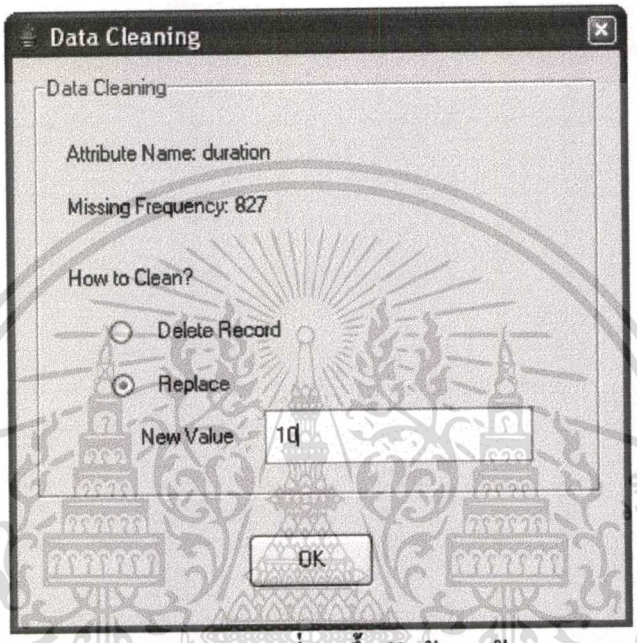
1) **Data Cleaning** เป็นการกำจัดความไม่สมบูรณ์ หรือความผิดปกติที่เกิดขึ้นในแต่ละเซตข้อมูลที่เลือกมา วิธีใช้งานก็คือให้เลือกที่รายการเมนูคำสั่ง Preprocess จากนั้นเลือกคำสั่ง Clean Data ส่วนของหน้าจอวิธีเรียกใช้งานสามารถแสดงได้ดังภาพ



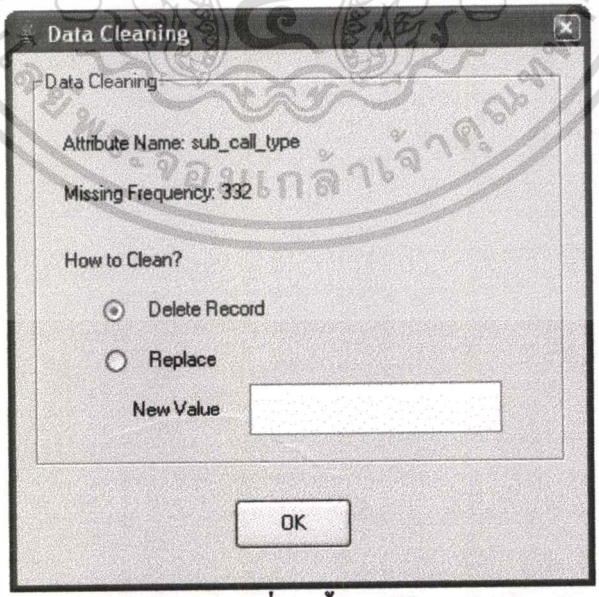
รูปที่ 6.8 แสดงวิธีเรียกใช้งานการกำจัดความไม่สมบูรณ์ข้อมูล

หลังจากเรียกใช้งานระบบก็จะทำการวิเคราะห์เซตข้อมูลที่เลือกออกมาว่ามีความผิดปกติอะไรเกิดขึ้นบ้าง พร้อมทั้งแสดงวิธีการในการจัดการแก้ไขความผิดปกติที่เกิดขึ้นเหล่านั้นอีกด้วย

สำหรับวิธีแก้ไขที่ระบบนำเสนอจะมีอยู่ด้วยกัน 2 แบบคือ การตัดชุดข้อมูลที่มีความไม่สมบูรณ์อยู่เหล่านั้นทิ้งไป และการทดแทนค่าที่ขาดหายไป ส่วนของหน้าจอที่แสดงวิธีการแก้ไขสามารถแสดงได้ดังภาพ

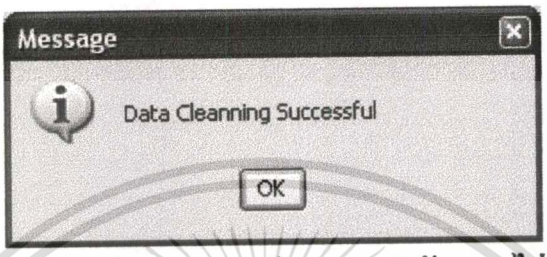


รูปที่ 6.9 แสดงวิธีในการแก้ไขความผิดปกติที่เกิดขึ้นกับข้อมูลด้วยการทดแทนค่าที่ขาดหายไป



รูปที่ 6.10 แสดงวิธีในการแก้ไขความผิดปกติที่เกิดขึ้นกับข้อมูลด้วยการตัดชุดข้อมูลเหล่านั้นทิ้งไป

หลังจากเลือกวิธีที่ใช้ในการจัดการแก้ไขความผิดปกติของแต่ละข้อมูลเป็นที่เรียบร้อยให้เลือกที่ปุ่ม OK ตอบตกลงยืนยันเพื่อระบบจะได้ดำเนินการปรับปรุงแก้ไขข้อมูลที่ได้เลือกมาจนเสร็จสิ้นครบทุกเขตข้อมูล และรายงานผลการทำงานออกมาดังภาพ



รูปที่ 6.11 ระบบแสดงรายงานผลความสำเร็จของการกำจัดความไม่สมบูรณ์ของข้อมูล

เมื่อเราได้กำจัดความผิดปกติที่เกิดขึ้นในแต่ละเขตข้อมูลเป็นที่เรียบร้อยแล้วขั้นตอนต่อไปก็คือการแปลงข้อมูลนั่นเอง

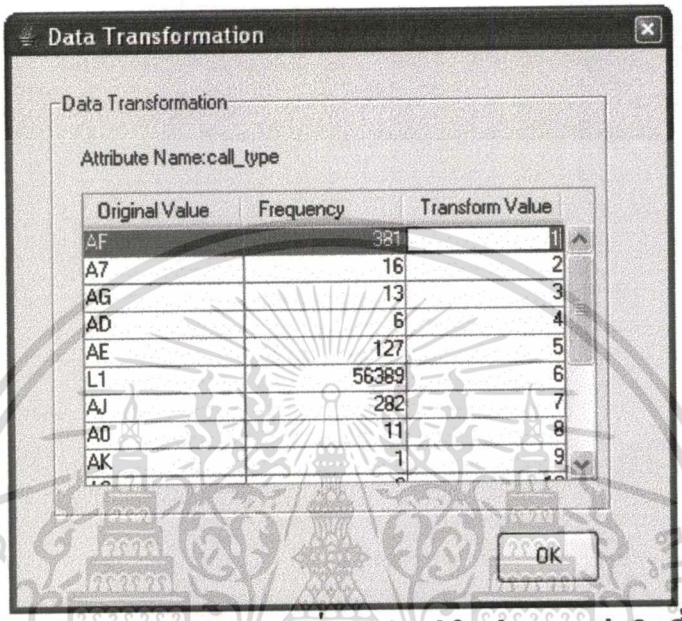
2) **Data Transformation** ในส่วนของขั้นตอนย่อยนี้นั้นจะประกอบด้วย 2 ขั้นตอนที่สำคัญดังต่อไปนี้

- **Transform** เป็นการปรับปรุง และเปลี่ยนแปลงค่าที่อยู่ในแต่ละเขตข้อมูล เพื่อให้เกิดความเหมาะสมกับอัลกอริทึมที่เราได้เลือกนำมาใช้ ซึ่งในที่นี้อัลกอริทึมที่ใช้นั้นคือ Fuzzy C-Means โดยจะสามารถทำงานได้กับค่าของข้อมูลที่เป็นตัวเลขเท่านั้น ดังนั้นจึงต้องมีการปรับเปลี่ยนค่าของข้อมูลก่อนการประมวลผลให้ถูกต้องเสียก่อน วิธีการในการปรับปรุง และเปลี่ยนแปลงค่าที่อยู่ในเขตข้อมูลของโปรแกรมก็คือให้เลือกที่รายการเมนูคำสั่ง Preprocess จากนั้นให้เลือก Data Transform แล้วจึงเลือกคำสั่ง Transform ซึ่งสามารถแสดงได้ดังภาพ

รูปที่ 6.12 แสดงวิธีการใช้งานการปรับปรุง และเปลี่ยนแปลงค่าในเขตข้อมูล (Transform Data)

เมื่อเลือกรายการคำสั่งเป็นที่เรียบร้อยแล้ว ระบบก็จะทำการวิเคราะห์ว่าเขตข้อมูลใดบ้างที่มีชนิดข้อมูลที่ไม่ใช่ตัวเลข หรือไม่สามารถคำนวณได้ หลังจากนั้นระบบก็จะแสดงหน้าจอเพื่อให้

ผู้ใช้สามารถทำการกำหนด และปรับปรุงเปลี่ยนแปลงค่าใหม่ลงไปแทนค่าเดิมที่มีอยู่ได้นั่นเอง ซึ่งสามารถแสดงหน้าจอการปรับปรุง และเปลี่ยนแปลงค่าต่างๆได้ดังภาพข้างล่าง



รูปที่ 6.13 แสดงตัวอย่างการปรับปรุง และเปลี่ยนแปลงค่าใหม่ทดแทนค่าเดิมที่มีอยู่ให้เหมาะสม

หลังจากที่ได้ทำการปรับปรุง และเปลี่ยนแปลงค่าต่างๆเป็นที่เรียบร้อยแล้ว ระบบก็จะแสดงรายงานผลความสำเร็จในการทำงานออกมาซึ่งแสดงได้ดังภาพ

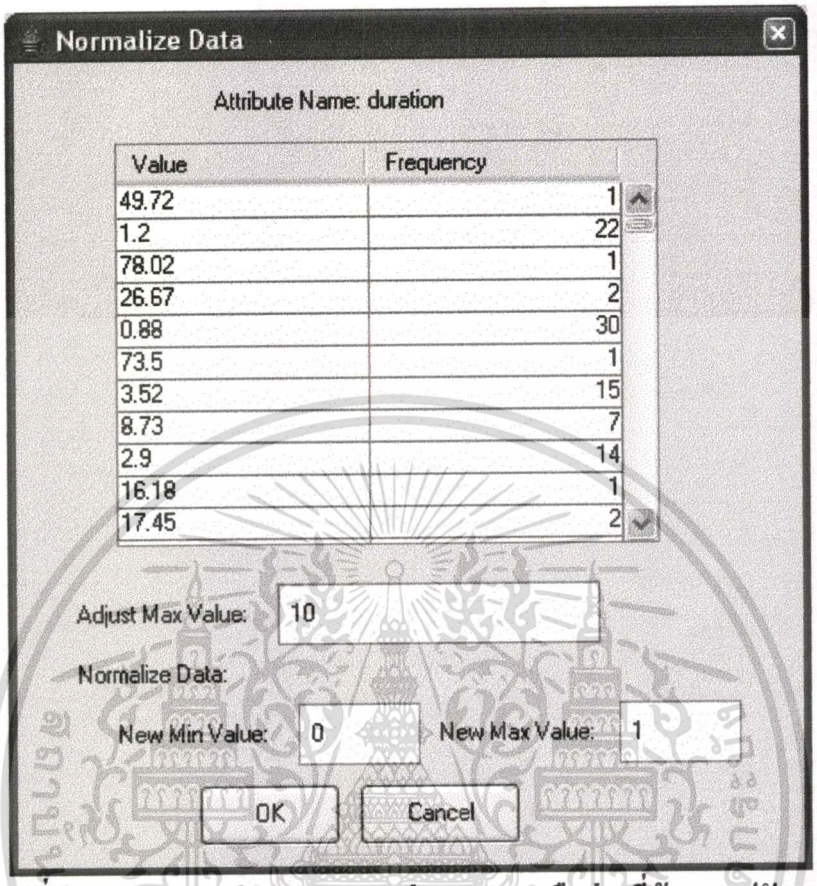
รูปที่ 6.14 แสดงผลรายงานความสำเร็จของการปรับปรุง และเปลี่ยนแปลงค่า (Transform Data)

- **Normalize** สำหรับขั้นตอนนี้จะเป็นการปรับปรุงค่าให้อยู่ในช่วงที่เราต้องการ เพราะเนื่องจากค่าในบางเซตข้อมูลมีช่วงของค่าที่มากเกินไป ซึ่งอาจจะทำให้ไม่เหมาะสมต่อการนำไปประมวลผล หรือการทำไมนิ่งข้อมูลต่อภายในอัลกิริทึมนั้นๆ ยกตัวอย่างเช่น ค่าของเงินเดือน ซึ่งบางกรณีลูกค้าแต่ละคนอาจจะมีค่าเงินเดือนที่แตกต่างกันมาก เช่น ลูกค้าคนหนึ่งมีเงินเดือนเพียง 3,000 บาท แต่ลูกค้าอีกคนหนึ่งกลับมีเงินเดือนถึง 100,000 บาท ซึ่งจะเห็นได้ว่าลูกค้ามีค่าของเงินเดือน

อยู่ในช่วงที่แตกต่างกันมาก ดังนั้นจึงต้องมีการปรับลดค่าให้อยู่ในช่วงเดียวกัน เพื่อที่ว่าการจัดแบ่งลูกค้าจะได้สามารถทำงานได้อย่างถูกต้อง แต่ในบางกรณีที่ค่าของเขตข้อมูลอาจจะมีความไม่แตกต่างกันมากนัก ดังนั้นขั้นตอนนี้จึงอาจไม่มีความจำเป็นที่เราจะต้องทำก็เป็นได้ ซึ่งจะทำให้เราสามารถข้ามขั้นตอนนี้ไปสู่ขั้นตอนไมนิ่งข้อมูลต่อไป วิธีการในการปรับลดค่าของเขตข้อมูลให้อยู่ในช่วง สามารถทำได้ดังนี้ ให้เลือกรายการเมนูคำสั่ง Preprocess จากนั้นให้เลือก Data Transform แล้วจึงเลือก Normalize หน้าจอวิธีการใช้งานสามารถแสดงได้ดังภาพ

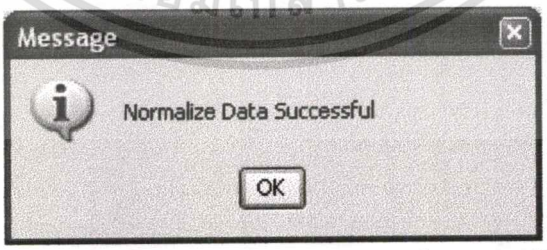
รูปที่ 6.15 แสดงวิธีการใช้งานการปรับค่าของเขตข้อมูลให้อยู่ในช่วง (Normalize)

ระบบจะแสดงรายละเอียดเกี่ยวกับความถี่แต่ละค่าของข้อมูลภายในเขตข้อมูลนั้นๆ พร้อมทั้งให้เราสามารถกำหนดค่าต่ำสุด และค่าสูงสุดซึ่งเป็นขอบเขตของช่วงที่เราต้องการปรับได้ นอกจากนี้ยังสามารถกำหนดค่าสูงสุดของข้อมูลได้อีกด้วย เพื่อเป็นการลดความแตกต่างของข้อมูลลงก่อนจะปรับค่าให้อยู่ในช่วงที่เราต้องการ ซึ่งการกำหนดค่าสูงสุดของข้อมูลนี้อาจไม่จำเป็นต้องระบุก็ได้ถ้าความแตกต่างของข้อมูลมีไม่มากนัก เมื่อเรากำหนดขอบเขตของช่วงเสร็จเรียบร้อยแล้วให้เลือกที่ปุ่มตกลง (OK) แต่ถ้าไม่ต้องการปรับในเขตข้อมูลนั้นก็ให้เลือกที่ปุ่มยกเลิก (Cancel) ซึ่งระบบจะไม่ทำการปรับลดค่าของเขตข้อมูลนั้นให้ แต่จะข้ามไปดำเนินการกับเขตข้อมูลถัดไปนั่นเอง ส่วนของหน้าจอที่แสดงผลในการกำหนดค่าขอบเขต หรือช่วงที่เราต้องการสามารถแสดงได้ดังภาพ



รูปที่ 6.16 แสดงตัวอย่างการกำหนดค่าขอบเขต หรือช่วงที่ต้องการปรับลด

หลังจากที่ระบบดำเนินการปรับค่าในเขตข้อมูลให้อยู่ในช่วงเป็นที่เรียบร้อย ระบบก็จะแสดงผลรายงานความสำเร็จอีกด้วย ซึ่งสามารถแสดงได้ดังภาพ

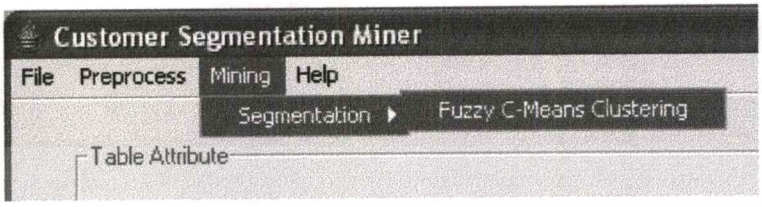


รูปที่ 6.17 แสดงผลรายงานความสำเร็จในการปรับลดค่าของเขตข้อมูลให้อยู่ในช่วง

❖ **<u>ขั้นตอนที่ 4</u>** จัดแบ่งกลุ่มข้อมูลด้วยวิธี Fuzzy C-Means Clustering Algorithm

เป็นกระบวนการของการทำไมนิ่งข้อมูลอย่างแท้จริง ซึ่งโปรแกรมระบบที่ใช้งานอยู่นี้จะมีรูปแบบเฉพาะการจัดแบ่งกลุ่มข้อมูล (Data Segmentation) กับอัลกอริทึม Fuzzy C-Means Clustering Algorithm ที่ได้เลือกไว้เพียงเท่านั้น โดยมีวิธีการใช้งานคือให้เลือกที่รายการเมนูคำสั่ง

Mining จากนั้นเลือก Segmentation แล้วจึงเลือกอัลกอริทึมที่ได้เลือกไว้นั่นคือ Fuzzy C-Means Clustering ส่วนของวิธีการใช้งานสามารถแสดงได้ดังภาพ

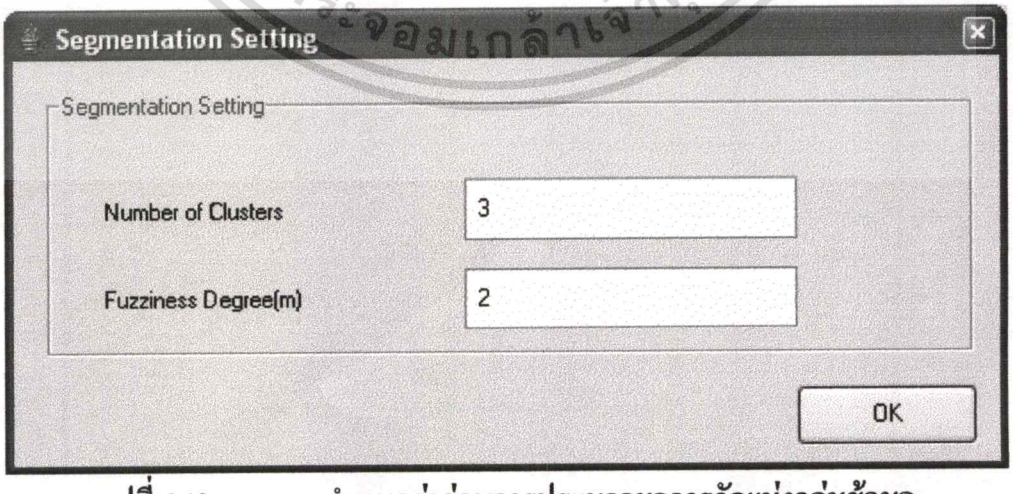


รูปที่ 6.18 แสดงวิธีการใช้งานระบบของการจัดแบ่งกลุ่มข้อมูลด้วยวิธี Fuzzy C-Means Clustering

เมื่อสั่งให้ระบบดำเนินการจัดแบ่งกลุ่มข้อมูลแล้ว แถบสถานะสีเขียวของโปรแกรมก็จะเลื่อนไปอยู่ ณ ตำแหน่งในส่วนของขั้นตอนการไมนิ่งข้อมูลนั่นเอง (Segmentation) และก่อนที่ระบบจะเริ่มทำงานจริงๆ ระบบจะเปิดหน้าจอเพื่อให้ทำการกำหนดค่าต่างๆในการจัดแบ่งกลุ่มของข้อมูลเสียก่อน เพื่อให้อัลกอริทึมสามารถทำงานได้อย่างถูกต้อง โดยสิ่งที่ระบบจะให้กำหนดนั้นมีรายละเอียดดังต่อไปนี้

1) จำนวนกลุ่มที่ต้องการในการจัดแบ่ง (Cluster)
2) Fuzziness Degree (m)

ทั้งสองค่านี้จะมีผลอย่างยิ่งในการจัดแบ่งกลุ่มข้อมูล ดังนั้นผู้ใช้ควรจะระมัดระวังตรงจุดนี้ด้วย การกำหนดทั้งสองค่านี้จึงไม่ควรที่จะกำหนดให้มากจนเกินไป หรือน้อยจนเกินไปนั่นเอง ส่วนของหน้าจอในการกำหนดค่าต่างๆในการจัดแบ่งกลุ่มข้อมูลสามารถแสดงได้ดังภาพข้างล่าง



รูปที่ 6.19 แสดงการกำหนดค่าก่อนการประมวลผลการจัดแบ่งกลุ่มข้อมูล

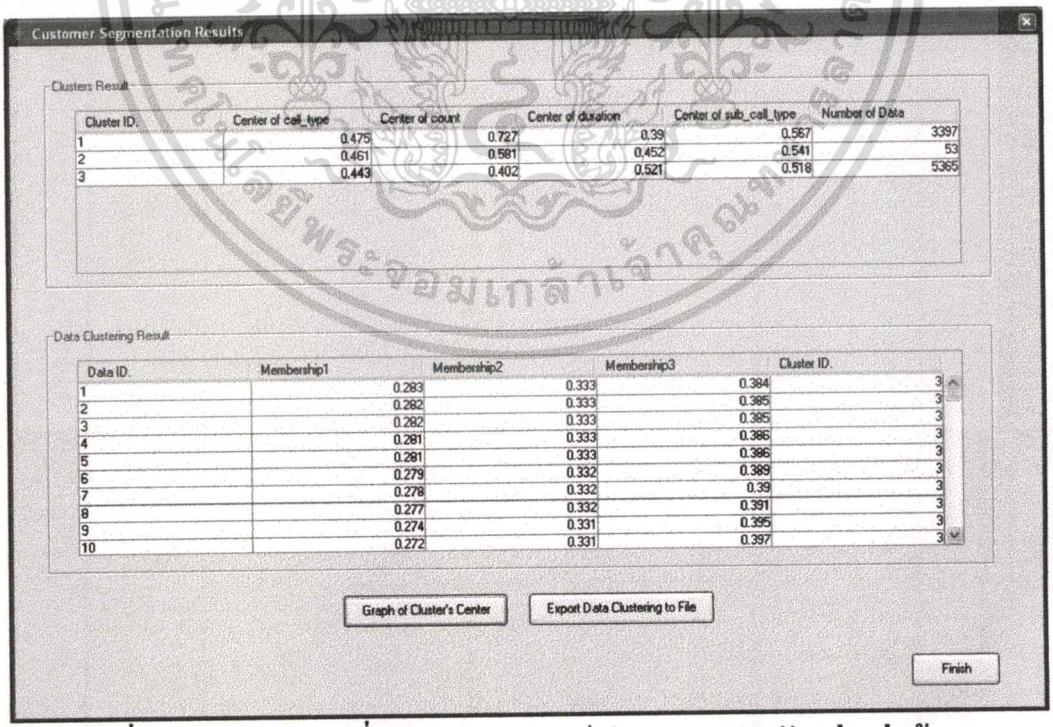
หลังจากที่กำหนดค่าเป็นที่เรียบร้อยแล้วให้ทำการกดปุ่มตกลง (OK) เพื่อเป็นการยืนยันการกำหนดค่า จากนั้นระบบจะทำการประมวลผล วิเคราะห์ และจัดแบ่งกลุ่มข้อมูลของลูกค้าให้

❖ **ขั้นตอนที่ 5** การแสดงผลลัพธ์การจัดแบ่งกลุ่มข้อมูลของลูกค้า

เมื่อระบบทำการประมวลผล วิเคราะห์ และจัดแบ่งกลุ่มข้อมูลลูกค้าเป็นที่เรียบร้อยแล้ว แถบสถานะสีเขียวของโปรแกรมก็จะถูกเลื่อนไปอยู่ ณ ตำแหน่งสุดท้ายเพื่อเป็นการบอกให้ผู้ใช้ทราบว่าระบบได้ทำงานเสร็จสิ้นแล้ว และจะแสดงผลลัพธ์ที่ได้จากการประมวลผลจากการจัดแบ่งกลุ่มข้อมูลแบ่งแยกออกเป็นสองส่วนด้วยกันซึ่งมีรายละเอียดดังต่อไปนี้

1) ส่วนแรกจะเป็นการแสดงเกี่ยวกับผลลัพธ์ที่ได้ในแต่ละกลุ่ม ซึ่งมีรายละเอียดต่างๆอันได้แก่ กลุ่มที่เท่าไหร่ ค่าจุดศูนย์กลางของกลุ่มนั้นๆในแต่ละเซตข้อมูล จำนวนข้อมูลที่อยู่ในแต่ละกลุ่ม
2) ส่วนที่สองเป็นการแสดงรายละเอียดเกี่ยวกับผลลัพธ์ของแต่ละชุดข้อมูล ซึ่งจะบ่งบอกค่าระดับความเป็นสมาชิกในแต่ละกลุ่มของชุดข้อมูลแต่ละชุด ซึ่งจะทำให้เราทราบได้ว่าข้อมูลแต่ละชุดนั้นจัดอยู่ในกลุ่มใด

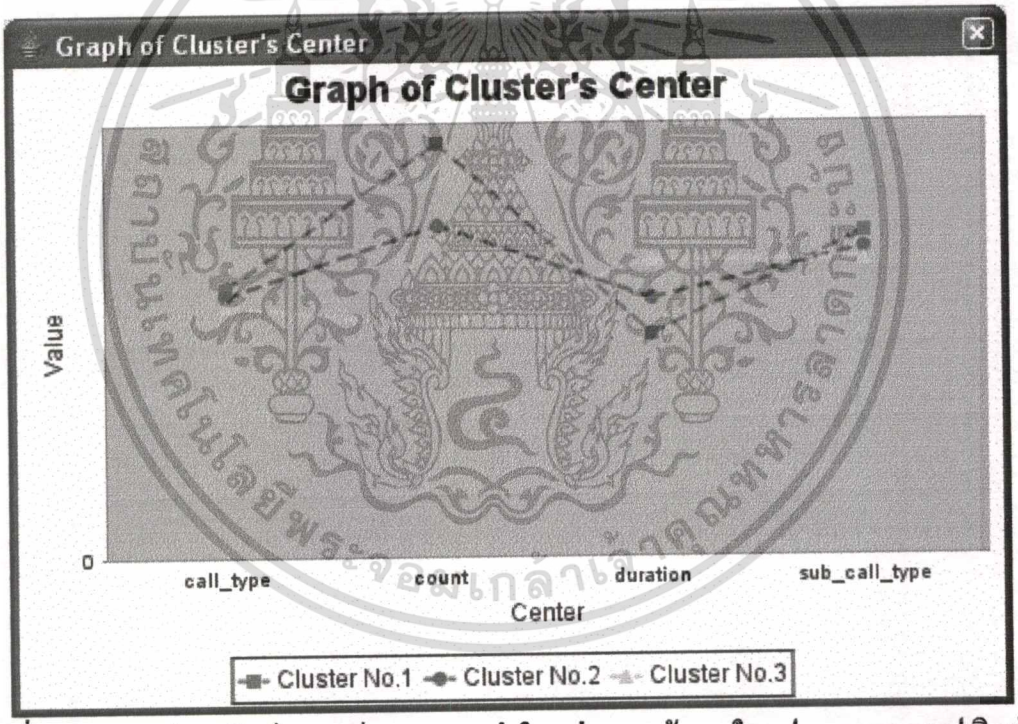
สำหรับหน้าจอในการแสดงผลลัพธ์ของการจัดแบ่งกลุ่มข้อมูลนั้นสามารถแสดงได้ดังภาพ



รูปที่ 6.20 แสดงผลลัพธ์ที่ได้จากการวิเคราะห์ ประมวลผลการจัดแบ่งกลุ่มข้อมูล

จากหน้าจอแสดงผลลัพธ์ที่ได้จากการประมวลผล วิเคราะห์ จัดแบ่งกลุ่มข้อมูลลูกค้านั้น จะเห็นได้ว่าผู้ใช้สามารถที่จะดูผลลัพธ์ในรูปแบบอื่นๆ ได้อีก นอกจากการแสดงผลลัพธ์ที่เป็นตารางข้อมูลเหล่านั้นแล้ว อันได้แก่

1) การแสดงค่าผลลัพธ์จุดศูนย์กลางของกลุ่มในแต่ละเซตข้อมูล ซึ่งมีลักษณะเป็นกราฟเชิงเส้น โดยกราฟแต่ละเส้นจะหมายถึงกลุ่มแต่ละกลุ่ม และค่าในแกนตั้งก็คือค่าจากจุดศูนย์กลางของกลุ่มในแต่ละเซตข้อมูล ส่วนแกนนอนก็คือเซตข้อมูลที่เราเลือกมานั่นเอง วิธีการดูผลลัพธ์ในรูปแบบกราฟเชิงเส้นสามารถทำได้โดยเลือกที่ปุ่ม Graph of Cluster's Center จากหน้าจอแสดงผลลัพธ์ ระบบก็จะแสดงค่าของจุดศูนย์กลางในรูปแบบของกราฟเชิงเส้น เราสามารถแสดงกราฟเชิงเส้นได้ดังรูป

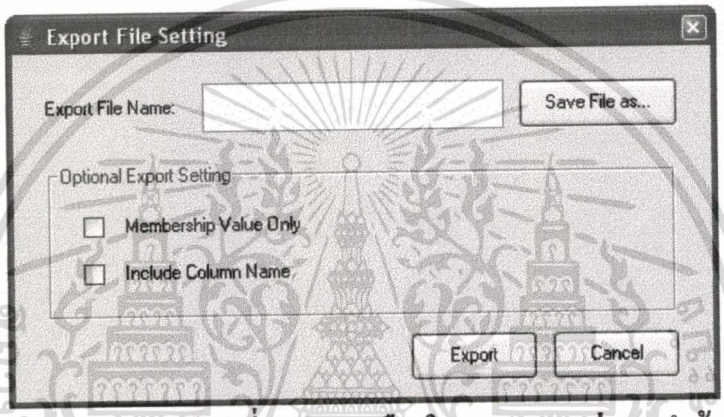


รูปที่ 6.21 แสดงค่าผลลัพธ์จุดศูนย์กลางของกลุ่มในแต่ละเซตข้อมูล ในรูปแบบของกราฟเชิงเส้น

2) ผู้ใช้สามารถที่จะนำผลลัพธ์ที่ได้จากข้อมูลแต่ละชุดออกเป็นไฟล์ในรูปแบบต่างๆ ได้ถึง 3 ประเภทด้วยกันดังต่อไปนี้

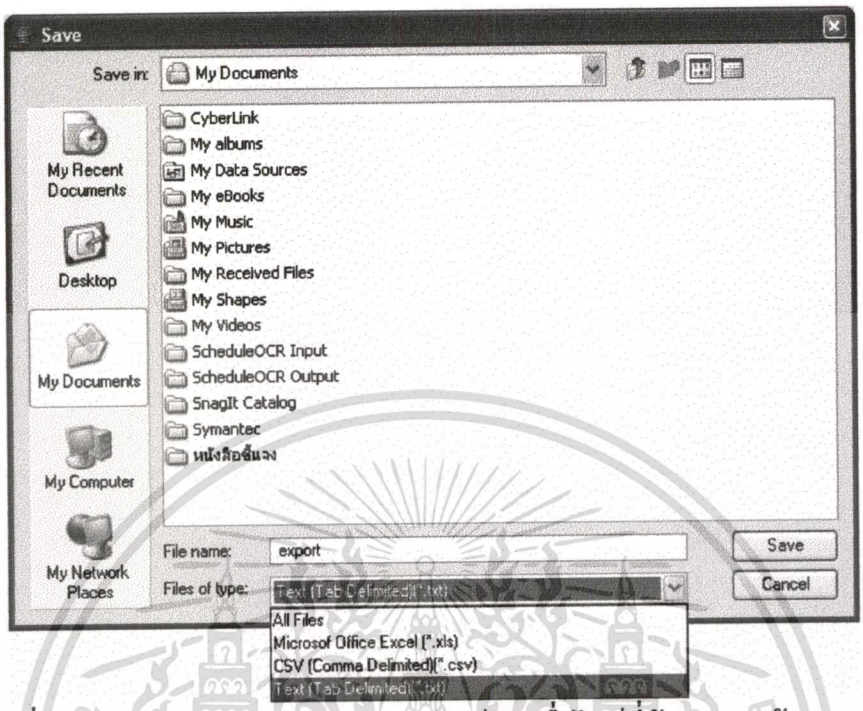
- รูปแบบของ Microsoft Office Excel
- รูปแบบของ CSV File (Comma Delimited)
- รูปแบบของ Text File (Tab Delimited)

วิธีการในการนำผลลัพธ์ที่ได้จากข้อมูลแต่ละชุดออกเป็นไฟล์สามารถทำได้ดังนี้ จากหน้าจอแสดงผลลัพธ์ให้เลือกที่ปุ่ม Export Data Clustering to File ระบบก็จะแสดงผลหน้าจอเกี่ยวกับกำหนดทางเลือกต่างในการนำข้อมูลออกได้แก่ ชื่อไฟล์ที่จะนำออก เป็นต้น นอกจากนี้ยังมีทางเลือกอื่นๆอีกอย่างเช่น การกำหนดความต้องการในการนำชื่อคอลัมน์ออกหรือไม่ และความต้องการในการนำค่าระดับความเป็นสมาชิกออกอย่างเดียวหรือไม่ ซึ่งสามารถแสดงหน้าจอในการกำหนดค่าของการนำไฟล์ออกดังภาพ



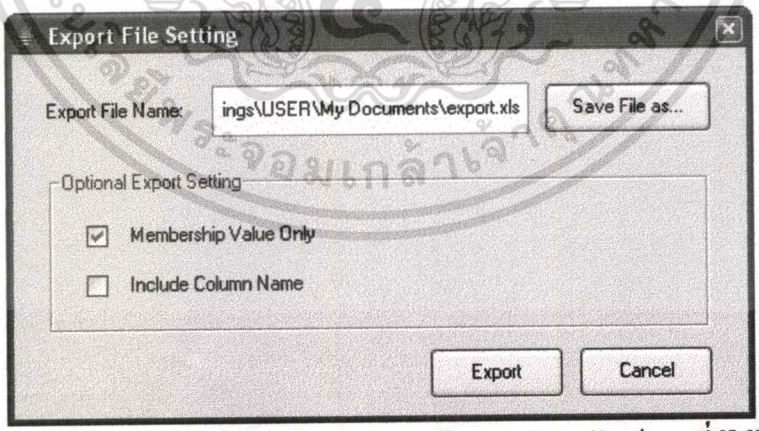
รูปที่ 6.22 แสดงรายละเอียดเกี่ยวกับทางเลือกในการกำหนดค่าการนำข้อมูลออก

หลังจากที่กำหนดความต้องการของทางเลือกในการนำข้อมูลออกแล้ว ให้กดเลือกที่ปุ่ม Save File as เพื่อทำการเปิดหน้าจอในการกำหนดประเภทของไฟล์ และชื่อไฟล์ที่จะนำออก โดยประเภทของไฟล์ที่สามารถเลือกได้จะมีอยู่ด้วยกัน 3 ประเภทตามที่กล่าวมาแล้วข้างต้น เมื่อเลือกประเภทของไฟล์ที่ต้องการนำออกเสร็จสิ้นแล้วก็ให้กำหนดชื่อไฟล์ให้เรียบร้อย จากนั้นให้กดปุ่ม Save เพื่อทำการยืนยัน ส่วนของหน้าจอในการกำหนดประเภทของไฟล์ และชื่อไฟล์สามารถแสดงได้ดังภาพข้างล่าง



รูปที่ 6.23 แสดงการกำหนดประเภทของไฟล์ และชื่อไฟล์ที่ต้องการนำข้อมูลออก

เมื่อได้กำหนดรายละเอียดของไฟล์ที่จะนำข้อมูลออกเสร็จสิ้นแล้ว ระบบก็จะกลับมาสู่หน้าจอของการกำหนดค่าต่างๆในการนำข้อมูลอีกครั้งหนึ่ง ซึ่งสามารถแสดงได้ดังภาพ



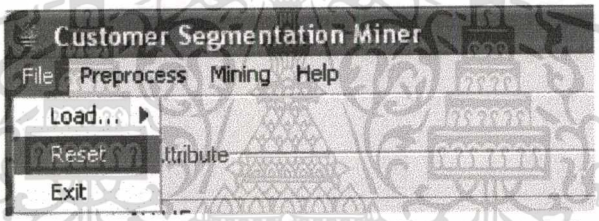
รูปที่ 6.24 แสดงรายละเอียดต่างๆของการนำข้อมูลออกไปยังไฟล์ตามที่ผู้ใช้กำหนด

หลังจากนั้นให้เลือกที่ปุ่มส่งออกไปยังไฟล์ (Export) ระบบก็จะดำเนินการนำผลลัพธ์ของข้อมูลที่ได้จากการจัดแบ่งกลุ่มเรียบร้อยแล้วนั้นนำออกไปยังไฟล์ที่ได้กำหนดไว้ พร้อมทั้งทำงานตามทางเลือกที่ผู้ใช้ได้ระบุไว้อีกด้วย ดังนั้นเราก็จะได้ไฟล์ข้อมูลที่นำออกมา ซึ่งสามารถที่จะนำไปใช้ประโยชน์ หรือใช้งานทางด้านอื่นๆได้ ยกตัวอย่างเช่น อาจสามารถนำไปใช้ใน

กระบวนการดาต้าไมนิ่งในรูปแบบอื่นเพื่อทำการวิเคราะห์หาแนวโน้ม (Predictive Modeling) เป็นต้น

สำหรับผู้ใช้ที่ต้องการจะออกจากหน้าจอการแสดงผลลัพธ์ของระบบก็ให้เลือกปุ่ม Finish ระบบก็จะออกจากหน้าจอการแสดงผลลัพธ์การจัดแบ่งกลุ่มข้อมูล เพื่อเข้าสู่หน้าจอหลักของโปรแกรมต่อไป

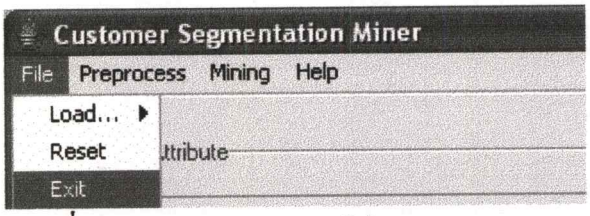
ในกรณีที่ผู้ใช้มีความต้องการที่จะเปลี่ยน หรือเลือกแหล่งข้อมูลที่มาใหม่ เพื่อที่จะนำข้อมูลจากที่อื่นมาวิเคราะห์ดูบ้างนั้นก็สามารถที่จะทำได้เช่นเดียวกัน เพียงแต่ผู้ใช้จะต้องทำการเคลียร์หรือจัดการข้อมูลอันเก่าที่แสดงผลอยู่บนหน้าจอหลักนั้นทิ้งไปเสียก่อนจึงจะสามารถเริ่มการทำงานครั้งใหม่ได้ วิธีการที่ใช้ในการเคลียร์ หรือจัดการข้อมูลอันเก่าบนหน้าจอหลักทิ้งนั้นสามารถทำได้โดยการเลือกที่รายการเมนูคำสั่ง File จากนั้นให้เลือก คำสั่ง Reset ดังภาพ



รูปที่ 6.25 แสดงวิธีการเคลียร์ หรือจัดการข้อมูลเก่าบนหน้าจอหลัก

เมื่อผู้ใช้จัดการเคลียร์ข้อมูลเก่าทิ้งไปจากหน้าจอหลักของโปรแกรมแล้วผลที่ได้จะทำให้หน้าจอหลักส่วนที่สองจะเกิดเป็นพื้นที่ว่างเปล่า และแถบสถานะสีเขียวก็จะย้อนกลับไปอยู่ ณ ตำแหน่งจุดเริ่มต้น หรือขั้นตอนแรกนั่นคือการระบุแหล่งที่มาของข้อมูลอีกครั้งหนึ่ง เหมือนกับตอนที่เพิ่งเริ่มทำงานกับระบบครั้งแรกนั่นเอง

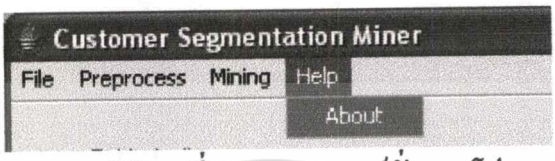
สำหรับผู้ใช้ที่ต้องการออกจากโปรแกรม เมื่อการทำงานเสร็จสิ้นก็ให้เลือกที่รายการเมนูคำสั่ง File จากนั้นเลือก Exit ระบบก็จะจบการทำงาน และปิดหน้าต่างของโปรแกรม วิธีการออกจากโปรแกรมสามารถแสดงได้ดังภาพ



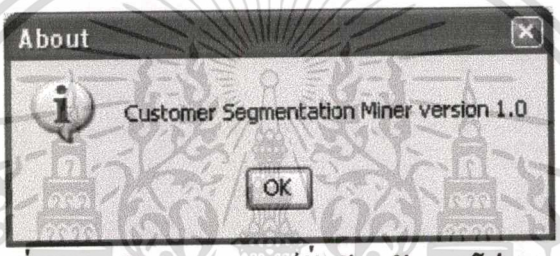
รูปที่ 6.26 แสดงการออกจากโปรแกรมของระบบ

❖ <u>รายการเมนูคำสั่งอื่นๆเพิ่มเติม</u>

สำหรับผู้ใช้ที่ต้องการทราบเกี่ยวกับเวอร์ชั่นของโปรแกรมสามารถเลือกเมนู Help จากนั้นให้เลือก About ระบบก็จะแสดงเวอร์ชั่นของโปรแกรมปัจจุบันให้ทราบได้ดังภาพข้างล่าง



รูปที่ 6.27 แสดงการใช้งานเกี่ยวกับการดูเวอร์ชั่นของโปรแกรมปัจจุบัน



รูปที่ 6.28 หน้าจอแสดงเวอร์ชั่นปัจจุบันของโปรแกรม

6.3 การวิเคราะห์ผลลัพธ์ที่ได้จากการทำดาต้าไมนิ่ง และการนำความรู้ไปใช้

หลังจากที่ได้เรียนรู้วิธีการใช้งานในการทำไมนิ่งจากโปรแกรมแล้วเป็นที่เรียบร้อย สุดท้ายเราก็จะนำผลลัพธ์ที่ได้จากระบบมาทำการวิเคราะห์ และตีความว่าผลลัพธ์ที่ได้มีคุณค่า และประโยชน์ตรงตามวัตถุประสงค์มากน้อยแค่ไหน

สำหรับการจัดแบ่งกลุ่มข้อมูลลูกค้าที่ใช้โทรศัพท์พื้นฐานด้วยกระบวนการดาต้าไมนิ่งที่ได้ทดลองทำมานั้น เราได้เริ่มจากการเตรียมข้อมูลโดยทำการจัดเก็บข้อมูลการใช้งานโทรศัพท์พื้นฐานของลูกค้าในรูปแบบของตาราง (Table) บนระบบฐานข้อมูลซึ่งมีจำนวนข้อมูลทั้งสิ้น 65,536 รายการ แต่ส่วนของข้อมูลที่ได้นำมาใช้จริงมีเพียง 9,147 รายการ เนื่องจากได้คัดเลือกข้อมูลของลูกค้าเฉพาะที่มีการใช้งานโทรศัพท์พื้นฐานทางด้านบริการ 1900 (บริการข้อมูลเสียง) เท่านั้นมาทดลองจัดแบ่งกลุ่ม ซึ่งจะมีประเภทของการโทรที่ขึ้นต้นด้วยตัวอักษรเอ (A) เมื่อได้คัดเลือกข้อมูลเสร็จสิ้นแล้วพบว่าเกิดความไม่สมบูรณ์ขึ้นในเขตข้อมูลของ duration และ sub_call_type จึงดำเนินการกำจัดความไม่สมบูรณ์ของข้อมูลเหล่านั้นออกไป ด้วยการทดแทนค่า พร้อมทั้งตัดข้อมูลบางส่วนทิ้งไปเนื่องจากค่าในเขตข้อมูลของบางรายการนั้นขาดหายไปจึงทำให้เหลือข้อมูลที่จะนำมาใช้ในการไมนิ่ง หรือการวิเคราะห์จัดแบ่งกลุ่มเพียงแค่ 8,815 รายการ หลังจากนั้นจึงนำไปผ่านขั้นตอนของการปรับเปลี่ยนแปลงค่าตัวอักษรต่างๆให้กลายเป็นตัวเลข (Transform) เพื่อให้สามารถทำงานได้กับอัลกอริทึม Fuzzy C-Means โดยจะปรับเปลี่ยนค่าเฉพาะในเขตข้อมูลที่มีชนิดของ

ข้อมูลเป็นตัวอักษร หรือข้อความเท่านั้นอันได้แก่เซตข้อมูล call_type และ sub_call_type จากนั้นก็ปรับลดค่าสูงสุดของค่าในเซตข้อมูล count และ duration ให้มีค่าเท่ากับ 2 และ 10 ตามลำดับ แล้วจึงปรับช่วงของข้อมูลเสียใหม่ให้ทุกๆเซตข้อมูลอยู่ในช่วงระหว่างค่าต่ำสุดคือ 0 และค่าสูงสุดคือ 1 จึงค่อยนำไปทำการจัดแบ่งกลุ่มต่อไป

ในการแบ่งกลุ่มครั้งนี้ได้กำหนดให้มีการแบ่งข้อมูลทั้งหมดออกเป็น 3 กลุ่มด้วยกัน พร้อมทั้งกำหนดค่าของตัวแปร m (Fuzziness Degree) เป็น 2 ซึ่งเป็นค่ามาตรฐานของการดำเนินการอยู่แล้ว ผลลัพธ์ของทั้ง 3 กลุ่มสามารถอธิบายได้ดังนี้

<u>กลุ่มที่ 1</u> จำนวนข้อมูลที่ถูกจัดแบ่งอยู่ในกลุ่มนี้มีทั้งหมด 3,397 รายการ ซึ่งนับได้ว่ามีจำนวนมากเป็นอันดับสองรองจากกลุ่มที่ 3 และมีค่าจุดศูนย์กลางในแต่ละเซตข้อมูลดังต่อไปนี้

- call_type = 0.475
- count = 0.727
- duration = 0.39
- sub_call_type = 0.567

<u>กลุ่มที่ 2</u> จำนวนข้อมูลที่ถูกจัดแบ่งไว้ในกลุ่มนี้มีทั้งหมด 53 รายการ ซึ่งมีจำนวนน้อยที่สุดเมื่อเปรียบเทียบกับกลุ่มอื่นๆ ส่วนค่าจุดศูนย์กลางในแต่ละเซตข้อมูลต่างๆแสดงได้ดังนี้

- call_type = 0.461
- count = 0.581
- duration = 0.452
- sub_call_type = 0.541

<u>กลุ่มที่ 3</u> สำหรับกลุ่มสุดท้ายมีจำนวนข้อมูลทั้งหมด 5,365 รายการ และมีจำนวนข้อมูลมากที่สุดเมื่อเปรียบเทียบกับกลุ่มอื่นๆ และมีค่าจุดศูนย์กลางของแต่ละเซตข้อมูลแสดงได้ดังนี้

- call_type = 0.443
- count = 0.402
- duration = 0.521
- sub_call_type = 0.518

จากผลลัพธ์ที่ได้ในแต่ละกลุ่ม เราสามารถวิเคราะห์ และสรุปได้ 2 ประการซึ่งสามารถอธิบายได้ดังต่อไปนี้

<u>ประการแรก</u> เราสามารถทราบเกี่ยวกับลักษณะของลูกค้าในแต่ละกลุ่มได้ โดยสังเกตุจากค่าจุดศูนย์กลางในแต่ละเซตข้อมูลของแต่ละกลุ่ม ซึ่งเป็นตัวแทน และบ่งบอกว่าสมาชิกที่อยู่ภายในนั้นเป็นอย่างไร ยกตัวอย่างเช่นในกลุ่มที่ 3 ซึ่งพบว่ามีลูกค้าถูกจัดอยู่ในกลุ่มนี้มากที่สุดนั้นมีลักษณะ

การใช้งานโทรศัพท์พื้นฐานทางด้านบริการ 1900 ประเภท A2 อยู่มากซึ่งมีราคา 3 บาท/นาที สาเหตุอันเนื่องมาจากค่าจุดศูนย์กลางของเขตข้อมูล call_type ในกลุ่มนี้มีค่าเท่ากับ 0.443 ซึ่งใกล้เคียงกับค่าของประเภทการโทรแบบ A2 ที่มีค่าเท่ากับ 0.5 นอกจากนี้ลูกค้าส่วนใหญ่ที่อยู่ในกลุ่มที่สาม ก็จะมีอัตราการใช้เวลาโทรอยู่ที่ 5.21 นาที หรือประมาณ 5 นาทีนั่นเอง ส่วนลูกค้าที่ถูกจัดอยู่ในกลุ่มที่หนึ่ง และกลุ่มที่สองนั้นส่วนใหญ่ก็จะมีการใช้งานโทรศัพท์พื้นฐานไปยังบริการ 1900 เช่นเดียวกันกันกลับกลุ่มที่สาม แต่จะแตกต่างกันตรงที่อัตราเวลาที่ใช้ในการโทรจะอยู่ที่ 3.9 นาทีสำหรับกลุ่มที่หนึ่ง และ 4.52 นาทีสำหรับกลุ่มที่สองตามลำดับ

ประการที่สอง เมื่อสังเกตุจากกราฟของค่าจุดศูนย์กลางในแต่ละเขตข้อมูลของแต่ละกลุ่มจะทำให้เราทราบได้ว่า เขตข้อมูลไหนบ้างที่มีความสำคัญ และเหมาะสมต่อการนำมาใช้งานในการจัดแบ่งกลุ่มข้อมูลของลูกค้า โดยเราจะพิจารณา และเปรียบเทียบจากค่าจุดศูนย์กลางของแต่ละกลุ่มในหนึ่งเขตข้อมูลถ้าหากค่าของจุดศูนย์กลางในเขตข้อมูลนั้นๆมีความแตกต่างกันมากในแต่ละกลุ่มก็หมายความว่าเขตข้อมูลนั้นมีความเหมาะสมต่อการที่จะนำไปใช้ในการพิจารณา หรือเป็นเกณฑ์หลักที่ใช้ในการจัดแบ่งกลุ่มลูกค้าต่อไปได้ แต่ถ้าหากค่าของจุดศูนย์กลางของแต่ละกลุ่มในเขตข้อมูลนั้นๆที่ได้มีค่าไม่แตกต่างกันมากนัก หรือแทบจะรวมกันเป็นจุดๆเดียวก็หมายความว่าเขตข้อมูลนั้นไม่มีความเหมาะสมที่จะนำมาใช้เป็นหลักเกณฑ์ในการพิจารณาการจัดแบ่งกลุ่มลูกค้าได้นั่นเอง ซึ่งจะทำให้ผลของการวิเคราะห์ที่ได้มีโอกาสผิดพลาด นอกจากนี้ยังไม่สามารถนำไปประเมินผลต่อ หรือเป็นประโยชน์ต่อการตัดสินใจต่างๆได้อีกด้วย

ปัจจัยที่มีผลกระทบต่อการทำงานของอัลกอริทึม Fuzzy C-Means ได้แก่

1) ถ้าข้อมูลมีความไม่สมบูรณ์ หรือความผิดปกติมากๆ จะทำให้การทำงานด้วยอัลกอริทึมนี้ได้ประสิทธิภาพที่ไม่ดี
2) การกำหนดจำนวนของกลุ่ม (Cluster)
3) การกำหนดค่าระดับความเป็นฟัซซี่ หรือค่าตัวแปร m (Fuzziness Degree)
4) การกำหนดค่าเริ่มต้นของ Matrix U ถ้ามีการกำหนดที่ไม่ดีจะทำให้จำนวนรอบในการทำงานเกิดการวนซ้ำหลายครั้ง ส่งผลต่อความเร็วในการประมวลผลการทำงาน

อย่างไรก็ตามผลลัพธ์ที่ได้จากการวิเคราะห์นี้อาจจะมีข้อผิดพลาดไปบ้างอันเนื่องมาจากเหตุปัจจัยหลายประการด้วยกัน ได้แก่ การเตรียมข้อมูล (Data Preparation) ตลอดไปจนถึงการกำจัดความไม่สมบูรณ์ของข้อมูล (Data Cleaning) การปรับเปลี่ยนค่าของข้อมูล และการปรับลดค่าให้อยู่ในช่วงที่เราต้องการ (Transform and Normalize Data) กระบวนการเหล่านี้ถ้าหากมีการปรับค่าที่ไม่เหมาะสมลงไปแล้วจะส่งผลกระทบก่อให้เกิดผลเสียต่อการไมนิ่งข้อมูลเป็นอย่างมาก หรืออาจ

ก่อให้เกิดความลำเอียงต่อข้อมูล ซึ่งจะทำให้ผลลัพธ์ที่ได้จากการวิเคราะห์จัดแบ่งกลุ่มข้อมูลของลูกค้าเสียหายได้ และไม่สามารถนำไปทำให้เกิดคุณค่าต่อธุรกิจ หรือองค์กรอีกด้วย

บทที่ 7

สรุปผลการศึกษา และข้อเสนอแนะ

โครงการพัฒนาระบบวิเคราะห์จัดแบ่งกลุ่มลูกค้าด้วยวิธี Fuzzy C-Means นี้ถูกจัดทำขึ้นเพื่อมุ่งเน้นการนำความรู้ทั้งภาคทฤษฎีของดาต้าไมนิ่ง และทฤษฎีของฟัซซี่เซตมาประยุกต์ใช้ให้เกิดประโยชน์กับการวิเคราะห์ข้อมูลการใช้งานโทรศัพท์พื้นฐานของลูกค้า นอกจากนี้ยังช่วยเพิ่มประสิทธิภาพในการศึกษาเกี่ยวกับพฤติกรรมการโทรของลูกค้าในแต่ละประเภทว่ามีลักษณะอย่างไรด้วย ผลลัพธ์ที่ได้จากการจัดแบ่งกลุ่มลูกค้าจะช่วยทำให้เรามองเห็นภาพรวมของการใช้งานโทรศัพท์ของลูกค้ามากยิ่งขึ้น และสามารถที่จะตอบสนองต่อความต้องการของลูกค้าด้วยโปรโมชั่นที่ดีได้อีกด้วย ซึ่งส่งผลให้องค์กรมีผลกำไรเพิ่มมากขึ้น และสามารถแย่งส่วนแบ่งทางการตลาดได้อีกทางหนึ่ง

7.1 สรุปผลการดำเนินงานของโครงการพัฒนาระบบ

ดาต้าไมนิ่งเป็นกระบวนการในการค้นหา และวิเคราะห์ข้อมูล เพื่อหาความสัมพันธ์ภายในข้อมูลนั้น หรือเพื่อให้ได้ข้อสารสนเทศบางประการที่เราสนใจออกมา โดยลักษณะของข้อสารสนเทศที่ได้นั้นจะต้องไม่ทราบมาก่อน และจะต้องตรงกับความเป็นจริง หรือสามารถเชื่อถือได้นั่นเอง ซึ่งขั้นตอนในการทำดาต้าไมนิ่งนั้นเริ่มจากการกำหนดวัตถุประสงค์ การเตรียมข้อมูล การไมนิ่งข้อมูล ตลอดจนการนำผลลัพธ์ที่ได้มาวิเคราะห์ และตีความเพื่อจะได้สามารถนำไปใช้ประกอบการตัดสินใจในเชิงธุรกิจ หรือเพื่อคุณประโยชน์ต่อการทำธุรกิจในด้านอื่นๆอีกมากมาย อาทิเช่น การบริหารจัดการ และการแย่งส่วนแบ่งทางการตลาด เป็นต้น

สำหรับโครงการศึกษานี้เป็นการพัฒนาระบบวิเคราะห์จัดแบ่งกลุ่มลูกค้าด้วยอัลกอริทึม Fuzzy C-Means ซึ่งเป็นเทคนิคที่ใช้ในการจัดแบ่งกลุ่มข้อมูลได้อย่างมีประสิทธิภาพวิธีหนึ่งของรูปแบบการแบ่งกลุ่มข้อมูล (Database Segmentation) ในกระบวนการของดาต้าไมนิ่ง โดยอาศัยหลักเกณฑ์ในการทำงานร่วมกันกับทฤษฎีฟัซซี่เซต การพิจารณาว่าชุดข้อมูลใดที่มีลักษณะเหมือนกัน หรือมีส่วนที่คล้ายคลึงกันจะถูกจัดแบ่งเข้าไว้ในกลุ่มเดียวกันนั้นจะสามารถพิจารณาได้จากค่าระดับความเป็นสมาชิก (Membership Grade) ในแต่ละกลุ่มของข้อมูลชุดนั้นๆ ซึ่งถ้าค่าระดับความเป็นสมาชิกในกลุ่มใดมีค่ามากที่สุดนั่นก็หมายความว่าข้อมูลชุดนั้นจะถูกจัดแบ่งไว้ในกลุ่มดังกล่าวนั่นเอง

ค่าที่บ่งบอกระดับความเป็นสมาชิกเหล่านี้ในความเป็นจริงแล้วก็คือ ค่าความน่าจะเป็นของชุดข้อมูลที่จะอยู่ในกลุ่มนั้นๆนั่นเอง การหาค่าระดับความเป็นสมาชิกในแต่ละกลุ่มเราสามารถคำนวณได้จาก Cost Function ของอัลกอริทึม Fuzzy C-Means ซึ่งจะมีลักษณะเป็นเมทริกซ์ และค่าที่อยู่ภายในแต่ละคอลัมน์จะบอกถึงระดับความเป็นสมาชิกในแต่ละกลุ่มกับชุดข้อมูลในแต่ละแถวนั่นเอง

ผลลัพธ์ที่ได้จากการวิเคราะห์ และจัดแบ่งกลุ่ม โดยมีข้อมูลนำเข้าเป็นชุดข้อมูลของยอดการใช้งานโทรศัพท์ในแต่ละวันตามประเภทการโทร และประเภทย่อยของการโทรไม่รวมประเภทการใช้งานโทรศัพท์ที่เกี่ยวข้องกับทางด้านอินเทอร์เน็ต และอินเทอร์เน็ตความเร็วสูง ช่วยทำให้เราได้ทราบถึงลักษณะการใช้งานโทรศัพท์พื้นฐานของกลุ่มลูกค้าในแต่ละกลุ่มว่าเป็นอย่างไร และปัจจัยอะไรบ้างที่มีผลต่อการใช้งานของลูกค้า สิ่งเหล่านี้จะช่วยให้เราสามารถเข้าใจพฤติกรรมของลูกค้า อีกทั้งยังสามารถตอบสนองต่อความต้องการของลูกค้าที่เพิ่มขึ้นได้อีกทางด้วย นอกจากนี้ก็ยังช่วยในเรื่องของการประหยัดเวลาในการวิเคราะห์ข้อมูลจำนวนมากๆ ซึ่งจะส่งผลให้เกิดประโยชน์ต่อธุรกิจขององค์กรนั้นๆต่อไป

อย่างไรก็ตามผลลัพธ์ที่ได้จากการวิเคราะห์ และการไมนิ่งข้อมูลนั้นอาจเกิดความผิดพลาดขึ้นได้เสมอ ทั้งนี้ทั้งนั้นอาจเกิดขึ้นจากหลายเหตุปัจจัยด้วยกันไม่ว่าจะเป็นการไม่เข้าใจถึงปัญหาอย่างแท้จริง จึงทำให้เราไม่สามารถกำหนดวัตถุประสงค์ได้อย่างชัดเจนตรงตามความต้องการ นอกจากนี้ขั้นตอนในกระบวนการทำดาต้าไมนิ่งก็มีส่วนสำคัญเป็นอย่างยิ่งในการทำงานเริ่มตั้งแต่การคัดเลือกข้อมูล ตลอดไปจนถึงการเตรียมข้อมูลก่อนการไมนิ่งข้อมูล สิ่งเหล่านี้ถ้าไม่สามารถควบคุมการทำงานได้ก็จะก่อให้เกิดผลลัพธ์จากการวิเคราะห์ที่ผิดเพี้ยนไป และจะส่งผลให้การทำดาต้าไมนิ่งนั้นไม่เกิดคุณค่าต่อการแก้ปัญหาทางธุรกิจได้

7.2 ข้อเสนอแนะ

สำหรับโครงงานพัฒนาระบบนี้เป็นเพียงแค่พื้นฐานในการจัดแบ่งกลุ่มลูกค้า (Customer Segmentation) ซึ่งจัดได้ว่ายังมีข้อบกพร่องอยู่พอสมควรที่ควรจะปรับปรุงการทำงานให้มีประสิทธิภาพ และตรงกับความต้องการของผู้ใช้มากยิ่งขึ้น ทางผู้พัฒนาระบบจึงมีข้อเสนอแนะดังต่อไปนี้

1. การระบุแหล่งที่มาของข้อมูล (Specify the Data Source) ควรจะรองรับการนำเข้าแบบไฟล์ข้อมูลหลายๆรูปแบบได้ก็จะเป็นการดียิ่งขึ้น เพื่อการประมวลผลที่หลากหลายรูปแบบ

2. การคัดเลือกข้อมูลอาจจะเพิ่มในส่วนของการยอมให้ผู้ใช้สามารถเขียนภาษา SQL เพื่อคัดเลือกข้อมูลได้ด้วยตนเองจะช่วยส่งผลให้การทำงานมีประสิทธิภาพมากยิ่งขึ้น
3. การติดต่อกับตารางเพื่อที่จะใช้ทำการเลือกเขตข้อมูลควรจะปรับปรุงให้สามารถทำการเลือกเป็นรายการได้มากกว่าการพิมพ์ เพราะอาจจะทำให้เกิดข้อผิดพลาด และการเสียเวลาอีกด้วย
4. ข้อมูลของลูกค้าที่รวบรวมมาควรจะพิจารณาให้ตรงกับวัตถุประสงค์ที่ได้กำหนดไว้ตอนแรกมากที่สุด เพื่อโปรแกรมของระบบจะได้สามารถวิเคราะห์ และประมวลผลได้อย่างแม่นยำ ถูกต้องมากยิ่งขึ้น
5. โปรแกรมอาจจะสามารถทำงานแบบย้อนกลับทีละขั้นตอนได้ เพราะในความเป็นจริงของกระบวนการทำดาต้าไมนิ่งนั้นเป็นไปได้ที่จะย้อนกลับไปทำขั้นตอนใดขั้นตอนหนึ่งใหม่ ดังนั้นจึงควรที่จะต้องเก็บสถานะของการทำงานในแต่ละขั้นตอนด้วย

บรรณานุกรม

Cabena Peter. et al. 1997. **Discovering data mining: from concept to implementation.** New Jersey: Prentice-Hall.

Deitel H.M and Deitel P.J. 2005. **JAVA HOW TO PROGRAM.** 6th. New Jersey: Pearson Education.

Gordan Mihaela. 2002. **A NEW FUZZY C-MEANS BASED SEGMENTATION STRATEGY. APPLICATIONS TO LIP REGION IDENTIFICATION.** Technical University of Cluj-Napoca.

Han Jiawei and Kamber Micheline. 2001. **Data Mining Concept and Techniques.** n.p.

Jiang Hong. n.p. **Generalized Fuzzy Clustering Model with Fuzzy C-Means.** Computer Science and Engineering: University of South Carolina.

Keller Annette and Klawornn Flank. n.p. **Fuzzy Clustering with weighting of data variables.** Ostfriesland University.

Lampinen. et al. n.p. **PROFILING NETWORK APPLICATIONS WITH FUZZY C-MEANS CLUSTERING AND SELF-ORGANIZING MAP.** Tampere: University of Technology.

Leski Jacek M. n.p. **An ε-Insensitive Fuzzy C-Mean clustering.** Silessian: University of Technology.

Masulli F. et al. n.p. **Fuzzy Clustering Methods for the Segmentation of Multivariate Medical Images.** University of Genova.

Michalopoulos M. et al. n.p. **Decision Making Using Fuzzy C-means and Inductive Machine Learning for Managing Bank Branches Performance.** Technical University of Crete.

Shing Jyh. et al. 1997. **Neuro-Fuzzy and Soft Computing.** United States of America: Printice-Hall

Xu Bugao and Lin Sheng. 2002. **AUTOMATIC COLOR IDENTIFICATION IN PRINTED FABRIC IMAGES BY A FUZZY-NEURAL NETWORK.** Texas: University of Texas at Austin

ประวัติผู้เขียน

ชื่อผู้เขียน นาย ต่อพงษ์ โลหะรังสิกุล

วันเดือนปีเกิด 8 ธันวาคม 2522

สถานที่เกิด กรุงเทพมหานคร

ประวัติการศึกษา

- สำเร็จการศึกษาระดับประถมศึกษาจาก โรงเรียนลาซาลโชติรวินครสวรรค์
- สำเร็จการศึกษาระดับมัธยมศึกษาจาก โรงเรียนสตรีวิทยา 2
- สำเร็จการศึกษาระดับปริญญาตรี วิทยาศาตรบัณฑิต สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยธรรมศาสตร์

ประวัติการทำงาน

- Benchmark Vision (ปี พ.ศ. 2544 - 2545)

ตำแหน่ง : Programmer Analyst

- Binary Neighborhood (ปี พ.ศ. 2545)

ตำแหน่ง : Web Programmer

- Contracted with Quality System (ปี พ.ศ. 2545 - 2547)

ตำแหน่ง : System Integration ทำงานให้กับ HP Thailand

- Contracted with Nemera International (ปี พ.ศ. 2548)

ตำแหน่ง : Web Programmer ทำงานให้กับ Orange