ห้องสมุดคณะเทคโนโลยีสารสนเทศ สจล.

การพัฒนาระบบเพื่อการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม CLARANS

System Development for Data Segmentation with CLARANS Algorithm

โดย

นางสาวดลินธา แก้วโกมลมาลย์

รหัส 45066017

H002084

อาจารย์ที่ปรึกษา

ผศ.ดร.วรพจน์ กรีสุระเดช

วัน เดือน ปี 20 มิ.ย. 2550
เลขทะเบียน 02084
เลขเรียกหนังสือ วพ. ด1441 2546
"ห้องสมุดคณะเทคโนโลยีสารสนเทศ สจล."

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน

หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

ภาคเรียนที่ 2 ปีการศึกษา 2546

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ชื่อหัวข้อ	การพัฒนาระบบเพื่อการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม CLARANS
นักศึกษา	นางสาวดลินธา แก้วโกมลมาลย์
อาจารย์ที่ปรึกษา	ผศ.ดร. วรพจน์ กรีสุระเดช
ระดับการศึกษา	วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2546

บทคัดย่อ

ในปัจจุบันการดำเนินงานทางธุรกิจเป็นยุคที่มีการแข่งขันกันเป็นอย่างสูง เราจึงจำเป็นจะต้องทำการพัฒนาเครื่องมือและเทคนิคต่างๆขึ้นมา เพื่อช่วยในการสนับสนุนการตัดสินใจ ดาต้าไมนิ่งเป็นเทคนิคอย่างหนึ่งที่ถูกนำมาใช้งาน เนื่องจากมีเทคนิคและวิธีการที่ช่วยนำมาวิเคราะห์ข้อมูลในฐานข้อมูลได้เป็นอย่างดี

ในโครงการศึกษานี้ จะเป็นการจัดการงานทางด้านการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม CLARANS ซึ่งการจัดกลุ่มให้กับข้อมูลนั้นเป็นวิธีการวิเคราะห์ที่สำคัญของดาต้าไมนิ่ง เพื่อที่จะนำไปใช้ในการวิเคราะห์ทางการตลาดหรือการทำลูกค้าสัมพันธ์ได้ ซึ่งโดยหลักการของอัลกอริทึม CLARANS นั้นเป็นอัลกอริทึมที่สามารถจัดการกับกลุ่มข้อมูลขนาดใหญ่ได้เป็นอย่างดี ที่จะทำให้เราได้ผลลัพธ์ที่น่าพอใจจากการจัดกลุ่มข้อมูล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Title	System Development for Data Segmentation with CLARANS Algorithm
Student	Miss Dalintha Kaewkomonman
Advisor	Asst.Prof.Dr.Worapoj Kreesuradej
Level of Study	Master of Science in Information Technology
Major	Information Science
Academic Year	2003

ABSTRACT

As highly business competitive era, many methods have been used to analyze information in order to make the business decision. Data mining is a method used to manage database by using CLARANS algorithm.

In this project, managed about Database segmentation by CLARANS algorithm. This method can be used with large database by separating data into cluster which will be easy to use in term of analyze the whole market or make customer relationship management.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

กิตติกรรมประกาศ

ข้าพเจ้าขอขอบพระคุณ ดร.วรพจน์ กรีสุระเดช อาจารย์ที่ปรึกษาวิชาโครงการพัฒนาระบบงาน ที่ได้กรุณาให้ความรู้ คำปรึกษาและคำแนะนำต่างๆ ที่เป็นประโยชน์ต่อการพัฒนาระบบ และสละเวลาในการตรวจสอบแก้ไขข้อบกพร่อง

นอกจากนี้ข้าพเจ้าขอขอบพระคุณบิดา มารดา และบุคคลภายในครอบครัว ที่ได้ให้การส่งเสริม สนับสนุน และเป็นกำลังใจในการศึกษาเล่าเรียนตลอดมา อีกทั้ง ขอบคุณเพื่อนๆทุกคนที่มีส่วนให้ความช่วยเหลือในการพัฒนาระบบงานครั้งนี้ และเพื่อนๆที่คอยสนับสนุน เป็นกำลังใจที่ทำให้งานชิ้นนี้บรรลุผลสำเร็จได้เป็นอย่างดี

ข้าพเจ้าหวังเป็นอย่างยิ่งว่าโครงการพัฒนาระบบงานนี้ จะเป็นประโยชน์แก่ผู้ที่สนใจในงานทางด้านดาต้าไมนิ่ง ไม่มากก็น้อย อีกทั้งในโครงการพัฒนาระบบงานชิ้นนี้ หากมีข้อผิดพลาดประการใด ข้าพเจ้าขอรับไว้เพื่อนำไปปรับปรุงแก้ไขในคราวต่อไป

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ

หน้า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ (ต่อ)

หน้า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ (ต่อ)

หน้า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญตาราง

หน้า

ตารางที่

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญภาพ

หน้า

รูปที่

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญภาพ (ต่อ)

หน้า

รูปที่

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 1
บทนำ

1.1 หลักการและเหตุผล

ในธุรกิจยุคปัจจุบันถือได้ว่าเป็น ยุคที่ข้อมูลข่าวสารมีความสำคัญเป็นอย่างมาก ดังนั้นผู้ที่มีข้อมูลอยู่เป็นจำนวนมากก็จะมีโอกาสเป็นผู้ชนะมากกว่า เนื่องจากมีทางเลือกและโอกาสในการตัดสินใจที่มากกว่าผู้ที่มีข้อมูลอยู่ในปริมาณที่ไม่มากนัก และจากการเพิ่มขึ้นอย่างรวดเร็วของจำนวนฐานข้อมูล การขยายขนาดของแต่ละฐานข้อมูล ความรู้ในการวิเคราะห์ฐานข้อมูลขนาดใหญ่เพื่อหาความสัมพันธ์ของข้อมูลที่มีอยู่ รวมไปถึงสารสนเทศในระดับสูง ล้วนแล้วแต่เป็นสิ่งที่ทำให้ข้อมูลได้เกิดการแปรเปลี่ยนไปเป็นข้อมูลที่มีความสำคัญเป็นอย่างมาก ในการสนับสนุนการตัดสินใจและทำนายเพื่อหาลักษณะที่จะเกิดขึ้นภายในอนาคต ในการทำธุรกิจนั้นจำเป็นที่จะต้องศึกษาถึงความต้องการของลูกค้า เพื่อที่จะนำมาปรับปรุงกลยุทธ์ต่างๆขององค์กร เพื่อให้การทำธุรกิจนั้นประสบผลสำเร็จ และให้เกิดความได้เปรียบทางการค้ากับคู่แข่ง

การศึกษาในเรื่องดาต้าไมนิ่ง (Data mining) ถือได้ว่าเป็นเทคนิคอย่างหนึ่งที่สามารถวิเคราะห์ข้อมูลต่างๆได้เป็นอย่างดี เพราะจำนวนข้อมูลในปัจจุบันที่เก็บอยู่ภายในฐานข้อมูลมักมีขนาดใหญ่จนเกินความสามารถที่จะทำการวิเคราะห์ด้วยคน จึงเริ่มใช้การวิเคราะห์ที่เป็นอัตโนมัติ นอกจากนั้น ดาต้าไมนิ่งยังเป็นเครื่องมือที่ดี ในการพยากรณ์แนวโน้มและพฤติกรรมของข้อมูล รวมไปถึงเก็บความรู้และข้อมูลต่างๆไว้ช่วยในการตัดสินใจ เนื่องจากดาต้าไมนิ่งสามารถที่จะตอบคำถามทางธุรกิจได้เป็นอย่างดี

กระบวนการในการทำดาต้าไมนิ่งมีอยู่หลายเทคนิคด้วยกัน ซึ่งในเรื่องของการจัดกลุ่มข้อมูลเพื่อการวิเคราะห์นั้น ในโครงการศึกษาฉบับนี้เลือกที่จะใช้เทคนิคที่เรียกว่า Database Segmentation หรือที่เรียกอีกอย่างหนึ่งว่า Clustering ซึ่งเป็นกระบวนการที่ทำการจัดแบ่งกลุ่มข้อมูลเพื่อดูว่าข้อมูลทั้งหมดถูกจัดแบ่งออกเป็นกี่กลุ่ม เป็นการกำหนดโดยดูจากธรรมชาติของตัวข้อมูลเอง Database Segmentation นั้นบ่อยครั้งที่ถือได้ว่าเป็นเทคนิคที่ดีที่สุดที่จะนำมาใช้เป็นเทคนิคแรกๆเมื่อมีข้อมูลปริมาณมากๆ ข้อมูลที่มีความซับซ้อนสูงกับข้อมูลที่มีตัวแปรจำนวนมาก

Database Segmentation เป็นกระบวนการที่ซึ่งกลุ่มของ ออบเจ็คจะถูกแบ่งออกเป็นหลายๆ Cluster ที่ซึ่งแต่ละสมาชิกจะมีความเหมือนและแตกต่างกันจากสมาชิกใน Cluster อื่นๆ จากคุณสมบัติที่เด่นชัดของการวิเคราะห์โดยใช้วิธี Database Segmentation นั้นคือการทำงานได้เป็น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

อย่างดีกับฐานข้อมูลที่มีขนาดใหญ่มากๆ ที่ประกอบไปด้วยออบเจ็คจำนวนมาก ที่ในแต่ละออบเจ็คนั้นยังมี ฟิลด์ ที่มีชนิดแตกต่างกันออกไปอีก จากข้อมูลดังกล่าวนั้นแสดงถึงว่าอัลกอริทึมของ Database Segmentation นั้นสามารถที่จะทำการปรับเปลี่ยนได้และยังจัดการกับข้อมูลที่มี ฟิลด์ ที่มีหลายชนิดได้ ในโครงการศึกษาฉบับนี้ได้ทำการศึกษาการพัฒนาระบบเพื่อการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม CLARANS ที่มีหลักการทำงานที่ว่า จะหาค่า medoid ที่ดีที่สุดขึ้นมานั่นเอง

1.2 วัตถุประสงค์ของการพัฒนาระบบงาน

โครงการศึกษานี้มีวัตถุประสงค์ในการพัฒนาระบบดังนี้

1. ศึกษาดาต้าไมนิ่งโดยใช้เทคนิคการวิเคราะห์ข้อมูลแบบ Database Segmentation หรือเรียกอีกอย่างหนึ่งว่า Clustering ในการพัฒนาระบบงานเพื่อการจัดกลุ่มข้อมูล และวิเคราะห์ถึงอัลกอริทึมที่มีความเหมาะสมกับกลุ่มข้อมูลที่นำมาทำการวิเคราะห์
2. เกิดความเข้าใจอย่างลึกซึ้งหลังจากที่ได้ศึกษาหลักการดาต้าไมนิ่ง และสามารถที่จะนำไปประยุกต์ใช้กับระบบงานต่างๆในทางธุรกิจได้เป็นอย่างดี
3. ระบบที่พัฒนาขึ้นมานั้นสามารถที่จะนำไปใช้เป็นข้อมูลอย่างหนึ่งในการสนับสนุนการตัดสินใจทางธุรกิจได้เป็นอย่างดี

1.3 ขอบเขตของการพัฒนาระบบงาน

โครงการศึกษานี้เป็นการนำเอาเทคนิค Database Segmentation ของดาต้าไมนิ่งมาประยุกต์ใช้งาน ซึ่งในโครงการศึกษานี้เลือกที่จะใช้อัลกอริทึมในการจัดแบ่งกลุ่มข้อมูลบนพื้นฐานของการค้นหาแบบสุ่มที่เรียกว่า อัลกอริทึม CLARANS มาทำการวิเคราะห์เพื่อที่จะทำการศึกษาการจัดกลุ่มของข้อมูล

1.4 ขั้นตอนการดำเนินงานการพัฒนาระบบงาน

ในโครงการศึกษานี้มีขั้นตอนในการดำเนินงาน ดังนี้

1. กำหนดวัตถุประสงค์ในการจัดทำโครงการและเก็บรวบรวมข้อมูลที่เกี่ยวข้อง
2. ศึกษาค้นคว้าข้อมูลเกี่ยวกับดาต้าไมนิ่งที่ใช้เทคนิค Database Segmentation โดยเลือกอัลกอริทึม CLARANS เพื่อใช้ในการจัดกลุ่มข้อมูล
3. กำหนดแหล่งข้อมูลที่จะนำมาใช้ในการศึกษาในโครงการศึกษานี้ โดยในโครงการศึกษานี้เลือกใช้ข้อมูลเกี่ยวกับข้อมูลของลูกค้าของธุรกิจหนึ่ง
4. พัฒนาระบบเพื่อจัดการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม CLARANS

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

5. ประเมินผลและวิเคราะห์ผลที่ได้จากการพัฒนาระบบ

1.5 ประโยชน์ที่คาดว่าจะได้รับจากการพัฒนาระบบงาน

หลังจากที่ได้ทำการศึกษาและพัฒนาระบบงาน คาดว่าจะได้รับประโยชน์ดังต่อไปนี้

1. ทำให้ได้รับความรู้และมีความเข้าใจถึงขั้นตอนต่างๆของการทำดาต้าไมนิ่งได้เป็นอย่างลึกซึ้ง
2. จากหลักการทำงานของดาต้าไมนิ่งโดยใช้เทคนิค Database Segmentation ด้วยอัลกอริทึม CLARANS ทำให้สามารถจัดการงานทางด้านการแบ่งกลุ่มข้อมูลได้เป็นอย่างดี และนำข้อมูลที่ได้จากการจัดกลุ่มนั้นไปใช้ในการวิเคราะห์งานทางด้านต่างๆ ไม่ว่าจะเป็นงานทางด้านธุรกิจ หรือ งานทางด้านลูกค้าสัมพันธ์ รวมไปถึงงานทางด้านต่างๆอีกมากมาย
3. จากการที่ธุรกิจในยุคปัจจุบัน การทำงานที่นำคอมพิวเตอร์เข้ามาใช้งานมีบทบาทอย่างกว้างขวาง อีกทั้งข้อมูลที่จัดเก็บอยู่ในแต่ละงานมีมากขึ้นเรื่อยๆ การที่จะหาความสัมพันธ์หรือความเกี่ยวข้องของข้อมูลนั้นบางครั้งยากที่จะจัดการได้ แต่เมื่อเรานำเทคนิคของดาต้าไมนิ่งเข้ามาจัดการแล้ว จะทำให้เราทราบถึงความสัมพันธ์ที่มีอยู่ภายในข้อมูลนั้นได้อย่างชัดเจนมากยิ่งขึ้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 2
ดาต้าไมนิ่งและทฤษฎีที่เกี่ยวข้อง

2.1 ดาต้าไมนิ่ง (Data Mining)

ดาต้าไมนิ่งเป็นวิธีการที่ใช้ในการวิเคราะห์ข้อมูลจำนวนมาก เพื่อหาแนวโน้มหรือความสัมพันธ์ของข้อมูลที่มีอยู่ ข้อมูลที่ได้มาจากทำดาต้าไมนิ่งไม่ได้เกิดจากการคาดคะเน หรือจากการสมมติฐาน แต่เป็นข้อมูลที่มีความสัมพันธ์ที่ซ่อนอยู่ภายใต้ข้อมูลที่เรามีอยู่ ดังนั้นในการทำดาต้าไมนิ่งจึงไม่ได้เป็นการตั้งสมมติฐานแต่เป็นการดูผลลัพธ์ที่ได้จากการทำงานมากกว่า จะเห็นได้ว่าการทำ ดาต้าไมนิ่งนั้นเป็นวิธีการที่แตกต่างไปจากวิธีการวิเคราะห์ข้อมูลทางสถิติในแบบอื่นๆ ในการทำดาต้าไมนิ่งนั้น ผลลัพธ์ที่เกิดขึ้นถือได้ว่าเป็นข้อมูลที่มีประโยชน์เป็นอย่างมาก โดยสามารถที่จะนำข้อมูลเหล่านี้ไปใช้เป็นแนวทางในการตัดสินใจที่ก่อให้เกิดผลดีในการทำธุรกิจ

โดยทั่วไปแล้วในการทำ ดาต้าไมนิ่งนั้นมีด้วยกันอยู่สองบรรทัดฐานด้วยกัน คือ การค้นหาความรู้ (Knowledge Discovery , KD) และการสร้างแบบจำลองการคาดการณ์ (Predictive Modeling , PM) ในทางปฏิบัติแล้วจะทำการประยุกต์ใช้ AI หรือ เทคโนโลยีในการเรียนรู้ของเครื่องจักร ในการวิเคราะห์ข้อมูลในฐานข้อมูลขนาดใหญ่ จุดประสงค์ของทั้งสองบรรทัดฐานนี้ก็คือ พยายามที่จะสร้างกระบวนการที่เป็นแบบอัตโนมัติให้มากที่สุดเท่าที่จะเป็นไปได้ ซึ่งในทางปฏิบัติแล้ว การทำดาต้าไมนิ่งไม่ใช่ระบบอัตโนมัติอย่างสมบูรณ์ทั้งหมด แต่เป็นกระบวนการแบบกึ่งอัตโนมัติเท่านั้น ตัวอย่างดาต้าไมนิ่งเช่น ผู้ใช้ระบบโทรศัพท์มือถือนั้น สามารถที่จะแบ่งออกได้เป็นทั้งสิ้น 3 กลุ่มด้วยกันคือ กลุ่มที่มีแนวโน้มว่ามีความต้องการที่จะเปลี่ยนผู้ให้บริการหรือรูปแบบในการใช้บริการสูง ปานกลาง และ ต่ำ จากการวิเคราะห์ข้อมูลพบว่า ผู้ใช้งานโทรศัพท์มือถือที่ได้รับการโทรเข้ามากกว่าสองครั้งต่อวันมีแนวโน้มต่ำ ที่จะเปลี่ยนแปลงผู้ให้บริการหรือรูปแบบการใช้บริการ

ปัจจัยที่ทำให้ดาต้าไมนิ่งเป็นที่ใช้งานกันอย่างกว้างขวางนั้นเป็นผลอันเนื่องมาจาก

1. ขนาดของข้อมูลมีขนาดใหญ่ได้ถูกผลิตและขยายตัวอย่างรวดเร็ว การสืบค้นข้อมูลจะมีประโยชน์ก็ต่อเมื่อฐานข้อมูลที่ใช้มีขนาดใหญ่มาก ปัจจุบันมีจำนวนและขนาดข้อมูลขนาดใหญ่ที่ขยายตัวอย่างรวดเร็ว โดยผ่านทางอินเตอร์เน็ต ดาวเทียม และแหล่งผลิตข้อมูลในรูปแบบอื่นๆ

2. ข้อมูลถูกจัดเก็บเพื่อนำไปสร้างระบบการสนับสนุนการตัดสินใจ เพื่อเป็นการง่ายต่อการนำข้อมูลมาใช้ในการวิเคราะห์เพื่อการตัดสินใจ ส่วนมากข้อมูลจะถูกจัดเก็บโดยอยู่ในรูปของ Data Warehouse ซึ่งเป็นการง่ายต่อการนำเอาไปใช้ในการสืบค้นความรู้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3. เทคนิคดาต้าไมนิ่งประกอบไปด้วยอัลกอริทึมที่มีความซับซ้อนจึงจำเป็นต้องใช้งานกับระบบคอมพิวเตอร์ที่มีประสิทธิภาพสูงด้วย ซึ่งในปัจจุบันระบบคอมพิวเตอร์ที่มีประสิทธิภาพสูงในท้องตลาดก็มีราคาถูกลงมาก จึงเป็นสาเหตุให้หันมานิยมใช้ ดาต้าไมนิ่งกันมากขึ้น

4. เนื่องจากในปัจจุบันมีการแข่งขันสูงทางด้านอุตสาหกรรมและการค้า จึงมีข้อมูลที่เกิดขึ้นจำนวนมากแต่ไม่ได้นำมาใช้ให้เกิดประโยชน์ จึงเป็นการจำเป็นอย่างยิ่งที่จะทำการสืบค้นข้อมูลที่ซ่อนอยู่ภายในฐานข้อมูลมาทำการวิเคราะห์เพื่อนำไปใช้เป็นแนวทางในการตัดสินใจในการจัดการกับระบบต่างๆ ซึ่งจะเห็นได้ว่าข้อมูลเหล่านี้ถือเป็นผลผลิตอีกชิ้นหนึ่งเลยทีเดียว

2.2 ขั้นตอนในการสืบค้นความรู้จากฐานข้อมูล

ขั้นตอนในการสืบค้นความรู้จากฐานข้อมูลนั้นเป็นขั้นตอนที่มีความสำคัญเป็นอย่างมากในการหารูปแบบที่มีความน่าสนใจที่ซ่อนอยู่ภายในออกมาแสดงจากฐานข้อมูลทั้งหมดที่มีอยู่

2.2.1 การกำหนดวัตถุประสงค์ทางธุรกิจ (Business Objective Determination)

เป็นการกำหนดวัตถุประสงค์ทางธุรกิจขึ้นมา โดยผู้ที่ทำการกำหนดวัตถุประสงค์นั้นจะต้องเข้าใจถึงปัญหาของงานและความต้องการของตัวธุรกิจ เพราะในการเข้าใจถึงปัญหาและกำหนดวัตถุประสงค์ให้ชัดเจนนั้นจะเป็นสิ่งที่กำหนดทิศทางในการทำดาต้าไมนิ่งต่อไป โดยที่ถ้าวัตถุประสงค์ไม่ชัดเจนแล้ว อาจทำให้ผลที่ได้เกิดความคลุมเครือไม่สามารถที่จะนำไปใช้ได้ ต้องกลับไปเริ่มต้นใหม่อีกครั้งหนึ่ง ซึ่งทำให้เสียเวลาไปอย่างเปล่าประโยชน์

2.2.2 การเตรียมข้อมูล (Data Preparation)

ในขั้นตอนของการเตรียมข้อมูลถือได้ว่าเป็นขั้นตอนที่มีความสำคัญมาก และใช้เวลาในการทำงานมากกว่าการทำงานในขั้นตอนอื่นๆของการสืบค้นความรู้จากฐานข้อมูล ซึ่งในขั้นตอนนี้ยังแบ่งออกเป็นขั้นตอนย่อยได้อีก 3 ขั้นตอนดังนี้

1. **การเลือกข้อมูล (Data Selection)**

เป็นการวิเคราะห์เลือกสิ่งที่น่าสนใจและมีความสำคัญออกมาฐานข้อมูล เพื่อทำงานในขั้นตอนต่อไป ซึ่งการเลือกข้อมูลนั้นก็ขึ้นอยู่กับวัตถุประสงค์ขององค์กรที่ได้กำหนดไว้ตั้งแต่แรกแล้ว และในการเลือกข้อมูลจำเป็นที่จะเข้าใจความหมายของข้อมูล โดยข้อมูลที่ถูกเลือกขึ้นมานั้นจะเปลี่ยนแปลงไปตามวัตถุประสงค์ทางธุรกิจที่ถูกกำหนดไว้ในตอนแรกแล้ว โดยตัวแปรที่ถูกเลือกขึ้นมานั้นจะมีการกำหนด ชนิด ค่า รูปแบบ และลักษณะที่ชัดเจนเอาไว้

2. **การเตรียมข้อมูลก่อนการประมวลผล (Data Preprocessing)**

ในขั้นตอนนี้เป็นการนำข้อมูลที่ได้จากขั้นตอนที่หนึ่ง มาทำการตรวจสอบในขั้นตอนต่างๆว่า ข้อมูลที่ถูกเลือกขึ้นมานั้น เป็นข้อมูลที่มีคุณภาพ มีความถูกต้องและมีความเป็น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ปัจจุบันมากที่สุด เหมาะสมที่จะใช้ในการทำงานในขั้นตอนต่อๆไป โดยข้อมูลที่ผ่านการเลือกมานั้นจะต้องผ่านการตรวจสอบดังนี้

- การตรวจสอบข้อมูล (Data Cleansing)

ข้อมูลที่ได้มาในตอนแรกนั้นเป็นข้อมูลที่ยังไม่มีความสมบูรณ์ ที่จะสามารถนำไปผ่านกระบวนการดาต้าไมนิ่งได้ จึงต้องมีการจัดเตรียมข้อมูลก่อน ซึ่งข้อมูลที่ได้รับมานั้นอาจจะมีบางค่าของข้อมูลที่ขาดหายไป (Missing Data) เราจึงต้องนำมาผ่านกระบวนการบางอย่างเพื่อกำจัดปัญหาที่ค่าของข้อมูลหายไป เช่น ตัดข้อมูลเหล่านั้นทิ้งไปในกรณีที่ข้อมูลเหล่านั้นมีจำนวนไม่มากนัก หรือบางข้อมูลเป็นข้อมูลที่เราเรียกกันว่า Noisy Data คือเป็นข้อมูลที่มีความคลาดเคลื่อนไป วิธีการแก้ไขข้อมูลที่ผิดพลาดเหล่านี้มีวิธีการแก้ไขหลายวิธีด้วยกันไม่ว่าจะเป็น วิธีการ Binning การจัดกลุ่ม หรือการทำ Regression

- การรวบรวมข้อมูล (Data Integration)

เป็นการนำข้อมูลจากหลายๆแหล่งมารวมเข้าไว้ด้วยกัน

- การลดจำนวนข้อมูล (Data Reduction)

เป็นการลดขนาดข้อมูลให้น้อยลง มีวิธีการทำหลายวิธีด้วยกัน เช่น Dimensionality reduction ที่เป็นการลดขนาดตามแนวคอลัมน์ หรือ Data size reduction ที่เป็นการลดขนาดตามแนวแถวนั่นเอง

3. **การแปลงข้อมูล (Data Transformation)**

เป็นการแปลงข้อมูลหรือรวบรวมข้อมูลเข้าไว้ด้วยกัน ซึ่งเป็นการแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมพร้อมที่จะนำไปใช้ในการวิเคราะห์ได้ ซึ่งจะส่งผลกับข้อมูลทำให้ข้อมูลที่ผ่านการแปลงเป็นข้อมูลที่มีคุณภาพมากและยังสอดคล้องกับเทคนิคที่จะนำไปใช้งานอีกด้วย

2.2.3 การทำดาต้าไมนิ่ง (Data Mining)

เป็นกระบวนการสำคัญในการเลือกเทคนิคดาต้าไมนิ่งที่เหมาะสม เพื่อหารูปแบบที่ซ่อนอยู่ออกมา ซึ่งเป็นการประมวลผลข้อมูลตามอัลกอริทึมที่ได้กำหนดไว้

2.2.4 การวิเคราะห์ผลลัพธ์ที่ได้ (Analysis of Knowledge)

เป็นการเลือกรูปแบบที่เหมาะสมจากงานที่ได้จากขั้นตอนดาต้าไมนิ่งเพื่อนำไปสู่ความรู้ใหม่ๆ หรือเป็นการนำผลที่ได้จากการทำดาต้าไมนิ่งนั้นมาตีความเพื่อหารูปแบบที่ซ่อนอยู่ภายในโดยถ้างานที่ทำออกมานั้นยังไม่ตรงตามวัตถุประสงค์ที่วางไว้ ก็จะย้อนกลับไปแก้ไขใหม่อีกครั้ง

2.2.5 การนำความรู้มาใช้งาน (Assimilation of Knowledge)

เป็นเทคนิคที่ใช้ในการแสดงความรู้ที่ได้จากการทำดาต้าไมนิ่งไปสู่ผู้ใช้งาน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

โดยจากขั้นตอนต่างๆข้างต้น จะเห็นว่า ดาต้าไมนิ่งเป็นเพียงแค่ขั้นตอนหนึ่งในกระบวนการทั้งหมดที่มีอยู่เท่านั้น ส่วนขั้นตอนที่มีความสำคัญในการสืบค้นความรู้จากฐานข้อมูลนั้นเป็นขั้นตอนในการเตรียมข้อมูลสำหรับทำดาต้าไมนิ่ง และเป็นขั้นตอนที่ใช้เวลาในการทำงานมากที่สุดอีกด้วย เนื่องมาจากอาจจะต้องมีการรวบรวมข้อมูลมาจากหลายๆแหล่งด้วยกันเพื่อที่จะดูความสัมพันธ์ของข้อมูล ซึ่งข้อมูลที่ได้จากขั้นตอนการเตรียมจะต้องมีความชัดเจน และมีความถูกต้องด้วย

2.3 เทคนิคในการวิเคราะห์ข้อมูลของดาต้าไมนิ่ง

ในโครงการศึกษานี้ เราจะเน้นถึงเทคนิคในการทำ Database Segmentation เพื่อใช้ในการจัดกลุ่มข้อมูลเพื่อทำการวิเคราะห์ ซึ่งเทคนิคที่ใช้ในการวิเคราะห์ข้อมูลของดาต้าไมนิ่งแบ่งออกเป็น 4 ประเภทหลัก ดังนี้

2.3.1 Predictive Modeling

เป็นการสร้างแบบจำลองเพื่อทำนายหาค่าความเป็นไปได้ โดยใช้หลักการสังเกตจากข้อมูลที่มีอยู่ ซึ่งแบบจำลองในการพัฒนาจะแบ่งออกเป็นสองช่วงด้วยกัน คือ

1. Training เป็นช่วงในการสร้างแบบจำลองขึ้นมาใหม่โดยใช้ข้อมูลในอดีต และใช้ข้อมูลในปริมาณมาก
2. Testing เป็นการตรวจสอบประสิทธิภาพของแบบจำลองใหม่ที่สร้างขึ้นมา ซึ่งใช้ข้อมูลในปริมาณที่ไม่มากนัก

ในPredictive Modeling ยังแบ่งออกเป็นอีก 2 เทคนิคที่ใช้คือ

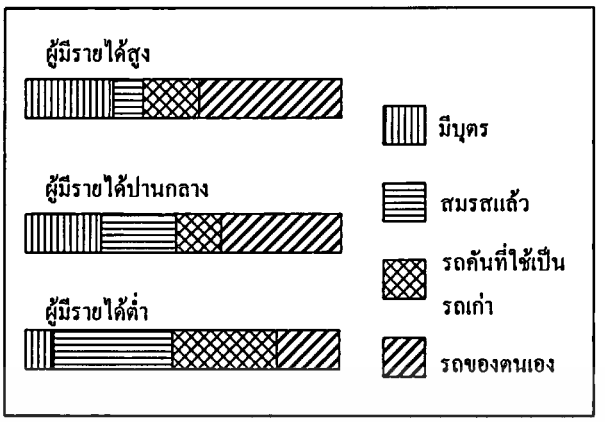
1. Classification เป็นการทำนายว่าสิ่งที่พิจารณาควรที่จะอยู่ภายในกลุ่มใด ซึ่งสามารถแบ่งกลุ่มได้อย่างชัดเจน
2. Value prediction เป็นการทำนายค่าที่เป็นตัวเลขค่าความต่อเนื่องของข้อมูล

2.3.2 Database Segmentation (Clustering)

บางครั้งวิธีนี้จะถูกเรียกว่า Clustering จุดประสงค์ของวิธีนี้ก็คือ การแบ่งกลุ่มของฐานข้อมูลให้กลายเป็น Cluster ย่อยๆโดยที่แต่ละ Cluster ย่อยนั้นจะประกอบไปด้วยออบเจ็คที่เหมือนกัน เช่น การแบ่งตามอายุ รายได้ ซึ่งในการแบ่งกลุ่มข้อมูลนั้นไม่สามารถที่จะกำหนดได้ว่าข้อมูลควรที่จะอยู่ในกลุ่มใด แต่จะเป็นการกำหนดกลุ่มจากธรรมชาติของข้อมูลเองมากกว่า โดยที่ไม่มีการใช้อคติหรือประสบการณ์มาช่วยในการตัดสินใจ

ใน Database Segmentation ยังแบ่งเทคนิคได้อีกหลายวิธี โดยมีวิธีที่นิยมใช้กันมากเช่น Demographic Clustering, Neural Clustering

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 2.1 ตัวอย่างการทำ Database Segmentation

ตัวอย่างในการทำ Database Segmentation บริษัทรถยนต์ได้แยกกลุ่มลูกค้าออกเป็น 3 กลุ่มด้วยกัน คือ กลุ่มผู้ที่มีรายได้สูง รายได้ปานกลาง และรายได้ต่ำ และภายในแต่ละกลุ่มยังแบ่งออกเป็น มีบุตร สมรสแล้ว รถคันที่ใช้เป็นรถเก่า รถของตนเอง

จากข้อมูลข้างต้นทำการวิเคราะห์ได้ว่า เมื่อมีลูกค้าเข้ามาที่บริษัทควรจะทำการเสนอขายรถประเภทใด เช่น ในกลุ่มผู้ที่มีรายได้สูงควรจะเสนอขายรถใหม่ ซึ่งเป็นรถครอบครัวขนาดใหญ่พอสมควร แต่ถ้าเป็นผู้ที่มีรายได้ค่อนข้างต่ำควรเสนอขายรถยนต์มือสองที่มีขนาดค่อนข้างเล็ก

2.3.3 Link Analysis

เป็นการศึกษาถึงความสัมพันธ์ของข้อมูลเพื่อที่จะดูว่า กลุ่มของข้อมูลมีความสัมพันธ์กันในลักษณะใด ซึ่งความสัมพันธ์นี้ถูกเรียกว่า Associations Link Analysis เป็นแบบจำลองที่ได้รับความนิยมเป็นอย่างมากในการวิเคราะห์เพื่อหาความสัมพันธ์ระหว่าง ลูกค้ากับสินค้าซึ่ง Link Analysis ยังแบ่งออกเป็นอีก 3 เทคนิคคือ Associations discovery, Sequential pattern discovery และ Similar time sequence discovery

2.3.4 Deviation Detection

เป็นวิธีการในการวิเคราะห์หาสิ่งที่มีความแตกต่างในข้อมูล การตรวจจับสิ่งผิดปกติต่างๆ เป็นกรรมวิธีในการหาค่าที่แตกต่างไปจากค่ามาตรฐานหรือค่าที่คาดคิดไว้ว่าต่างไปมากน้อยเพียงใด ซึ่งเป็นการสรุปข้อมูลออกมาในรูปแบบการแสดงผลทางกราฟิก ซึ่ง Deviation Detection เป็นแบบจำลองที่ใช้เทคนิคทางสถิติหรือการแสดงให้เห็นภาพ เทคนิคนี้ถูกใช้ในงานทางด้านการตรวจสอบการปลอมลายเซ็น หรือบัตรเครดิตปลอม รวมทั้งการตรวจหาจุดบกพร่องของชิ้นงานในโรงงานอุตสาหกรรม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 3

Database Segmentation

3.1 ประเภทของการทำ Database Segmentation (Clustering)

อัลกอริทึมที่ใช้ในการทำ Database Segmentation มีอยู่มากมายหลายประเภทด้วยกัน การเลือกอัลกอริทึมที่จะนำมาใช้งานนั้น ขึ้นอยู่กับชนิดของข้อมูลที่มีอยู่และนอกจากนั้นยังขึ้นกับจุดประสงค์ของงานนั้นอีกด้วย โดยในงานชิ้นนี้จะทำการแบ่งประเภทของ Database Segmentation ได้ดังนี้

3.1.1 วิธีการแบบ Partitioning

หลักการทำงานของวิธีนี้คือ มีฐานข้อมูลอยู่จำนวน n ออบเจ็คด้วยกัน ในวิธีการแบบ Partitioning จะทำการแบ่งข้อมูลออกเป็น k ส่วน ที่ซึ่งในแต่ละส่วนจะถูกเรียกว่า Cluster และ k จะมีค่าน้อยกว่าหรือเท่ากับ n จากตรงนี้เราจะเห็นได้ว่ามีการจัดแบ่งกลุ่มข้อมูลออกเป็น k กลุ่มด้วยกัน และมีข้อกำหนดดังต่อไปนี้

1. ในแต่ละกลุ่มจะต้องประกอบด้วยออบเจ็คอย่างน้อยหนึ่งออบเจ็ค
2. ในแต่ละออบเจ็คจะต้องมีกลุ่มเป็นของตัวมันเองอย่างน้อยหนึ่งกลุ่ม

จะเห็นได้ว่าในบางเทคนิคของการทำ Partitioning ที่มีความยุ่งยากมากๆ ก็ไม่จำเป็นที่จะต้องตรงตามข้อกำหนดข้างต้นก็ได้ แต่เป็นเฉพาะบางกรณีเท่านั้นที่มีความยุ่งยากมากจริงๆ

วิธีการแบบ Partitioning จะทำการสร้าง Partition เริ่มต้นขึ้นมา จากนั้นจะใช้เทคนิคที่เรียกกันว่า "Iterative Relocation" ซึ่งเป็นความพยายามที่จะทำการปรับปรุงการ Partitioning โดยการย้ายออบเจ็คจากกลุ่มหนึ่งไปยังอีกกลุ่มหนึ่ง ซึ่งมีหลักเกณฑ์การทำ Partitioning ที่ดีอยู่ว่าออบเจ็คที่อยู่ภายในกลุ่มเดียวกันจะมีลักษณะที่คล้ายคลึงกัน ส่วนออบเจ็คที่อยู่ต่างกลุ่มกัน ก็จะมีลักษณะที่แตกต่างกันออกไป

ซึ่งในวิธีการแบบ Partitioning มีอัลกอริทึมที่เหมาะกับงานประเภทนี้เป็นจำนวนมาก เช่น อัลกอริทึม K-means โดยมีหลักการทำงานคือ ในแต่ละ Cluster จะแสดงค่า mean ของ ออบเจ็คในแต่ละ Cluster ขึ้นมา อัลกอริทึม K-medoids หลักการทำงานคือ ในแต่ละ Cluster จะแสดงโดยออบเจ็คหนึ่งที่อยู่ใกล้กับศูนย์กลางของ Cluster ซึ่งอัลกอริทึมทั้งสองที่กล่าวมานั้นเหมาะกับการจัดการฐานข้อมูลที่มีขนาดเล็กจนถึงขนาดกลาง โดยจะทำงานได้ดีใน Cluster ที่มีรูปร่างแบบทรงกลม แต่ไม่เหมาะที่จะจัดการกับฐานข้อมูลที่มีขนาดใหญ่มากๆหรือมีรูปทรงที่มีความซับซ้อน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

แต่อย่างไรก็ตามการทำ Partitioning ได้มีการพัฒนาขึ้นมามากจนมีอัลกอริทึมที่สามารถจะจัดการกับฐานข้อมูลขนาดใหญ่ได้แล้ว เช่น อัลกอริทึม PAM อัลกอริทึม CLARA และอัลกอริทึม CLARANS ในโครงการศึกษานี้จะกล่าวถึงหัวข้อนี้อย่างละเอียดอีกครั้งหนึ่ง

3.1.2 วิธีการแบบ Hierarchical

ในวิธีนี้จะทำการสร้างข้อมูลให้แยกออกเป็นลำดับชั้นจากข้อมูลทั้งหมดที่มีอยู่ ซึ่งในที่นี้จะทำการแบ่งออกเป็น 2 ประเภทด้วยกันคือ Agglomerative และ Divisive โดยที่ Agglomerative หรือจะเรียกว่าวิธีการแบบ Bottom-up หลักการคือ การรวมแต่ละออบเจ็คที่กระจายกลุ่มกันเข้าไว้ด้วยกัน จนท้ายที่สุดจะได้เป็นกลุ่มใหญ่ที่สุดเพียงกลุ่มเดียว ส่วนวิธี Divisive หรือที่เรียกว่าวิธีการแบบ Top-down ในตอนเริ่มแรกออบเจ็คนั้นจะอยู่ภายใน Cluster เดียวกัน เมื่อเกิดกระบวนการทำซ้ำอย่างต่อเนื่อง Cluster จะแตกตัวออกเป็น Cluster ที่เล็กลง จนในที่สุดแต่ละออบเจ็คจะอยู่ภายใน Cluster เพียง Cluster เดียว

อัลกอริทึมที่น่าสนใจในวิธีการแบบ Hierarchical ได้แก่ BIRCH, CURE, ROCK และ Chameleon

3.1.3 วิธีการแบบ Density-based

วิธีการอื่นๆมักจะมีหลักการอยู่ที่การวัดระยะทางระหว่างออบเจ็ค และสามารถที่จะทำงานได้เพียง Cluster ที่มีลักษณะเป็นทรงกลม ซึ่งวิธีในการทำ Database Segmentation โดยทั่วไปมักมีการพัฒนามาจากความเข้าใจทางด้านความหนาแน่น (Density) ดังนั้นแนวคิดของวิธีนี้จึงอยู่ที่ว่าจะปล่อยให้ Cluster ขยายไปเรื่อยๆจนกว่าค่าความหนาแน่น (จำนวนของออบเจ็ค) ของ Cluster ที่อยู่ใกล้กันมีค่ามากกว่าค่าที่กำหนดเอาไว้ในตอนแรก จะเห็นได้ว่าในวิธีนี้ สามารถที่จะกำจัดสิ่งรบกวนต่างๆออกไปได้ แล้วทำงานกับ Cluster ที่มีรูปทรงแบบใดก็ได้

อัลกอริทึมที่น่าสนใจในวิธีการแบบ Density-based ได้แก่ DBSCAN, OPTICS, DENCLUE

3.1.4 วิธีการแบบ Grid-based

ในวิธีนี้จะทำการ Quantize ที่ว่างของออบเจ็คลงในเซลล์จำนวนหนึ่ง ที่ซึ่งอยู่ในรูปของโครงสร้างแบบตาราง (Grid) ซึ่งในทุกๆขั้นตอนในการทำ Database Segmentation จะถูกจัดการลงบนโครงสร้างแบบตาราง ข้อดีของวิธีนี้คือ ใช้เวลาในการทำงานไม่มากนัก โดยไม่ขึ้นกับจำนวนของข้อมูล จะขึ้นกับเพียงจำนวนเซลล์ในแต่ละส่วนในที่ว่างของ Quantize

อัลกอริทึมที่น่าสนใจในวิธีการแบบ Grid-based ได้แก่ STING, CLIQUE และ Wave Cluster

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.1.5 วิธีการแบบ Model-based

สมมติฐานของวิธีนี้คือ สร้างโมเดลสำหรับแต่ละ Cluster ขึ้นมาและทำการหาข้อมูลที่เหมาะสมที่สุดให้กับโมเดล อัลกอริทึมในวิธีนี้อาจจะค้นหา Cluster โดยการสร้างฟังก์ชันความหนาแน่นที่ส่งผลกับระยะการกระจายของตำแหน่งข้อมูล ซึ่งมันจะนำไปสู่การทำงานที่มีความเป็นอัตโนมัติมากขึ้น

อัลกอริทึมที่น่าสนใจในวิธีการแบบ Model-based ได้แก่ COBWEB

3.2 อัลกอริทึมในการจัดแบ่งกลุ่มข้อมูลโดยวิธีการแบบ Partitioning

ในหัวข้อนี้จะนำเสนอถึงอัลกอริทึมในการจัดแบ่งกลุ่มของข้อมูล ซึ่งได้แก่ อัลกอริทึม PAM, อัลกอริทึมCLARA โดยจะมีการกล่าวถึงตัวแปรต่างๆเพื่อให้เกิดความเข้าใจในอัลกอริทึมมากยิ่งขึ้นจึงได้สรุปค่าตัวแปรต่างๆไว้ในตารางที่ 3.1 ดังนี้

ตารางที่ 3.1 ตารางสัญลักษณ์และคำจำกัดความ

Symbol	Definitions
D	Data set to be Clustered
n	Number of objects in D
O_i	Object i in D
k	Number of Clusters
S	A sample of D
s	Size of S

3.2.1 อัลกอริทึม PAM

หลักการทำงานจะทำการสมมติว่ามีออบเจ็คอยู่ n ออบเจ็ค อัลกอริทึม PAM ต้องการที่จะหาจำนวนของ Cluster จำนวน k Cluster โดยในการหาครั้งแรกนั้นจะแสดงถึงออบเจ็คที่อยู่ในแต่ละ Cluster จากผลที่ได้ตรงนี้ ส่วนที่อยู่ตรงกลางมากที่สุดใน Cluster นั้นเราจะเรียกว่า medoid จากการที่แบ่ง Cluster เป็น k Cluster แสดงว่าจะมี medoid อยู่ k medoid ด้วย จากนั้นจะทำซ้ำในอัลกอริทึมเพื่อหา medoid ที่เป็นทางเลือกที่ดีกว่า medoid เดิม โดยในทุกๆคู่ของออบเจ็คที่ทำการวิเคราะห์นั้น ออบเจ็คหนึ่งจะต้องเป็น medoid และอีกออบเจ็คเป็นออบเจ็คที่ไม่ได้เป็น medoid ซึ่งจากการทำซ้ำระยะทางโดยรวมระหว่างออบเจ็คที่ไม่ได้เป็น medoid และ medoid นั้นอาจจะมีค่า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ลดลง โดยการสลับค่าของ medoid กับออบเจ็คที่ไม่ได้เป็น medoid เราสามารถที่จะอธิบายการทำงานของอัลกอริทึม PAM ได้เป็นลำดับขั้นตอนดังนี้

ขั้นที่ 1 เลือก k medoid ขึ้นมาจาก n (เป็นการเลือก medoid ขึ้นมาแบบสุ่ม)

ขั้นที่ 2 คำนวณหาค่า Cost ของ $D_i' - D_i$ สำหรับแต่ละการสลับค่าของ medoid กับออบเจ็คอื่น ที่ซึ่ง D_i เป็นระยะทางรวมก่อนที่จะมีการสลับและ D_i' เป็นระยะทางรวมภายหลังจากที่มีการสลับแล้ว

ขั้นที่ 3 ถ้าค่า Cost เป็นค่าติดลบ แสดงว่าค่านั้นเป็นค่า cost ที่ดีที่สุดและให้ทำซ้ำในขั้นตอนที่ 2 แต่ถ้าค่า cost เป็นบวกก็จะจัดการเก็บค่า medoid นั้นไว้และสิ้นสุดการทำงาน

3.2.2 อัลกอริทึม CLARA

อัลกอริทึม CLARA (Clustering LARge Applications) หลักการทำงานของอัลกอริทึมนี้คือการสุ่มกลุ่มตัวอย่างขึ้นมาเพื่อจัดการกับกลุ่มข้อมูลขนาดใหญ่ทั้งหมดที่มีอยู่แทนที่จะเป็นการค้นหา medoid สำหรับกลุ่มข้อมูลทั้งหมด CLARA จะทำการสุ่มกลุ่มข้อมูล ซึ่งมีขนาดเล็กๆขึ้นมาจากกลุ่มข้อมูลทั้งหมดที่มีอยู่และทำการประยุกต์ให้เข้ากับอัลกอริทึม PAM เพื่อหาค่าที่ดีที่สุดของกลุ่ม medoid สำหรับกลุ่มตัวอย่าง โดยสิ่งที่ใช้วัดถึงประสิทธิภาพของ medoidที่ได้ คือ การวัดโดยอาศัยค่าเฉลี่ยของความไม่เหมือนกันระหว่างทุกๆออบเจ็คในข้อมูลทั้งหมด (D) และ medoid ของ Cluster มันเอง สามารถเขียนให้อยู่ในรูปของสมการได้ดังนี้

$$Cost(M,D) = \frac{\sum_{i=1}^{n} dissimilar\ ity\ (O_i, rep(M, O_i))}{n}$$

โดยที่ M เป็น กลุ่มของ medoid ที่ถูกเลือกขึ้นมา

$dissimilar\ ity\ (O_i, O_j)$ เป็น ความแตกต่างระหว่างออบเจ็ค O_i และ O_j

$rep(M_i, O_i)$ เป็น การคืนค่า medoid ใน M ที่ซึ่งติดกับ O_i

จากกลุ่มตัวอย่าง CLARA จะทำการสุ่มตัวอย่างซ้ำและทำการ Database Segmentation ข้อมูลหลายๆครั้ง จนกระทั่งเลือกกลุ่มที่ดีที่สุดของ medoid ขึ้นมาที่มีค่า $Cost(M,D)$ น้อยที่สุด โดยให้ q เป็นจำนวนครั้งในการสุ่ม อัลกอริทึม CLARA แสดงได้ดังรูปที่ 3.1

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

```
Set mincost to a large number;
Repeat q times
        Create S by drawing s objects randomly from D;
        Generate the set of medoids M from S by applying
        the PAM algorithm;
        If Cost(M, D) <mincost
        Then
                mincost = Cost(M, D);
                bestset = C;
        End-if;
End-repeat;
Return bestset;
```

รูปที่ 3.1 การทำงานของอัลกอริทึม CLARA

เนื่องจากอัลกอริทึมCLARAรับเอาวิธีการสุ่มเข้ามาใช้งานประสิทธิภาพของการ Database Segmentation จึงขึ้นกับขนาดของกลุ่มตัวอย่างเป็นสำคัญ

3.3 อัลกอริทึมในการจัดแบ่งกลุ่มข้อมูลบนพื้นฐานของการค้นหาแบบสุ่ม

3.3.1 อัลกอริทึม CLARANS

ถ้ากลุ่มตัวอย่างที่ทำการสุ่มขึ้นมามีขนาดใหญ่ไม่เพียงพอในการทำงานของอัลกอริทึม CLARAแล้ว ประสิทธิภาพของอัลกอริทึม CLARA ก็จะทำงานได้ไม่ดีนัก แต่อย่างไรก็ตาม ประสิทธิภาพก็จะไม่ดีถ้ากลุ่มตัวอย่างมีขนาดใหญ่มากจนเกินไป โดยเฉพาะถ้า medoid ที่ดีที่สุดนั้นไม่ได้อยู่ภายในกลุ่มตัวอย่างที่ทำการสุ่มขึ้นมา จากที่ต้องการรักษาประสิทธิภาพของงานเอาไว้และต้องการกระทำให้อยู่ในรูปของค่าเฉลี่ยระยะทางต่อออบเจ็ค อัลกอริทึม CLARANS จึงเป็นอีกทางเลือกหนึ่งขึ้นมา

CLARANS (Clustering Large Applications based on RANdomized Search) อยู่ในรูปของกระบวนการในการหา medoid ออกมาจำนวน k medoid โดยเป็นการหาออกมาจากกราฟ ในกราฟนี้แต่ละโหนดจะถูกแสดงโดยกลุ่มของออบเจ็คจำนวน k ออบเจ็ค โดยโหนดสองโหนดที่อยู่ใกล้กันถ้าถูกแยกความแตกต่างโดยออบเจ็คเพียงออบเจ็ค คือ medoid ที่แตกต่างกัน มันเป็นสิ่งที่บ่งชี้ได้ว่า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ในแต่ละโหนดที่มีอยู่ในกราฟมีโหนดที่อยู่ใกล้เคียง k(n-k) เพราะว่าโหนดจะแสดงถึงการรวมเข้าด้วยกันของ k medoid ในแต่ละโหนดที่มีลักษณะเช่นเดียวกัน ในการทำ Database Segmentation ที่เป็นไปได้และมีประสิทธิภาพสามารถวัดได้โดยใช้ค่า Cost function

แทนที่จะมีการใช้กลยุทธ์ในการค้นหาอย่างละเอียด CLARANS จะทำการค้นหาแบบที่เรียกว่า serial randomized search นั่นก็คือ CLARANS จะเริ่มจากการสุ่มเลือกโหนดใดๆที่อยู่ภายในกราฟและทำการสุ่มเลือกโหนดที่อยู่ใกล้เคียงกันขึ้นมาอีกหนึ่งโหนด โดยถ้าค่า Cost ของโหนดใกล้เคียงที่ถูกเลือกขึ้นมามีค่าน้อยกว่าโหนดที่เลือกอยู่ในตอนแรก CLARANS จะเลือกใช้โหนดที่อยู่ใกล้เคียงนี้แทนและจะทำการเลือกโหนดใกล้เคียงโหนดต่อๆไปขึ้นมา เพื่อจะทำการเปรียบเทียบไปเรื่อยๆ ดังนั้น CLARANS จะทำการเลือกสุ่มโหนดใกล้เคียงไปเรื่อยๆ จนกระทั่งได้โหนดที่พอใจแล้ว หรืออาจมีการกำหนดถึงจำนวนที่มากที่สุดของโหนดที่ต้องการที่จะตรวจสอบเอาไว้ก่อน ต่อมาโหนดที่ได้ถูกเลือกไว้จะมีการระบุว่าเป็น local minimum เพื่อการหลีกเลี่ยงถึงข้อผิดพลาดต่างๆในการหาผลลัพธ์ที่ดีที่สุด CLARANS จะทำงานในขั้นตอนต่างๆในการค้นหา local minimum ซ้ำอีกครั้งจากโหนดตอนเริ่มต้นที่ต่างกันออกไปเพื่อการกำหนดถึงจำนวนครั้ง ภายหลังจากนั้นโหนดกับค่า Cost ที่ต่ำที่สุดที่ถูกเลือกขึ้นมานั้นถือได้ว่าเป็น Clustering สุดท้าย โดยให้ maxneighbor เป็นจำนวนที่มากที่สุดของโหนดที่อยู่ใกล้เคียงเพื่อทำการตรวจสอบ และให้ numlocal เป็นจำนวนของ local minima ที่ได้ รายละเอียดของอัลกอริทึม CLARANS จะแสดงดังรูปที่ 3.2

```
Set mincost to a large number;
For i = 1 to numlocal do
        Randomly select a node as the current node C in the graph;
        Initialize j to 1;
        Repeat
                Randomly select a neighbor N of C;
                If Cost(N, D) < Cost(C, D) Then
                        Assign N as the current node C;
                        Reset j to 1;
                Else increment j by 1;
                End-if;
        Until j > maxneighbor;
```

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

```
        If Cost(C, D) < mincost Then
                mincost = Cost(C, D);
                bestnode = C;
        End-if;
    End-for;
    Return bestnode;
```

รูปที่ 3.2 การทำงานของอัลกอริทึม CLARANS

ประสิทธิภาพของ CLARANS จะขึ้นกับค่าของ maxneighbor และ numlocal เป็นสำคัญ

3.3.2 ผลลัพธ์จากการทดลอง : การปรับใช้อัลกอริทึม CLARANS

1. รายละเอียดของการทดลอง

จากการสังเกตพฤติกรรมและประสิทธิภาพของอัลกอริทึม CLARANS เราใช้อัลกอริทึม CLARANS กับกลุ่มข้อมูลที่สร้างขึ้นมา โดยหลักการกว้างๆแล้ว เราจะใช้ Cluster 2 ชนิดที่มีคุณสมบัติที่มีความแตกต่างกันทีเดียว

1. ชนิดแรกของCluster จะมีลักษณะเป็นสี่เหลี่ยมมุมฉากและ ออบเจ็ค ที่อยู่ภายในแต่ละ Cluster จะถูกกำหนดขึ้นมาอย่างสุ่ม เพื่อให้เกิดความชัดเจนยิ่งขึ้นเราจะทำการอธิบายดังนี้ ถ้ากำหนดให้มีกลุ่มข้อมูลอยู่ 3,000 ออบเจ็ค ใน 20 Cluster แล้ว ก่อนอื่นเลยเราต้องทำการสร้างกล่องขึ้นมา 20 กล่องที่มีขนาดเดียวกัน จัดการให้แต่ละ Cluster แยกออกจากกันอย่างชัดเจน ที่ทำแบบนี้เพราะเราจะประยุกต์ใช้กับ Spatial Data Mining ออบเจ็คที่อยู่ในการทดลองนี้จะมีคู่ลำดับ คือ คู่ลำดับ x , y สำหรับในแต่ละกล่อง เราจะทำการ สร้างคู่ลำดับขึ้นมาแบบสุ่มจำนวน 150 คู่ที่อยู่ภายในกล่อง ในทางเดียวกัน เราจะสร้างกลุ่มข้อมูลในชนิดเดียวกันอีกแต่จะมีการเปลี่ยนแปลงจำนวนของออบเจ็ค และ Cluster จากรูปภาพทางด้านล่าง (รูปที่ 3.3) เครื่องหมาย rn-k (เช่น r3000-20) จะแสดงแทนกลุ่มข้อมูลของกลุ่มที่มี n จุดและ k Cluster

2. ชนิดที่สองของ Cluster เราจะทำการทดลองโดยไม่ใช้วิธีบรรจุจุดแบบสุ่ม จุดที่อยู่ภายในแต่ละ Cluster ค่อนข้างจะถูกเรียงลำดับในรูปแบบสามเหลี่ยม ตัวอย่างเช่น คู่ลำดับดังต่อไปนี้ (0,0),(1,0),(0,1),(2,0),(1,1),และ(0,2) จากรูปแบบนี้จะเห็นว่า Cluster รูปสามเหลี่ยมมีขนาดเท่ากับ 6 การสร้าง Cluster ถัดมา เราจะใช้การแปลงจากจุดเดิม(เช่นจุด (10,10) , (11,10) , (10,11) , (12,10) , (11,11) , และ(10,12)) จากรูปที่ 3.3 เครื่องหมาย tn-k (เช่น t3000-20) จะแสดงแทนกลุ่มข้อมูลที่ถูก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

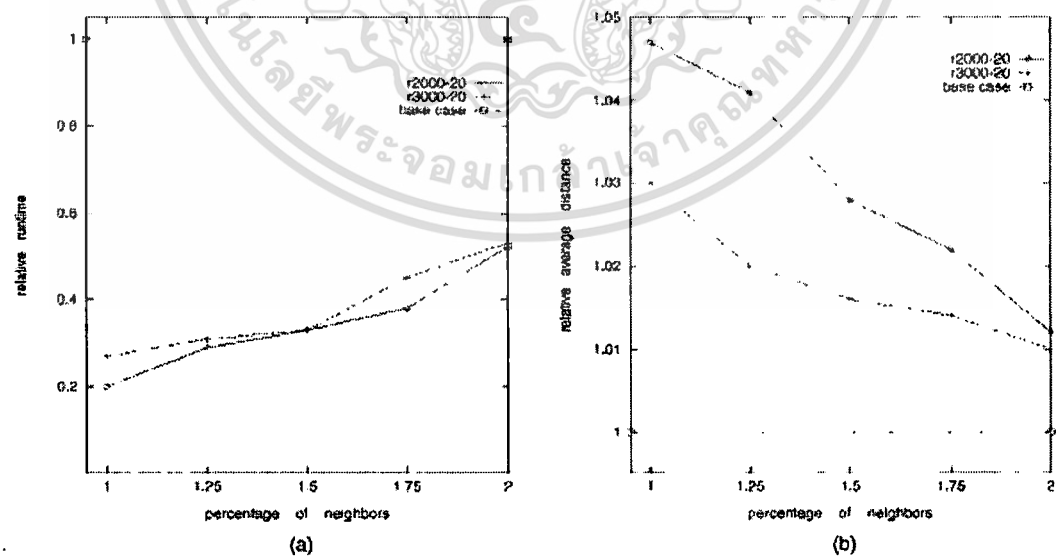
สร้างขึ้นที่มี n จุดและ k Cluster

ทุกรายงานการทดลองในโครงการศึกษาชิ้นนี้ถูกทำในการทำงานแบบที่เรียกว่า time-sharing SPARC-LX workstation เพราะว่าธรรมชาติของ CLARANS คือการถูกสุ่มเลือกขึ้นมา รูปภาพทั้งหมดที่เกี่ยวข้องกับ CLARANS คือรูปค่าเฉลี่ยที่ได้จากการทดสอบการทดลองเป็นจำนวน 10 ครั้ง (ด้วยจำนวนที่แตกต่างกันของจำนวนในการสุ่ม)

2. การกำหนดจำนวนที่มากที่สุดของโหนดที่อยู่ติดกัน

ในรุ่นแรกของการทดลอง เราจะประยุกต์อัลกอริทึม CLARANS ให้เข้ากับพารามิเตอร์ maxneighbor ดังนี้ 250, 500, 750, 1,000และ 10,000 บนกลุ่มข้อมูล rn-k และ tn-k ที่ n อยู่ในช่วง 100-3,000 และ k อยู่ในช่วง 5-20 เราจะสรุปเพียง 2 วิธีการหลักๆที่ใช้ในการหาเพื่อให้เข้ากับการทดลองข้างต้น

- เมื่อจำนวนที่มากที่สุดของโหนดที่อยู่ติดกัน (maxneighbor) ถูกกำหนดให้เท่ากับ 10,000 คุณภาพของการClustering ถูกกำหนดโดยอัลกอริทึม CLARANS ในขณะที่เราจะอธิบายเหตุการณ์นี้อย่างคร่าวๆ ซึ่งเราจะใช้ผลลัพธ์สำหรับค่า maxneighbor เท่ากับ 10,000 เพื่อความชัดเจนยิ่งขึ้นค่า runtime ของกราฟแรกและค่าเฉลี่ยระยะทางของกราฟที่ 2 ในรูปที่ 3.3 ที่อยู่ภายใต้การทำให้เป็นปกติโดยสิ่งที่สร้างด้วยการกำหนดค่า maxneighbor เท่ากับ 10,000 นี้เป็นการอธิบายได้ว่า เส้นแนวนอน 2 เส้นที่ค่า y เท่ากับ 1 ในกราฟทั้งคู่



รูปที่ 3.3 การกำหนดจำนวนที่มากที่สุดของโหนดที่อยู่ใกล้เคียงกัน โดยที่ (a) Relative efficiency (b) Relative quality

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ห้องสมุดคณะเทคโนโลยีสารสนเทศ สจล.

- อย่างที่คาดการณ์ไว้ว่า ค่าที่น้อยกว่าค่า maxneighbor จะเป็นการสร้างการจัดแบ่งกลุ่มที่มีคุณภาพไม่ดีนัก คำถามที่มักถามกันเสมอคือค่าของ maxneighbor มีขนาดเล็กมากแค่ไหนก่อนที่คุณภาพของการจัดแบ่งกลุ่มข้อมูลจะเป็นค่าที่ยากจะยอมรับได้ จากการทดลองในรุ่นแรกนั้น เราทำการหาค่าที่เป็นค่าที่มีความสำคัญนั้นได้โดยเหมือนจะได้สัดส่วนกับค่า k(n-k) นี้เป็นสิ่งที่กระตุ้นให้เกิดการกระทำกับการทดลองรุ่นอื่นๆเพื่อให้ได้มาซึ่งสูตรสำหรับการกำหนดค่า maxneighbor โดยที่ค่า minmaxneighbor เป็นค่าที่น้อยที่สุดที่ผู้ใช้กำหนดขึ้นมาสำหรับค่า maxneighbor

โดยถ้า k(n-k) ≤ minmaxneighbor แล้ว maxneighbor = k(n-k) อีกกรณีหนึ่ง maxneighbor จะเท่ากับค่าที่มากกว่าระหว่าง p% ของ k(n-k) และ minmaxneighbor

จากสูตรด้านบนยอมให้อัลกอริทึม CLARANS ทำการตรวจสอบทุกๆโหนดที่อยู่ติดกันตราบเท่าที่จำนวนรวมของโหนดที่อยู่ติดกัน มีค่าต่ำกว่าค่าที่กำหนดเอาไว้ (threshold) ของ minmaxneighbor นอกจากค่าที่ได้ทำการกำหนดเอาไว้แล้ว อัตราร้อยละของโหนดที่อยู่ติดกันจะตรวจสอบอย่างช้าๆโดยค่อยๆลดจาก100% ไปจนถึงค่าน้อยที่สุด p% จากกราฟในรูปที่ 3.3 จะแสดงถึงค่า relative runtime และคุณภาพของอัลกอริทึม CLARANS กับ minmaxneighbor ที่เท่ากับ 250 และ p ที่อยู่ในช่วง 1-2% ถึงแม้ว่ากราฟจะแสดงเพียงผลลัพธ์ของกลุ่มข้อมูลสี่เหลี่ยมมุมฉากที่ 2,000และ 3,000 จุดใน 20 Cluster กราฟเหล่านี้จะบรรยายถึง ลักษณะของกราฟสำหรับกลุ่มข้อมูลขนาดเล็กและขนาดกลางและสำหรับกลุ่มข้อมูลที่มีลักษณะเป็นสามเหลี่ยมที่มีความคล้ายคลึงกัน ในรูปที่ 3.3(a) แสดงถึง p ที่มีค่าน้อย ซึ่งมีค่าน้อยกว่าค่าโดยรวมของค่า runtime ที่ CLARANS ต้องการ และอย่างที่คาดว่าในรูปที่ 3.3(b) จะแสดงถึงค่า p ที่มีค่าน้อย ที่สร้างการจัดแบ่งกลุ่มข้อมูลที่มีคุณภาพต่ำออกมา (มีค่าเฉลี่ยระยะทางสูง) แต่สิ่งที่น่าแปลกใจในรูปที่ 3.3(b) คือคุณภาพที่ยังคงอยู่ภายใน 5% จากสิ่งที่สร้างโดยการกำหนดค่า maxneighbor = 10,000 ตัวอย่างเช่น ถ้าค่าสูงสุดของ p = 1.5% ของจำนวนโหนดที่อยู่ติดกันที่ถูกตรวจสอบ คุณภาพจะอยู่ใน 3% ขณะที่ค่า runtime แค่ 40% หมายความว่าอย่างไรที่มีการตรวจสอบ 98.5% มากกว่าโหนดที่อยู่ติดกัน และมีเพียงการสร้างที่บริเวณขอบที่ให้ผลลัพธ์ที่ดีมากยิ่งขึ้น เป็นสิ่งที่ตรงกันกับรายการตอนแรกที่อัลกอริทึม CLARANS กำหนดค่า maxneighbor = 10,000 ที่ให้คุณภาพเหมือนกับอัลกอริทึม PAM โดยประสิทธิภาพจะเหมือนกับการกำหนดค่า maxneighbor = k(n-k) = 20(3000-20) =59,600

เราจะเลือกเก็บค่าที่เหมาะสมระหว่างค่า Runtime และ คุณภาพ เราเชื่อว่าค่า p ที่อยู่ระหว่าง 1.25% และ 1.5% เป็นค่าที่เหมาะสมที่สุด สำหรับทุกๆการทดลองภายหลังที่ใช้อัลกอริทึม CLARANS เราจะเลือกใช้ค่า p ที่ 1.25%

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3. การกำหนดจำนวนของ Local Minima

การเรียกใช้อัลกอริทึม CLARANS ต้องพิจารณาพารามิเตอร์ 2 ตัวคือ maxneighbor และ numlocal ซึ่งในตอนนี้เราจะพิจารณาแค่ค่า numlocal ในผลการทดลองของรุ่นนี้ เราจะประมวลผลอัลกอริทึม CLARANS ด้วย numlocal ที่มีค่า 1,...,5 บนกลุ่มข้อมูล m-k และ tn-k สำหรับค่าของ n และ k ในขนาดเล็ก กลางและใหญ่ ในแต่ละการประมวลผล เราจะทำการบันทึกค่า runtime และคุณภาพของการแบ่งกลุ่มของข้อมูลเอาไว้ โดยในตารางที่ 3.2 (ที่ซึ่งเป็นตัวอย่างของกลุ่มข้อมูลทั้งหมด) จะแสดงถึง relative runtime และคุณภาพสำหรับกลุ่มข้อมูล r2000-20 ในที่นี้ค่าทั้งหมดได้กลายเป็นค่ามาตรฐานโดยค่า numlocal = 5

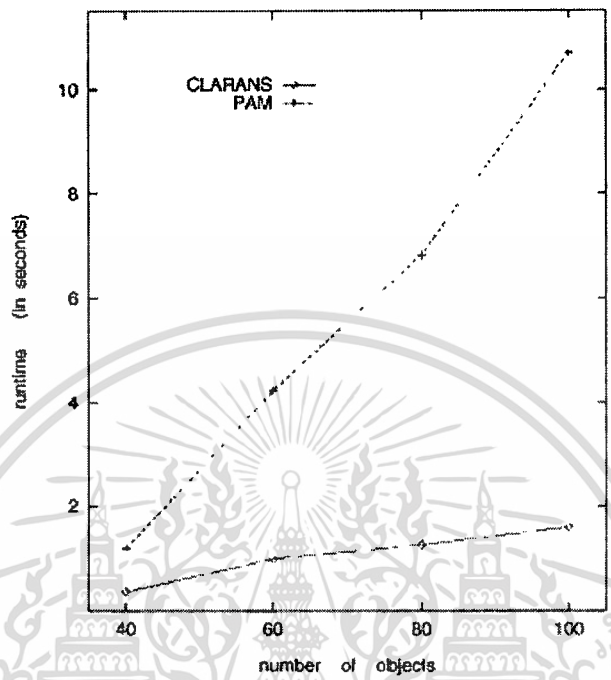
ตารางที่ 3.2 แสดงความสัมพันธ์ระหว่าง Runtime และคุณภาพของกลุ่มข้อมูล r2000-20

numlocal	1	2	3	4	5
relative runtime	0.19	0.38	0.6	0.78	1
Relative average distance	1.029	1.009	1	1	1

อย่างที่คาดว่า ค่า runtime เป็นสัดส่วนกับจำนวนของ Local minima ที่ได้รับ สำหรับค่า relative quality นี้เป็นการปรับปรุงจาก numlocal = 1 ไปยัง numlocal = 2 การดำเนินการในการค้นหาครั้งที่ 2 สำหรับการหาค่า local minima ดูเหมือนเป็นการลดผลกระทบของการสุ่มที่ไม่ดีที่ซึ่งอาจจะเกิดในการค้นหาเพียงครั้งเดียว แต่อย่างไรก็ตาม การกำหนดให้ค่า numlocal มากกว่า 2 นั้นถือได้ว่าไม่มีประสิทธิภาพเลย อย่างที่มีการเพิ่มขึ้นคุณภาพมากขึ้นเล็กน้อย นี่เป็นสิ่งที่บ่งชี้ว่าตัวอย่างของค่า local minimum เป็นแบบอย่างที่มีคุณภาพสูงมากๆ สำหรับทุกๆการทดลองภายหลังกับอัลกอริทึม CLARANS เราจะเลือกใช้ค่า local minima จากการหาในครั้งที่สอง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4. การเปรียบเทียบอัลกอริทึม CLARANS กับอัลกอริทึม PAM

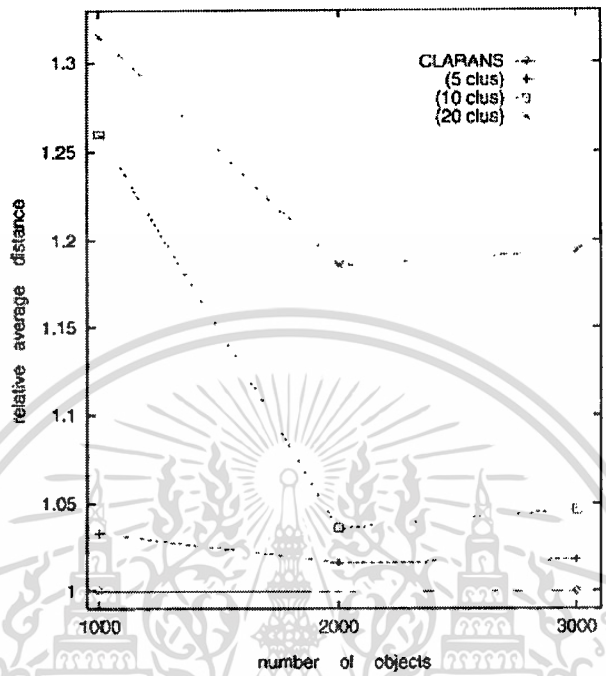


รูปที่ 3.4 การเปรียบเทียบประสิทธิภาพระหว่าง CLARANS กับ PAM

จากการทดลองที่ผ่านมา เราจะทำการเปรียบเทียบ CLARANS กับ PAM จากหัวข้อที่ 2 อัลกอริทึม CLARANS สำหรับกลุ่มข้อมูลที่มีขนาดกลางและขนาดใหญ่ ขณะที่มีการจัดแบ่งกลุ่มข้อมูลโดยการเปรียบเทียบจากคุณภาพ ซึ่งจะเห็นได้ว่ามีประสิทธิภาพมากกว่าอัลกอริทึม PAM ดังนั้นเป้าหมายในตอนนี้คือจะทำการเปรียบเทียบอัลกอริทึมสองอัลกอริทึมบนกลุ่มข้อมูลที่มีขนาดเล็ก เราจะทำการประยุกต์ทั้งสองอัลกอริทึมนี้บนกลุ่มข้อมูลที่มีขนาด 40,60,80และ 100 จุดใน 5 Cluster ในรูปที่ 3.4 จะแสดงถึง runtime ที่เกิดขึ้นโดยทั้งสองอัลกอริทึม กล่าวได้ว่า สำหรับกลุ่มข้อมูลทั้งหมด การจัดแบ่งกลุ่มข้อมูลจะถูกสร้างโดยอัลกอริทึมทั้งสองที่มีคุณภาพเหมือนกัน (เช่น มีค่าเฉลี่ยระยะทางที่เท่ากัน) ดังนั้นความแตกต่างระหว่างสองอัลกอริทึมจะถูกกำหนดจากประสิทธิภาพของมันเป็นสำคัญ ซึ่งจะเห็นได้จากในรูปที่ 3.4 ที่แม้แต่ในกลุ่มข้อมูลที่มีขนาดเล็ก CLARANS ก็ให้ผลลัพธ์ที่ดีกว่า PAM จากการคาดว่าประสิทธิภาพของสองอัลกอริทึมจะแตกต่างกันก็เนื่องจากการเพิ่มขึ้นของขนาดของกลุ่มข้อมูลนั่นเอง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

5. การเปรียบเทียบอัลกอริทึม CLARANS กับอัลกอริทึม CLARA



รูปที่ 3.5 แสดงความสัมพันธ์ของคุณภาพ : ในเวลาเดียวกันสำหรับ CLARANS และ CLARA

จากการทดลองที่ผ่านมา เราจะทำการเปรียบเทียบ CLARANS กับ CLARA ได้ อัลกอริทึม CLARA ไม่ได้ถูกออกแบบมาสำหรับกลุ่มข้อมูลที่มีขนาดเล็ก ดังนั้นเราจะทำการทดลองนี้ได้บนกลุ่มข้อมูลที่มีจำนวนของออบเจ็คมากกว่า 100 และออบเจ็คจะถูกจัดการในจำนวนของCluster ที่แตกต่างกัน เช่นเดียวกับ Cluster ทั้งสองชนิดที่อธิบายไว้ในหัวข้อที่ 1

เมื่อเรานำการทดลองชุดดังกล่าวมาทดสอบโดยอัลกอริทึม CLARA และ CLARANS แล้วนั้น CLARANS สามารถที่จะทำการจัดแบ่งกลุ่มข้อมูลได้มีประสิทธิภาพมากกว่า CLARA แต่อย่างไรก็ตามในบางกรณี CLARA อาจจะใช้เวลาน้อยกว่า CLARANS ดังนั้นมันจึงเป็นสิ่งที่น่าแปลก เมื่อ CLARA สามารถทำการจัดแบ่งกลุ่มข้อมูลที่มีคุณภาพเหมือนกันถ้ามีการให้จำนวนของเวลาที่เท่ากัน สิ่งนี้เองจะนำไปสู่การทดลองรุ่นถัดไปที่เราจะให้ทั้ง CLARA และ CLARANS มีจำนวนเวลาที่เท่ากัน ในรูปที่ 3.5 จะแสดงถึง คุณภาพของการจัดแบ่งกลุ่มข้อมูล ที่สร้างโดย CLARA ค่าปกติคือค่าที่ตรงกันที่สร้างโดย CLARANS

เมื่อให้จำนวนรวมของเวลาเท่าเดิม แน่นอนว่า CLARANS จะให้ผลลัพธ์ที่ดีกว่า CLARA ในทุกๆกรณี ความแตกต่างระหว่างอัลกอริทึมทั้งสองจะเพิ่มจาก 4% เมื่อ k (จำนวนของCluster) มีค่า 5 ถึง 20% เมื่อ k มีค่าเท่ากับ 20 นี้เป็นการขยายความแตกต่าง เมื่อ k เพิ่มขึ้นสามารถอธิบายได้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

โดยดูจากการวิเคราะห์ที่มีความซับซ้อนของ CLARA และ CLARANS ดังนั้นการเพิ่มขึ้นของ k จะกำหนดให้มีค่า cost ของ CLARA ที่มากกว่า CLARANS

การเปรียบเทียบที่มีความซับซ้อนในด้านบน อธิบายได้ว่าทำไมจำนวนของออบเจ็คจึงมีจำนวนมากสำหรับจำนวนของ Cluster ที่กำหนดไว้ ความแตกต่างระหว่างสองอัลกอริทึมจะน้อยลง เช่น เมื่อจำนวนของออบเจ็คมีค่าเป็น 1,000 ความแตกต่างจะสูงถึง 30% แต่ความแตกต่างจะลดน้อยลงประมาณ 20% เมื่อจำนวนของออบเจ็คเพิ่มขึ้นเป็น 2,000 ออบเจ็ค เพราะว่าในแต่ละการทำซ้ำของ CLARA มีค่า $O(k^3 + nk)$ ในกลุ่ม k^3 เป็นส่วนสำคัญกว่าส่วนที่สอง ดังนั้น CLARA จะมีผลกระทบน้อยมากจากการเพิ่มขึ้นของ n ในอีกกรณีหนึ่ง เนื่องจาก Cost ของ CLARANS ขึ้นอยู่กับค่า n ดังนั้นการเพิ่มขึ้นของ n เป็นการกำหนดให้ค่า Cost ของ CLARANS มากกว่า CLARA ในการอธิบายนี้ที่มีการกำหนดค่า k ที่แน่นอน จะทำให้ความแตกต่างลดน้อยลง ในขณะที่จำนวนของออบเจ็คจะเพิ่มขึ้น อย่างไรก็ตาม เส้นที่อยู่ข้างล่างสุดในรูปที่ 3.5 นั้นแสดงว่า CLARANS ดีกว่า CLARA ในทุกๆกรณี

จากหลักฐานทางการทดลองแสดงว่า CLARANS มีประสิทธิภาพที่ดีกว่า PAM และ CLARA สำหรับกลุ่มข้อมูลที่มีขนาดเล็กขนาดกลางและขนาดใหญ่

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 4
การประยุกต์การใช้ดาต้าไมนิ่งเพื่อการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม CLARANS

ในบทนี้จะกล่าวถึงรายละเอียดทั้งหมดของการประยุกต์ใช้ดาต้าไมนิ่งเพื่อการจัดกลุ่มข้อมูล โดยจะอธิบายถึงขั้นตอนต่างๆในการทำการพัฒนาระบบ ไม่ว่าจะเป็น การเตรียมข้อมูล การทำงานของระบบโดยใช้อัลกอริทึม CLARANS รวมไปถึงการนำผลลัพธ์ที่ได้จากการจัดแบ่งกลุ่มข้อมูลไปใช้งาน

4.1 กำหนดวัตถุประสงค์

ดาต้าไมนิ่งเป็นอีกวิธีหนึ่งที่นิยมนำมาใช้ในการจัดแบ่งกลุ่มของข้อมูล โดยในการพัฒนาระบบงานของโครงการศึกษานี้มีวัตถุประสงค์หลักเพื่อ ทำการจัดกลุ่มข้อมูลต่างๆ โดยใช้อัลกอริทึม CLARANS ออกมา แล้วนำผลที่ได้จากการจัดแบ่งกลุ่มนั้นมาวิเคราะห์ดูความแตกต่างของข้อมูลในแต่ละกลุ่ม

4.2 เครื่องมือที่ใช้ในการพัฒนาระบบ

ในการพัฒนาระบบงานครั้งนี้ได้เลือกใช้ Microsoft Visual Basic 6.0 ในการพัฒนาระบบงาน เนื่องมาจาก Microsoft Visual Basic 6.0 เป็นเครื่องมือที่สามารถพัฒนาระบบงานบนระบบปฏิบัติการ Windows ได้ อีกทั้งสามารถที่จะทำการติดต่อกับระบบฐานข้อมูลได้ ซึ่งตรงกับการพัฒนาระบบงานชิ้นนี้ที่จะต้องมีการติดต่อกับระบบฐานข้อมูลเพื่อนำข้อมูลออกมาทำการประมวลผล

ระบบฐานข้อมูลที่เลือกใช้ในโครงการศึกษานี้เพื่อทำการติดต่อกับข้อมูลนั้น เลือกที่จะใช้ Microsoft Access เป็นฐานข้อมูลในการติดต่อสื่อสารกับข้อมูลออกมา เพื่อทำการประมวลผลต่อไป

4.3 รายละเอียดของระบบงาน

การพัฒนาระบบของงานชิ้นนี้แบ่งออกเป็นสองส่วนด้วยกันคือ

1. ขั้นตอนการเตรียมข้อมูล(Preparation) ในขั้นตอนนี้จะประกอบไปด้วย การเลือกข้อมูล (Selection Data) ที่จะทำการจัดกลุ่มข้อมูลขึ้นมา การ Cleaning ข้อมูล(Cleaning Data) การแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสม(Transformation Data)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2. ขั้นตอนการจัดกลุ่มข้อมูล เป็นการนำข้อมูลที่ได้มาจากการเตรียมข้อมูลในขั้นตอนที่ 1 มาผ่านกระบวนการแบ่งกลุ่มข้อมูลโดยใช้อัลกอริทึม CLARANS ผลที่ได้จากการทำงานในขั้นตอนนี้ ก็คือ ข้อมูลจะถูกจัดแบ่งกลุ่มตามจำนวนที่ต้องการ โดยที่ในแต่ละกลุ่มข้อมูลก็จะมีข้อมูลที่มีความแตกต่างกันออกไปในแต่ละกลุ่ม ส่วนข้อมูลภายในแต่ละกลุ่มนั้นจะมีความคล้ายคลึงกัน

4.4 การเตรียมข้อมูล

ข้อมูลที่นำเข้ามาใช้ในการพัฒนาระบบในโครงการศึกษานี้ จะทำการยกตัวอย่างให้เป็นข้อมูลของลูกค้ากลุ่มหนึ่ง โดยมีรายละเอียดดังนี้

ตารางที่ 4.1 ตารางแสดงข้อมูลของลูกค้า

ชื่อข้อมูลเป็นภาษาอังกฤษ	ชื่อข้อมูลเป็นภาษาไทย	ชนิดของข้อมูล
sex	เพศ	Integer
age	อายุ	Integer
title_id	อาชีพของลูกค้า	Integer
type_id	การจ่ายเงินของลูกค้า	Integer
zone_id	รหัสภาคที่ลูกค้าอาศัยอยู่	Integer
province_id	รหัสจังหวัดที่ลูกค้าอยู่	Integer
inv_language	ภาษาที่ใช้ในการติดต่อสื่อสาร	Integer
credit_limit	กำหนดวงเงินของลูกค้า	Currency
credit_time	กำหนดระยะเวลาการจ่ายเงินของลูกค้า	Integer
pretransferbill	ลูกค้าที่จ่ายเงินก่อน/หลังกำหนดเวลา	Integer

ตารางที่ 4.2 ตารางแสดงรายละเอียดของ sex

sex	รายละเอียด
0	เพศชาย
1	เพศหญิง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.3 ตารางแสดงรายละเอียดของ title_id

title_id	รายละเอียด
1	ผู้บริหาร ข้าราชการระดับสูง
2	ผู้จัดการ ข้าราชการระดับกลาง
3	รองผู้จัดการ ผู้ช่วย หัวหน้า
4	พนักงานทั่วไป
5	ข้าราชการทั่วไป รับจ้าง
6	อื่นๆ

ตารางที่ 4.4 ตารางแสดงรายละเอียดของ type_id

type_id	รายละเอียด
1	ชำระเงินเป็นเงินสด
2	ชำระเงินผ่านบัตรเครดิต
3	อื่นๆ

ตารางที่ 4.5 ตารางแสดงรายละเอียดของ zone_id

zone_id	รายละเอียด
1	ภาคเหนือ
2	ภาคกลาง
3	ภาคตะวันออก
4	ภาคตะวันออกเฉียงเหนือ
5	ภาคตะวันตก
6	ภาคใต้

ตารางที่ 4.6 ตารางแสดงรายละเอียดของ province_id

province_id	รายละเอียด
1	กรุงเทพมหานคร
2	กระบี่

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

province_id	รายละเอียด
3	กาญจนบุรี
4	กาฬสินธุ์
5	กำแพงเพชร
6	ขอนแก่น
7	จันทบุรี
8	ฉะเชิงเทรา
9	ชลบุรี
10	ชัยนาท
11	ชัยภูมิ
12	ชุมพร
13	เชียงราย
14	เชียงใหม่
15	ตรัง
16	ตราด
17	ตาก
18	นครนายก
19	นครปฐม
20	นครพนม
21	นครราชสีมา
22	นครศรีธรรมราช
23	นครสวรรค์
24	นนทบุรี
25	นราธิวาส
26	น่าน
27	บุรีรัมย์
28	ปทุมธานี
29	ประจวบคีรีขันธ์
30	ปราจีนบุรี

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

province_id	รายละเอียด
31	ปัตตานี
32	พระนครศรีอยุธยา
33	พะเยา
34	พังงา
35	พัทลุง
36	พิจิตร
37	พิษณุโลก
38	เพชรบุรี
39	เพชรบูรณ์
40	แพร่
41	ภูเก็ต
42	มหาสารคาม
43	แม่ฮ่องสอน
44	มุกดาหาร
45	ยะลา
46	ยโสธร
47	ร้อยเอ็ด
48	ระนอง
49	ระยอง
50	ราชบุรี
51	ลพบุรี
52	ลำปาง
53	ลำพูน
54	เลย
55	ศรีสะเกษ
56	สกลนคร
57	สงขลา
58	สตูล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

province_id	รายละเอียด
59	สมุทรปราการ
60	สมุทรสงคราม
61	สมุทรสาคร
62	สระบุรี
63	สระแก้ว
64	สิงห์บุรี
65	สุโขทัย
66	สุพรรณบุรี
67	สุราษฎร์ธานี
68	สุรินทร์
69	หนองคาย
70	หนองบัวลำภู
71	อ่างทอง
72	อุดรธานี
73	อุตรดิตถ์
74	อุทัยธานี
75	อุบลราชธานี
76	อำนาจเจริญ

ตารางที่ 4.7 ตารางแสดงรายละเอียดของ inv_language

Inv_language	รายละเอียด
1	ภาษาไทย
2	ภาษาอังกฤษ
3	อื่นๆ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

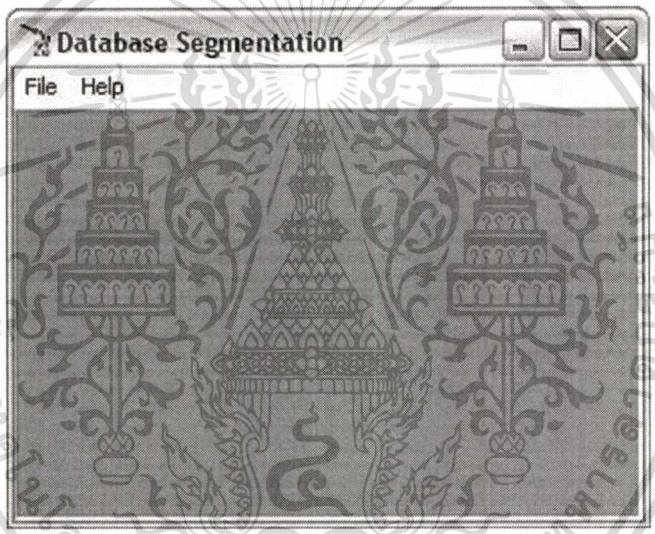
ตารางที่ 4.8 ตารางแสดงรายละเอียดของ pretransferbill

pretransferbill	รายละเอียด
0	จ่ายเงินในระยะเวลาที่กำหนด
1	จ่ายเงินเลยระยะเวลาที่กำหนด

4.5 ขั้นตอนและรายละเอียดการใช้งาน

4.5.1 การติดต่อกับข้อมูลที่จะนำมาวิเคราะห์

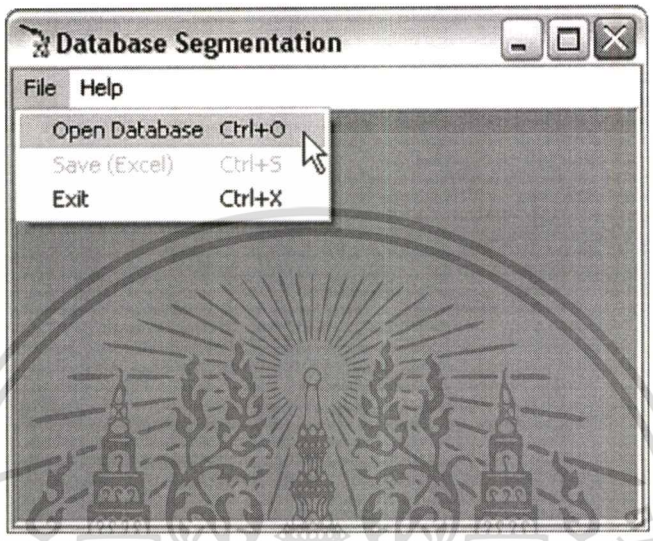
เมื่อเข้าสู่โปรแกรมจะปรากฏหน้าจอเมนูหลักดังรูปที่ 4.1



รูปที่ 4.1 หน้าจอหลักของระบบงาน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากนั้นเมื่อทำการกดที่ปุ่ม File จะปรากฏเมนูออกมาให้เลือก ในที่นี้ต้องการที่จะเปิดฐานข้อมูลขึ้นมา จึงทำการเลือก Open Database เพื่อทำการเปิดฐานข้อมูลที่ต้องการขึ้นมา



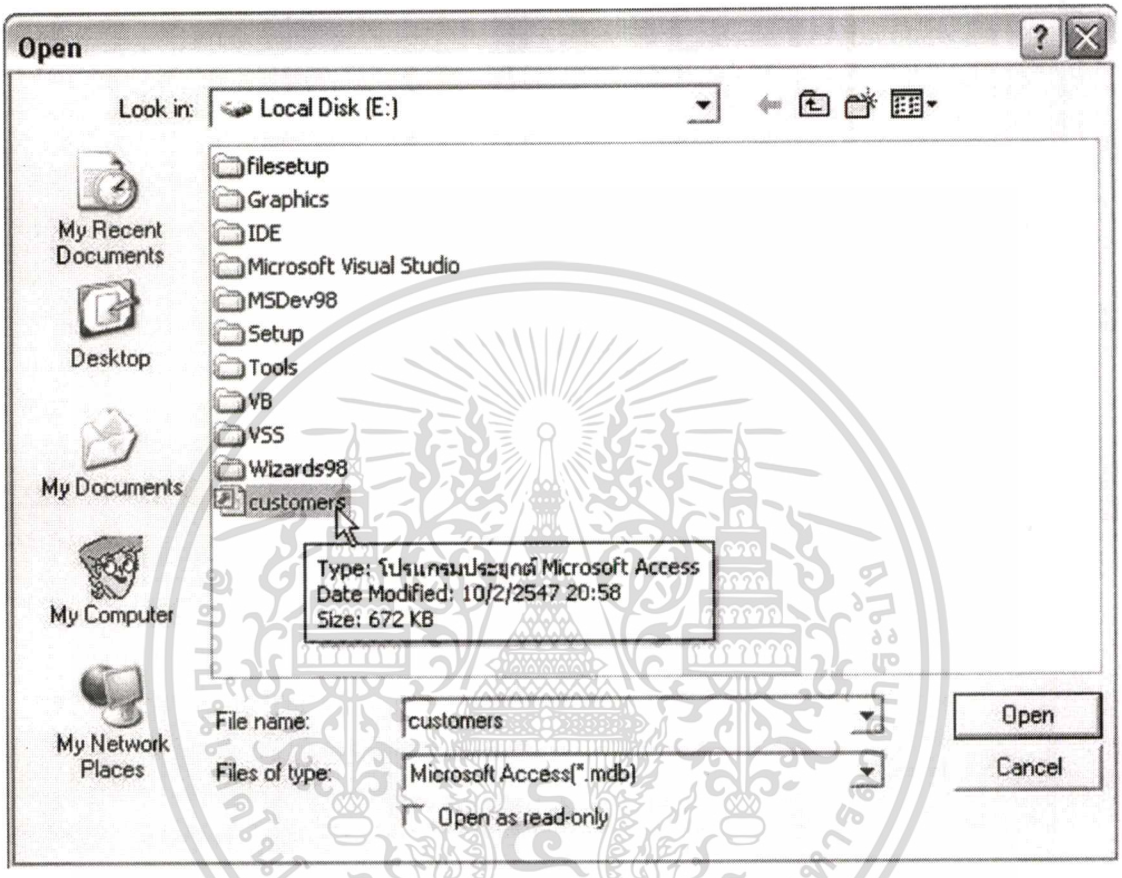
รูปที่ 4.2 หน้าจอแสดงการเลือกเมนูเพื่อทำการเปิดฐานข้อมูล

ในขั้นตอนต่อไปนี้ เป็นการเลือกฐานข้อมูลที่ต้องการทำงานขึ้นมา โดยทำการกดที่ปุ่ม Browse... ในช่อง Database Name ขึ้นมา เพื่อทำการเลือกฐานข้อมูลที่ต้องการขึ้นมา

รูปที่ 4.3 หน้าจอแสดงการเลือกฐานข้อมูล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ภายหลังจากที่กดปุ่ม Browse... แล้ว จะปรากฏหน้าจอดังรูปที่ 4.4 ขึ้นมา ให้ผู้ใช้ทำการเลือกไฟล์ฐานข้อมูลที่ต้องการขึ้นมาจากไฟล์ในคอมพิวเตอร์



รูปที่ 4.4 หน้าจอแสดงการเลือกฐานข้อมูล

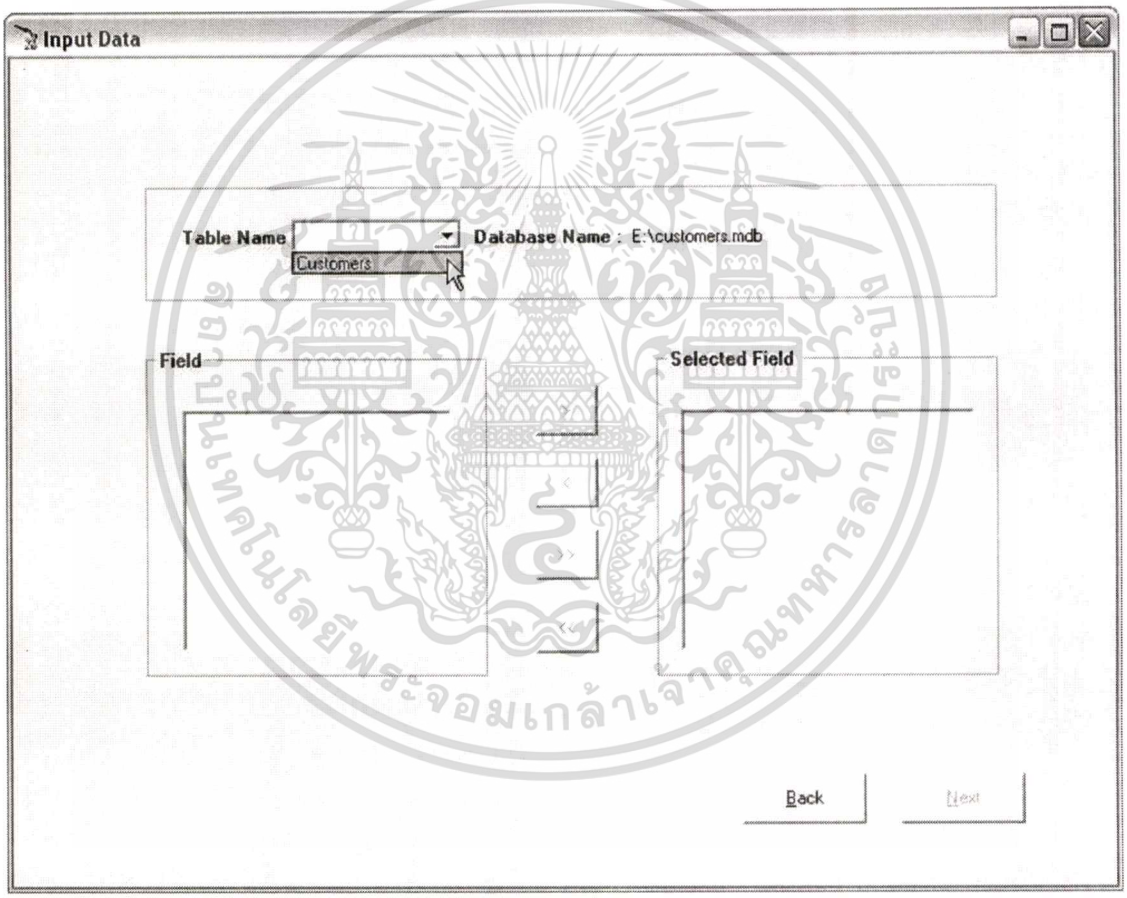
หลังจากเลือกฐานข้อมูลที่ต้องการได้แล้ว โปรแกรมจะกลับไปสู่หน้าจอ Open Database ดังรูปที่ 4.3 จากนั้นทำการกดปุ่ม Open ที่หน้าจอ Open Database เพื่อนำฐานข้อมูลที่ต้องการเข้ามาใช้งาน ในที่นี้ฐานข้อมูลที่นำเข้ามาใช้งานชื่อฐานข้อมูล customers.mdb

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.5.2 การเลือกฟิลด์ข้อมูลเข้ามาใช้ในการวิเคราะห์

ในขั้นตอนนี้เป็นการเลือกฟิลด์ข้อมูลที่ต้องการวิเคราะห์เข้ามาใช้งาน โดยผู้ใช้สามารถที่จะทำการกดปุ่มเพื่อเลือกฟิลด์ต่างๆเข้ามาทำงาน เมื่อทำการเลือกฟิลด์แล้ว ฟิลด์ที่ถูกเลือกจะเข้ามาอยู่ในช่อง Selected Field

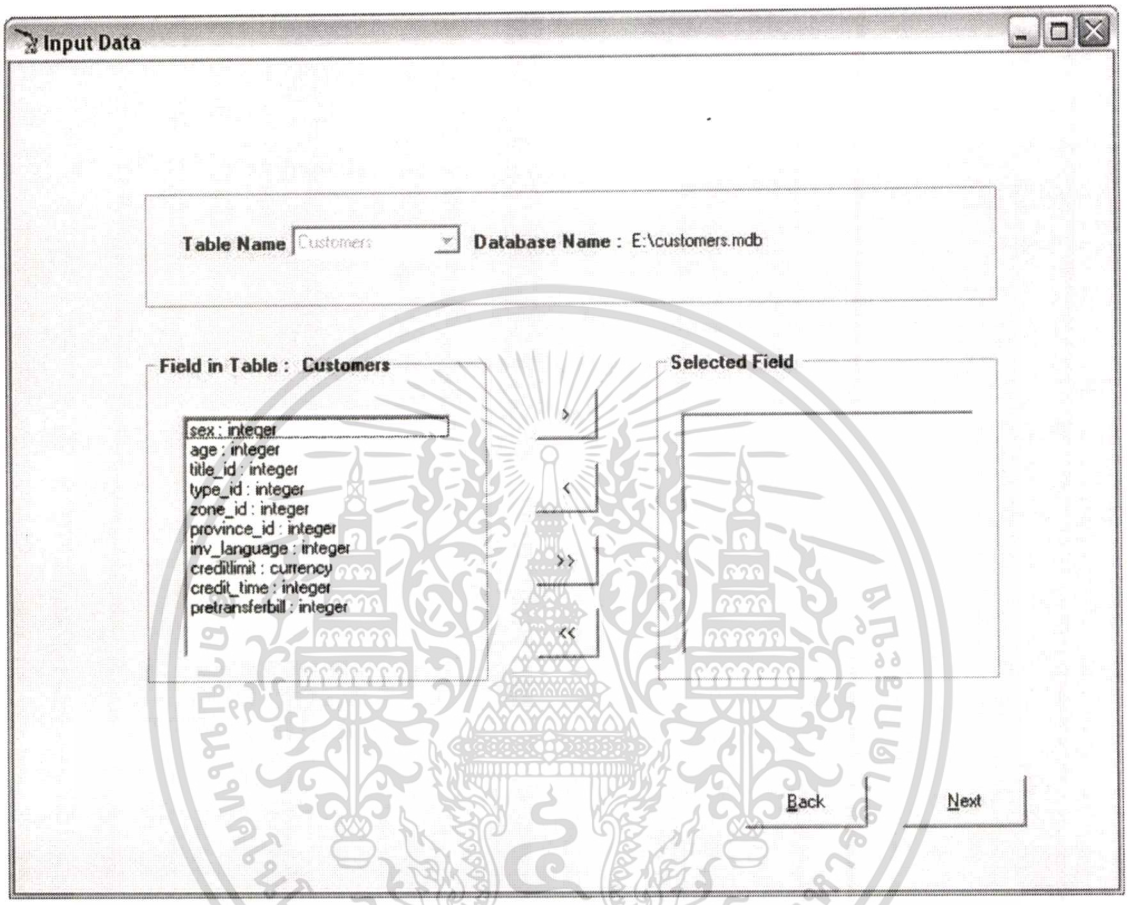
ในขั้นแรกต้องทำการเลือกตารางที่ต้องการก่อน ในที่นี้เลือกตารางที่ชื่อว่า Customers โดยจะคลิกที่ช่อง Table Name เพื่อเลือกตารางที่ต้องการขึ้นมา โดยจะเห็นว่าทางขวามือนั้นจะปรากฏชื่อของฐานข้อมูลที่ถูกเลือกขึ้นมาด้วย



รูปที่ 4.5 หน้าจอแสดงการเลือกตาราง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

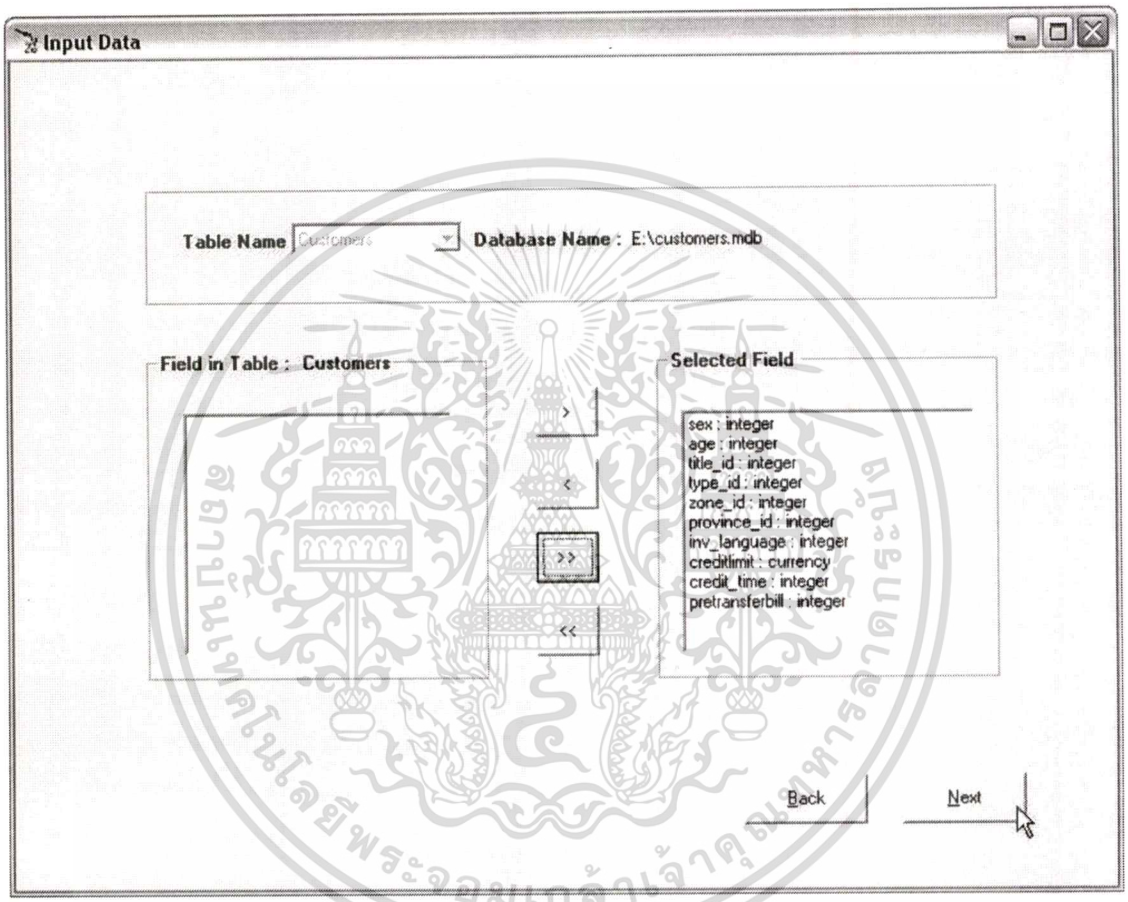
จากนั้นเมื่อได้ตารางที่ต้องการแล้ว ซึ่งในที่นี้คือตาราง Customers ที่ช่อง Field in Table: Customers จะแสดงฟิลด์ที่อยู่ภายในตาราง Customers ขึ้นมา



รูปที่ 4.6 หน้าจอแสดงข้อมูลภายในตาราง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ขั้นตอนถัดมา เป็นการเลือกฟิลด์ที่ต้องการนำมาวิเคราะห์ โดยการกดปุ่ม [>] หรือ [>>] เพื่อเลือกฟิลด์ที่ต้องการ ภายหลังจากที่กดปุ่มแล้ว ฟิลด์ที่ถูกเลือกจะไปปรากฏอยู่ที่ช่อง Selected Field



รูปที่ 4.7 หน้าจอแสดงการเลือกฟิลด์ที่ต้องการนำมาวิเคราะห์

เมื่อเลือกฟิลด์ที่ต้องการจะนำมาวิเคราะห์ได้แล้ว จากนั้นให้ทำการกดปุ่ม [Next] เพื่อทำงานในขั้นตอนถัดไป

ในที่นี้ทำการเลือกทุกฟิลด์ที่อยู่ในตาราง Customers ขึ้นมาใช้งาน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.5.3 การ Cleaning ข้อมูล

หลังจากที่เลือกฟิลด์ที่ต้องการนำมาวิเคราะห์แล้ว จากนั้นจะต้องนำข้อมูลที่อยู่ภายในฐานข้อมูล มาผ่านกระบวนการเตรียมข้อมูลเสียก่อน จึงจะสามารถนำข้อมูลไปทำการวิเคราะห์ได้

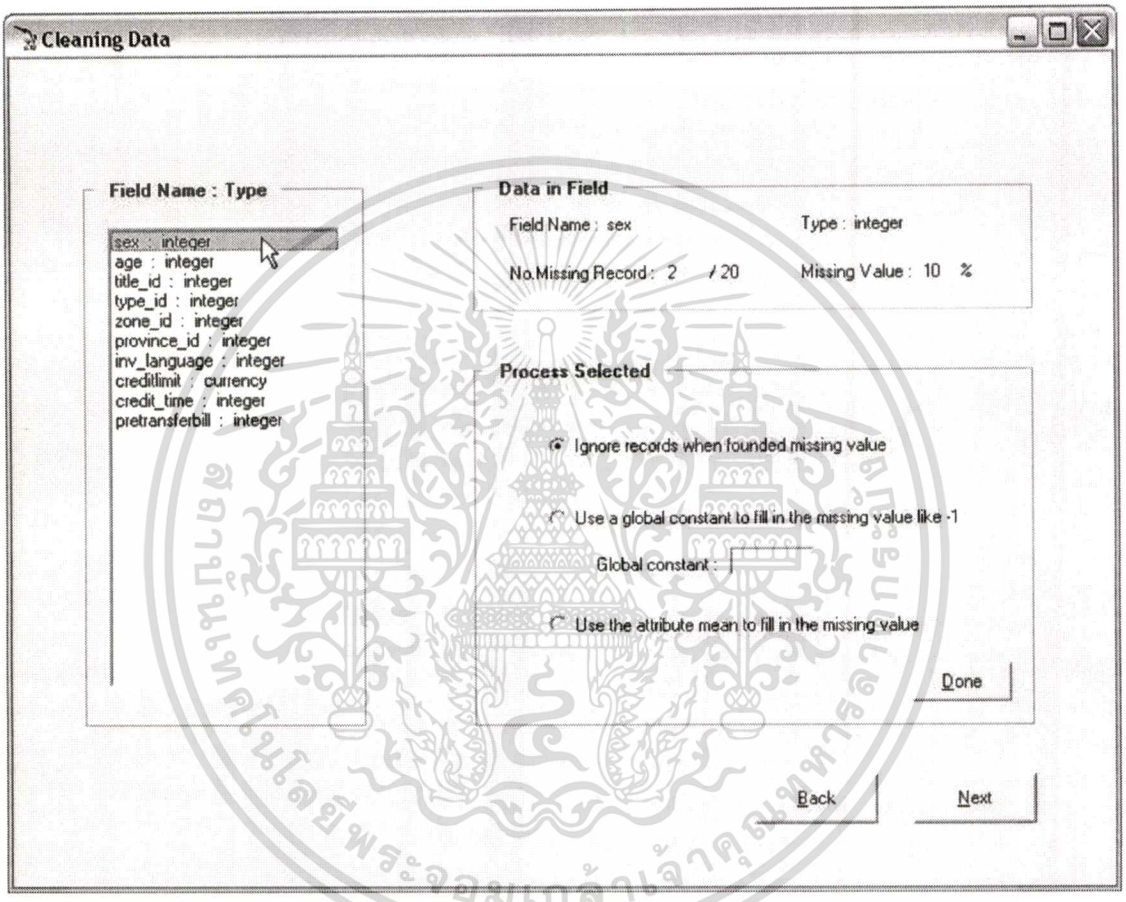
ในขั้นตอนนี้เป็นการตรวจสอบข้อมูลว่าภายในฐานข้อมูลที่ถูกเลือกขึ้นมาในตอนแรกนั้นมีค่า Missing Value เกิดขึ้นหรือไม่ในแต่ละฟิลด์ของข้อมูล โดยมีขั้นตอนดังต่อไปนี้

เมื่อเข้าสู่หน้าจอการ Cleaning Data ที่ช่อง Field Name: Type จะปรากฏฟิลด์ที่ได้ทำการเลือกไว้ในขั้นตอนที่แล้วขึ้นมาดังรูปที่ 4.8

รูปที่ 4.8 หน้าจอแสดงข้อมูลที่ได้จากการเลือกฟิลด์ข้อมูล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากนั้นเมื่อทำการเลือกโดยการคลิกลงไปที่แต่ละฟิลด์แล้ว จะเห็นว่าข้อมูลในแต่ละฟิลด์จะปรากฏขึ้นที่ช่อง Data in Field โดยข้อมูลที่ปรากฏในช่องนี้คือ ชื่อฟิลด์ (Field Name) ประเภทของฟิลด์ (Type) จำนวนของเรคอร์ดที่ไม่มีข้อมูลในแต่ละฟิลด์ (No. Missing Record) และ เปอร์เซ็นต์ของค่าMissing Value ในแต่ละฟิลด์ (Missing Value)



รูปที่ 4.9 หน้าจอแสดงข้อมูลของฟิลด์ที่ถูกเลือก

ในที่นี้ทำการเลือกที่ฟิลด์ที่ชื่อว่า sex โดยมีชนิดของข้อมูลคือ integer โดยมีจำนวนของเรคอร์ดที่ไม่มีข้อมูล จำนวน 2 เรคอร์ดจาก 20 เรคอร์ด เมื่อคิดเป็นเปอร์เซ็นต์แล้วมีค่าเท่ากับ 10 %

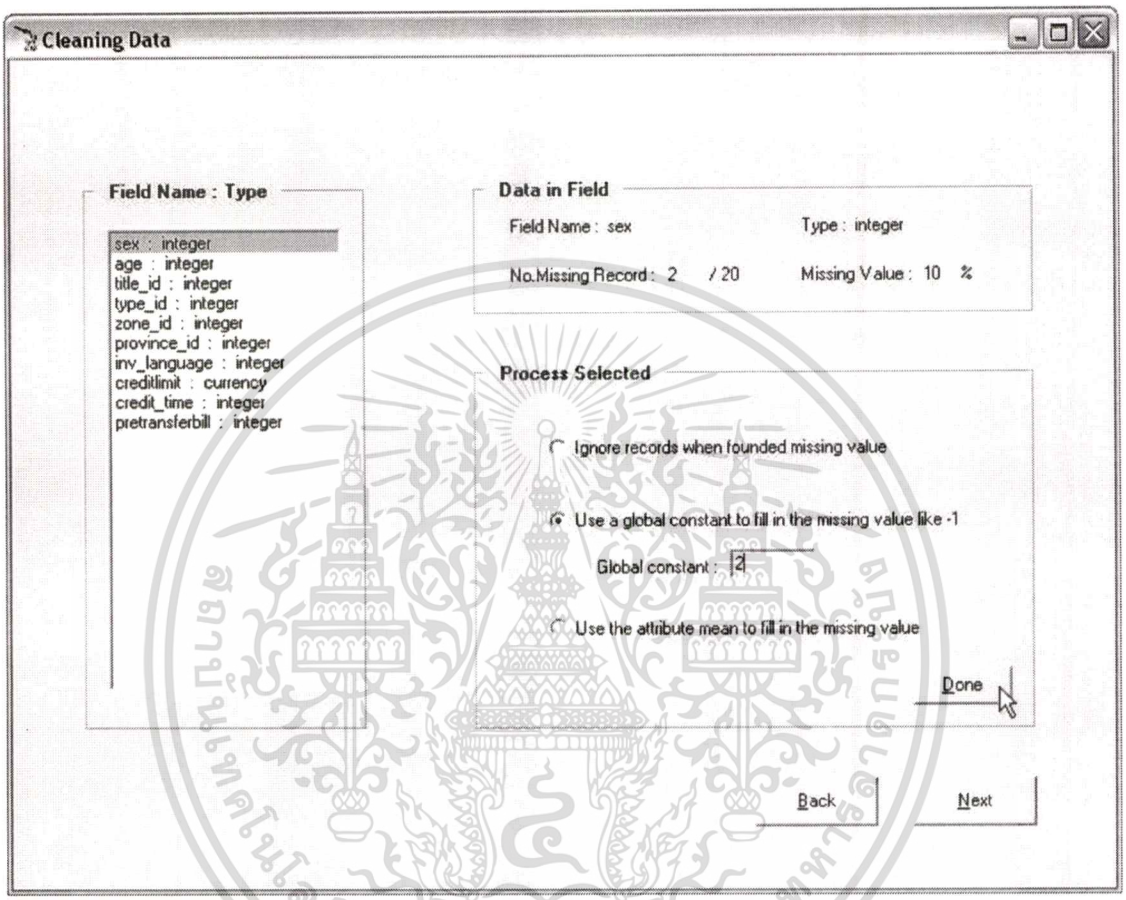
ในขั้นตอนต่อมา เป็นการตรวจสอบว่าในแต่ละฟิลด์มีฟิลด์ใดบ้างที่มีค่า Missing Value โดยในโครงการศึกษานี้ จะมีวิธีการกำจัดค่า Missing Value ให้เลือกด้วยกันทั้งหมด 3 วิธี ดังปรากฏในช่อง Process Selected ซึ่งจะประกอบไปด้วย

1. ตัดเรคอร์ดที่มีค่า Missing Value อยู่ออกไปจากฐานข้อมูล
2. ให้ผู้ใช้ทำการใส่ค่าที่ต้องการเข้าไป แทนค่าเรคอร์ดที่มี Missing Value

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3. เติมค่าเรคอร์ดที่เป็น Missing Value ด้วยค่า Mean ของข้อมูลทั้งหมด

หลังจากเลือกวิธีการที่จะกำจัดค่า Missing Value ได้แล้ว ให้เราทำการกดปุ่ม Done



รูปที่ 4.10 หน้าจอแสดงการกำจัดค่า Missing Value ในฟิลด์ sex

ในที่นี้ที่ฟิลด์ sex นั้นทำการแทนค่า Missing Value ด้วยวิธีการเติมค่าที่ต้องการลงไป โดยในที่นี้เลือกที่จะเติมค่า 2 เข้าไปแทนที่เรคอร์ดที่เป็นค่าว่างอยู่

จากนั้นเราจะทำการกำจัดค่า Missing Value ในทุกๆฟิลด์ จนกว่าฟิลด์ที่อยู่ในช่องทางซ้ายมือ ไม่มีฟิลด์ใดที่มีค่า Missing Value อยู่อีก

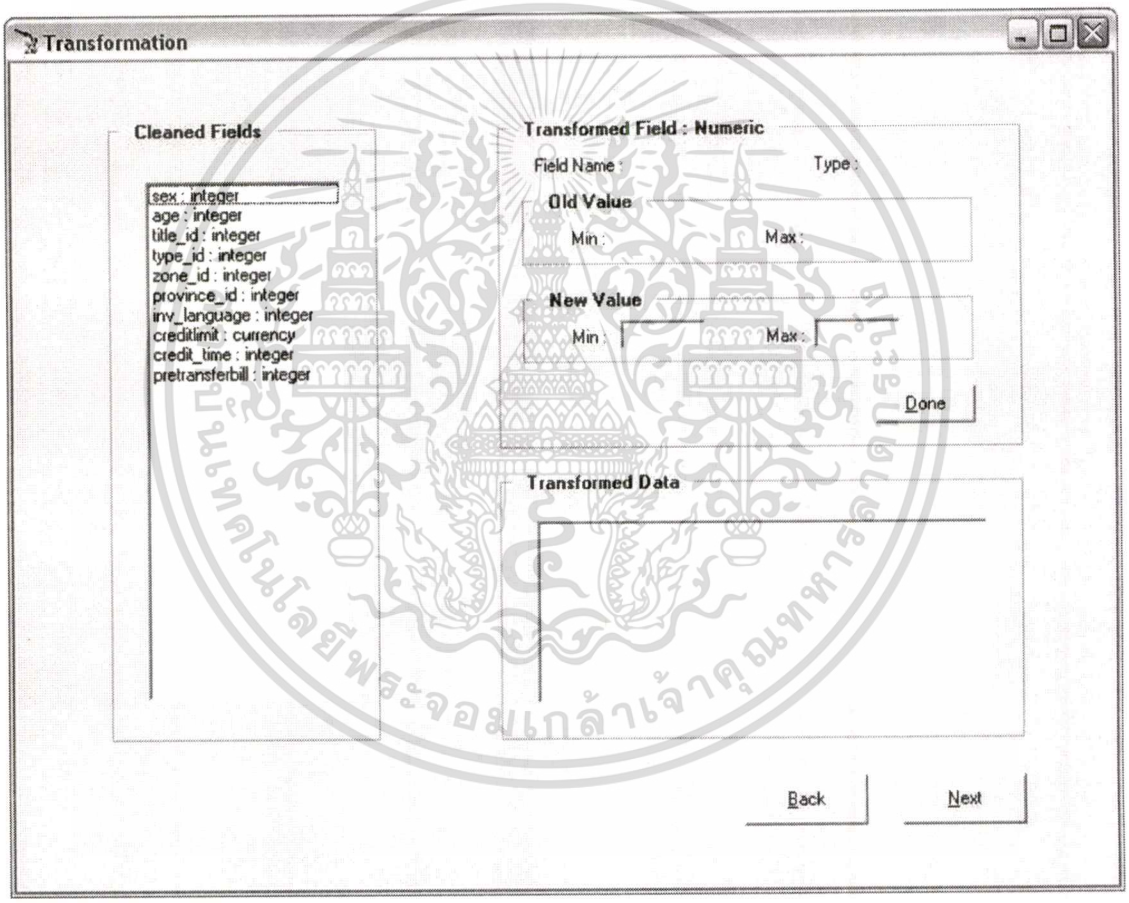
ขั้นตอนถัดมา เราจะทำการกดปุ่ม Next เพื่อทำงานในขั้นตอนถัดไป

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.5.4 การแปลงค่าของข้อมูล

ขั้นตอนนี้เป็นการแปลงค่าของข้อมูลให้มีความเหมาะสม โดยที่ผู้ใช้จะทำการแปลงค่าของข้อมูลในส่วนนี้ก็ได้ หรือจะเลือกใช้ข้อมูลเดิมก็ได้ ขึ้นกับความพอใจและความเหมาะสมของข้อมูล โดยมีขั้นตอนดังต่อไปนี้

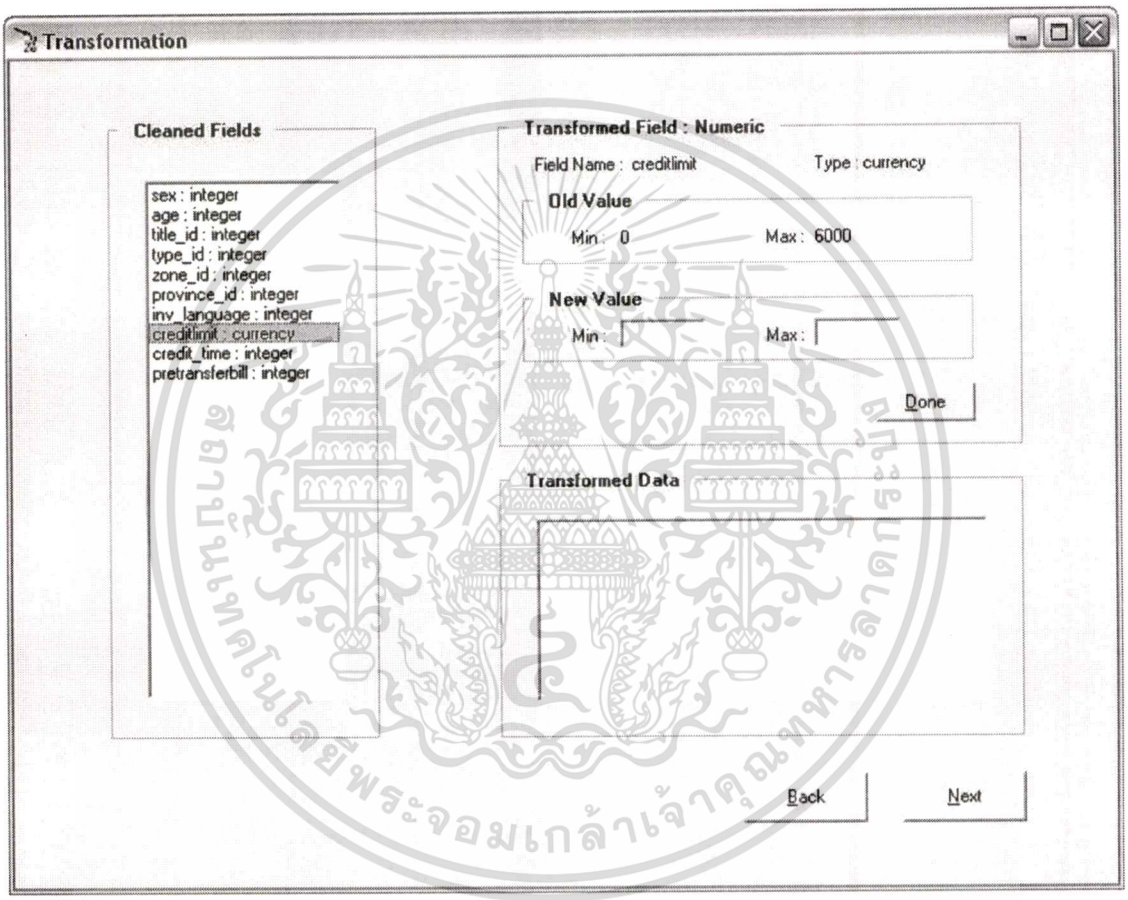
หลังจากข้อมูลที่ผ่านการทำงานในขั้นตอนการ Cleaning Data มาแล้ว ข้อมูลที่ปรากฏในช่อง Cleaned Fields คือข้อมูลที่เราเลือกขึ้นมาในตอนแรก และข้อมูลเหล่านี้ผ่านขั้นตอนของการ Cleaning Data มาแล้ว



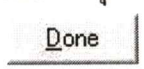
รูปที่ 4.11 หน้าจอแสดงข้อมูลหลังจากผ่านการ Cleaning Data

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

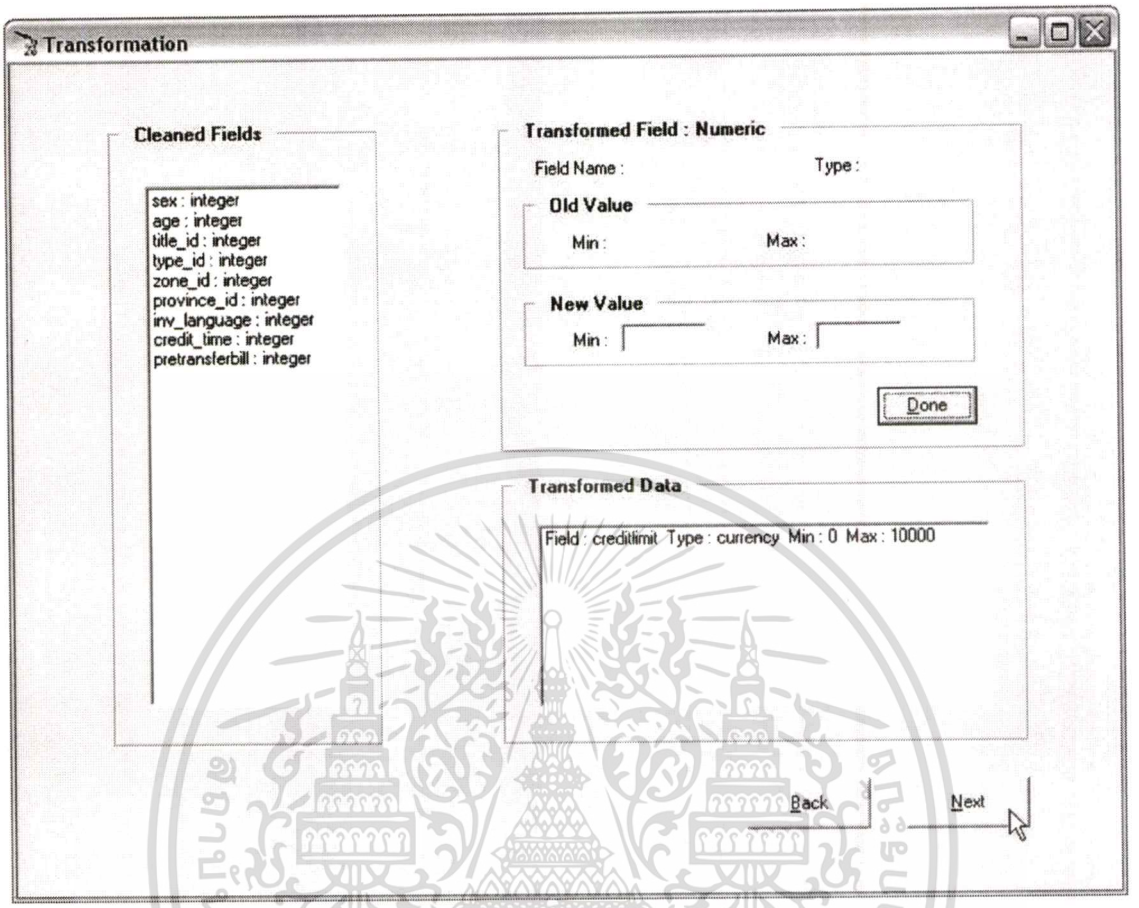
ขั้นตอนต่อมา ถ้าต้องการแปลงข้อมูล ก็ให้เลือกฟิลด์ข้อมูลที่จะทำการแปลงค่าข้อมูล โดยคลิกเลือกฟิลด์ที่ช่อง Cleaned Fields จากนั้นเมื่อดูที่ช่อง Transformed Field : Numeric จะพบว่ามีข้อมูลต่างๆ ปรากฏขึ้นมาหลังจากที่เลือกฟิลด์ในช่อง Cleaned Fields ซึ่งประกอบไปด้วย ชื่อฟิลด์ (Field Name) ประเภทของฟิลด์ (Type) ค่าเดิมในแต่ละฟิลด์(Old Value) ที่ประกอบไปด้วย ค่าต่ำสุด (Min) และค่าสูงสุด(Max)



รูปที่ 4.12 หน้าจอแสดงข้อมูลเมื่อเลือกฟิลด์ที่ต้องการแปลงข้อมูล

โดยถ้าต้องการที่จะแปลงค่าของข้อมูล ให้ทำการใส่ค่าใหม่ที่ต้องการลงไปในช่อง New Value โดยใส่ค่าต่ำสุดที่ต้องการลงในช่อง Min และค่าสูงสุดใหม่ที่ต้องการในช่อง Max จากนั้นทำการกดปุ่ม  เพื่อทำการแปลงข้อมูล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.13 หน้าจอแสดงข้อมูลเมื่อทำการแปลงค่าข้อมูลในฟิลด์ที่ต้องการ

หลังจากกดปุ่ม Done แล้วจะเห็นว่าข้อมูลที่ผ่านการแปลงค่าข้อมูลแล้ว จะแสดงรายละเอียดไว้ที่ช่อง Transformed Data เช่นในที่นี้ทำการแปลงค่าฟิลด์ที่ชื่อ credit_limit ให้มีค่าใหม่อยู่ระหว่าง 0-10000

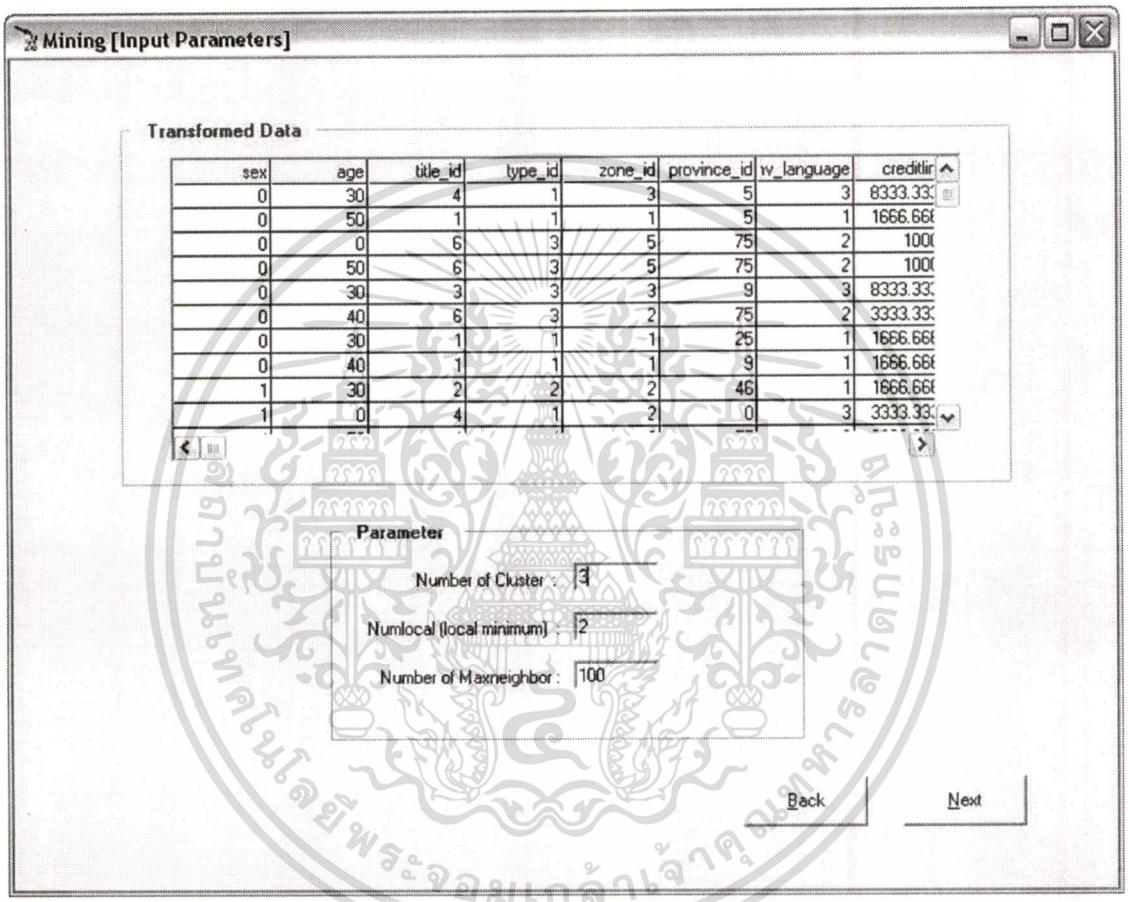
เมื่อทำการแปลงค่าข้อมูลที่ต้องการเรียบร้อยแล้ว จากนั้นให้กดปุ่ม Next เพื่อทำงานในขั้นตอนถัดไป

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.5.5 การทำดาต้าไมนิ่งและการแสดงผลลัพธ์

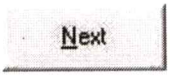
หลังจากข้อมูลผ่านการเตรียมข้อมูลในขั้นตอนต่างๆ แล้ว จากนั้นจะเข้าสู่กระบวนการทำดาต้าไมนิ่ง เพื่อวิเคราะห์ผลลัพธ์ออกมา โดยมีขั้นตอนต่างๆ ดังต่อไปนี้

หลังจากผ่านขั้นตอนการ Transformation Data แล้วจะปรากฏหน้าจอดังนี้



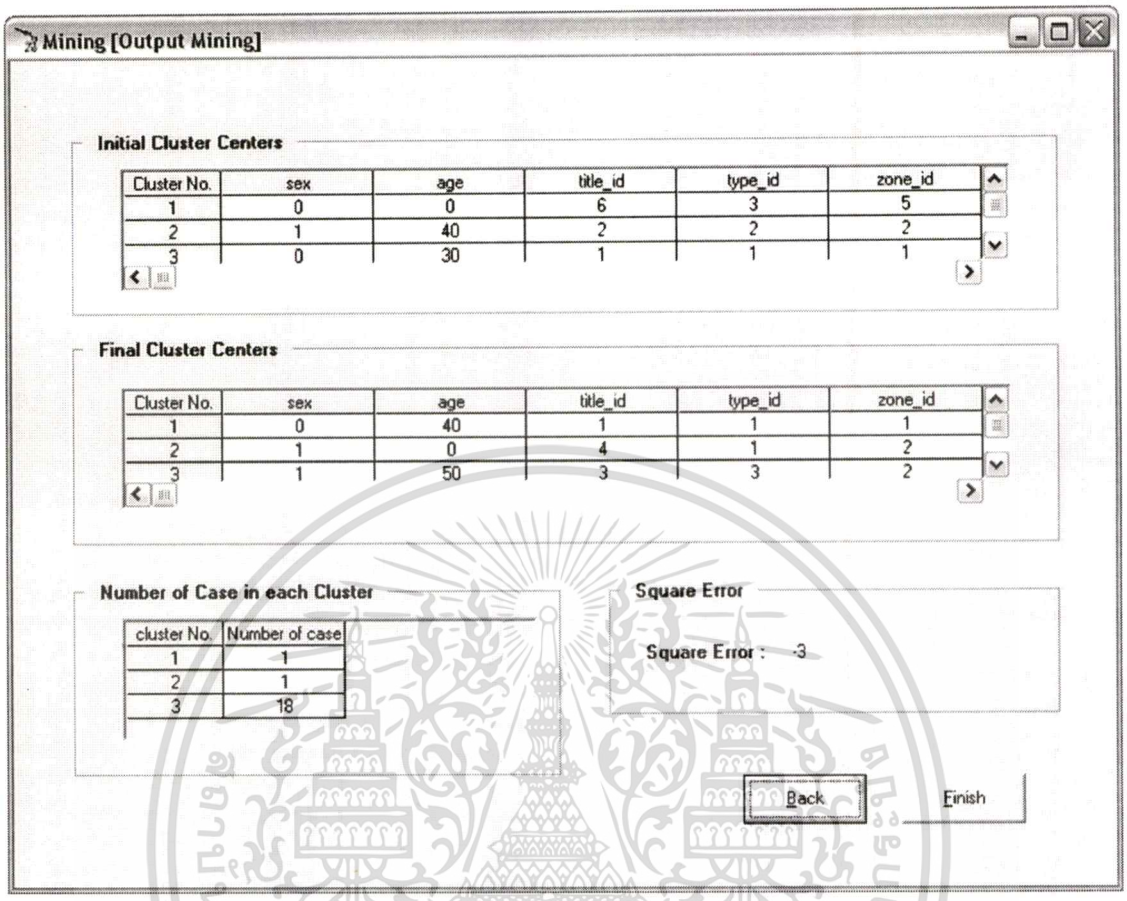
รูปที่ 4.14 หน้าจอแสดงข้อมูลหลังจากผ่านการแปลงข้อมูล

จะเห็นว่าข้อมูลที่อยู่ในช่อง Transformed Data คือข้อมูลที่ผ่านกระบวนการต่างๆในขั้นตอนแรกๆที่ได้ผ่านการทำงานมาแล้ว ในขั้นตอนนี้ ผู้ใช้จะต้องทำการกำหนดค่า Parameter ต่างๆที่จำเป็นสำหรับการทำงานลงไป โดยในที่นี้ต้องกำหนดด้วยกัน 3 ค่า คือค่า Number of Cluster คือจำนวนของ Cluster ที่ต้องการจะแบ่ง ค่า Numlocal และ ค่า Number of Maxneighbor

โดยหลังจากที่ทำการใส่ค่าให้กับ Parameter ทั้งสามตัวแล้ว ให้ทำการกดปุ่ม 

เพื่อทำงานในขั้นตอนถัดไป

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

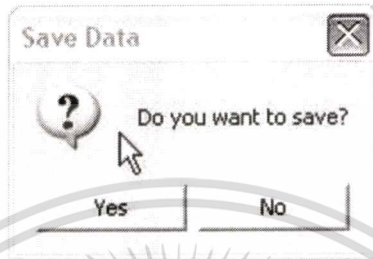


รูปที่ 4.15 หน้าจอแสดงข้อมูลหลังจากผ่านกระบวนการดาต้าไมนิ่ง

ที่หน้าจอ Mining [Output Mining] นั้นเป็นการแสดงผลลัพธ์ออกมาหลังจากที่ผ่านกระบวนการทำงานของอัลกอริทึม CLARANS แล้ว โดยที่หน้าจอนี้จะมีการแสดงผลลัพธ์ออกมา 4 รูปแบบด้วยกัน คือ Initial Cluster Centers ซึ่งเป็นการแสดงถึง medoid เริ่มต้นของแต่ละ Cluster ขึ้นมา ผลลัพธ์ถัดมาคือ Final Cluster Centers ก็คือข้อมูลหลังจากที่ผ่านกระบวนการของอัลกอริทึม CLARANS แล้วนั่นเอง ซึ่งจะแสดงถึงค่า medoid ใน Cluster ต่างๆ ซึ่งในแต่ละฟิลด์ก็จะมีค่า medoid ที่แตกต่างกันออกไป ตามข้อมูลของแต่ละฟิลด์ ถัดมาคือ Number of Case in each Cluster ในส่วนนี้จะแสดงถึง จำนวนของเรคอร์ดที่อยู่ในแต่ละ Cluster ว่ามีจำนวนเท่าไหร่บ้าง ในส่วนสุดท้ายคือ ค่า Square Error ที่เกิดขึ้นหลังจากทำการจัดกลุ่มข้อมูลแล้วว่าในการจัดกลุ่มนั้นมีความผิดพลาดเกิดขึ้นมากน้อยเพียงใด

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากนั้นเมื่อผู้ใช้ต้องการจบการทำงานของข้อมูลชุดนี้แล้ว ให้ทำการกดปุ่ม Finish จากนั้นจะมีข้อความขึ้นมาถามว่า Do you want to save? โดยถ้าต้องการที่จะ Save งานก่อนให้ทำการกดปุ่ม Yes แต่ถ้าต้องการที่จะจบการทำงานเลยให้กดปุ่ม No



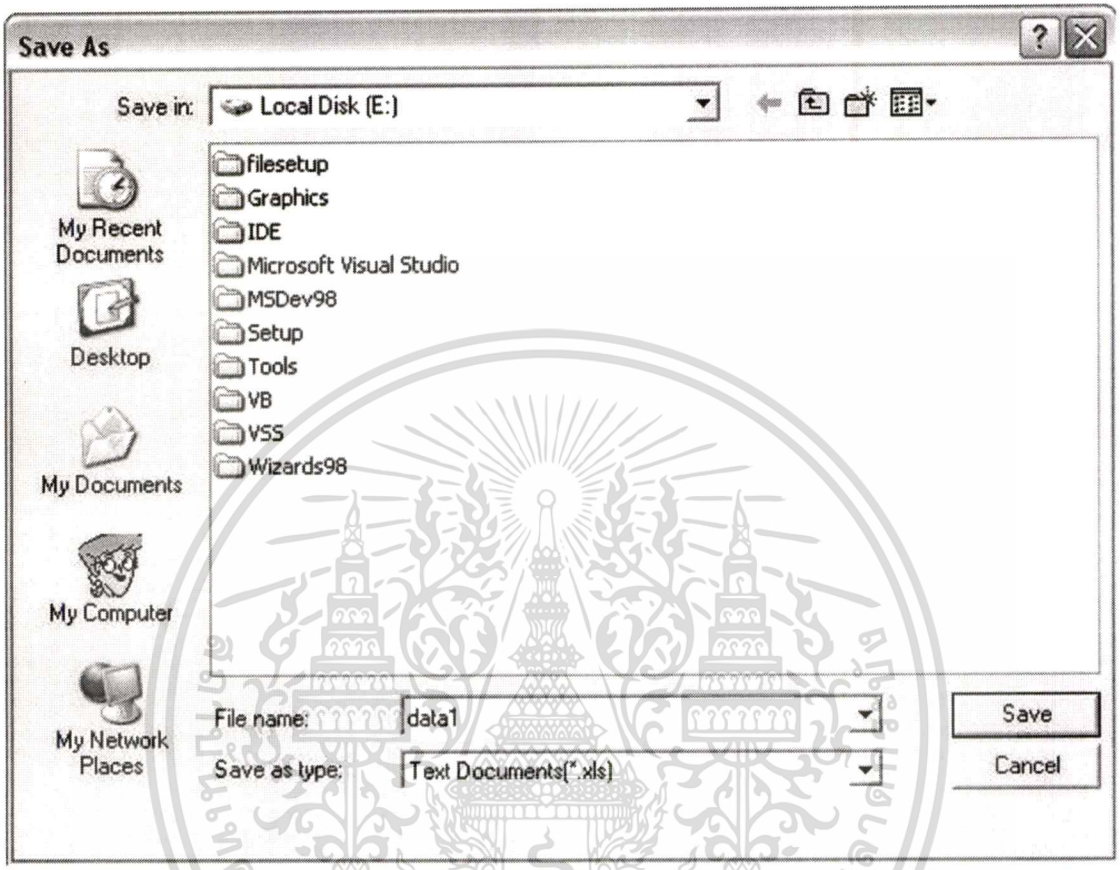
รูปที่ 4.16 หน้าจอแสดงข้อความถามความต้องการว่าต้องการที่จะบันทึกงานหรือไม่

ถ้าผู้ใช้ต้องการที่จะบันทึกงานให้ทำการเลือกที่เมนู โดยเลือก Save (Excel) ดังรูปที่ 4.17

รูปที่ 4.17 หน้าจอแสดงการเลือกเมนู Save(Excel)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากนั้นจะปรากฏหน้าจอดังรูปที่ 4.18

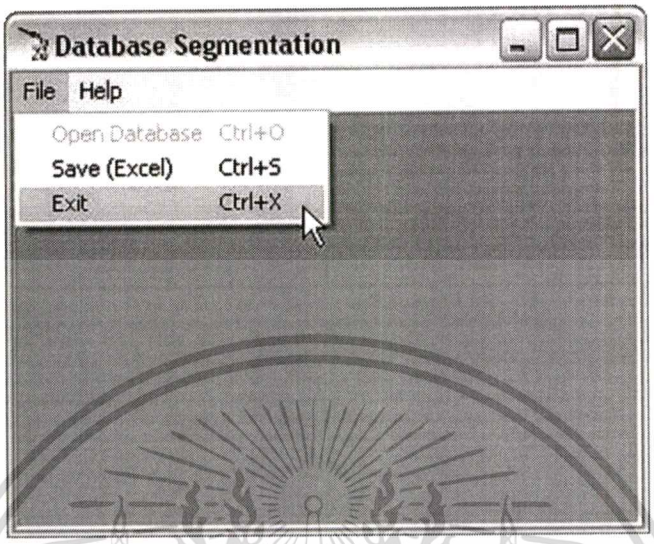


รูปที่ 4.18 หน้าจอแสดงการบันทึกผลลัพธ์

ที่หน้าจอนี้ ให้ทำการใส่ชื่อไฟล์ของข้อมูลที่จะทำการบันทึกลงไปในช่อง File name: โดยในที่นี้ใส่ชื่อไฟล์ว่า data1 โดยหลังจากทำการบันทึกข้อมูลแล้ว ไฟล์ข้อมูลที่ได้จะอยู่ในรูปไฟล์แบบ Excel

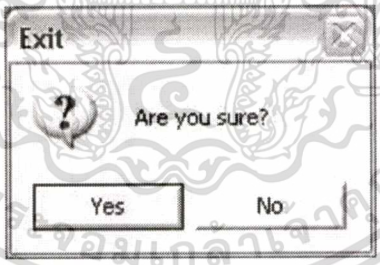
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากนั้นถ้าต้องการจบการทำงาน ให้เลือกที่เมนูโดยเลือกที่ Exit ดังรูปที่ 4.19



รูปที่ 4.19 หน้าจอแสดงการเลือกเมนู Exit

โปรแกรมจะขึ้นข้อความว่า Are you sure?



รูปที่ 4.20 หน้าจอแสดงข้อความถามความต้องการว่าต้องการที่จะจบการทำงานหรือไม่

โดยถ้าตอบ Yes จะเป็นการปิดโปรแกรมการใช้งานลงทันที แต่ถ้าตอบ No จะเป็นการกลับไปสู่หน้าจอเมนูหลักของโปรแกรม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 5
สรุปผลการศึกษาและข้อเสนอแนะ

5.1 สรุปผลการดำเนินการ

โครงการพัฒนาระบบงานเพื่อการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม CLARANS นี้เป็นโครงการที่จัดทำขึ้นเพื่อนำเสนอให้เห็นถึงประโยชน์ของการความรู้ทางด้านดาต้าไมนิ่งมาใช้งานให้เป็นประโยชน์ โดยในโครงการศึกษานี้ได้นำมาใช้ในการจัดกลุ่มให้กับข้อมูล เพื่อนำข้อมูลที่ได้นั้นไปประยุกต์ใช้กับงานทางด้านต่างๆ ไม่ว่าจะเป็นงานทางด้านธุรกิจ หรือ งานทางด้านลูกค้าสัมพันธ์ เพื่อให้เกิดประโยชน์อย่างสูงสุดกับแต่ละธุรกิจ

โครงการศึกษานี้เป็นการพัฒนาระบบโดยใช้อัลกอริทึม CLARANS (CLARANS: Clustering Large Applications based on RANdomized Search) ซึ่งเป็นอัลกอริทึมที่อยู่ในประเภทของการทำ Database Segmentation แบบ Partitioning โดยมีรูปแบบการทำงานที่ยังอยู่ในรูปของการรักษาค่าเฉลี่ยของระยะทางต่อออบเจ็คเอาไว้ โดยที่ตัวอัลกอริทึม CLARANS เองอยู่ในรูปของกระบวนการในการหา medoid ออกมาจำนวน k medoid ด้วยกัน (k เป็นจำนวนของ Cluster ที่ต้องการแบ่งออกมา) ซึ่งหลักการทำงานของอัลกอริทึมนี้เป็นการทำงานแบบสุ่มขึ้นมา โดยเป้าหมายของการสุ่มก็เพื่อเลือก medoid ที่ดีที่สุดขึ้นมานั่นเอง

ในระบบที่ได้ทำการพัฒนาขึ้นมานี้ สามารถที่จะทำการรับข้อมูลเข้ามาใช้งานได้ โดยที่ข้อมูลต้องอยู่ในรูปแบบของ .mdb โดยสามารถนำระบบที่ทำการพัฒนาขึ้นมานี้ไปใช้ในการวิเคราะห์ข้อมูลต่างๆได้ เช่น ข้อมูลของลูกค้าในบริษัทแห่งหนึ่ง มาทำการวิเคราะห์ว่าเมื่อทำการแบ่งกลุ่มลูกค้าออกมาแล้ว สามารถที่จะแบ่งลูกค้าได้เป็นกี่กลุ่ม และในแต่ละกลุ่มลูกค้ามีลักษณะเช่นไรบ้าง มีความแตกต่างกันมากน้อยแค่ไหน ในข้อมูลแต่ละกลุ่ม และนำข้อมูลที่ได้จากการแบ่งกลุ่มนี้ไปทำการวิเคราะห์เพื่อวางแผนทางการตลาด เพื่อรักษากลุ่มลูกค้าที่ดีขององค์กรเอาไว้ หรือทำการวางแผนสำหรับกลุ่มลูกค้าที่มีแนวโน้มว่าจะเป็นลูกค้าที่ดีให้ทำการซื้อขายกับองค์กรของเราต่อไป

ซึ่งผลลัพธ์ที่ได้จากการทำดาต้าไมนิ่งนั้น บางครั้งอาจจะไม่สามารถนำมาวิเคราะห์อะไรได้เลย อาจจะเป็นสาเหตุอันเนื่องมาจาก การกำหนดวัตถุประสงค์ของงานที่ไม่ชัดเจน หรือมีความคลุมเครือ อีกทั้งอาจเกิดจากความเปลี่ยนแปลงของฐานข้อมูลที่ลูกค้าบางคนอาจจะมีพฤติกรรมที่เปลี่ยนแปลงออกไปจากเดิม ทำให้การวิเคราะห์อาจเกิดความคลาดเคลื่อนได้ แต่อย่างไรก็ดี การจัด

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

แบ่งกลุ่มข้อมูลโดยใช้ดาต้าไมนิ่งนั้น ก็ยังคงเป็นแนวทางหนึ่งที่ช่วยทำให้เกิดโอกาสที่จะประสบความสำเร็จในการทำงานและการวางแผนได้เป็นอย่างดี

5.2 ข้อเสนอแนะ

1. ในระบบที่ทำการพัฒนาขึ้นมานี้สามารถที่จะนำไปวิเคราะห์ข้อมูลในด้านต่างๆ ได้ เนื่องมาจากระบบถูกออกแบบมาให้สามารถที่จะทำการติดต่อกับฐานข้อมูลอะไรก็ได้ โดยที่ผู้ใช้จะต้องทำการกำหนดวัตถุประสงค์ของงานเสียก่อน เพื่อให้ข้อมูลที่ผ่านการวิเคราะห์ออกมานั้นมีประโยชน์และใช้งานได้จริง อีกทั้งยังตรงตามความต้องการของผู้ใช้ จากนั้นผู้ใช้ก็ทำการเลือกตารางที่ต้องการทำงาน และเลือกฟิลด์ข้อมูลต่างๆที่จะนำมาวิเคราะห์ได้ด้วยตัวเอง ซึ่งจากตรงนี้เองทำให้ระบบมีความยืดหยุ่นในการใช้งานได้เป็นอย่างดี

2. ในระบบที่ทำการพัฒนาขึ้นมา สามารถที่จะรับข้อมูลเข้ามาทำการวิเคราะห์ได้ แต่ข้อมูลนั้นต้องอยู่ในรูปแบบที่เป็น Numeric เท่านั้น เนื่องมาจากข้อจำกัดทางด้านอัลกอริทึม ที่สามารถจะรับข้อมูลเข้าไปในแบบ Numeric ได้เพียงรูปแบบเดียว ระบบที่ทำการพัฒนาขึ้นจึงไม่รองรับข้อมูลที่เป็นแบบ Categorical

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บรรณานุกรม

Cabena et al. 1998. **Discovery Data Mining From Concept to Implementation.** New Jersy : Prentice Hall.

Chu, Shu-Chuan. Roddick, F John and Pan, J.S. **An Efficient K-medoids-based Algorithm Using Previous Medoid Index, Triangular Inequality Elimination Criteria and Partial Distance.** [Online]. Available: Http://www.ifs.univie.ac.at/~ww/dawak2002.htm

Gary Saarenvirta. 1998. **Mining Customer Data.** [Online]. Available: Http://www.db2mag.com/db_area/archives/1998/q3/98fsaar.shtml

Groth, Robert. 1998. **Data Mining a Hand-On Approach for Business Professionals.** New Jersy : Prentice Hall.

Han, Jiawei. and Kamber, Micheline. 2001. **Data Mining: Concepts and Techniques.** United States of America: Academic Press

Krzysztof (Kris) Koperski. 1997. **Methods Using Clustering.** [Online]. Available: Http://db.cs.sfu.ca/GeoMiner/survey/html/node9.html

Ng, R.T. and Han, Jiawei. 2002. **CLARANS: A Method for Clustering Objects for Spatial Data Mining.** IEEE Transactions on knowledge and data engineering. 14(5):1003-1016

Wei,Chih-Ping. Lee,Yen-Hsien. and Hsu,Che-Ming. 2000. **Empirical Comparison of Fast Clustering Algorithms for Large Data.** [Online]. Available: Http://www.computer.org/proceedings/hicss/0493/04932/049320/3abs.htm

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ประวัติผู้เขียน

ชื่อผู้เขียน	นางสาวดลินธา แก้วโกมลมาลย์
วันเดือนปีเกิด	5 มกราคม 2523
สถานที่เกิด	กรุงเทพมหานคร
การศึกษาระดับประถมศึกษา	โรงเรียนเซนต์โยเซฟ บางนา
การศึกษาระดับมัธยมศึกษา	โรงเรียนเซนต์โยเซฟ บางนา
วุฒิการศึกษาระดับปริญญาตรี	วิทยาศาสตรบัณฑิต สาขาวิชาคณิตศาสตร์ประยุกต์
สถานที่สำเร็จการศึกษา	คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ปีที่สำเร็จการศึกษา	ปีการศึกษา 2544

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้