ห้องสมุดคณะเทคโนโลยีสารสนเทศ สจล.

การพัฒนาระบบจัดกลุ่มข้อมูลโดยใช้ Algorithm CLARA

Clustering System Using CLARA Algorithm

โดย

นาย นิธิสิทธิ์ สุขแสง

รหัส 45061504

H002192

อาจารย์ที่ปรึกษา

ผศ.ดร.วรพจน์ กรีสุระเดช

b 11697957
i 1884077x

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน

หลักสูตรวิทยาศาตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

ภาคเรียนที่ 1 ปีการศึกษา 2547

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

วัน เดือน ปี 08 ก.พ. 2550
เลขทะเบียน 02192
เลขเรียกหนังสือ วพ. พ 6174ก 2547
"ห้องสมุดคณะเทคโนโลยีสารสนเทศ สจล."

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ชื่อหัวข้อ	การพัฒนาระบบจัดกลุ่มข้อมูลโดยใช้ Algorithm CLARA
นักศึกษา	นายนิธิสิทธิ์ สุขแสง
อาจารย์ที่ปรึกษา	ผศ.ดร.วรพจน์ กรีสุระเดช
ระดับการศึกษา	วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2547

บทคัดย่อ

การทำดาต้าไมนิ่งถูกนำมาใช้วิเคราะห์ข้อมูล และจัดกลุ่มข้อมูลเพื่อให้ได้ความรู้ใหม่ๆที่ซ่อนอยู่ในฐานข้อมูล ปัจจุบันข้อมูลเหล่านี้มีจำนวนมากขึ้น ทำให้เราสามารถนำข้อมูลเหล่านี้ไปใช้ประโยชน์ได้ดียิ่งขึ้น ดังนั้นในปัจจุบันจึงได้มีการคิดค้นวิธีการต่างๆในการนำข้อมูลเหล่านี้มาใช้ให้เกิดประโยชน์ ในโครงงานฉบับนี้ได้ศึกษาถึงวิธีการวิเคราะห์ข้อมูลลูกค้าเพื่อนำมาแบ่งกลุ่มลูกค้าโดยใช้ CLARA Algorithm มาวิเคราะห์ข้อมูล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Title	Clustering System Using CLARA Algorithm
Student	Mr. Nitisit Sooksang
Advisor	Asst. Prof. Dr.Worapoj Kreesuradej
Level of Study	Master of Science in Information Technology
Major	Information Science
Academic Year	2004

ABSTRACT

Data mining is used for data analysis and supports knowledge discovery by finding hiden data in databases and these data are clustered or grouped based.

In present there are large amounts of data which need to bring to analysis for the most useful. Therefore, this project presents the method for customers data analysis for group that are similar data to one another within the same cluster and are dissimilar to the objects in other clusters by using CLARA algorithm

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

กิตติกรรมประกาศ

สำหรับการพัฒนาระบบงานครั้งนี้ ข้าพเจ้าขอขอบพระคุณ ผศ.ดร.วรพจน์ กรีสุระเดช อาจารย์ที่ปรึกษาวิชาโครงการพัฒนาระบบงานเป็นอย่างสูง ที่ได้กรุณาให้คำแนะนำและคำปรึกษาในด้านต่างๆเป็นอย่างดี ทำให้การพัฒนาระบบสำเร็จลุล่วงด้วยดี

นอกจากนี้ข้าพเจ้าขอขอบพระคุณคุณพ่อคุณแม่ที่เป็นกำลังใจในการเรียนมาโดยตลอด และเพื่อนๆ คณะเทคโนโลยีสารสนเทศ รุ่น 14.1 และ 13.1 ที่ให้คำแนะนำติชมและข้อมูลอันเป็นประโยชน์ในการพัฒนางานชิ้นนี้

ข้าพเจ้าหวังเป็นอย่างยิ่งว่า โครงการพัฒนาระบบงานชิ้นนี้ จะเป็นประโยชน์และให้ความรู้แก่ผู้สนใจและบุคคลทั่วไป หากมีข้อบกพร่องประการใด ข้าพเจ้าขอน้อมรับไว้เพื่อพัฒนาระบบให้ดีขึ้นในโอกาสต่อไป

นายนิธิสิทธิ์ สุขแสง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ

หน้า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ (ต่อ)

หน้า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญตาราง

หน้า

ตารางที่

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญภาพ

หน้า

ภาพที่

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญภาพ (ต่อ)

หน้า

ภาพที่

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 1
บทนำ

1.1 หลักการและเหตุผล

ในปัจจุบันนี้ธุรกิจต่างๆได้มีการแข่งขันกันอย่างสูงซึ่งทำให้นักการตลาดและผู้บริหารต้องวางแผนกลยุทธ์ทางการตลาดของตนให้มีประสิทธิภาพสูงที่สุด เพื่อที่จะเพิ่มผลกำไรหรือแย่งส่วนแบ่งทางการตลาด โดยการตัดสินใจและการวางแผนกลยุทธ์ของนักการตลาดหรือผู้บริหารนั้นก็จะอาศัยปัจจัยหลายๆด้านมาช่วยไม่ว่าจะเป็นประสบการณ์ที่ได้เรียนรู้มาของตัวผู้บริหารเองหรืออาศัยข้อมูลที่ได้เก็บไว้ในฐานข้อมูลหรือจากที่อื่นๆที่ได้เก็บรวบรวมไว้มาวิเคราะห์หาความต้องการของลูกค้าเพื่อที่จะนำมาปรับปรุงกลยุทธ์ต่างๆขององค์กรเพื่อให้ได้ผลตอบแทนทางธุรกิจมากที่สุดเท่าที่จะมากได้ และเนื่องจากข้อมูลในฐานข้อมูลที่จะนำมาใช้วิเคราะห์,วางแผนทางกลยุทธ์นั้นนับวันจะมีจำนวนมากขึ้นเรื่อยๆจนเกินความสามารถของคนที่จะทำการวิเคราะห์ข้อมูลเหล่านี้ด้วยตัวเอง จึงมีการนำเทคโนโลยีดาต้าไมนิ่งมาช่วยในการวิเคราะห์ข้อมูล

โครงการพัฒนาระบบงานนี้ จึงได้นำเอาดาต้าไมนิ่งมาช่วยในการแบ่งกลุ่มข้อมูลของลูกค้าโดยนำเอาข้อมูลที่มีอยู่มาผ่านกระบวนการให้กลายเป็นความรู้ ซึ่งจะช่วยให้ผู้บริหารสามารถวางแผนการดำเนินธุรกิจได้อย่างมีประสิทธิภาพ ช่วยลดต้นทุน อีกทั้งยังทำให้ทราบถึงความต้องการของลูกค้าซึ่งจะช่วยให้รักษาลูกค้าไว้และเพิ่มลูกค้าใหม่ ซึ่งจะนำไปสู่ช่องทางที่จะทำให้ธุรกิจประสบความสำเร็จมากยิ่งขึ้น

1.2 วัตถุประสงค์ในการพัฒนาระบบ

การศึกษาโครงการพัฒนาระบบงานนี้มีวัตถุประสงค์เพื่อศึกษาเทคนิคของดาต้าไมนิ่งมาใช้ในการแบ่งกลุ่มข้อมูลโดยใช้ CLARA(Clustering LARge Applications)อัลกอริทึม เพื่อจัดกลุ่มลูกค้าที่มีลักษณะใกล้เคียงกัน จะช่วยให้ผู้บริหารสามารถวางแผนกลยุทธ์และกำหนดเป้าหมายทางการตลาดในการนำเสนอรายการส่งเสริมการขายหรือสิทธิพิเศษและบริการต่างๆ ให้ตรงตามกลุ่มเป้าหมายได้ชัดเจนมากขึ้น

1.3 ขอบเขตของการพัฒนาระบบ

โครงการพัฒนาระบบงานนี้ ได้กำหนดขอบเขตของการศึกษาไว้ ดังนี้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

1. ศึกษาถึงการนำเอาเทคนิคดาต้าไมนิ่งมาประยุกต์ใช้ โดยอาศัยหลักการของ Database Segmentation (Clustering Analysis) ด้วยอัลกอริทึม CLARA(Clustering LARge Applications)ในการแบ่งกลุ่มข้อมูลลูกค้า
2. ข้อมูลที่นำมาใช้ในการวิเคราะห์นั้น เป็นข้อมูลของลูกค้าที่ได้สมัครลงทะเบียนเรียนพิเศษที่สถาบันอบรมคอมพิวเตอร์แห่งหนึ่ง โดยจะนำข้อมูลมาผ่านขั้นตอนต่างๆของดาต้าไมนิ่ง แล้วจัดกลุ่มลูกค้าที่มีความเหมือนหรือใกล้เคียงกัน

1.4 ขั้นตอนและวิธีการดำเนินงาน

ในการศึกษาโครงการพัฒนาระบบงานนี้ เพื่อให้ศึกษาคลอบคลุมวัตถุประสงค์และขอบเขตในการพัฒนาระบบ จึงได้กำหนดขั้นตอนในการศึกษาไว้ดังนี้

1. กำหนดวัตถุประสงค์ในการแบ่งกลุ่มข้อมูลของลูกค้า
2. ศึกษาแนวคิด ขั้นตอน และกระบวนการในการทำดาต้าไมนิ่งโดยเลือกใช้ CLARA(Clustering LARge Applications)อัลกอริทึม
3. เก็บรวบรวม และศึกษาข้อมูลที่จะนำมาใช้ในการแบ่งกลุ่ม โดยเลือกใช้ข้อมูลลูกค้าที่ได้สมัครลงทะเบียนเรียนพิเศษที่สถาบันอบรมคอมพิวเตอร์แห่งหนึ่ง
4. เลือกข้อมูล และเตรียมข้อมูลให้อยู่ในรูปแบบที่เหมาะสมกับอัลกอริทึม
5. ออกแบบและพัฒนาระบบงานเพื่อวิเคราะห์ข้อมูล

1.5 ประโยชน์ที่คาดว่าจะได้รับ

จากการศึกษาและพัฒนาระบบงานเพื่อแบ่งกลุ่มลูกค้าที่มีลักษณะใกล้เคียงกัน คาดว่าจะให้ประโยชน์ในการวางแผนทางธุรกิจ ดังนี้

1. ช่วยทำให้เข้าใจความต้องการของลูกค้ามากขึ้น
2. ช่วยให้สามารถกำหนดเป้าหมายทางการตลาดได้ชัดเจนมากขึ้น เช่น สร้างและเสนอรายการส่งเสริมการขายที่ตรงกับความต้องการของลูกค้าในแต่ละกลุ่ม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 2
ดาต้าไมนิ่งและทฤษฎีที่เกี่ยวข้อง

ข้อมูลที่จัดเก็บภายในคลังข้อมูลนั้น ถึงแม้ว่าจะถูกจัดเก็บอย่างมีระบบและมีประสิทธิภาพสูงก็ตาม แต่ถ้าข้อมูลเหล่านั้นไม่มีกระบวนการในการทำสารสนเทศที่ดีแล้ว ข้อมูลที่มีก็จะเป็นเพียงข้อมูล (Data) ที่ถูกจัดเก็บไว้ซึ่งจะไม่มีประโยชน์เลย แต่หากเรานำข้อมูลเหล่านั้นมาผ่านกระบวนการในการทำสารสนเทศที่ถูกวิธีแล้วข้อมูลเหล่านั้นก็จะกลายเป็นสารสนเทศ (Information) เกิดเป็นฐานความรู้ (Knowledge Base) และนำความรู้ที่ได้นั้นไปประยุกต์ใช้ในการทำธุรกิจ และเมื่อพิจารณาถึงความสามารถของคอมพิวเตอร์ในปัจจุบันที่มีสมรรถนะสูงแต่ในราคาที่ต่ำ สามารถรองรับเทคนิคของดาต้าไมนิ่งที่ประกอบด้วยอัลกอริทึมที่มีความซับซ้อนและความต้องการการคำนวณสูงได้นั้น ยิ่งทำให้ดาต้าไมนิ่งจึงเป็นเทคโนโลยีที่ได้รับความนิยมเพื่อใช้ในสร้างระบบสนับสนุนการตัดสินใจ (Decision Support) มากขึ้นด้วย

2.1 ความหมายของดาต้าไมนิ่ง (Data Mining)

ดาต้าไมนิ่งเป็นขั้นตอนหรือกระบวนการในการดึงข้อมูลสารสนเทศจากฐานข้อมูลขนาดใหญ่โดยข้อมูลสารสนเทศที่ดึงออกมานั้นต้องเป็นข้อมูลที่ไม่ทราบมาก่อนว่ามีประโยชน์หรือมีความสัมพันธ์กันอย่างไร นอกจากนี้ยังจะต้องมีความถูกต้องและสามารถนำไปใช้งานได้จริงเพื่อนำไปช่วยในการตัดสินใจในการทำธุรกิจต่อไป

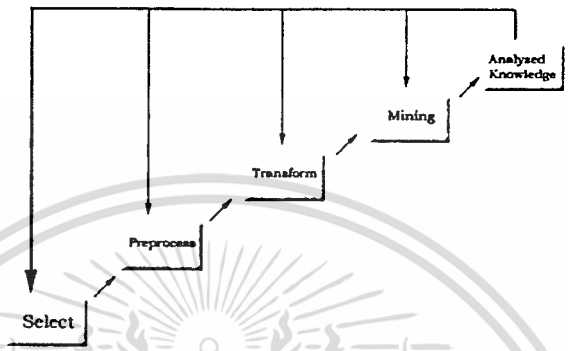
สาเหตุที่ต้องนำดาต้าไมนิ่งมาช่วยในการวิเคราะห์ข้อมูลก็เพราะว่าในองค์กรต่างๆจะมีการเก็บข้อมูลในฐานข้อมูลอยู่เยอะมากแต่ไม่สามารถนำข้อมูลเหล่านี้มาใช้ให้เกิดประโยชน์ได้มากนักจึงมีการนำกระบวนการทำดาต้าไมนิ่งมาช่วยค้นหาข้อมูลในฐานข้อมูลที่สามารถนำไปใช้ให้เกิดประโยชน์ได้

กระบวนการในการทำดาต้าไมนิ่งหรือเรียกอีกอย่างหนึ่งว่าการทำ Knowledge Discovery in Database (KDD) ประกอบไปด้วย ขั้นตอนดังต่อไปนี้

- การกำหนดวัตถุประสงค์ (Business Objective Determination)
- การเลือกข้อมูลที่จะนำมาใช้ในการทำดาต้าไมนิ่ง (Data Selection)
- การทำให้ข้อมูลมีคุณภาพที่ดี(Data Preprocessing)
- การแปลงรูปแบบข้อมูล (Data Transformation)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

- การทำดาต้าไมนิ่ง (Data Mining)
- การวิเคราะห์ผลลัพธ์ (Analysis of result)
- การนำไปประยุกต์ใช้งาน (Assimilation of knowledge)



รูปที่ 2.1 แสดงกระบวนการในการทำดาต้าไมนิ่ง

2.2 กระบวนการทำงานของดาต้าไมนิ่ง (Data Mining Process)

Database Segmentation หรือ Clustering Analysis เป็นเทคนิคหนึ่งในการทำดาต้าไมนิ่ง โดยหลักการทำงานจะทำการแบ่งกลุ่มของข้อมูลในฐานข้อมูล โดยที่ข้อมูลในแต่ละกลุ่มจะมีความเหมือนหรือคล้ายคลึงกัน ซึ่งเมื่อได้แบ่งข้อมูลออกเป็นกลุ่มแล้วก็จะนำข้อมูลเหล่านี้มาวิเคราะห์ดูว่าข้อมูลในแต่ละกลุ่มมีลักษณะอย่างไร ซึ่งเทคนิคในการทำ Clustering ก็แบ่งออกเป็นหลายๆ อัลกอริทึม ในสัมมนาฉบับนี้จะกล่าวถึง การนำ CLARA(Clustering LARge Applications) อัลกอริทึมมาใช้ในการทำ Database Segmentation เพื่อที่จะได้นำผลลัพธ์ที่ได้จากการทำดาต้าไมนิ่งมาวางแผนกลยุทธ์ทางการตลาดต่อไป

ขั้นตอนที่ 1 การกำหนดวัตถุประสงค์ของงาน (Business Objective Determination)

เป็นการกำหนดวัตถุประสงค์ของการทำดาต้าไมนิ่ง ซึ่งขั้นตอนนี้เป็นขั้นตอนที่จำเป็นที่จะต้องหาวัตถุประสงค์ออกมาให้ได้ หากไม่สามารถกำหดวัตถุประสงค์ที่ต้องการได้ ก็จะไม่สามารถทำงานในกระบวนการต่อๆไป ตัวอย่างของการกำหนดวัตถุประสงค์ เช่นการหาพฤติกรรมของผู้บริโภคหรือหาความสัมพันธ์ของสินค้าที่ผู้บริโภคมักจะซื้อควบคู่กันไป เช่น เมื่อผู้บริโภคซื้อขนมปังก็จะซื้อนมไปด้วย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ขั้นตอนที่ 2 การเลือกข้อมูล (Data Selection)

กระบวนการนี้เป็นการเลือกข้อมูลจากฐานข้อมูลที่จะนำไปใช้ในการทำดาต้าไมนิ่ง สาเหตุที่ต้องทำการเลือกข้อมูลก็เพราะว่า โดยปกติ เราจะไม่นำข้อมูลทุกๆอย่างที่เก็บไว้มาใช้ในการทำงาน เราจะเลือกข้อมูลที่คิดว่าเกี่ยวข้องกับสิ่งที่เราต้องการทำการวิเคราะห์มาใช้ในการทำงานเท่านั้น

ขั้นตอนที่ 3 การทำให้ข้อมูลมีคุณภาพดี (Data Preprocessing)

จุดประสงค์ของกระบวนการนี้เป็นการนำข้อมูลที่จะนำมาใช้ในการทำดาต้าไมนิ่ง มาทำให้เป็นข้อมูลที่มีคุณภาพดีก่อนที่จะนำไปใช้งานต่อไป

เนื่องจากข้อมูลที่เรามีอยู่ในปัจจุบันอาจจะมีบางส่วนที่ผิดพลาดหรือไม่สมบูรณ์เช่น ข้อมูลบางอย่างขาดหายไป, ข้อมูลมีความซ้ำซ้อน,ข้อมูลไม่สอดคล้องกัน โดยถ้าหากเรานำข้อมูลที่ไม่สมบูรณ์ไปใช้งานก็จะทำให้ผลลัพธ์ที่ได้จากการทำดาต้าไมนิ่งมีความผิดพลาด เพราะฉะนั้นก่อนที่จะนำข้อมูลไปใช้งานก็จะต้องผ่านกระบวนการของ Data Preprocessing ก่อน โดยที่กระบวนการทำ Data Preprocessing จะมีการนำเทคนิคต่างๆต่อไปนี้มาช่วยจัดการในเรื่องนี้

1. Data Cleaning

ในกรณีที่พบว่ามีข้อมูลในบางส่วนขาดหายไป ก็จะทำการ ตัด Field ที่ข้อมูลขาดหายเยอะๆ ทิ้งไป หรืออาจจะแทนค่าช่องที่ขาดหายไปด้วย "unknown" เป็นต้น

2. Data Integration

Data Integration เป็นวิธีการกำจัดความซ้ำซ้อนของข้อมูล การเกิดข้อมูลซ้ำซ้อนกันอาจจะเกิดจากการนำฐานข้อมูลหลายๆฐานข้อมูลมารวมกัน โดยที่แต่ละฐานข้อมูลก็จะมี attribute ที่มีชื่อแตกต่างกันไป ถึงแม้ว่า attribute บาง attribute ในแต่ละฐานข้อมูลจะเก็บข้อมูลที่เหมือนกันก็ตาม ยกตัวอย่างเช่น ฐานข้อมูล A เก็บข้อมูลอายุของพนักงาน โดยที่ตั้งชื่อของ attribute นี้ว่า AGE และฐานข้อมูล B ซึ่งเก็บข้อมูลอายุของพนักงานเหมือนกันแต่ตั้งชื่อ attribute ว่า AGE_PERSON เมื่อนำฐานข้อมูล A และฐานข้อมูล B มารวมกันก็จะทำให้เกิดความซ้ำซ้อนของข้อมูลขึ้น ซึ่งในกรณีนี้อาจจะใช้วิธี Correlational analysis มาช่วยแก้ปัญหาโดยมีสมการดังนี้

$$r_{a,b} = \frac{\sum(a - \bar{a})(b - \bar{b})}{(n - 1)\sigma_a \sigma_b}$$

$$\bar{a} = \frac{\sum a}{n}, \quad \sigma_a = \sqrt{\frac{\sum(a - \bar{a})^2}{n - 1}}$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

โดยที่ค่าของ $r_{a,b}$ มีค่าอยู่ระหว่าง –1 ถึง 1 ถ้าหากว่าค่าของ $r_{a,b}$ = 1 หมายความว่า attribute ทั้ง 2 attribute นี้มีข้อมูลที่ซ้ำซ้อนกัน

3. Data Reduction Strategies

Data Reduction Strategies เป็นวิธีในการลดขนาดข้อมูลโดยที่การลดขนาดของข้อมูลมีสิ่งที่ต้องระวังก็คือเมื่อทำการลดขนาดของข้อมูลแล้วต้องไม่ทำให้คุณสมบัติของข้อมูลที่เหลืออยู่ผิดเพี้ยนหรือเสียไป ซึ่งสาเหตุที่ต้องทำการลดขนาดของข้อมูลก็เพราะว่าบางครั้งเราอาจจะมีข้อมูลที่จะนำมาใช้งานเยอะเกินไปจนอาจจะทำให้เกิดความล่าช้าในการประมวลผลข้อมูล โดยที่เราอาจจะลดขนาดของข้อมูลที่จะนำมาใช้งานด้วยวิธีการสุ่มข้อมูลบางส่วนออกมาใช้งาน ซึ่งเราอาจจะใช้สูตรต่อไปนี้มาคำนวณจำนวนครั้งในการสุ่มข้อมูลขึ้นมาใช้งาน

$$n = \frac{N}{1+(N \cdot e^2)}$$

N = จำนวนประชากรทั้งหมด

E = error ที่ยอมรับได้

ขั้นตอนที่ 4 การแปลงรูปแบบของข้อมูล (Data Transformation)

เป็นขั้นตอนในการเปลี่ยนแปลงข้อมูลหรือรวบรวมข้อมูลให้อยู่ในรูปแบบที่เหมาะสำหรับการนำไปใช้งาน ซึ่งความเหมาะสมของข้อมูลก็ขึ้นอยู่กับโมเดลที่เราจะใช้งาน ตัวอย่างเช่น ถ้าโมเดลของเราไม่สามารถทำการคำนวณข้อมูลที่เป็นตัวอักษรได้เราก็จะต้องแปลงตัวอักษรไปเป็นตัวเลขก่อน เช่นการแปลงเกรด A, B, C, D ไปเป็นตัวเลข 1, 2, 3, 4 เพื่อให้สอดคล้องกับโมเดลที่เราจะใช้งาน

ขั้นตอนที่ 5 การทำดาต้าไม่นิ่ง (Data Mining)

โมเดลที่ใช้ในการวิเคราะห์ข้อมูลของดาต้าไมนิ่งมีหลายโมเดล ซึ่งโมเดลที่นิยมใช้ทั่วไปมีดังต่อไปนี้

- Predictive Modeling
- Database Segmentation
- Link Analysis
- Deviation Detection

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

โมเดลเหล่านี้เป็นเพียงอัลกอริทึมที่ใช้ในการวิเคราะห์ข้อมูล ดังนั้นก่อนที่เราจะเริ่มในการทำ Data Mining เราจะต้องทำการเลือกโมเดลที่จะนำมาช่วยในการวิเคราะห์ข้อมูลขึ้นมาเสียก่อน โดยที่แต่ละโมเดลมีรายละเอียดดังต่อไปนี้

1. Predictive Modeling

Predictive Modeling แบ่งออกได้เป็น 2 เทคนิคคือ Forecasting เป็นโมเดลที่ใช้การทำนายแนวโน้มของข้อมูลในอนาคต และ Classification เป็นการทำนายข้อมูลในอนาคตซึ่งสามารถแบ่งข้อมูลออกเป็นกลุ่มๆ ได้อีกด้วย

2. Database Segmentation

โมเดลนี้จะทำการแบ่งข้อมูลออกเป็นกลุ่มย่อยๆ โดยที่ข้อมูลภายในแต่ละกลุ่มจะมีลักษณะที่เหมือนกันหรือใกล้เคียงกัน ซึ่งเราจะไม่สามารถรู้มาก่อนได้เลยว่าข้อมูลที่เรามีอยู่จะสามารถแบ่งออกได้เป็นกี่กลุ่ม และเราจะต้องนำผลลัพธ์ของการแบ่งกลุ่มมาวิเคราะห์ตีความหมายอีกครั้งหนึ่ง

การทำ Database Segmentation เรียกได้อีกอย่างหนึ่งว่าการทำ Clustering โดยที่ Database Segmentation มีเทคนิคในการแบ่งกลุ่มของข้อมูลหลายวิธี แต่ที่นิยมใช้ก็คือ Partition Approach และ Neural Network Approach

- Partition Approach

Partition Approach ก็มีอีกหลายอัลกอริทึมที่สามารถเลือกนำมาใช้ในการแบ่งกลุ่มข้อมูล แต่ที่นิยมใช้ก็คือ K – means Clustering โดยที่หลักการทำงานของ K – means algorithm ก็คือจะทำการกำหนดว่าจะแบ่งข้อมูลออกเป็นกี่กลุ่ม แล้วทำการคำนวณหาจุดศูนย์กลางของข้อมูลแต่ละกลุ่ม หลังจากนั้นก็ลองนำข้อมูลไปใส่ในแต่ละกลุ่มโดยดูว่าเมื่อนำข้อมูลไปใส่ในแต่ละกลุ่มแล้วข้อมูลนี้อยู่ใกล้กับจุดศูนย์กลางของกลุ่มไหนมากที่สุดก็จะย้ายข้อมูลไปอยู่ในกลุ่มนั้น ต่อมาก็ทำการคำนวณหาจุดศูนย์กลางของกลุ่มที่เราเพิ่งย้ายข้อมูลไปใส่ ขั้นตอนต่อมาก็จะทำการลองย้ายข้อมูลตัวอื่นๆต่อไปจนกว่าจะพบเงื่อนไขที่จะหยุด ซึ่งข้อได้เปรียบของ K – means algorithm คือมีความเร็วสูงจึงเหมาะกับงานที่มี Data ที่ค่อนข้างเยอะ

- Neural Network Approach

หลักการของ Neural Network คือจะพยายามเลียนแบบโครงสร้างสมองของสิ่งมีชีวิตโดยที่โมเดลที่นิยมนำมาใช้ในการทำ Clustering ก็คือ Kohonen Neural Networks

การทำงานของ Neural Networks จะแบ่งเป็น 2 ขั้นตอนคือ

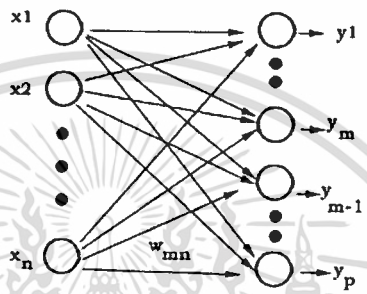
1.) Training (Learning) Mode

ก่อนที่เราจะใช้ Neural Networks เราจะต้องสอนให้มันเรียนรู้ ก่อนโดยแบ่งการเรียนรู้ออกเป็น 2 แบบ คือ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

- supervised Learning โดยวิธีนี้จะต้องสอนว่าถ้ามี input แบบนี้แล้ว output ที่ได้ควรจะเป็นอย่างไร
- Unsupervised Learning วิธีนี้จะไม่มีรูปแบบในการสอน ตัว Neural Networks จะต้องเรียนรู้ด้วยตัวเอง

2.) Deploying Mode คือการนำ Neural Networks ไปใช้งานนั่นเอง



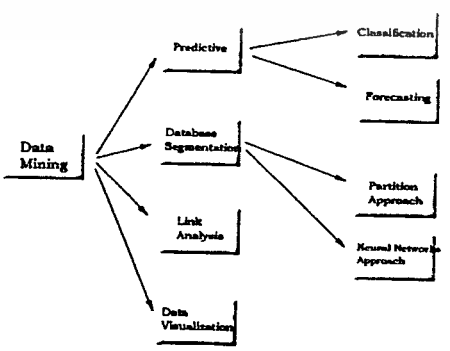
รูปที่ 2.2 สถาปัตยกรรมของ Kohonen Neural Networks

3. Link Analysis

เป็นโมเดลที่วิเคราะห์ข้อมูลที่มีความสัมพันธ์กัน ตัวอย่างเช่น การหาความสัมพันธ์ของสินค้าว่าโดยส่วนใหญ่แล้วลูกค้ามักจะซื้อสินค้าอะไรไปควบคู่กัน เพื่อที่จะนำมาวิเคราะห์และวางแผนทางด้านการตลาดต่อไป

4. Deviation Detection

เป็นโมเดลที่ใช้ในการวิเคราะห์ข้อมูลเพื่อหาสิ่งที่แตกต่างในข้อมูล ซึ่งมักจะนิยมใช้ในการตรวจจับความผิดปกติของข้อมูลเช่น การปลอมลายเซ็น



รูปที่ 2.3 แสดงเทคนิคแบบต่างๆในการทำดาต้าไมนิ่ง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ขั้นตอนที่ 6 การวิเคราะห์ผลลัพธ์ (Analysis of Result)

เมื่อผ่านขั้นตอนการทำดาต้าไมนิ่งแล้วเราจะต้องนำผลลัพธ์ที่ได้มาทำการวิเคราะห์ดูว่าเราสามารถนำผลลัพธ์ที่ได้มาใช้ประโยชน์ได้หรือไม่, อย่างไร ซึ่งถ้าหากว่าวิเคราะห์แล้วได้ผลลัพธ์เป็นที่ไม่น่าพอใจ ก็จำเป็นที่จะต้องย้อนกลับไปเริ่มทำกระบวนการในการทำดาต้าไมนิ่งใหม่ทั้งหมด ซึ่งการย้อนกลับไปเริ่มกระบวนการต่างๆใหม่ทั้งหมดถือว่าเป็นเรื่องปกติของการทำดาต้าไมนิ่ง

ขั้นตอนที่ 7 การประยุกต์ใช้งาน (Assimilation of Knowledge)

การประยุกต์ใช้งานจะนำผลลัพธ์ที่ได้มาจากการทำดาต้าไมนิ่งมาประยุกต์ใช้ให้เกิดประโยชน์กับองค์กรหรือธุรกิจต่อไป

จากกระบวนการในการทำดาต้าไมนิ่งทั้งหมดนี้ กระบวนการที่ใช้เวลาในการทำงานมากที่สุด คือ ขั้นตอนของการเตรียมข้อมูล (Data Preparation) จะใช้เวลาในการทำงานส่วนนี้ประมาณ 60% ของเวลาทั้งหมด ขั้นตอนในการเตรียมข้อมูลนี้จะประกอบด้วยการเลือกข้อมูล, การทำข้อมูลให้มีคุณภาพดี, การแปลงรูปแบบของข้อมูล ซึ่งขั้นตอนต่างๆเหล่านี้ถือว่ามีความสำคัญมากในการทำดาต้าไมนิ่งเพราะ ถ้าหากเรานำข้อมูลที่ไม่มีคุณภาพมาทำดาต้าไมนิ่งผลลัพธ์ที่ได้ก็จะไม่มีคุณภาพและไม่สามารถนำไปใช้งานได้

2.3 การนำ Database Segmentation มาประยุกต์ใช้สำหรับการตลาดในธุรกิจต่างๆ

1. การนำมาประยุกต์ใช้ในธุรกิจห้างสรรพสินค้า

เทคนิคในการทำ Database Segmentation ก็คือจะนำข้อมูลต่างๆที่เรามีอยู่มาแบ่งออกเป็นกลุ่มๆ โดยที่ข้อมูลในแต่ละกลุ่มจะมีความเหมือนหรือคล้ายเคียงกัน เราจึงนำเทคนิคนี้มาช่วยพัฒนา, วางแผนกลยุทธ์ทางด้านการตลาด

จากที่ได้กล่าวไปแล้วว่าเทคนิคของการทำ Database Segmentation ก็คือการจัดกลุ่มของข้อมูล เราจึงนำเทคนิคนี้มาประยุกต์ใช้ทางด้านการตลาดโดยการนำข้อมูลของลูกค้าที่โดยปกติมักจะเก็บรวบรวมไว้แต่ไม่ค่อยได้นำมาใช้ประโยชน์เท่าไหร่นักมาแบ่งออกเป็นกลุ่มๆ โดยที่การแบ่งกลุ่มอาจจะแบ่งตามความสนใจในสินค้าของลูกค้า เมื่อทำการแบ่งกลุ่มเรียบร้อยแล้วเราก็จะทราบว่าลูกค้ากลุ่มไหนสนใจสินค้าประเภทใด ในส่วนต่อไปก็ขึ้นอยู่กับองค์กรของเราแล้วว่าจะนำความรู้ที่ได้มาในส่วนนี้ไปประยุกต์ใช้อย่างไร ยกตัวอย่างเช่น อาจจะจัดทำรายการสินค้าบางประเภทแล้วส่งไปให้เฉพาะกลุ่มลูกค้าที่สนใจสินค้าประเภทนี้เท่านั้น ซึ่งจะเห็นได้ว่าวิธีนี้จะทำให้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เราไม่ต้องส่งรายการสินค้าของเราไปให้กับลูกค้าทุกๆราย ทำให้สามารถลดต้นทุนในส่วนนี้ลงไปได้ หรืออีกกรณีหนึ่งก็คือ เราอาจจะทำการแบ่งข้อมูลออกเป็นกลุ่มๆเพื่อศึกษาดูว่าช่วงใด(ฤดูกาลใด)สินค้าประเภทใดได้รับความนิยมสูง เพื่อที่จะได้จัดรายการส่งเสริมการขายให้เหมาะสมกับแต่ละช่วงฤดูกาล

2. การนำมาประยุกต์ใช้ในธุรกิจโรงเรียนสอนคอมพิวเตอร์

ในการนำ Database Segmentation มาช่วยทางด้านการตลาดในธุรกิจโรงเรียนสอนพิเศษนั้นก็จะเริ่มจากการนำข้อมูลในฐานข้อมูลที่ได้จัดเก็บประวัติของนักเรียนที่เคยมาเรียน แล้วนำมาทำการวิเคราะห์หาพฤติกรรมหรือความต้องการของนักเรียน โดยเราอาจจะนำข้อมูลมาวิเคราะห์จัดกลุ่มเพื่อช่วยให้ทางโรงเรียนสามารถจัดวิชาที่จะเปิดสอนได้เหมาะสมกับความต้องการของลูกค้า ยกตัวอย่างเช่น ลูกค้าที่มีวุฒิภาวะ หรือการศึกษาที่ค่อนข้างสูงอยู่แล้วมักจะต้องการเรียนวิชาที่ค่อนข้างยากเช่น การเขียนโปรแกรม และโดยส่วนใหญ่แล้วลูกค้ากลุ่มนี้มักจะอยู่ในวัยที่เป็นนักศึกษาหรือในวัยที่ทำงานแล้ว การที่จะเปิดสอนวิชานี้ถ้าหากว่าเป็นวันธรรมดาก็ควรจะเปิดสอนในช่วงหลังเลิกงาน หรือช่วงเย็นซึ่งจะทำให้ตรงกับความต้องการของลูกค้า จะเห็นได้ว่าถ้าหากจัดคอร์สที่จะสอนให้ถูกต้องตามเวลาที่เหมาะสมก็จะทำให้ได้ลูกค้าเยอะมากขึ้น หรืออาจจะแบ่งกลุ่มของลูกค้าเพื่อศึกษาดูว่าลูกค้าช่วงอายุประมาณเท่าไหร่สนใจที่จะเรียนในวิชาใดบ้าง ซึ่งการแบ่งกลุ่มลูกค้าตามช่วงอายุแล้ววิเคราะห์ดูว่าช่วงอายุนี้มักจะสนใจเรียนวิชาอะไรบ้าง ก็จะทำให้เมื่อเราจัดทำโปรโมชั่นหรือแนะนำรายละเอียดการเรียนการสอนแก่ผู้สนใจที่เข้ามาติดต่อสอบถามซึ่งจะทำให้สามารถนำเสนอในสิ่งที่ลูกค้าสนใจได้ตรงกับความต้องการ

3. การนำมาประยุกต์ใช้ในธุรกิจโรงพยาบาล หรือสถานพยาบาล

สำหรับโรงพยาบาลจะเห็นว่ามีการเก็บข้อมูลเป็นจำนวนมากเช่นข้อมูลประวัติคนไข้ ประวัติพนักงาน ยารักษาโรค และข้อมูลอื่นๆอีกเป็นจำนวนมาก ซึ่งสามารถนำมาทำ Database Segmentation เพื่อแบ่งข้อมูลออกเป็นกลุ่มๆ แล้วอาจจะนำมาวิเคราะห์เพื่อดูว่าคนไข้อายุในช่วงใดมักจะเป็นโรคอะไรบ้าง แล้วถ้าหากเป็นโรคนี้แล้วมักจะมีโรคอะไรแทรกซ้อนได้บ้าง ซึ่งถ้าหากแพทย์ทราบข้อมูลเหล่านี้แล้วก็จะทำให้สามารถตรวจสอบอาการและเตรียมรับมือกับโรคแทรกซ้อนได้ทันเวลา นอกจากนี้แล้วยังนำมาวิเคราะห์ดูว่าช่วงฤดูกาลใดผู้ป่วยมักจะเป็นโรคอะไรเป็นส่วนมาก จะได้จัดเตรียมยาและเครื่องมือต่างๆให้เพียงพอกับความต้องการของผู้ป่วยได้อย่างเหมาะสม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4. การนำมาประยุกต์ใช้ในธุรกิจด้านการเงิน การธนาคาร

สำหรับธุรกิจประเภทนี้สามารถนำ Database Segmentation มาใช้เพื่อระบุกลุ่มของลูกค้า และ สามารถระบุกลุ่มเป้าหมายได้ยกตัวอย่างเช่น เราสามารถจัดกลุ่มของลูกค้าที่มีพฤติกรรมคล้ายคลึงกัน กับการกู้ยืมไว้ด้วยกันโดยใช้เทคนิค Multidimensional Clustering ซึ่งจะทำให้สามารถระบุกลุ่มของลูกค้าและสามารถจัดกลุ่มให้กับลูกค้าที่เข้ามาใหม่ได้อย่างเหมาะสม ซึ่งจะช่วยให้เกิดความสะดวกในการวางแผนการตลาดให้ตรงกลุ่มเป้าหมาย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 3
เนื้อหาและหลักการของ CLARA อัลกอริทึม

Database Segmentation เป็นเทคนิคหนึ่งในการทำดาต้าไมนิ่งเพื่อนใช้ในการวิเคราะห์ข้อมูล โดยมีจุดประสงค์เพื่อแบ่งกลุ่มข้อมูลให้กลายเป็นกลุ่มย่อยๆ โดยที่แต่ละกลุ่มจะมีรูปแบบ ที่มีลักษณะเหมือนหรือคล้ายคลึงกัน ภายใน Database Segmentation Methods นั้นก็ยังสามารถแบ่งได้อีกหลายอัลกอริทึม สำหรับอัลกอริทึมที่จะศึกษานั้นชื่อว่า CLARA(Clustering LARge Applications)อัลกอริทึม การทำงานนั้นจะแบ่งจำนวนข้อมูลจำนวนn ออกเป็น กลุ่มๆ จำนวน k กลุ่ม แล้วกำหนดค่า k หรือจำนวนของกลุ่ม ว่าต้องการจัดกลุ่มเป็นกี่กลุ่ม แล้วนำเอาข้อมูลแต่ละตัวไปเปรียบเทียบกับจุดศูนย์กลางของกลุ่มต่างๆที่สร้างขึ้นทุกกลุ่ม จนกระทั่งได้ค่าที่เหมาะสมที่มีค่าเป็นศูนย์หรือเบนเข้าหาค่าที่ตั้งไว้

3.1 วัตถุประสงค์ของการนำ Database Segmentation มาใช้งาน

ในการนำ Database Segmentation มาช่วยทางด้านการตลาดสำหรับธุรกิจต่างๆนั้น ก็จะเริ่มจากการนำข้อมูลที่มีอยู่ในฐานข้อมูล ซึ่งได้จัดเก็บประวัติต่างๆของลูกค้าแล้วนำมาทำการวิเคราะห์หาพฤติกรรม หรือความต้องการของลูกค้า โดยเราอาจจะนำข้อมูลมาวิเคราะห์ จัดกลุ่มข้อมูลลูกค้า โดยในสัมมนาฉบับนี้จะวิเคราะห์แบ่งกลุ่มข้อมูลตามลักษณะของผู้ที่มาเรียนในวิชาต่างๆ ซึ่งจะทำให้สามารถนำเสนอในสิ่งที่ลูกค้าสนใจได้ตรงกับความต้องการมากยิ่งขึ้น

3.2 การจัดเตรียมข้อมูลเพื่อใช้ในการทำดาต้าไมนิ่ง

เนื่องจากข้อมูลต่างๆที่เก็บรวบรวมอยู่ในฐานข้อมูลนั้น บางส่วนยังเป็นข้อมูลที่ผิดพลาด หรือ ไม่สมบูรณ์ อันเนื่องมาจากหลายๆสาเหตุเช่น ผู้ที่ทำการจัดเก็บข้อมูลใส่ข้อมูลไม่ครบในบาง attribute หรือ ข้อมูลที่จัดเก็บมีความซ้ำซ้อนกัน เป็นต้น ซึ่งถ้าหากเรานำข้อมูลที่ไม่สมบูรณ์เหล่านี้ไปใช้ในการทำดาต้าไมนิ่งก็จะทำให้ผลลัพธ์ที่ได้จากการทำดาต้าไมนิ่งมีความผิดพลาด

ขั้นตอนในการจัดเตรียมข้อมูลยังสามารถแบ่งเป็นขั้นตอนย่อยๆได้ดังนี้

- การเลือกข้อมูล (Data Selection)
- การทำให้ข้อมูลมีคุณภาพดี (Data Preprocessing)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

- การแปลงรูปแบบของข้อมูล (Data Transformation)

1. การเลือกข้อมูล (Data Selection)

ขั้นตอนนี้จะเป็นการเลือกข้อมูลต่างๆจากฐานข้อมูลเพื่อที่จะนำมาใช้ในการทำดาต้าไมนิ่ง สำหรับข้อมูลที่จะนำมาใช้ในสัมมนาฉบับนี้เป็นข้อมูลจากโรงเรียนสอนคอมพิวเตอร์แห่งหนึ่ง ซึ่งมีตัวอย่างของข้อมูลที่คัดเลือกเพื่อที่จะนำมาทำดาต้าไมนิ่งดังนี้

ตารางที่ 3.1 ตารางข้อมูลห้องเรียน

ชื่อข้อมูล	ประเภทข้อมูล
Class_id	Text
เวลา รอบ	Text
วันเริ่ม	Text
วันปิด	Text
รหัสอาจารย์	Text
รหัสวิชาที่สอน	Text
รับ	Number
No	Number
Yes	Number
Detail	Text
รหัสสาขา	Text

ตารางที่ 3.2 ตารางประวัติอาจารย์

ชื่อข้อมูล	ประเภทข้อมูล
รหัสอาจารย์	Text
ชื่อ	Text
นามสกุล	Text
วัน/เดือน/ปี เกิด	Text
เลขที่บัตรประชาชน	Text
ชื่อเล่น	Text

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เพศ	Text
ที่อยู่	Text
รหัสไปรษณีย์	Text
โทรศัพท์	Text
อีเมล์	Text
คำถามส่วนตัว	Text
คำตอบ	Text
User Name	Text
Password	Text
Website	Text

ตารางที่ 3.3 ตารางสาขา

ชื่อข้อมูล	ประเภทข้อมูล
รหัสสาขา	Text
ชื่อสาขา	Text

ตารางที่ 3.4 ตารางประวัตินักเรียน

ชื่อข้อมูล	ประเภทข้อมูล
รหัสนักเรียน	Text
ชื่อ	Text
นามสกุล	Text
วัน/เดือน/ปี เกิด	Text
เลขที่บัตรประชาชน	Text
ชื่อเล่น	Text
เพศ	Text
ที่อยู่	Text
รหัสไปรษณีย์	Text
โทรศัพท์	Text

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

อีเมล์	Text
คำถามส่วนตัว	Text
คำตอบ	Text
User Name	Text
Password	Text
คะแนน	Number
อายุ	Text
Comment	Text

ตารางที่ 3.5 ตารางวิชาที่สอน

ชื่อข้อมูล	ประเภทข้อมูล
รหัสวิชา	Text
ชื่อวิชา	Text

ตารางที่ 3.6 ตารางนักเรียนที่เรียนในแต่ละห้อง

ชื่อข้อมูล	ประเภทข้อมูล
Class_id	Text
รหัสนักเรียน	Text

จากข้อมูลที่เก็บในฐานข้อมูลจะทำการเลือกข้อมูลเพียงบาง attribute ออกมาใช้งานเท่านั้น โดยที่หลักการในการเลือกข้อมูลออกมาทำงานจะดูจากสิ่งที่เราต้องการว่าต้องการจัดกลุ่มข้อมูลตามลักษณะของอะไร ยกตัวอย่างเช่น ในสัมมนาฉบับนี้จะทำการแบ่งกลุ่มข้อมูลตามลักษณะของข้อมูลทางกายภาพของนักเรียนที่มาเรียนกับทางโรงเรียน โดยมีจุดประสงค์ในการแบ่งกลุ่มข้อมูลนักเรียนเพื่อดูว่า นักเรียนช่วงอายุเท่าไหร่ และเพศอะไร มักจะสนใจเรียนในวิชาไหนบ้างและเลือกที่จะเรียนในช่วงเวลาไหนเป็นส่วนมาก ซึ่งคัดเลือก attribute ที่มีส่วนเกี่ยวข้องกับการแบ่งกลุ่มข้อมูลได้ดังนี้

ตารางที่ 3.7 แสดง attribute ที่เลือกเพื่อนำมาทำดาต้าไมนิ่ง

ชื่อข้อมูล	ประเภทข้อมูล
เพศ	Text
อายุ	Text

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ชื่อวิชา	Text
ชื่อสาขา	Text
เวลา รอบ	Text

2. การทำให้ข้อมูลมีคุณภาพดี (Data Preprocessing)

เนื่องจากข้อมูลที่เรามีอยู่ในปัจจุบันอาจจะมีบางส่วนที่ผิดพลาดหรือไม่สมบูรณ์เช่น ข้อมูลบางอย่างขาดหายไป ข้อมูลมีความซ้ำซ้อน ข้อมูลไม่สอดคล้องกัน โดยถ้าหากเรานำข้อมูลที่ไม่สมบูรณ์ไปใช้ก็จะทำให้ผลลัพธ์ที่ได้จากการทำดาต้าไมนิ่งมีความผิดพลาด เพราะฉะนั้นก่อนที่จะนำข้อมูลไปใช้งานก็จะต้องผ่านกระบวนการของ Data Preprocessing ก่อน โดยที่กระบวนการทำ Data Preprocessing จะมีการนำเทคนิคต่างๆมาช่วยจัดการในเรื่องนี้เช่น

- Data Cleaning จะเป็นการจัดการกับข้อมูลที่ขาดหายไปในบาง Field ซึ่งอาจจะแทนค่าใน Field ที่ขาดหายไปด้วย "unknown" เป็นต้น
- Data Integration เป็นวิธีการกำจัดความซ้ำซ้อนของของข้อมูล ซึ่งการเกิดข้อมูลซ้ำซ้อนกันอาจจะเกิดจากการนำฐานข้อมูลหลายๆแหล่งมารวมกัน โดยที่ฐานข้อมูลแต่ละตัวนั้นอาจจะทำการเก็บข้อมูลอย่างเดียวกันเป็นต้น
- Data Reduction Strategies เป็นวิธีในการลดขนาดข้อมูล โดยที่การลดขนาดของข้อมูลมีสิ่งที่ต้องระวังก็คือเมื่อทำการลดขนาดของข้อมูลแล้วต้องไม่ทำให้คุณสมบัติของข้อมูลที่เหลืออยู่ผิดเพี้ยนหรือเสียไป

เนื่องจากข้อมูลที่เลือกมา เพื่อนำมาทำดาต้าไมนิ่งเป็นข้อมูลที่สมบูรณ์อยู่แล้วไม่พบทั้ง Missing values, ไม่มีข้อมูลที่ซ้ำซ้อนกันเพราะเป็นข้อมูลที่นำมาจากฐานข้อมูลเพียงฐานข้อมูลเดียว และ ขนาดของข้อมูลมีไม่มากนัก จึงไม่ต้องทำการจัดการกับข้อมูลต่างๆนี้มากนัก

3. การแปลงข้อมูล (Data Transformation)

เป็นขั้นตอนในการเปลี่ยนแปลงข้อมูลหรือรวบรวมข้อมูลให้อยู่ในรูปแบบที่เหมาะสำหรับการนำไปใช้งาน ซึ่งความเหมาะสมของข้อมูลก็ขึ้นอยู่กับโมเดลที่เราจะใช้งาน เช่น การแปลงข้อมูล Text ให้กลายเป็นตัวเลขเพื่อนำไปใช้กับ K-Means Algorithm เป็นต้น

เนื่องจาก CLARA Algorithm ที่นำมาศึกษาในสัมมนาฉบับนี้จำเป็นจะต้องใส่ข้อมูลที่เป็น Numerical เท่านั้นเพราะฉะนั้นก่อนที่จะนำ attribute ต่างๆที่เราเลือกมาทำดาต้าไมนิ่งจะต้องทำการแปลงข้อมูลเหล่านี้ให้เป็น Numerical เสียก่อนโดยจะยกตัวอย่างการแปลงข้อมูลดังนี้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ห้องสมุดวิทยาเทคโนโลยีสารสนเทศ สจล.

ตารางที่ 3.8 แสดงการแปลงข้อมูล attribute เพศ

ข้อมูลเดิม	ข้อมูลใหม่
ชาย	1
หญิง	2

ตารางที่ 3.9 แสดงการแปลงข้อมูล attribute ชื่อวิชา

ข้อมูลเดิม	ข้อมูลใหม่
Web Design	1
Graphic Design	2
FLASH MX 2004	3
ASP	4
ASP.NET	5
PHP	6
JAVA	7
MAYA	8

ตารางที่ 3.10 แสดงการแปลงข้อมูล attribute ชื่อสาขา

ข้อมูลเดิม	ข้อมูลใหม่
สยามสแควร์	1
เมเจอร์ รัชโยธิน	2
ฟอร์จูนทาวน์	3

ตารางที่ 3.11 แสดงการแปลงข้อมูล attribute เวลา รอบ

ข้อมูลเดิม	ข้อมูลใหม่
เช้า	1
บ่าย	2
ค่ำ	3

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เนื่องจากข้อมูล อายุ ที่ใช้ในการทำคาต้าไมนิ่งเป็นข้อมูลประเภท Numerical อยู่แล้วจึงไม่ต้องทำการแปลงข้อมูลแต่อย่างใด

2.3 การทำงานของ **Database Segmentation**

โมเดลที่ใช้ในการวิเคราะห์ข้อมูลของคาต้าไมนิ่งมีหลายโมเดล โมเดลเหล่านี้เป็นเพียงอัลกอริทึมที่ใช้ในการวิเคราะห์ข้อมูล ดังนั้นก่อนที่เราจะเริ่มในการทำคาต้าไมนิ่งเราจะต้องทำการเลือกโมเดลที่จะนำมาช่วยในการวิเคราะห์ข้อมูลขึ้นมาเสียก่อน ในสัมมนาฉบับนี้จะกล่าวถึงโมเดล Database Segmentation โดยที่โมเดล Database Segmentation นั้นก็มีอีกหลายๆอัลกอริทึมซึ่งแต่ละอัลกอริทึมก็จะเหมาะกับลักษณะของข้อมูลแต่ละแบบแตกต่างกันไป ในสัมมนาฉบับนี้จะนำเสนอถึง CLARA Algorithm โดยที่พื้นฐานการทำงานของ CLARA Algorithm จะอาศัยการทำงานของ K-means Algorithm มาช่วยในการแบ่งกลุ่มข้อมูลและใช้ K-medoids Algorithm มาช่วยในการคำนวณหาจุดศูนย์กลางของแต่กลุ่มข้อมูล เพราะฉะนั้นในหัวข้อแรกจะกล่าวถึงอัลกอริทึมพื้นฐานของการทำ Database Segmentation ซึ่งเป็นที่รู้จักกันดีและเป็นพื้นฐานที่อัลกอริทึมอื่นๆนำไปประยุกต์ดัดแปลงก็คือ K-means Clustering หลังจากนั้นก็จะกล่าวถึง K-medoids Algorithm และ CLARA Algorithm ในหัวข้อถัดๆไปตามลำดับ

1. K–means Algorithm

จากที่กล่าวมาแล้ว K–means clustering ถือเป็นอัลกอริทึมพื้นฐานที่ใช้ในการทำคาต้าไมนิ่งโดยที่มีลักษณะการทำงานดังนี้

- K–means algorithm จะเริ่มจากการให้เราใส่ค่าพารามิเตอร์ K ลงไปในอัลกอริทึม ซึ่งอัลกอริทึมจะทำการแบ่งข้อมูลทั้งหมด จำนวน n ข้อมูลออกเป็นจำนวน K กลุ่ม โดยที่ข้อมูลที่อยู่ในแต่ละกลุ่มจะมีความคล้ายเคียงหรือเหมือนกัน
- หลังจากใส่พารามิเตอร์ K แล้ว อัลกอริทึมก็จะทำการสุ่มเลือกข้อมูลมาเป็นจำนวน K ข้อมูล เพื่อที่จะกำหนดให้ข้อมูลที่สุ่มเลือกขึ้นมานี้เป็นจุดศูนย์กลางของแต่ละกลุ่ม
- จากข้อมูลที่เหลืออยู่ทั้งหมด ข้อมูลแต่ละตัวจะถูกทำการเปรียบเทียบกับจุดศูนย์กลางของแต่ละกลุ่มว่ามีความใกล้เคียงกับกลุ่มไหนมากที่สุด ถ้าใกล้เคียงกับกลุ่มไหนมากที่สุดก็จะถือว่าเป็นข้อมูลในกลุ่มนั้น โดยที่ถ้าข้อมูลที่เราเพิ่มเข้าไปในกลุ่มไหน กลุ่มนั้นจะต้องทำการคำนวณหาจุดศูนย์กลางใหม่อีกครั้งหนึ่ง

โดยที่มีสมการที่ใช้ในการคำนวณหาจุดศูนย์กลางของข้อมูลแต่ละกลุ่มดังนี้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$$m^{(k)} = \frac{1}{n_k} \sum_{i=1}^{n_k} x_i^{(k)}$$

n_k คือ จำนวนข้อมูลทั้งหมดในกลุ่ม k

$x_i^{(k)}$ คือ ค่าของข้อมูลในกลุ่ม k

อีกสมการหนึ่งที่ใช้ในการคำนวณหาระยะห่างระหว่างจุดศูนย์กลางกับข้อมูลคือ

$$e_k^2 = \sum_{i=1}^{n_k} (x_i^{(k)} - m^{(k)})^T (x_i^{(k)} - m^{(k)})$$

$x_i^{(k)}$ คือ ค่าของข้อมูลในกลุ่ม k

$m^{(k)}$ คือ ค่าจุดศูนย์กลางของกลุ่ม k

โดยที่ประสิทธิภาพของอัลกอริทึมนี้คือ $O(tkn)$

n คือ จำนวนของ object

k คือ จำนวนของ Cluster

t คือ จำนวนรอบที่วนของขั้นตอน reallocate

จากการทำงานของ K-means Algorithm จะพบว่าถ้าหากทำการทดลองการแบ่งกลุ่มหลายๆ ครั้ง ผลลัพธ์จากการแบ่งกลุ่มสามารถมีผลลัพธ์ที่แตกต่างกันได้ถึงแม้ว่าจะใช้ข้อมูลชุดเดิมก็ตาม ซึ่งปัจจัยที่มีผลกระทบที่ทำให้ผลลัพธ์ที่ได้ออกมามีความแตกต่างกันมีหลายปัจจัยยกตัวอย่างเช่น จากการกำหนดพารามิเตอร์ K ซึ่งเป็น input ที่ใส่ลงไปในอัลกอริทึมเพื่อกำหนดว่าต้องการแบ่งข้อมูลทั้งหมดออกเป็นกี่กลุ่ม, วิธีการในการคำนวณหาความแตกต่างกันของข้อมูล และสมการที่ใช้ในการคำนวณหาจุดศูนย์กลางของแต่ละกลุ่ม เป็นต้น

ข้อจำกัดอีกอย่างหนึ่งของ K-means Algorithm ก็คือ ข้อมูลที่เราจะนำมาใส่เป็น input เพื่อที่จะทำการ Clustering ข้อมูลออกเป็นกลุ่มๆ ข้อมูลเหล่านี้จะต้องอยู่ในรูปแบบของ Numerical เท่านั้นไม่สามารถใส่ข้อมูลที่เป็น Categorical หรือ ตัวหนังสือได้เพราะฉะนั้นก่อนที่จะนำข้อมูลมาทำ Clustering จะต้องทำการแปลงข้อมูลเหล่านี้ให้อยู่ในรูปแบบของ Numerical เสียก่อน

2. K-Medoids Algorithm

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากที่กล่าวมาเกี่ยวกับการทำงานของ K-means Algorithm จะถูกรบกวนจากปัจจัยต่างๆได้ค่อนข้างง่ายไม่ว่าจะเป็น การกำหนดค่าเริ่มต้นของการแบ่งกลุ่ม การคำนวณหาจุดศูนย์กลางของแต่ละกลุ่ม ค่าของข้อมูลที่มีความแตกต่างกันมาก เป็นต้น ซึ่งเมื่อมีปัจจัยต่างๆเหล่านี้เกิดขึ้นก็จะทำให้ผลลัพธ์ในการทำงานเกิดการเปลี่ยนแปลงไปจึงได้มีการคิดค้นอัลกอริทึมใหม่ๆขึ้นโดยจะอาศัยหลักการที่คล้ายคลึงกับ K-means Algorithm

K-Medoids เป็นอีกอัลกอริทึมหนึ่งที่นำมาใช้ในการแบ่งกลุ่มข้อมูลในการทำดาต้าไมนิ่ง โดยที่หลักการทำงานพื้นฐานของ K-Medoids Algorithm จะใช้อัลกอริทึมของ K-means มาใช้ในการแบ่งกลุ่มของข้อมูล แต่จะแตกต่างจาก K-means ตรงที่การคำนวณหาจุดศูนย์กลางของข้อมูล (medoid) ซึ่งการคำนวณหาจุดศูนย์กลางของ K-medoids จะช่วยปรับปรุงคุณภาพของผมลัพธ์ของการแบ่งกลุ่มข้อมูล หลักการพื้นฐานของการหาจุดศูนย์กลางจะทำโดยการลดผลรวมของความไม่คล้ายคลึงกันระหว่างแต่ละข้อมูลกับจุดศูนย์กลาง โดยที่ K-Medoids มีลักษณะการทำงานดังนี้

- K–medoids algorithm จะเริ่มจากการให้เราใส่ค่าพารามิเตอร์ K ลงไปในอัลกอริทึม ซึ่งอัลกอริทึมจะทำการแบ่งข้อมูลทั้งหมด จำนวน n ข้อมูลออกเป็นจำนวน K กลุ่ม โดยที่ข้อมูลที่อยู่ในแต่ละกลุ่มจะมีความคล้ายเคียงหรือเหมือนกัน
- หลังจากใส่พารามิเตอร์ K แล้ว อัลกอริทึมก็จะทำการสุ่มเลือกข้อมูลมาเป็นจำนวน K ข้อมูล เพื่อที่จะกำหนดให้ข้อมูลที่สุ่มเลือกขึ้นมานี้คือจุดศูนย์กลาง(medoid)ของแต่ละกลุ่ม
- จากข้อมูลที่เหลืออยู่ทั้งหมด ข้อมูลแต่ละตัวจะถูกทำการกำหนดให้กับจุดศูนย์กลาง(medoid) ที่ใกล้ที่สุดโดยใช้ K-means อัลกอริทึมในการเลือกว่าจะให้ข้อมูลนี้ไปอยู่ในกลุ่มใด
- ทำการวัดค่าระยะห่างของข้อมูลนี้กับจุดศูนย์กลาง(medoid)ของกลุ่มนี้ โดยจะกำหนดให้เป็นค่า TDcurrent
- ทำการวัดค่าระยะห่างระหว่างแต่ละคู่ของ medoid กับ non-medoid โดยกำหนดให้เป็นค่า TD
- ทำการเลือกค่า TD ของคู่ที่ระยะห่างน้อยที่สุด
- ถ้าหากว่าค่า TD คู่ที่ระยะห่างที่น้อยที่สุด น้อยกว่าค่า TDcurrent จะทำการกำหนดให้จุดนี้เป็นจุดศูนย์กลางของกลุ่มนี้แทนแต่ถ้าหากว่ามากกว่าก็จะกำหนดให้ TDcurrent เป็นจุดศูนย์กลาง
- นำข้อมูลที่เหลืออยู่กำหนดให้กับจุด medoid ที่ใกล้ที่สุด แล้วย้อนกับไปทำขั้นตอนที่ 4 จนกว่าจะไม่เกิดการเปลี่ยนแปลงของจุดศูนย์กลาง(medoid)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากการทำงานของ อัลกอริทึมจะพบว่าการทำงานลักษณะนี้จะเกิดการวนรอบของการคำนวณเป็นจำนวนมากซึ่งผลกระทบที่ตามมาก็คือถ้าหากมีข้อมูลจำนวนมากๆ ก็จะทำให้ต้องใช้เวลาในการทำงานนานมาก จากข้อเสียที่พบจะเห็นว่าการทำงานของ K-medoids Algorithm จะเหมาะกับข้อมูลที่มีจำนวนไม่มากนัก

3. CLARA Algorithm

จากที่กล่าวไปแล้วเนื่องจาก K-medoids Algorithm เหมาะกับข้อมูลที่ไม่มากนัก เพราะฉะนั้นถ้าหากว่าต้องการทำงานกับข้อมูลจำนวนมากๆ จึงได้พัฒนา CLARA Algorithm ขึ้นมา ซึ่งการทำงานของ CLARA Algorithm จะยืนอยู่บนพื้นฐานการทำงานของ K-means Algorithm และ K-medoids Algorithm โดยที่การทำงานของ CLARA Algorithm จะเป็นการทำงานแบบ sampling-based method

การทำงานของ sampling-base method จะเป็นในลักษณะดังนี้คือ แทนที่จะนำข้อมูลทั้งหมดมาทำงาน เราจะทำการสุ่มเลือกข้อมูลเพียงแค่บางส่วน ซึ่งข้อมูลเพียงบางส่วนที่เลือกขึ้นมานี้จะทำหน้าที่เสมือนเป็นตัวแทนของข้อมูลทั้งหมด หลังจากนี้ก็จะนำข้อมูลที่สุ่มขึ้นมาไปทำงานกับ K-medoids Algorithm แต่ถ้าหากว่าเราทำการทดลองสุ่มข้อมูลขึ้นมาเพียงแค่กลุ่มเดียวอาจจะได้ผลลัพธ์ที่ไม่ถูกต้อง เพราะฉะนั้น CLARA Aglorithm จะใช้วิธีการสุ่มข้อมูลออกมาหลายๆกลุ่มแล้วก็นำข้อมูลที่สุ่มขึ้นมาแต่ละกลุ่มไปทำงานกับ K-medoids Algorithm เพื่อที่ให้ได้ผลลัพธ์ที่ดีที่สุดออกมา

จากที่กล่าวมาจะเห็นว่า CLARA สามารถจัดการกับข้อมูลที่มีจำนวนมากๆได้ดีกว่า K-medoids Algorithm ซึ่งความซับซ้อนของการทำงานจะเป็นดังสมการนี้

$$O(ks^2 + k(n - k))$$

s คือ ขนาดของข้อมูลที่สุ่มขึ้นมา

k คือ จำนวนของกลุ่มที่ต้องการแบ่ง

n คือ จำนวนของข้อมูลทั้งหมด

ประสิทธิภาพของ CLARA Algorithm จะขึ้นอยู่กับขนาดของข้อมูลที่ทำการสุ่มขึ้นมา จะสังเกตได้ว่า K-medoids Algorithm จะค้นหาการแบ่งกลุ่มที่ดีที่สุดจากข้อมูลทั้งหมด แต่ CLARA Algorithm จะค้นหาการแบ่งกลุ่มที่ดีที่สุดจากข้อมูลที่ทำการสุ่มเลือกขึ้นมา เพราะฉะนั้นจะเห็นได้ว่า CLARA Algorithm จะไม่สามารถหาการแบ่งกลุ่มที่ดีที่สุดได้ถ้าหากว่าข้อมูลที่ทำการสุ่มขึ้นมาไม่ใช่ข้อมูลที่เป็นข้อมูลที่ดีที่สุด ยกตัวอย่างเช่น ถ้าข้อมูลตัวที่ I เป็นข้อมูลที่เป็น medoid ในการ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

แบ่งกลุ่มที่ดีที่สุดของ K-medoids Algorithm แต่ในขณะที่ทำงานกับ CLARA Algorithm ข้อมูลตัวที่ 1 นี้ไม่ได้ถูกสุ่มเลือกขึ้นมาเพราะฉะนั้น CLARA จะไม่สามารถทำการแบ่งกลุ่มที่ดีที่สุดได้เลย ดังนั้นอาจจะกล่าวได้ว่า การแบ่งกลุ่มเพื่อให้ได้ผลลัพธ์ที่ดีที่สุดจะต้องขึ้นอยู่กับการสุ่มเลือกข้อมูลด้วย และการแบ่งกลุ่มที่ดีที่สุดไม่จำเป็นจะต้องทำงานกับข้อมูลทั้งหมดถ้าหากว่าทำการสุ่มเลือกข้อมูลออกมาได้เป็นอย่างดี

การทำงานของ CLARA Algorithm มีการทำงานดังต่อไปนี้

- ทำการสุ่มเลือกข้อมูลขึ้นมาเพียงบางส่วนจากทั้งหมดในฐานข้อมูล
- นำข้อมูลที่สุ่มเลือกขึ้นมาไปทำงานกับ K-medoids Algorithm
- ย้อนกลับไปทำงานขั้นตอนที่ 1 และ 2 อีกครั้งจนกว่าจะได้ผลลัพธ์ที่ดีที่สุด

เมื่อผ่านขั้นตอนการทำดาต้าไมนิ่งแล้วเราจะต้องนำผลลัพธ์ที่ได้มาทำการวิเคราะห์ดูว่าเราสามารถนำผลลัพธ์ที่ได้มาไปใช้ประโยชน์ได้หรือไม่ อย่างไร ซึ่งถ้าหากว่าวิเคราะห์แล้วได้ผลลัพธ์เป็นที่ไม่น่าพอใจ ก็จำเป็นที่จะต้องย้อนกลับไปเริ่มทำกระบวนการในการทำดาต้าไมนิ่งใหม่ทั้งหมด ซึ่งการย้อนกลับไปเริ่มกระบวนการต่างๆใหม่ทั้งหมดถือว่าเป็นเรื่องปกติของการทำดาต้าไมนิ่ง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 4
การประยุกต์ใช้ดาต้าไมนิ่งกับการจัดกลุ่มลูกค้า

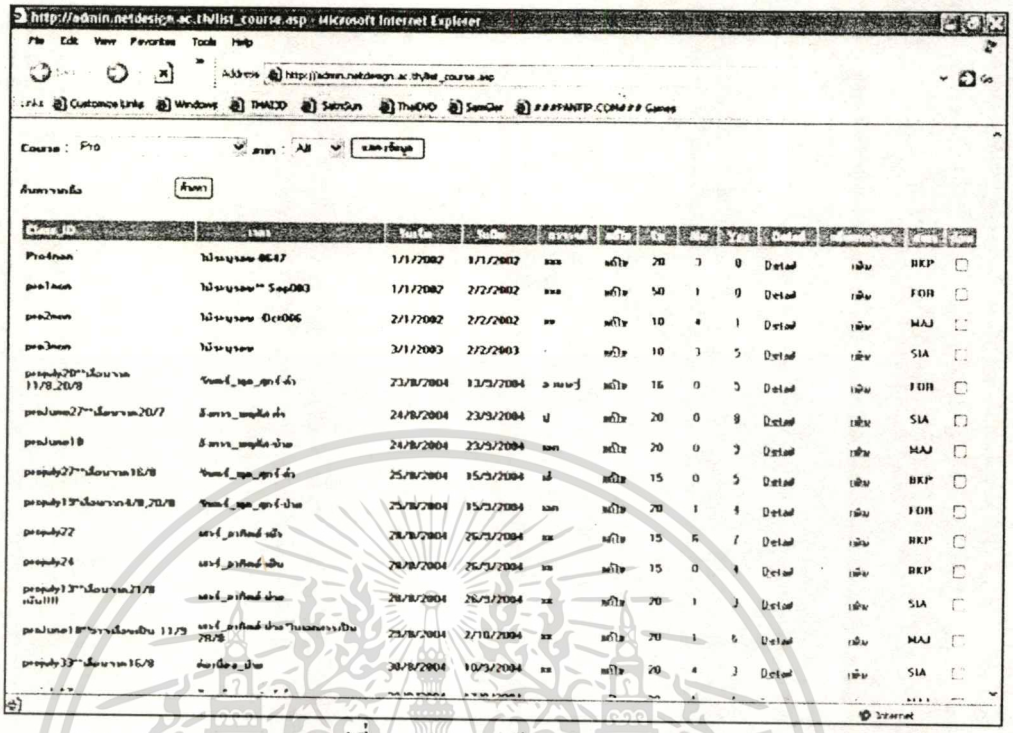
4.1 การกำหนดวัตถุประสงค์

เนื่องจากปัจจุบันข้อมูลลูกค้าได้ถูกจัดเก็บไว้ในฐานข้อมูลเป็นจำนวนมาก ซึ่งถ้าหากจัดเก็บข้อมูลเหล่านี้ไว้โดยที่ไม่นำไปใช้ประโยชน์อะไรก็จะเป็นที่น่าเสียดายอย่างยิ่ง ดังนั้นจึงได้นำเทคโนโลยีของดาต้าไมนิ่งเข้ามาช่วยในการวิเคราะห์ข้อมูล โดยมีวัตถุประสงค์เพื่อนำข้อมูลเหล่านี้มาวิเคราะห์จัดกลุ่มข้อมูลของลูกค้า โดยที่ข้อมูลที่มีความเหมือนหรือคล้ายคลึงกันจะถูกนำมาจัดกลุ่มไว้ด้วยกัน เพื่อที่จะได้นำมาพัฒนากลยุทธ์ทางธุรกิจ เพื่อให้สามารถตอบสนองได้ตรงกับความต้องการของลูกค้าได้มากยิ่งขึ้น

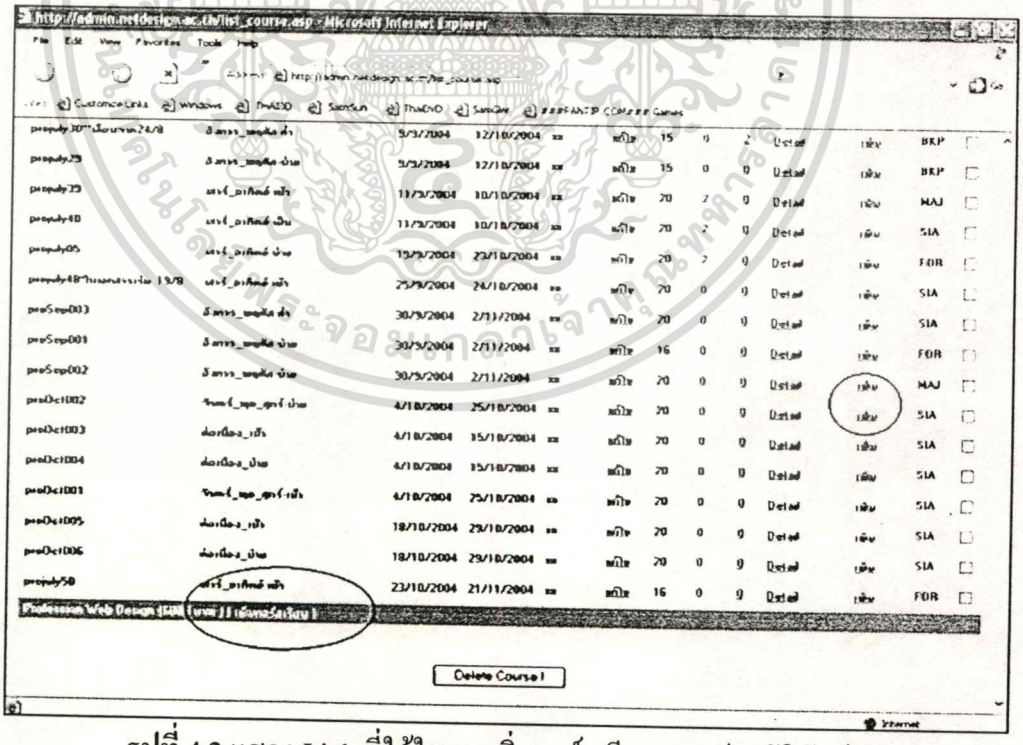
4.2 การเตรียมข้อมูลที่จะนำมาวิเคราะห์

ข้อมูลที่จะนำมาวิเคราะห์เพื่อแบ่งกลุ่มลูกค้าของโครงการนี้ ได้นำข้อมูลบางส่วนมาจากฐานข้อมูลที่ใช้เก็บข้อมูลต่างๆของสถาบันสอนคอมพิวเตอร์แห่งหนึ่ง ซึ่งลักษณะการจัดเก็บข้อมูลลูกค้าจะถูกทำเป็น web application ทุกๆสาขาจะเก็บข้อมูลลูกค้าก็จะต้องทำการ login เข้าสู่ระบบโดยผ่านหน้า web ของทางสถาบัน และทำการจัดเก็บข้อมูลต่างๆลงไปในฐานข้อมูล ดังรูปที่ 4.1,4.2 และ 4.3 ตามลำดับ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

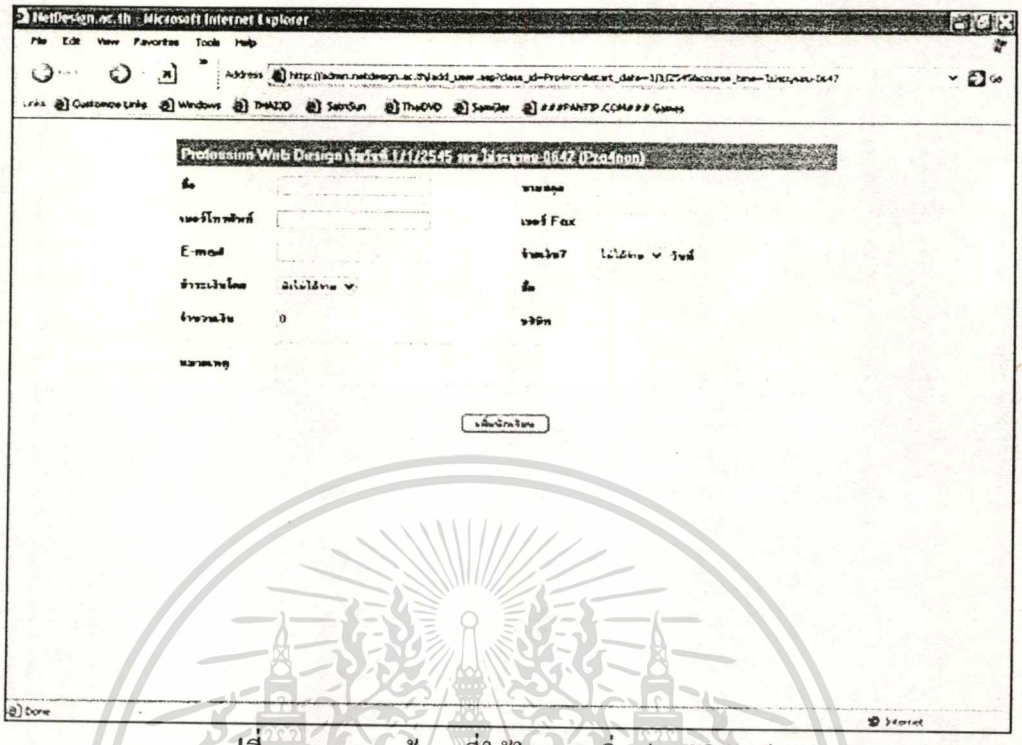


รูปที่ 4.1 แสดงหน้าหลักของ web



รูปที่ 4.2 แสดง Link ที่ใช้ในการเพิ่มคอร์สเรียน และประวัตินักเรียน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.3 แสดงหน้าจอที่ใช้ในการเพิ่มประวัตินักเรียน

จากวัตถุประสงค์ที่กำหนดไว้คือการจัดกลุ่มข้อมูลที่ความเหมือนหรือคล้ายคลึงกัน เพื่อที่จะได้เป็นแนวทางในการนำเสนอโปรโมชั่นได้ตรงความต้องการ และจะได้ทราบว่านักเรียนส่วนใหญ่มีความต้องการที่จะเรียนเวลาไหนบ้าง เนื่องจากว่าข้อมูลต่างๆที่ถูกจัดเก็บไว้มีเป็นจำนวนมาก จึงได้ทำการเลือกเฉพาะ Field ที่เกี่ยวข้องกับการวิเคราะห์มาดังนี้

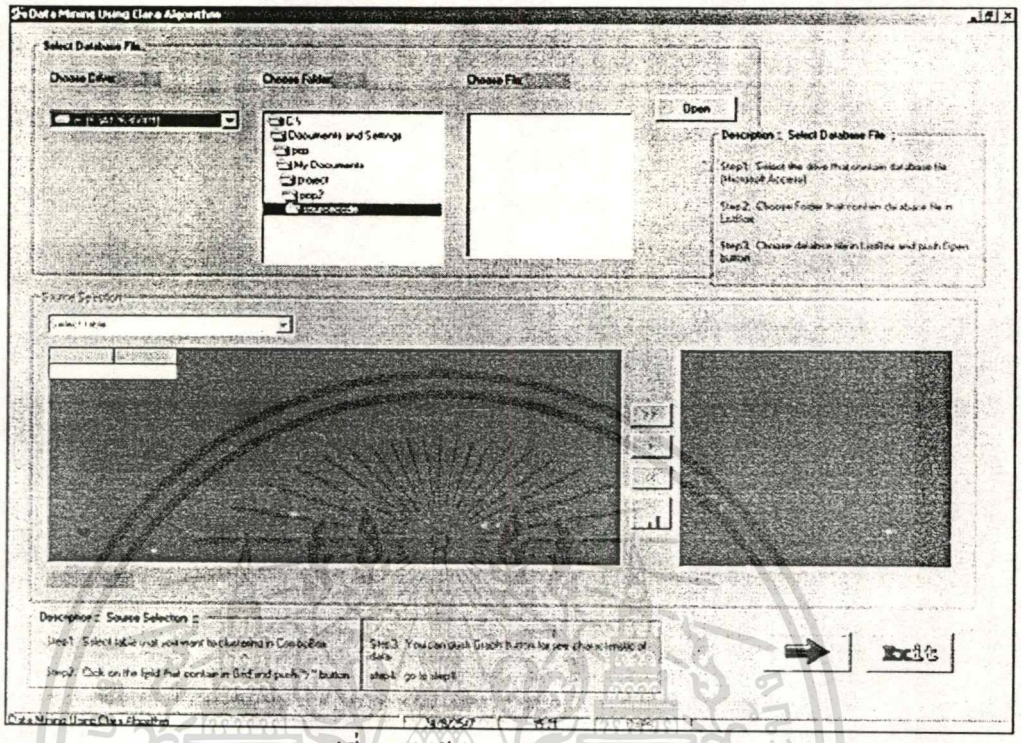
- เพศ
- อายุ
- ชื่อวิชา
- ชื่อสาขา
- เวลา/รอบ

4.3 การนำข้อมูลมาทำดาต้าไมนิ่ง

ในการทำดาต้าไมนิ่งจะใช้อัลกอริทึม CLARA Algorithm ในการแบ่งกลุ่มข้อมูลลูกค้า ซึ่งเป็นอัลกอริทึมที่มีความสามารถรองรับข้อมูลประเภท Numeric ดังนั้นจึงได้พัฒนาระบบงาน 'Clustering System Using CLARA Algorithm' โดยใช้ CLARA Algorithm เพื่อจัดกลุ่มข้อมูล แล้วนำมาวางแผนกลยุทธ์ทางธุรกิจต่อไป โดยมีลักษณะการทำงานของระบบดังนี้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เมื่อเข้าสู่โปรแกรมจะปรากฏหน้าจอดังรูปที่ 4.4



รูปที่ 4.4 หน้าจอหลักของโปรแกรม

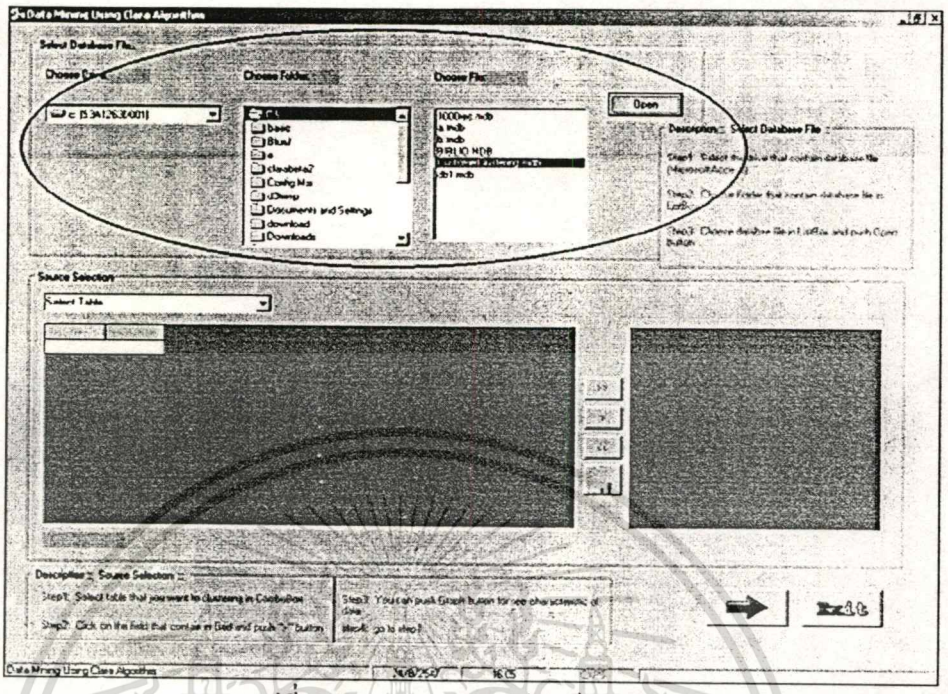
หน้าจอหลักเป็นหน้าจอที่ใช้ในการเลือกฐานข้อมูลและเลือก Field ในตารางที่ต้องการนำมาทำดาต้าไมนิ่ง โดยจะแบ่งออกเป็น 2 ส่วนคือ

- Select Database File ใช้ในการเลือกไฟล์ฐานข้อมูล
- Source Selection ใช้เลือก Field ในตาราง

4.3.1 การเลือกข้อมูลที่ต้องการนำมาวิเคราะห์

- เลือก Drive ที่จัดเก็บไฟล์ฐานข้อมูลจาก ComboBox ในส่วนของ Select Database File ดังรูปที่ 4.5
- เลือก Folder ที่จัดเก็บไฟล์ฐานข้อมูลจาก ListBox ในส่วนของ Select Database File
- เลือกไฟล์ฐานข้อมูล(*..mdb) ที่ต้องการนำมาทำดาต้าไมนิ่งจาก ListBox ในส่วนของ Select Database File
- กดปุ่ม Open เพื่อเปิดไฟล์ฐานข้อมูล

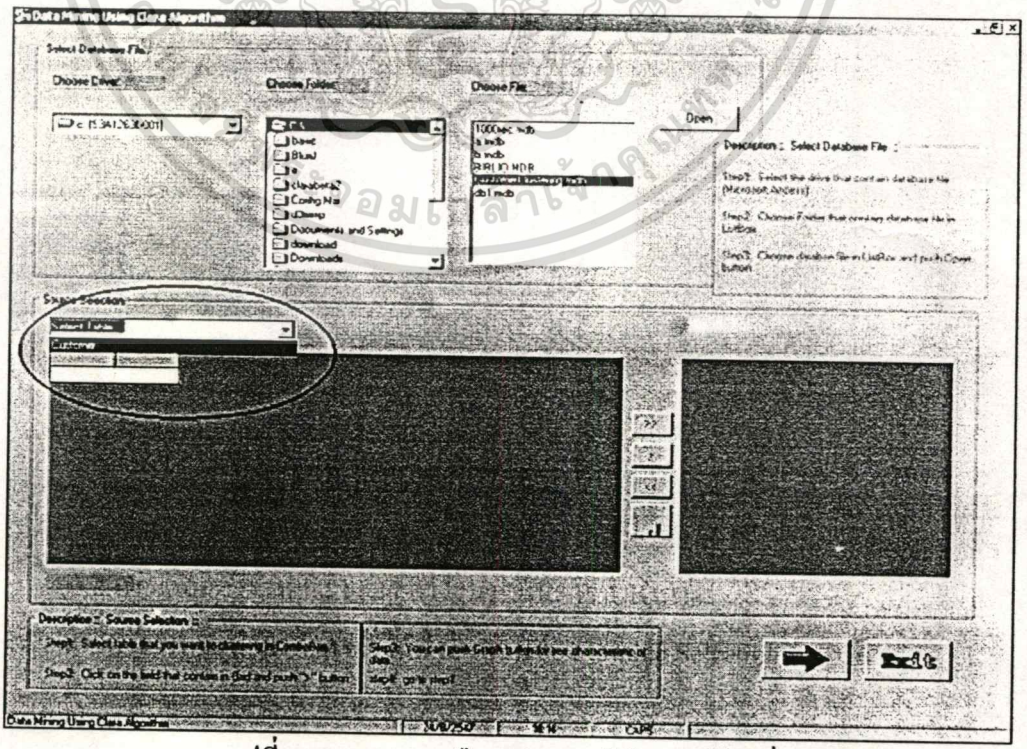
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.5 แสดงการติดต่อไฟล์ฐานข้อมูล

4.3.2 การเลือกตารางและ Field ต่างๆมาใช้ในการทำดาต้าไมนิ่ง

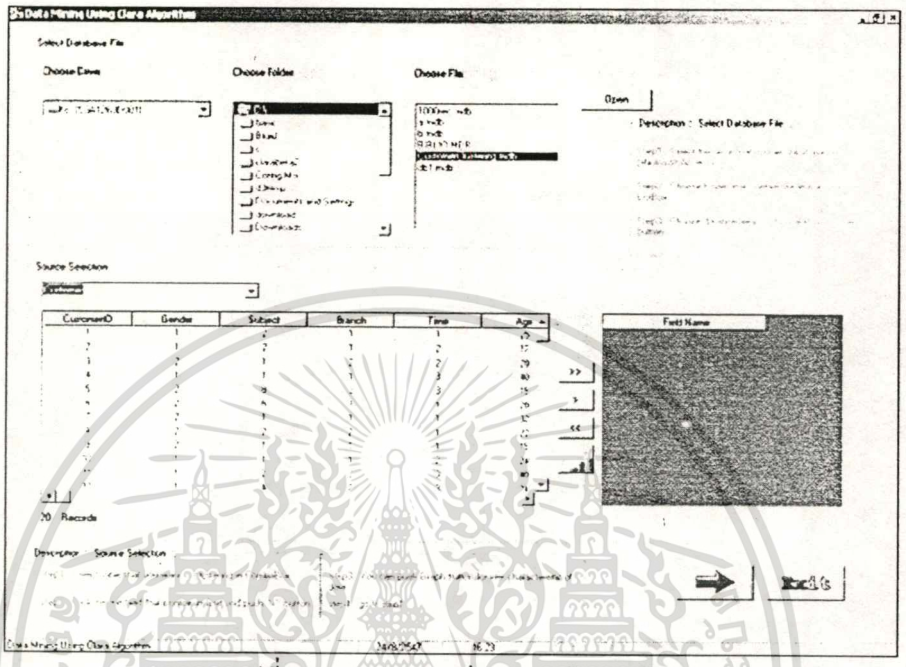
เลือกตารางที่จะนำมาทำดาต้าไมนิ่งในส่วนของ Source Selection ดังรูปที่ 4.6



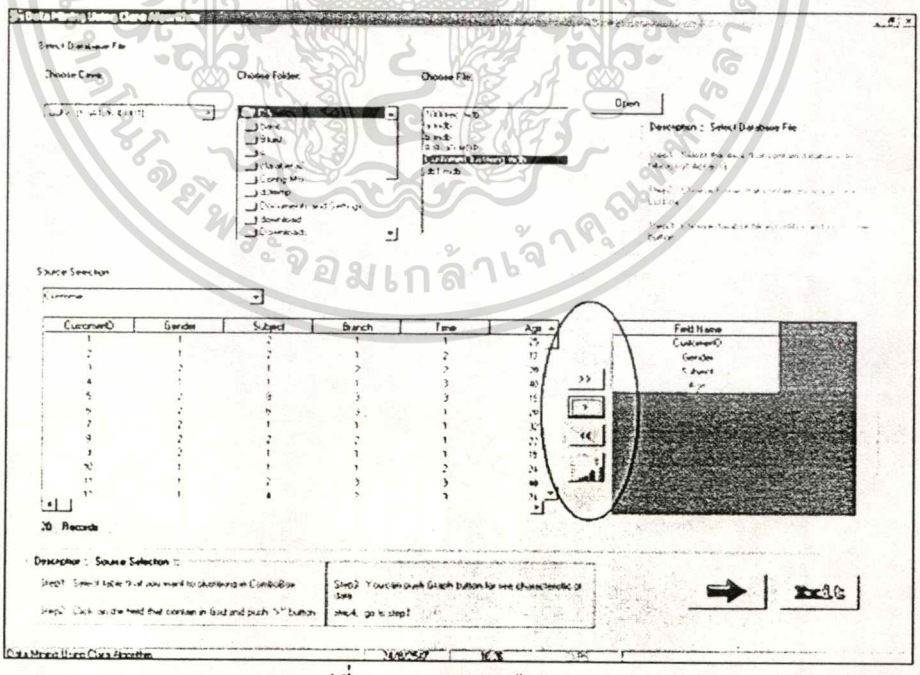
รูปที่ 4.6 แสดงการเลือกตารางมาใช้ทำดาต้าไมนิ่ง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

หลังจากเลือกตารางแล้ว โปรแกรมจะทำการแสดงข้อมูลที่ถูกจัดเก็บอยู่ภายในตารางขึ้นมา ดังรูปที่ 4.7



รูปที่ 4.7 แสดงผลลัพธ์จากการเลือกตาราง

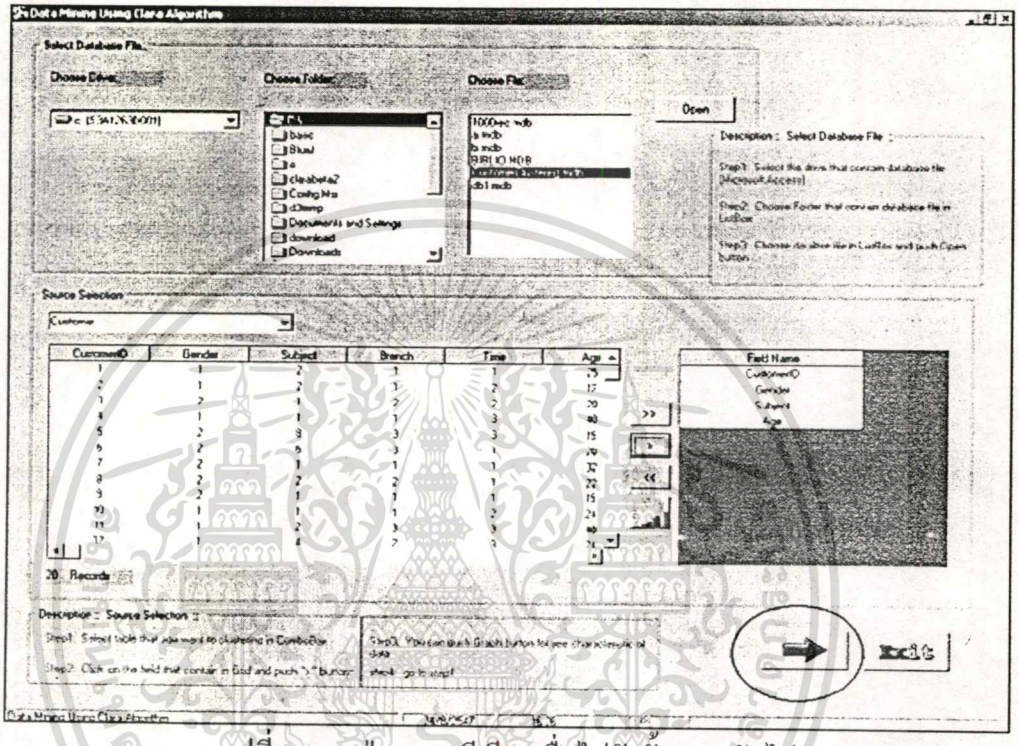


รูปที่ 4.8 แสดงการเลือก Field

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากนั้นให้ทำการเลือก Field ที่ต้องการทำดาต้าไมนิ่งจาก โดยทำการคลิ๊กที่ Field ที่เราต้องการ แล้วกดปุ่ม '>' ดังรูปที่ 4.8

เมื่อเลือก Field เสร็จแล้วให้กดปุ่มลูกศรสีเขียว ดังรูปที่ 4.9

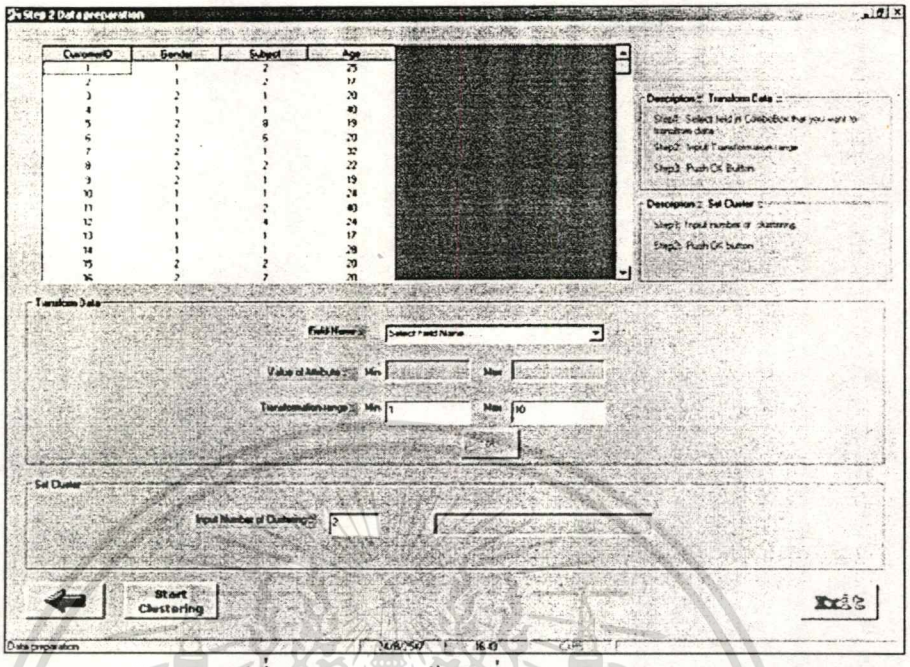


รูปที่ 4.9 กดปุ่มลูกศรสีเขียวเพื่อไปยังขั้นตอนถัดไป

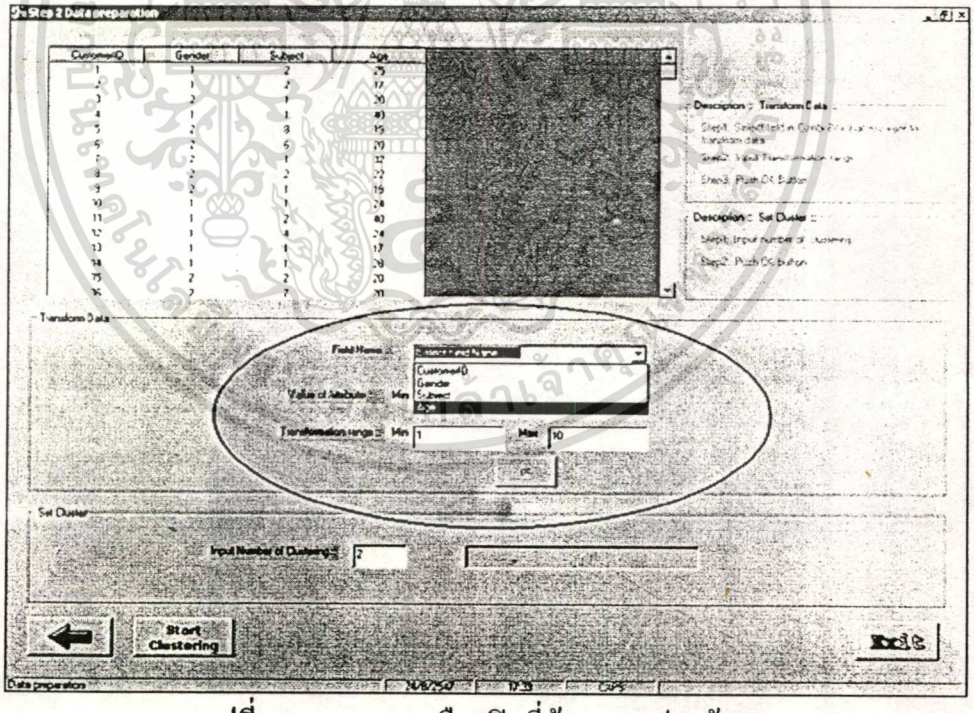
4.3.3 การแปลงข้อมูล

หน้าจอที่ 2 เป็นการแปลงข้อมูลให้อยู่ในขอบเขตที่ต้องการ เมื่อมายังหน้าจอที่ 2 โปรแกรมจะแสดงข้อมูลที่เราทำการเลือกจากข้อมูลที่ 1 ขึ้นมา ดังรูปที่ 4.10

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.10 แสดงหน้าจอที่ 2 ของโปรแกรม

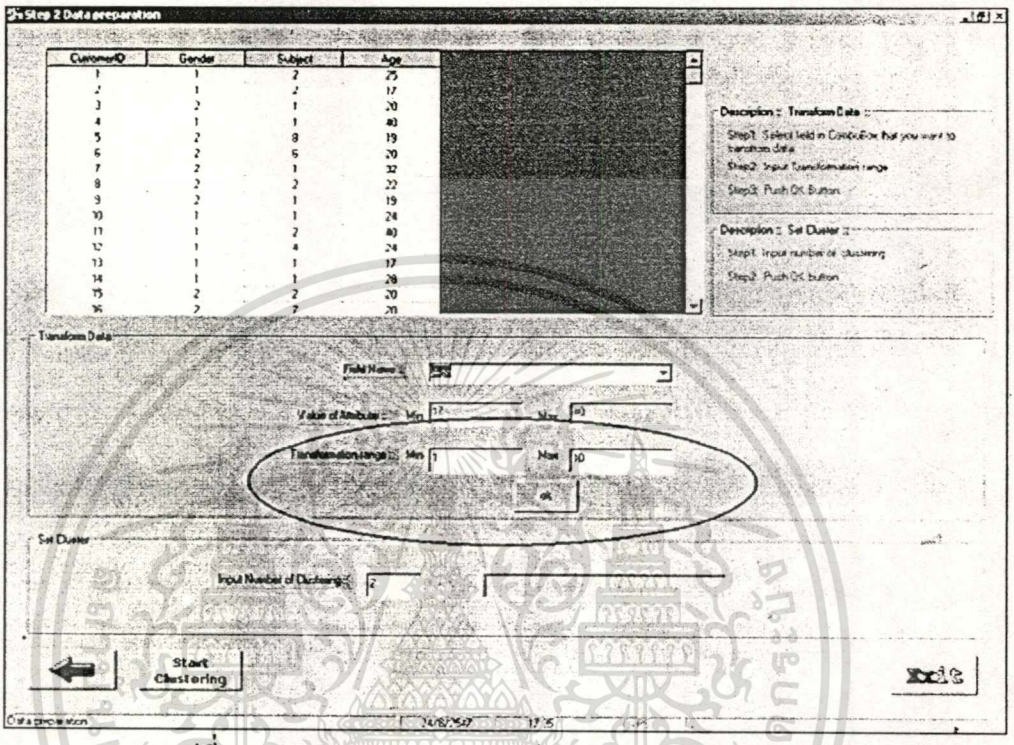


รูปที่ 4.11 แสดงการเลือกฟิลที่ต้องการแปลงข้อมูล

เมื่อเราเลือก Field ที่ต้องการแปลงข้อมูลจาก ComboBox โปรแกรมจะทำการแสดงค่า Minimum และ Maximum ของข้อมูล ดังรูปที่ 4.11

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

หลังจากที่เลือก Field ที่ต้องการแปลงข้อมูลแล้ว ให้เราทำการระบุ ช่วงของตัวเลขที่เราต้องการลงไปในช่อง Transformation range ดังรูปที่ 4.12

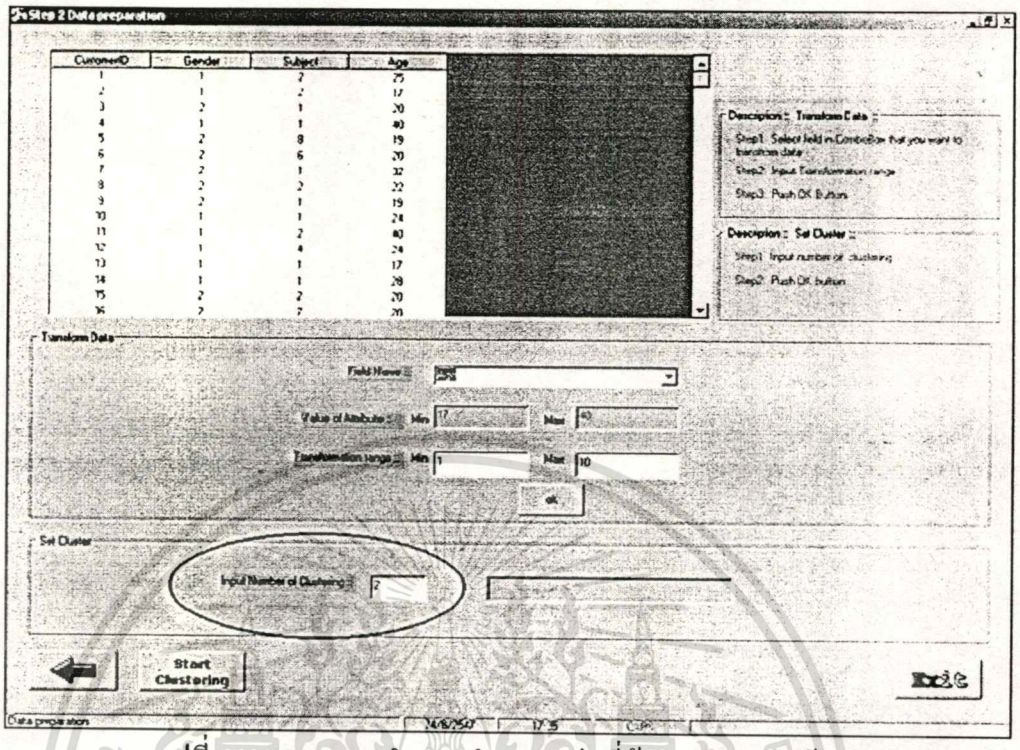


รูปที่ 4.12 แสดงการกำหนดค่าลงไปในช่อง Transform range

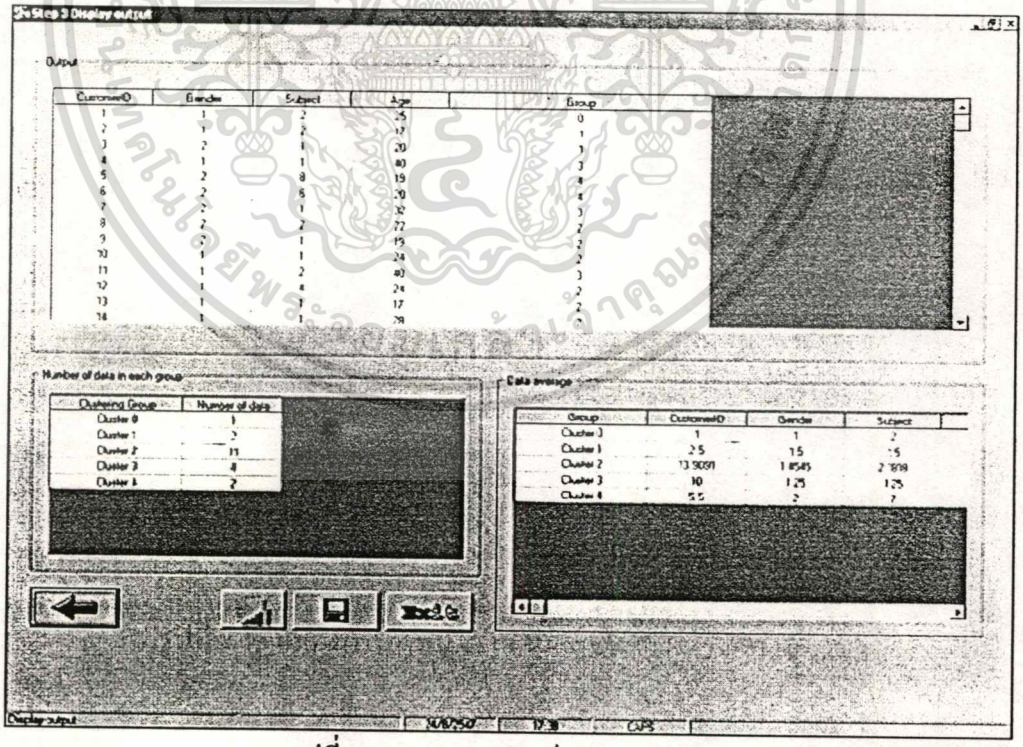
4.3.4 การกำหนดจำนวนกลุ่มข้อมูล

- ให้ทำการระบุจำนวนกลุ่ม ว่าต้องการแบ่งข้อมูลออกเป็นจำนวนกี่กลุ่ม
- จากนั้นกดปุ่ม Start Clustering เพื่อให้โปรแกรมเริ่มทำการแบ่งหลุ่มข้อมูล ดังรูปที่ 4.13

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.13 แสดงการกำหนดจำนวนกลุ่มที่ต้องการจัดกลุ่มข้อมูล



รูปที่ 4.14 แสดงผลลัพธ์จากการคำนวณ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.3.5 การแสดงผล

หลังจากที่กดปุ่ม Start Clustering แล้วระบบจะทำการคำนวณ และแสดงผลลัพธ์ของการคำนวณออกมาแสดงที่หน้าจอที่ 3 ดังรูปที่ 4.14

การแสดงผลลัพธ์จะแบ่งออกเป็น 3 ส่วนคือ

1. ในส่วนของ Output จะทำการแสดงข้อมูลใน Field ต่างๆที่เราเลือก และแสดงผลลัพธ์ออกมาว่าข้อมูลใดถูกแบ่งอยู่ในกลุ่มไหนบ้าง
2. ในส่วนของ Number of data in each group จะแสดงว่าในแต่ละกลุ่มประกอบด้วยข้อมูลจำนวนทั้งหมดกี่ตัว
3. ในส่วนของ Data Average จะแสดงว่าข้อมูลในแต่ละกลุ่มมีค่าเฉลี่ยเท่าไหร่

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 5
สรุปผลการศึกษาและข้อเสนอแนะ

5.1 สรุปผลการศึกษา

จากที่ได้ศึกษาการทำงานและทฤษฎีของอัลกอริทึมต่างๆในการทำดาต้าไมนิ่ง ทำให้ทราบว่าดาต้าไมนิ่งเป็นกระบวนการที่ใช้วิเคราะห์และค้นหาความรู้จากฐานข้อมูล ซึ่งฐานข้อมูลเหล่านี้มีการจัดเก็บข้อมูลอยู่เป็นจำนวนมาก จากกระบวนการดาต้าไมนิ่งทำให้เราสามารถนำความรู้ที่ค้นพบไปพัฒนา วางแผน และประยุกต์ใช้ในธุรกิจต่างๆได้เป็นอย่างดี เมื่อได้ทำการศึกษาหลักการทำงานต่างๆแล้ว จึงได้มีการพัฒนาโครงการโดยนำเสนอถึงการแบ่งกลุ่มข้อมูล โดยใช้ CLARA Algorithm ซึ่งเป็นอัลกอริทึมหนึ่งของการทำ Database Segmentation ที่สามารถนำข้อมูลเป็นที่เป็นตัวเลข(Numeric)มาวิเคราะห์ได้ ดังนั้นจึงได้วางแนวทางในการพัฒนาระบบด้วยโปรแกรม Microsoft Visual Basic 6.0 และใช้งานร่วมกับฐานข้อมูล Microsoft Access โดยมีจุดประสงค์ที่ต้องการศึกษาคือ

1. เพื่อจัดกลุ่มข้อมูลที่มีความเหมือนหรือความคล้ายคลึงกัน
2. นำผลลัพธ์มาวิเคราะห์หาความต้องการของลูกค้า เพื่อที่จะทำการเสนอโปรโมชั่นต่างๆได้ตรงกับความต้องการของลูกค้า

เมื่อได้ทำการพัฒนาระบบและศึกษาถึงผลลัพธ์ที่ได้นั้นพบว่า CLARA Algorithm เหมาะกับข้อมูลจำนวนไม่มากนัก เนื่องมาจากตัว CLARA Algorithm ได้ทำการพัฒนามาจาก K-means Algorithm ทำให้ CLARA Algorithm มีความซับซ้อนมากขึ้น จึงไม่เหมาะกับข้อมูลจำนวนมาก สิ่งที่สังเกตเห็นได้อีกอย่างหนึ่งก็คือ ผลลัพธ์จากการแบ่งกลุ่มพบว่าข้อมูลในแต่ละกลุ่มมีความใกล้เคียงกันมากยิ่งขึ้นเมื่อเปรียบเทียบกับ K-means Algorithm

5.2 ประโยชน์ที่ได้จากการศึกษาและพัฒนาระบบ

1. ทำให้เข้าใจทฤษฎี และหลักการทำงานของอัลกอริทึมต่างๆที่ใช้ในการทำดาต้าไมนิ่งได้มากยิ่งขึ้น
2. ทำให้ทราบถึงปัญหาต่างๆที่มักจะเกิดขึ้นในขั้นตอนต่างๆของการทำดาต้าไมนิ่ง
3. ทำให้ได้โปรแกรมที่ใช้ในการวิเคราะห์ข้อมูล ซึ่งสามารถนำไปใช้เป็นต้นแบบ ในการพัฒนาโปรแกรมในลักษณะอื่นๆได

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

5.3 ข้อเสนอแนะ

1. ระบบที่พัฒนาขึ้นมานี้สามารถนำไปประยุกต์ใช้กับฐานข้อมูลอื่นๆ ที่จัดเก็บในฐานข้อมูล Microsoft Access ได้

2. เพื่อให้ได้ผลลัพธ์จากการวิเคราะห์ข้อมูลที่มีความถูกต้องแม่นยำ ดังนั้นก่อนนำข้อมูลที่ต้องการจะทำดาต้าไมนิ่งมาใช้งานจึงควร แปลงข้อมูลต่างๆให้อยู่ในรูปแบบ Numeric ก่อน และทำความสะอาดข้อมูลโดยการ ตัด Field ที่มีข้อมูลไม่ครบถ้วนและไม่จำเป็นทิ้งไป เนื่องจากกรณีศึกษานี้ได้ทำการทำความสะอาดข้อมูลเรียบร้อยตั้งแต่ก่อนนำมาใช้งานแล้ว จึงไม่มีส่วนของการทำความสะอาดข้อมูลรวมอยู่ด้วย หากนำข้อมูลที่มีข้อบกพร่องต่างๆมาใช้งานอาจทำให้โปรแกรมคำนวณผิดพลาด และผลลัพธ์ที่ได้อาจมีความผิดพลาดและคลาดเคลื่อนได้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บรรณานุกรม

ศุภชัย สมพานิช. 2545. **สร้างระบบงานฐานข้อมูลด้วย Visual Basic ฉบับโปรแกรมเมอร์.** นนทบุรี: อินโฟเพรส

DATA MINING. [Online]. Available: http://www.cs.rpi.edu/~zaki/dmcourse/notes/10-30-03/10-30-03.doc

DATA MINING. [Online]. Available: http://www.internetweek.com/case/study071999-2.htm

Jiawei Han and Micheline Kamber. 2001. **Data Mining : Concepts and Techniqes.** :AcademicPress

Peter Cabena et al. 1997. **Discovering Data Mining From Concept to Implementation.** New Jersey: Prentice Hall.

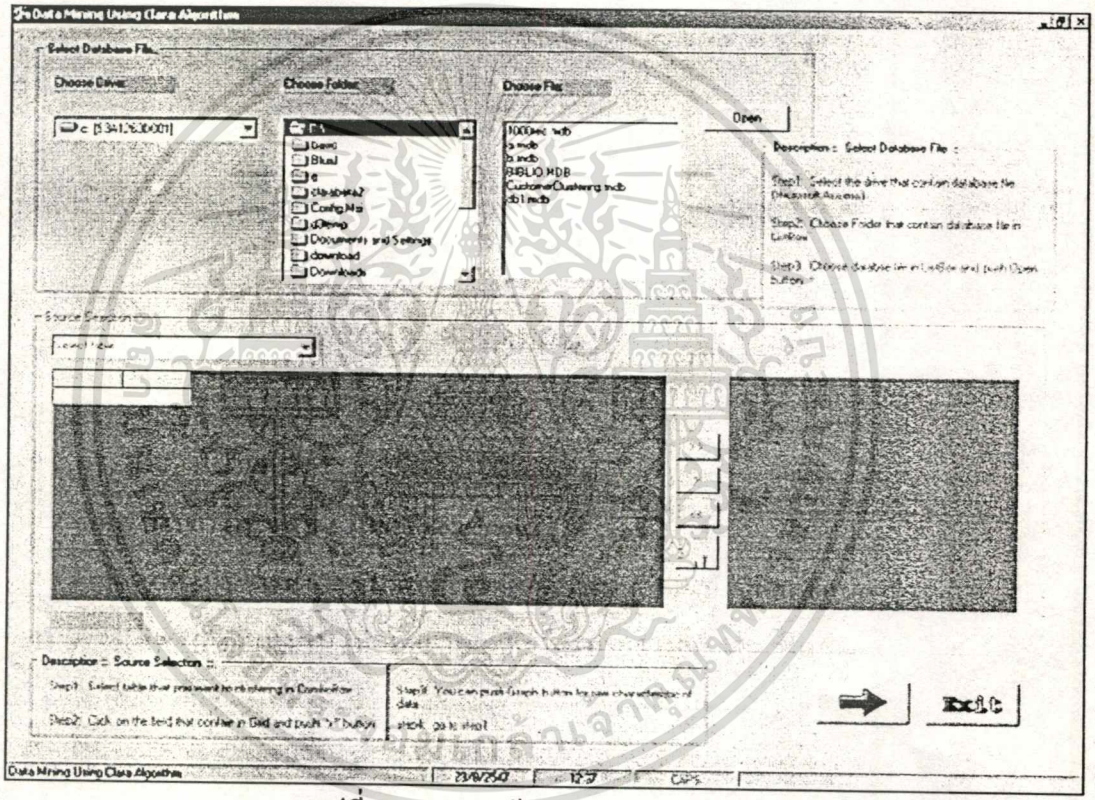
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ภาคผนวก

ก. การทำงานของโปรแกรม

ก.1 การทำงานขั้นที่ 1 (Data Selection)

เมื่อเข้าสู่โปรแกรม ขั้นตอนแรกที่เราต้องทำก็คือทำการเลือกข้อมูล เพื่อนำมาทำการวิเคราะห์โดยใช้ดาต้าไมนิ่ง ในขั้นตอนแรกเมื่อเข้าสู่โปรแกรมจะได้ผลลัพธ์ดังรูปที่ ก.1



รูปที่ ก.1 แสดงหน้าจอหลักของโปรแกรม

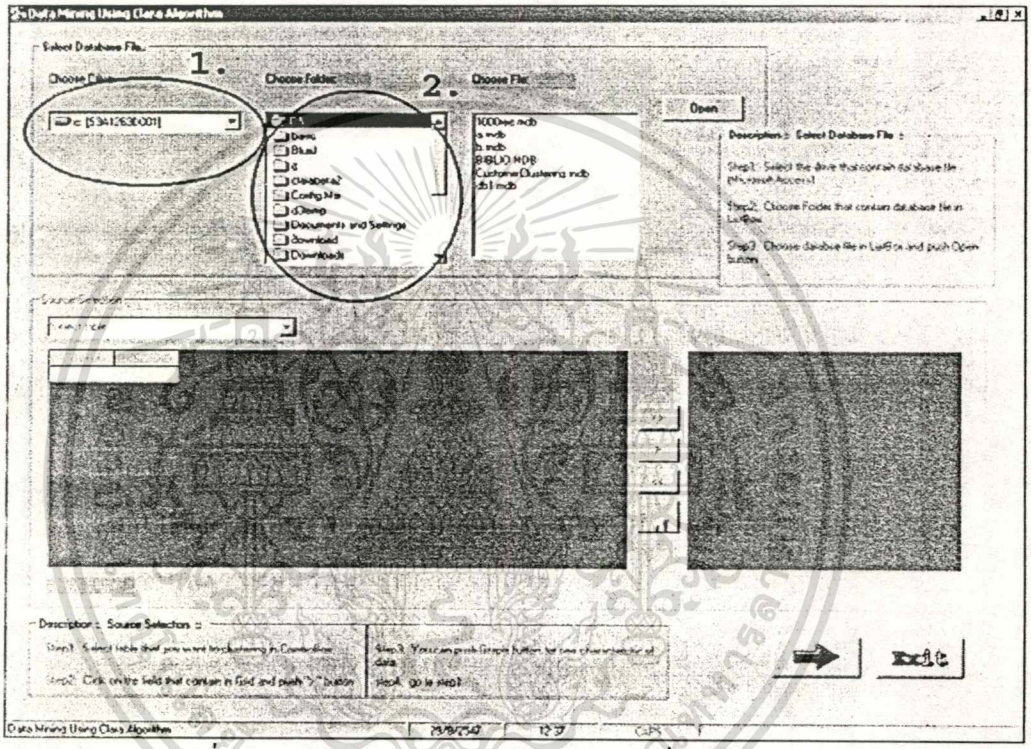
ในหน้าจอแรกของโปรแกรมจะแบ่งออกเป็น 2 ส่วนคือ

- Select Database File ในส่วนนี้ใช้ในการเลือกไฟล์ฐานข้อมูลที่ต้องการนำมาทำการวิเคราะห์
- Source Selection ใช้ในการเลือก Field ในฐานข้อมูล ที่ต้องการนำมาทำดาต้าไมนิ่ง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

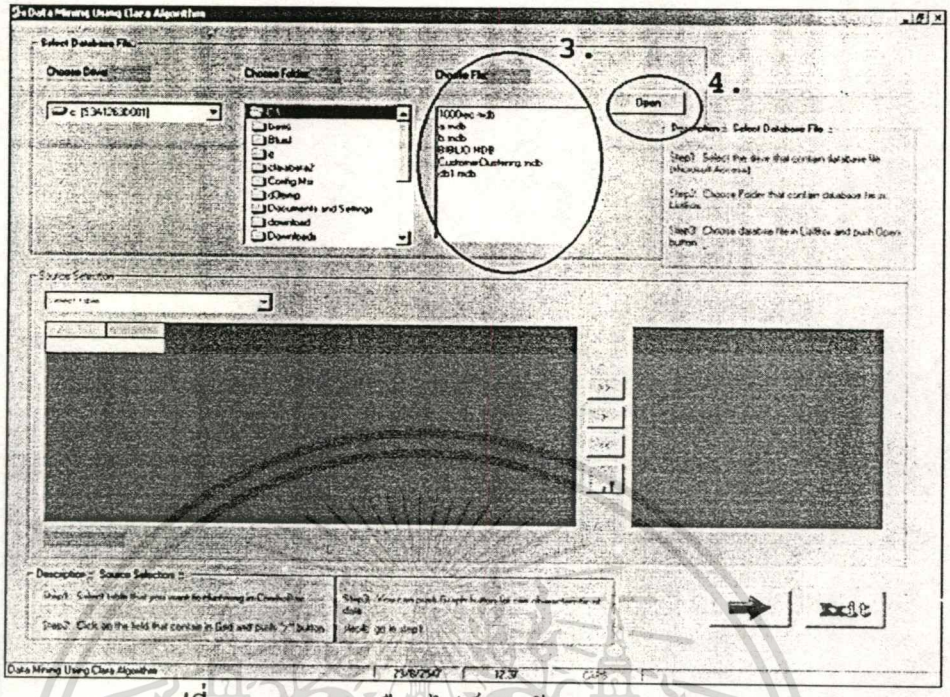
ก.1.1 การติดต่อกับฐานข้อมูล

1. ทำการเลือก Drive ที่เก็บไฟล์ฐานข้อมูล จาก ComboBox ที่ช่อง Choose Drive
2. เลือก Folder ที่เก็บไฟล์ฐานข้อมูลจาก ListBox ที่ช่อง Choose Folder
3. เลือกไฟล์ฐานข้อมูลที่ช่อง Choose File
4. กดปุ่ม Open เพื่อทำการเปิดฐานข้อมูลที่เราทำการเลือกขึ้นมาใช้งาน ดังรูปที่ ก.2



รูปที่ ก.2 แสดงการเลือก Drive และ Folder ที่จัดเก็บฐานข้อมูล

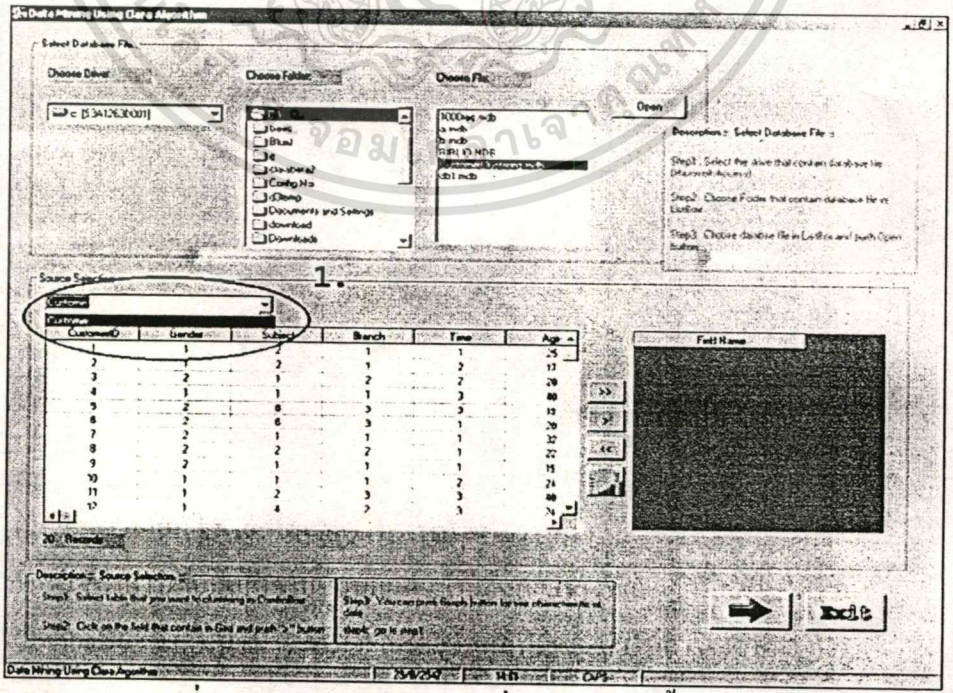
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ ก.3 แสดงการเลือกไฟล์ฐานข้อมูล และกดปุ่ม Open

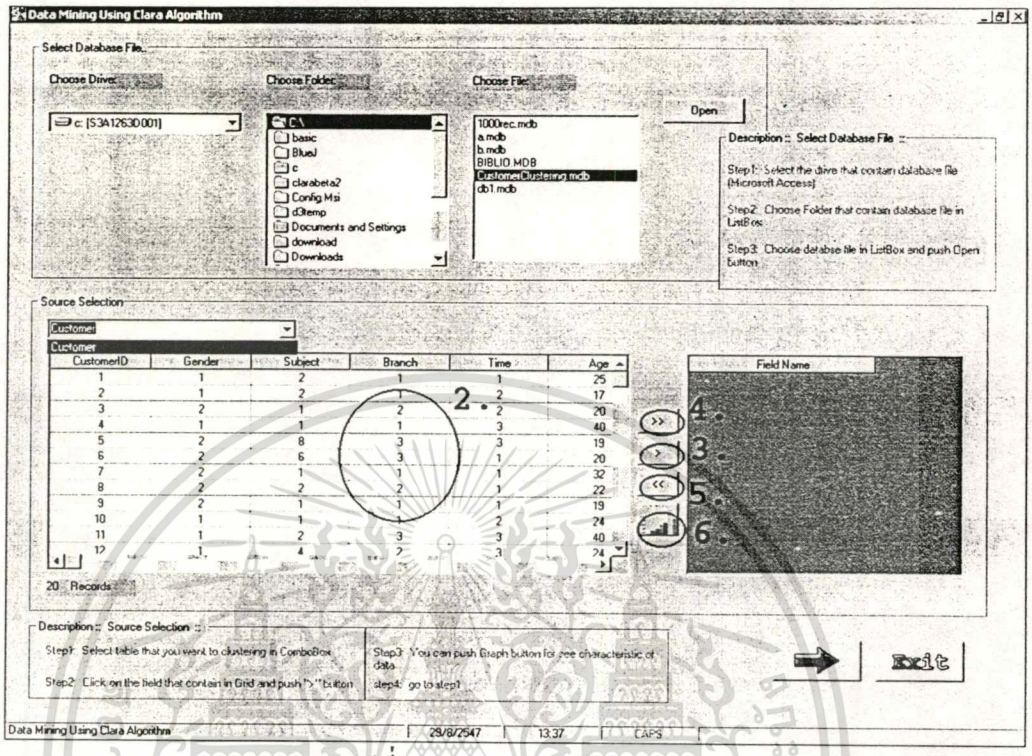
ก.1.2 การเลือก Field ที่ต้องการนำมาทำดาต้าไมนิ่ง

1. เมื่อทำการเลือกเปิดฐานข้อมูลแล้ว ให้เลือกตารางที่เก็บข้อมูลภายในฐานข้อมูลจาก ComboBox จากนั้นโปรแกรมจะแสดงข้อมูลที่มีอยู่ภายในตารางขึ้นมา ดังรูปที่ ก.4



รูปที่ ก.4 แสดงการเลือกตารางที่จัดเก็บข้อมูลขึ้นมาใช้งาน

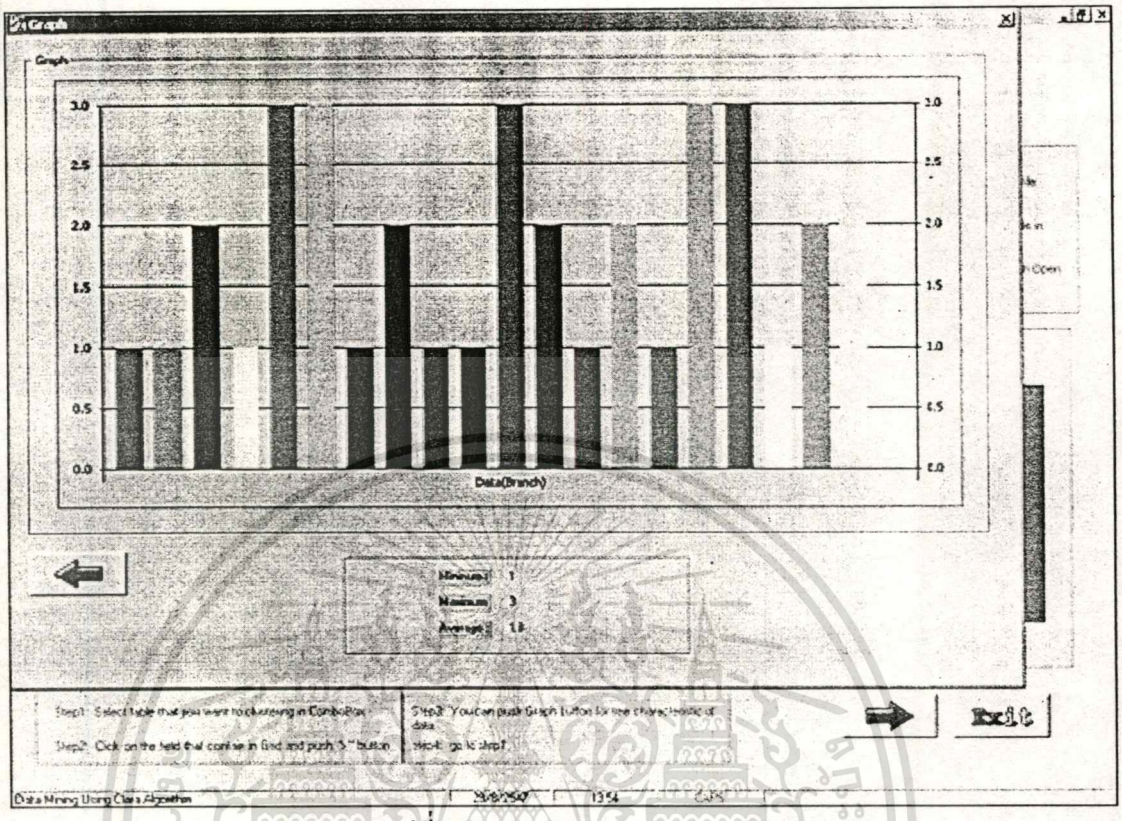
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ ก.5 แสดงการเลือก Field

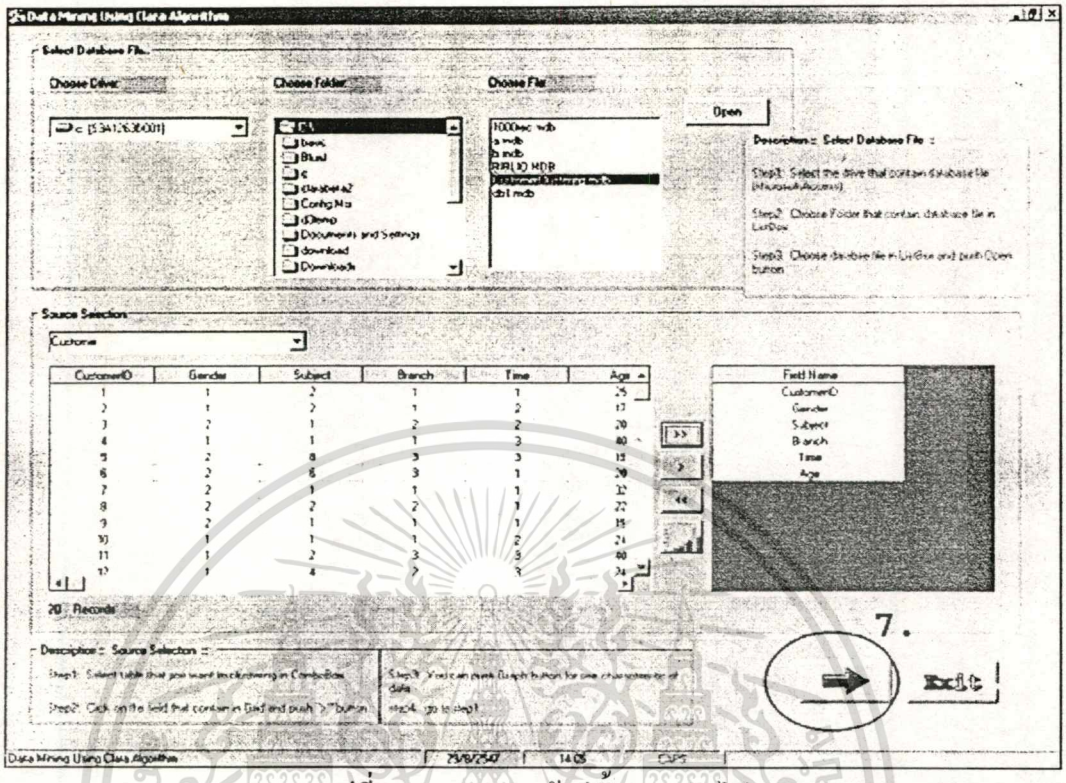
2. ให้เราทำการคลิ๊กที่ Field ที่เราต้องการจะเลือก ดังรูปที่ ก.5
3. จากนั้นกดปุ่ม '>' เพื่อทำการเลือก Field ที่เราได้ทำการคลิ๊กไว้ในตอนแรก
4. ถ้าหากต้องการเลือกทุกๆ Field ให้ทำการกดปุ่ม '>>'
5. ถ้าหากต้องการยกเลิก Field ที่เลือกไว้แล้ว ให้กดปุ่ม '<<'
6. สามารถทำการ plot กราฟ เพื่อดูรูปแบบของข้อมูลโดยการกดปุ่ม Graph ซึ่งจะได้ผลลัพธ์ของการ plot กราฟ ดังรูปที่ ก.6
7. เมื่อทำการเลือก Field ข้อมูลที่ต้องการนำมาทำดาต้าไมนิ่งเสร็จเรียบร้อยแล้ว ให้เราทำการกดปุ่มลูกศรชี้ขวาสีเขียว เพื่อไปยังขั้นตอนถัดไป ดังรูปที่ ก.7

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ ก.6 แสดงผลการ plot กราฟ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



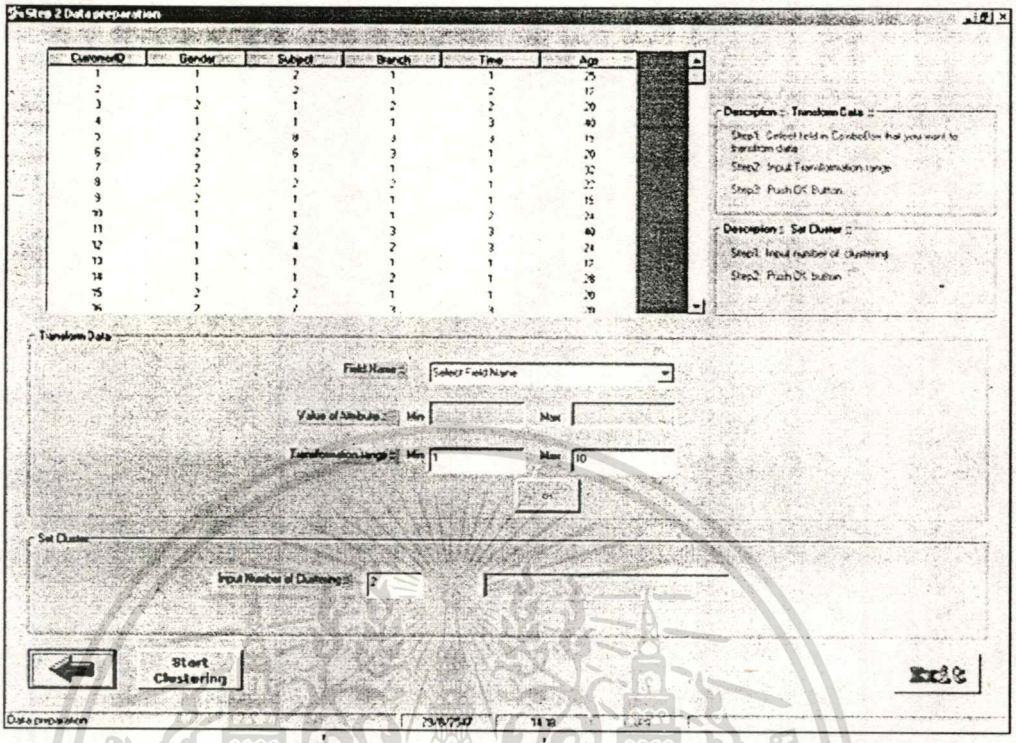
รูปที่ ก.7 แสดงการเข้าสู่ขั้นตอนถัดไป

ก.2 การทำงานขั้นที่ 2 (Data Transform)

ในขั้นตอนที่ 2 จะเป็นการ Transform ข้อมูล ให้อยู่ในช่วงที่เราต้องการ สาเหตุที่ต้องทำการ Transform ข้อมูลก็เพราะว่า ถ้าหากเรานำข้อมูลดิบๆที่จัดเก็บในฐานข้อมูลจริงๆมาใช้งานนั้น ตัวข้อมูลอาจจะทำการจัดเก็บอย่างไม่เป็นระเบียบ และอาจจะมีช่วงของข้อมูลที่ค่อนข้างกว้างมาก ยกตัวอย่างเช่น Field Salary ซึ่งจัดเก็บเงินเดือนของพนักงาน ซึ่งข้อมูลใน field นี้อาจจะอยู่ในช่วงตั้งแต่หลักพันจนถึงหลักแสน ถ้าหากว่าเรานำข้อมูลดิบๆนี้เข้ามาทำงานอาจจะทำให้การคำนวณของโปรแกรมเกิดความล่าช้า เพราะฉะนั้นจึงมีความจำเป็นที่ต้องทำการ Transform ข้อมูลในบาง Field ก่อนที่จะนำไปใช้งาน

เมื่อเข้าสู่หน้าจอที่ 2 ของโปรแกรมจะได้ผลลัพธ์ดังรูปที่ ก.8

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

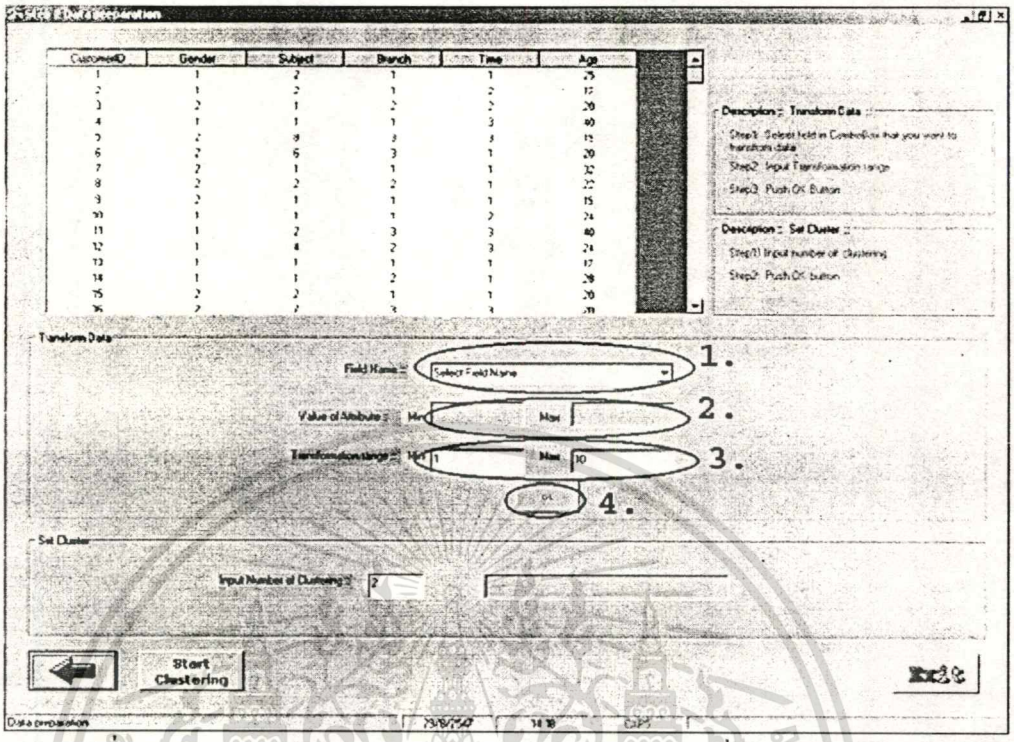


รูปที่ ก.8 แสดงหน้าจอที่ 2 ของโปรแกรม

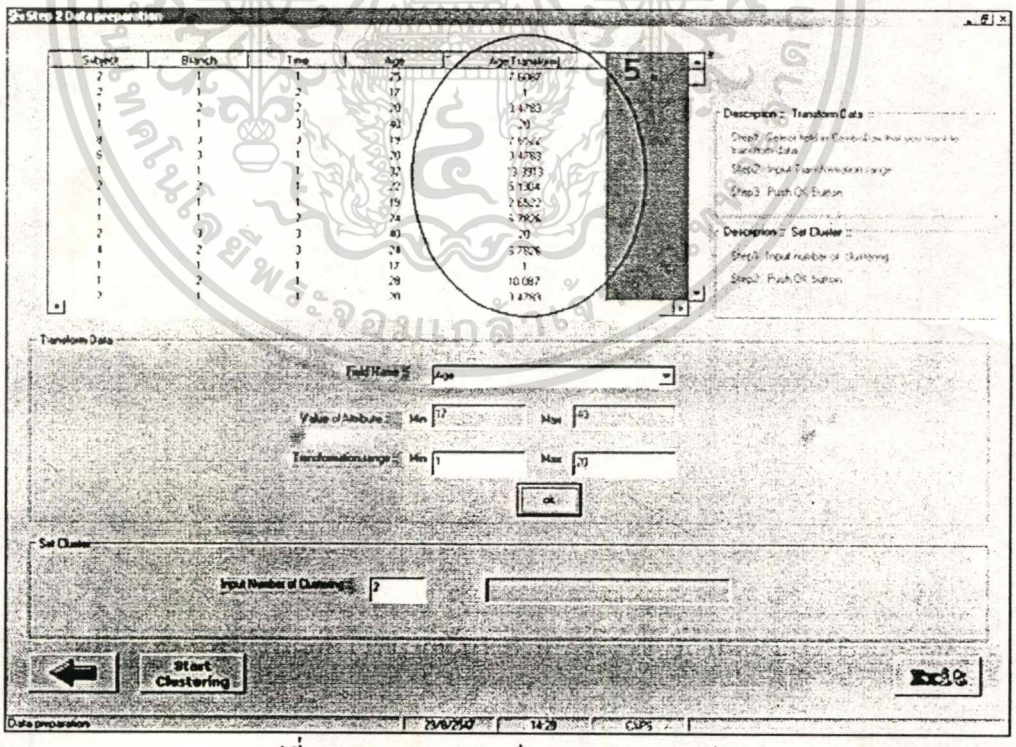
ก.2.1 การเลือก Field ที่ต้องการ Transform ข้อมูล

1. เลือก Field ที่ต้องการ Transform ในช่อง Field Name ดังรูปที่ ก.9
2. โปรแกรมจะทำการแสดงค่า Minimum และ Maximum Value ของ Field นั้นๆขึ้นมาแสดง
3. ให้ทำการกำหนดช่วงของข้อมูลที่ต้องการลองไปในช่อง Transformation Range
4. กดปุ่ม ok เพื่อให้โปรแกรมทำการคำนวณช่วงของข้อมูล
5. โปรแกรมจะแสดงผลลัพธ์ขึ้นมาดังรูปที่ ก.10

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ ก.9 แสดงการเลือก Field และการกำหนดช่วงของค่าที่ต้องการ Transform



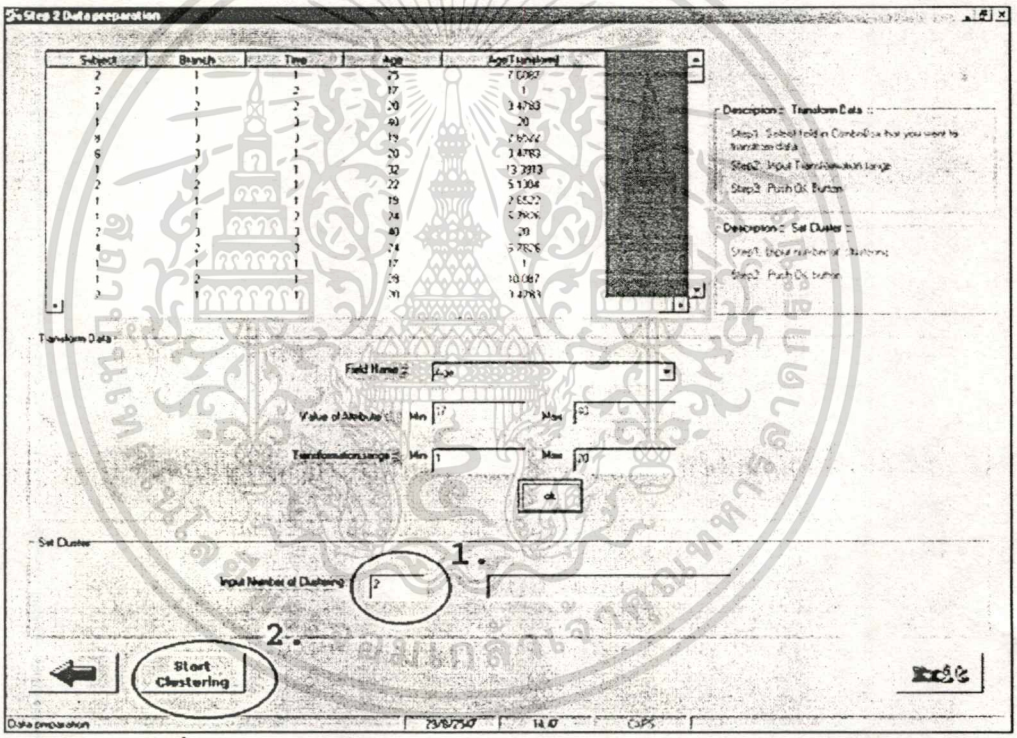
รูปที่ ก.10 แสดงผลลัพธ์การ Transform ข้อมูล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ก.3 การทำงานขั้นที่ 3 (Data Mining)

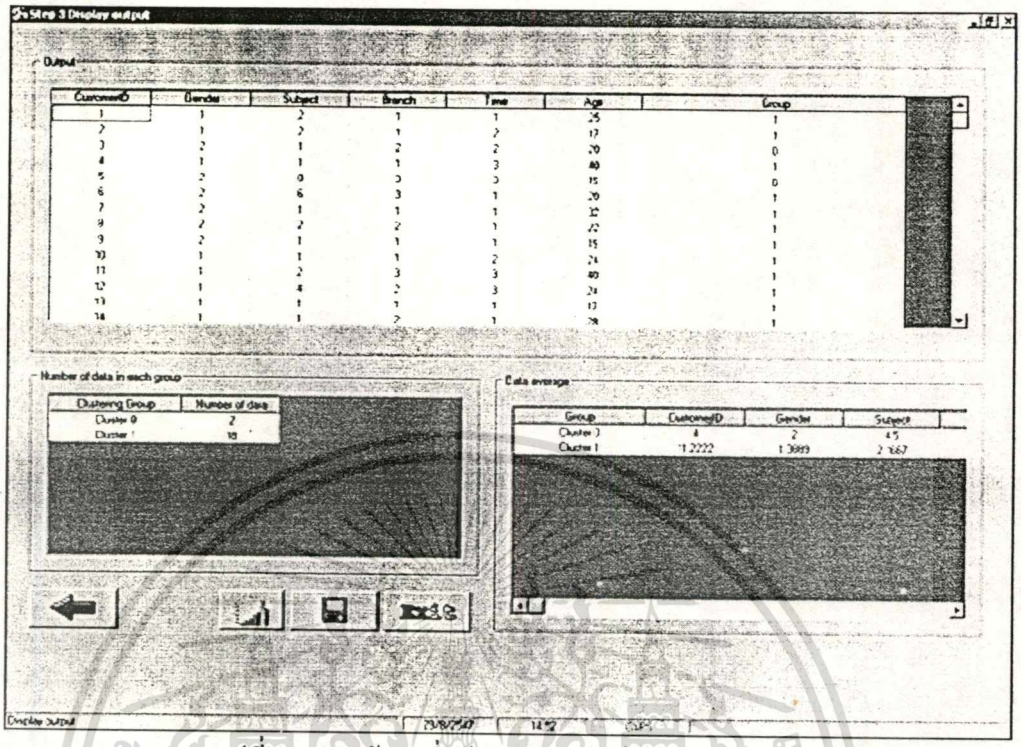
หลังจากทำการเลือกข้อมูลและ Transform ข้อมูลเป็นที่เรียบร้อยแล้ว ขั้นตอนต่อมาก็จะเป็นการนำข้อมูลทั้งหมดเข้าสู่อังกอริทึมของการทำดาต้าไมนิ่ง ซึ่งในระบบนี้จะใช้ CLARA Algorithm ในการคำนวณ โดยมีขั้นตอนในการทำงานดังต่อไปนี้

1. ในขั้นตอนนี้ให้ทำการกำหนดว่าต้องการจะทำการแบ่งข้อมูลออกเป็นกี่กลุ่ม โดยทำการกำหนดลงไปในช่อง Input Number of Clustering ดังรูปที่ ก.11
2. กดปุ่ม Start clustering เพื่อให้โปรแกรมเริ่มทำการแบ่งกลุ่มข้อมูล
3. โปรแกรมจะทำการแสดงผลลัพธ์ของการคำนวณ ดังรูปที่ ก.12



รูปที่ ก.11 แสดงการกำหนดว่าต้องการจะแบ่งข้อมูลออกเป็นกี่กลุ่ม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



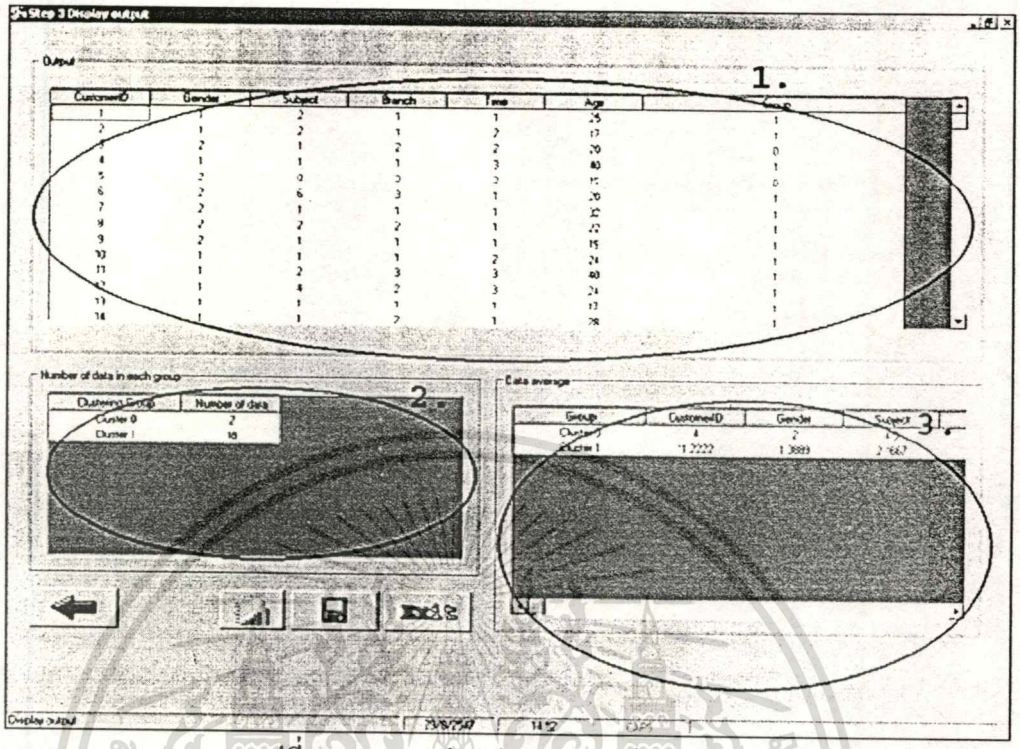
รูปที่ ก.12 หน้าจอที่ใช้แสดงผลลัพธ์ของการคำนวณ

ก.3.1 ส่วนต่างๆของหน้าจอผลลัพธ์

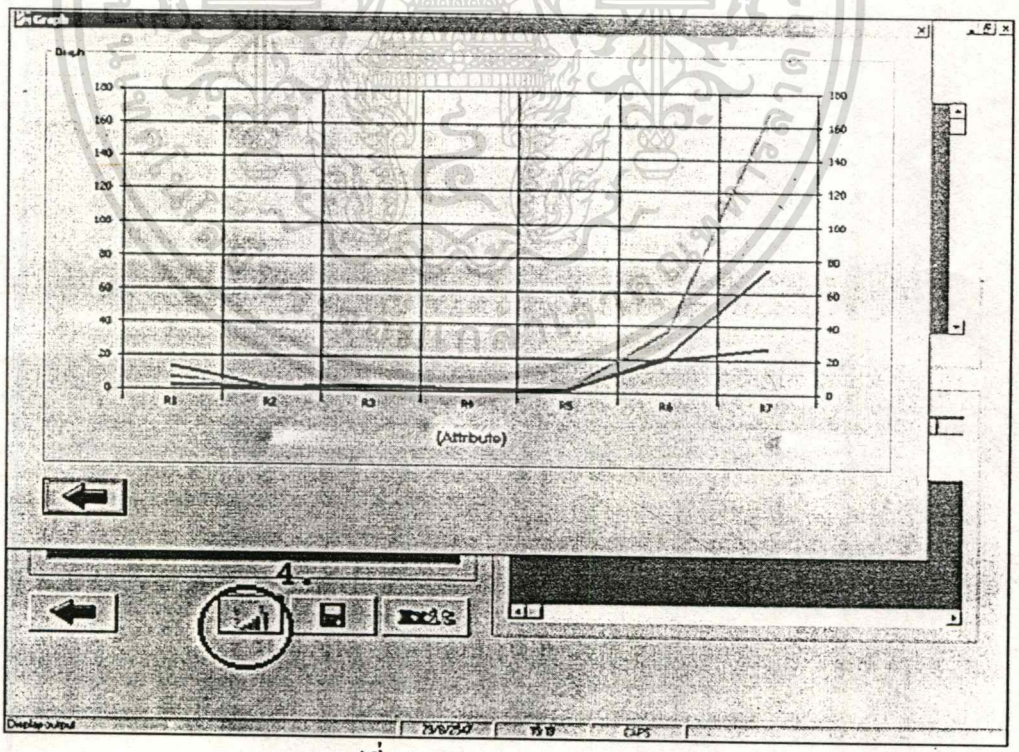
หลังจากโปรแกรมได้คำนวณหาผลัพธ์เรียบร้อยแล้ว ขั้นตอนต่อมาก็จะทำการวิเคราะห์ข้อมูลเพื่อนำมาวางแผนกลยุทธ์ต่อไป โดยหน้าจอผลลัพธ์แบ่งออกเป็น 3 ส่วนดังนี้

1. Output ในส่วนนี้จะทำการแสดงข้อมูลทั้งหมดในแต่ละ Field ที่เราเลือก และแสดงผลลัพธ์ออกมาว่า ข้อมูลแต่ละตัวถูกแบ่งไปในกลุ่มไหน ดังรูปที่ ก.13
2. Number of data in each group ในส่วนนี้เป็นการแสดงว่า แต่ละกลุ่มมีจำนวนข้อมูลทั้งหมดเท่าไหร่
3. Data Average ในส่วนนี้จะแสดงค่าเฉลี่ยของข้อมูลในแต่ละ Field ของแต่ละกลุ่ม และค่า Error ภายในแต่ละกลุ่ม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ ก.13 แสดงส่วนต่างๆของหน้าจอผลลัพธ์

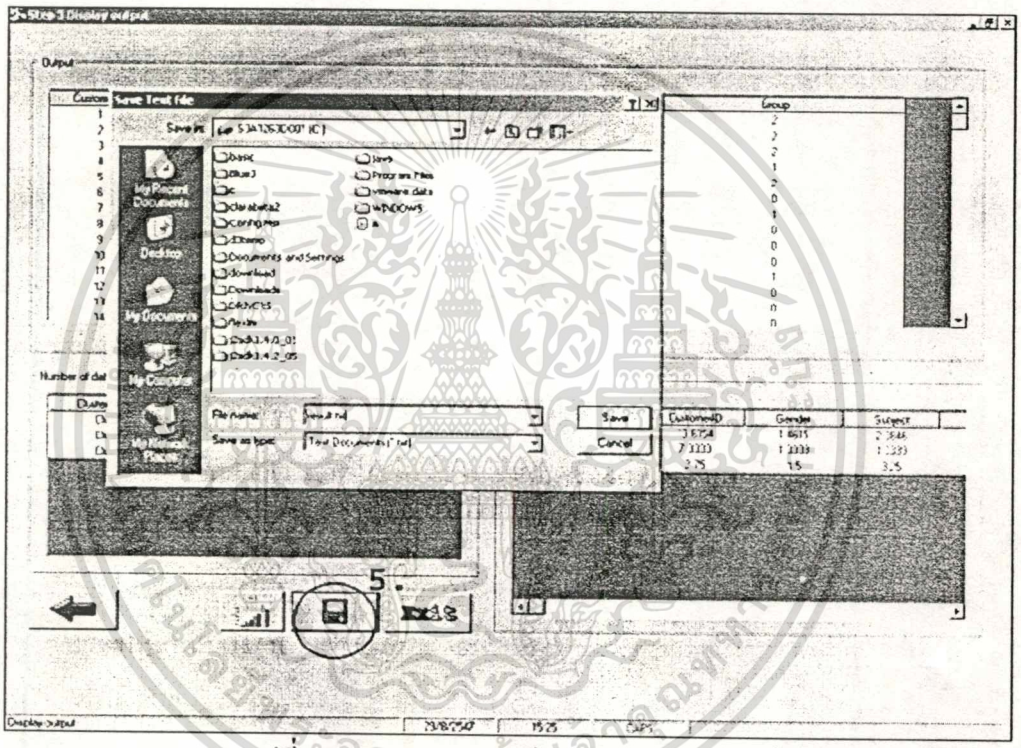


รูปที่ ก.14 แสดงการ plot กราฟ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4. เราสามารถดูกราฟของผลลัพธ์จากการคำนวณโดยการกดปุ่มรูปกราฟ โดยข้อมูลที่นำมาทำการ plot กราฟจะนำข้อมูลจาก ตาราง Data Average โดยในแนวแกนนอนหมายถึง Field ในแต่ละ Field และในแนวแกนตั้งแสดงถึงค่าของข้อมูล ซึ่งจะได้ผลลัพธ์ออกมาดังรูปที่ ก.14

5. สามารถทำการเซฟผลลัพธ์ของการคำนวณในรูปแบบของ Text ไฟล์ได้ โดยการกดปุ่มรูปแผ่นดิสเก็ต ดังรูปที่ ก.15



รูปที่ ก.15 แสดงการเซฟข้อมูลในรูป Text ไฟล์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ข. ความหมายของ Warning Massage และวิธีการแก้ไข

ในการทำงานของแต่ละขั้นตอน บางครั้งอาจจะเกิดข้อผิดพลาดขึ้นได้ ซึ่งในแต่ละขั้นตอนต่างๆถ้าหากว่ามีความผิดพลาดเกิดขึ้นจะปรากฏข้อความเตือนขึ้นมา ดังนั้นเพื่อให้ทราบถึงสาเหตุของความผิดพลาดที่เกิดขึ้น และคำแนะนำในการแก้ปัญหา เพื่อให้ผู้ใช้งานสามารถทำการแก้ไขได้ด้วยตัวเอง จึงได้ทำการแสดง Warning Message พร้อมคำอธิบายและวิธีการแก้ไขดังตารางด้านล่างดังต่อไปนี้

ตารางที่ ข.1 แสดง Warning Message ที่เกิดขึ้นในหน้าจอที่ 1

ขั้นตอนที่ 1 Data Selection	สาเหตุ	การแก้ไข
Please select database file	ยังไม่ได้เลือกไฟล์ฐานข้อมูล	ให้ทำการเลือกไฟล์ฐานข้อมูลจาก Listbox ในส่วนของ Select Database File ก่อน
No data in this drive	Drive ที่ทำการเลือกไม่มีข้อมูลอะไรอยู่เลย	ให้ทำการแก้ไขโดยเลือก Drive ที่เก็บไฟล์ฐานข้อมูลเอาไว้
Please select field that you want to mining	สาเหตุเนื่องมาจาก ยังไม่ได้เลือก Field อะไรเลยแล้วไปกดปุ่ม ลูกศรสีเขียว เพื่อไปยังขั้นตอนถัดไป	ให้ทำการเลือก Field ที่ต้องการใช้งานก่อน จากนั้นค่อยกดปุ่มลูกศรสีเขียว เพื่อไปยังขั้นตอนถัดไป
You can't choose all field because of some field contain back value	มีข้อมูลในบาง Field ขาดหายไป	ให้ทำการแก้ไขข้อมูลใน Field ต่างๆที่ขาดหายไป
You can't use database that do not have CustomerID field	ในตารางนี้ไม่มี Field ที่ชื่อว่า CustomerID	ให้แก้ไข Field ในฐานข้อมูลโดยที่จะต้องมี Field หลักชื่อว่า CustomerID

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ข.2 แสดง Warning Message ที่เกิดขึ้นในหน้าจอที่ 2

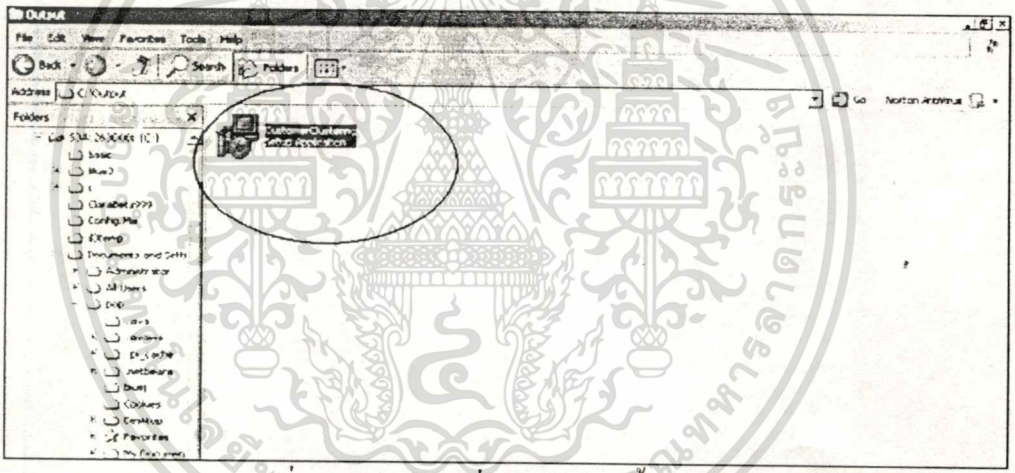
ขั้นตอนที่ 2 Transform data	สาเหตุ	การแก้ไข
You can not input number that less than 1	สาเหตุเกิดจาก ใส่ Minimun value ในช่อง Transform Range มีค่าน้อยกว่า 1	ให้ใส่ค่าที่มากกว่า 0
you can not input character in transform range	เกิดจากการที่ใส่ตัวอักษรแปลกๆลงไปในช่อง Transform range	ให้ใส่ตัวเลขลงไปในช่อง Transform range
You can not input number of clustering less than 2	ใส่ตัวเลขในช่อง Input Number of Clustering น้อยกว่า 2	ให้ใส่ตัวเลขลงไปในช่อง Input Number of Clustering ให้มากกว่า 1
You can not input character in number of clustering	ใส่ตัวอักษรแปลกๆ ลงไปในช่อง Input Number of Clustering	ห้ามใส่ตัวอักษรแปลกๆ ลงไปในช่อง Input Number of Clustering

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ค. การติดตั้งโปรแกรม

เนื่องจากว่าในการพัฒนาโปรแกรมนั้นได้มีการนำ component ต่างๆเข้ามาใช้งานภายในโปรแกรมเป็นจำนวนมาก และใช้เครื่องมือ Microsoft Visual Basic 6.0 ในการพัฒนา เพราะฉะนั้นถ้าหากต้องการจะนำโปรแกรมไปใช้งานนั้นจึงไม่สามารถนำแต่ไฟล์ .exe ไปใช้งานได้ เนื่องจากว่าการที่จะให้โปรแกรมทำงานได้จะต้องมีการติดตั้ง Microsoft Visual Basic Runtime Module และ component ต่างๆลงไปเสียก่อน ดังนั้นเพื่อลดความซับซ้อนในการติดตั้งโปรแกรม จึงได้ทำการรวบรวมไฟล์ต่างๆและทำเป็นไฟล์ CustomerClustering.exe ซึ่งเป็นไฟล์ที่ใช้ในการ setup โปรแกรม โดยมีวิธีการติดตั้งโปรแกรมดังต่อไปนี้

1. ให้ดับเบิ้ลคลิ้กที่ไฟล์ CustomerClustering.exe ซึ่งเป็นไฟล์ที่ใช้ทำการ setup โปรแกรม ดังรูปที่ ค.1



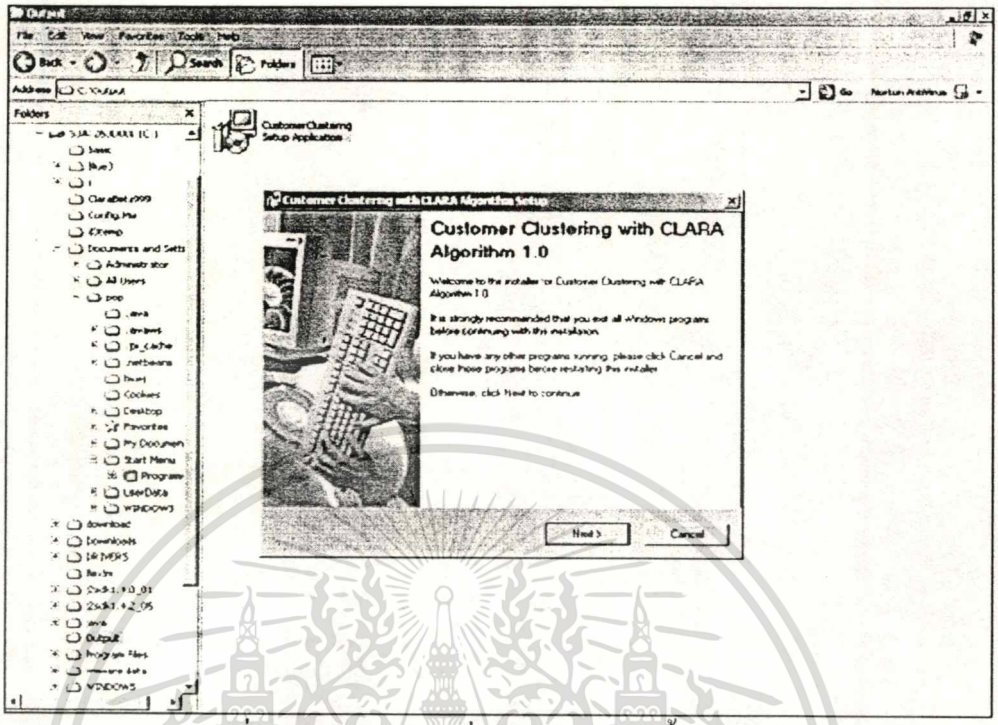
รูปที่ ค.1 แสดงไฟล์ที่ใช้ในการติดตั้งโปรแกรม

2. เมื่อดับเบิ้ลคลิ้กที่ไฟล์ CustomerClustering.exe แล้วจะเข้าสู่ขั้นตอนการติดตั้งโปรแกรม โดยจะมีหน้าจอแสดง ดังรูปที่ ค.2

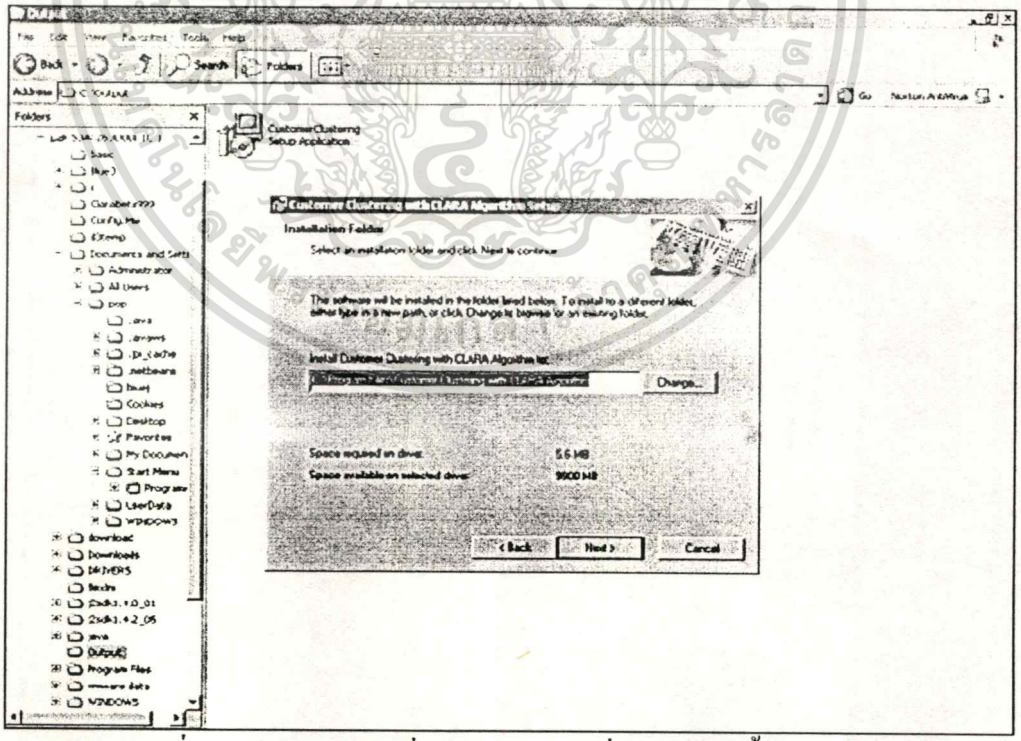
3. จากนั้นให้กดปุ่ม Next ที่ด้านล่างของหน้าจอ เพื่อเข้าสู่ขั้นตอนถัดไป ดังรูปที่ ค.3

4. ขั้นตอนนี้จะให้ User เลือกว่าต้องการจะติดตั้งโปรแกรมเก็บไว้ที่ Folder ไหน ถ้าหากว่าเราต้องการจะเปลี่ยน Folder ที่ต้องการติดตั้งโปรแกรมก็ให้ทำการกดปุ่ม Change.. แล้วเลือก Folder ที่ต้องการจัดเก็บโปรแกรม ดังรูปที่ ค.4 และ ค.5 ตามลำดับ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

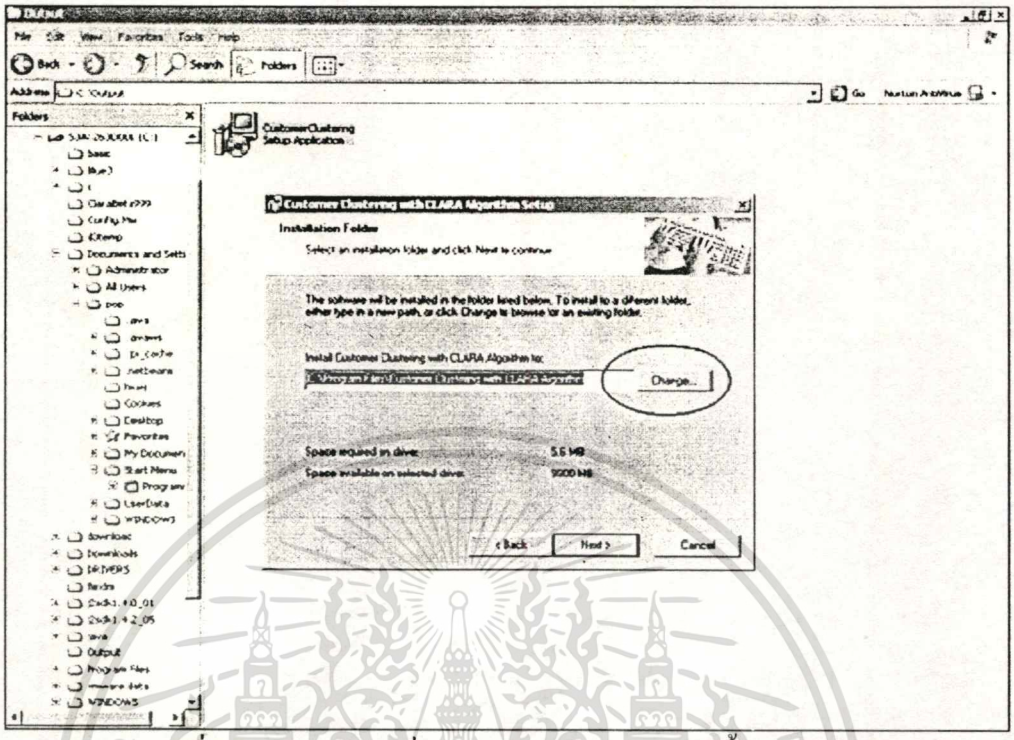


รูปที่ ค.2 แสดงหน้าจอเริ่มต้นของการติดตั้งโปรแกรม

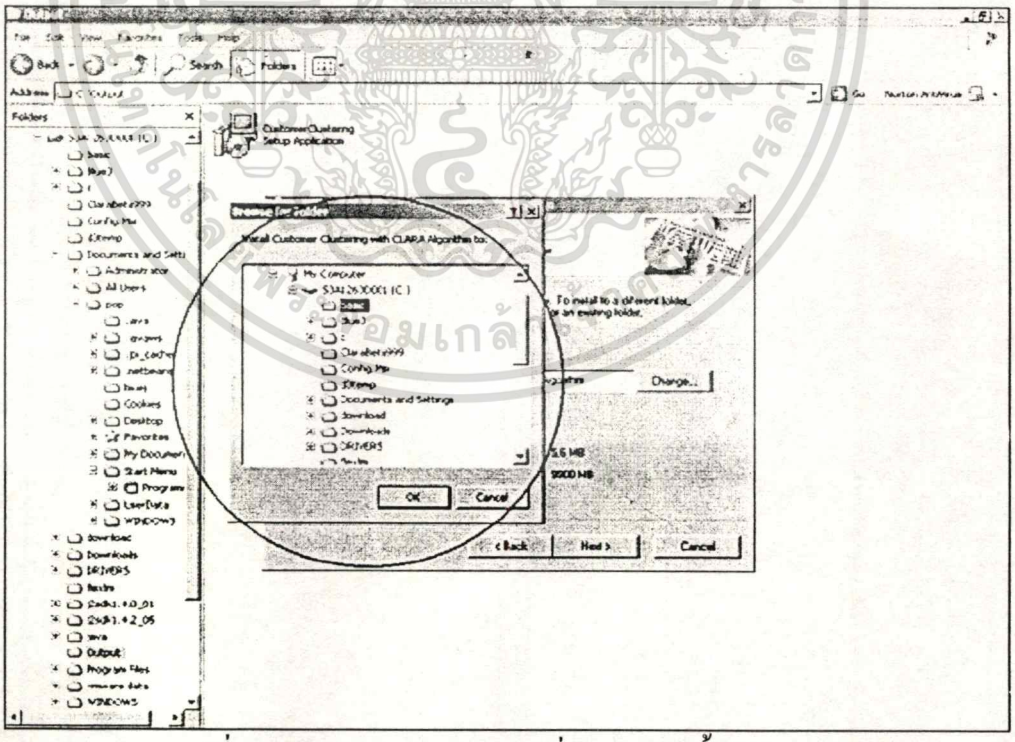


รูปที่ ค.3 แสดงหน้าจอที่ใช้เลือก Folder ที่ต้องการติดตั้งโปรแกรม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



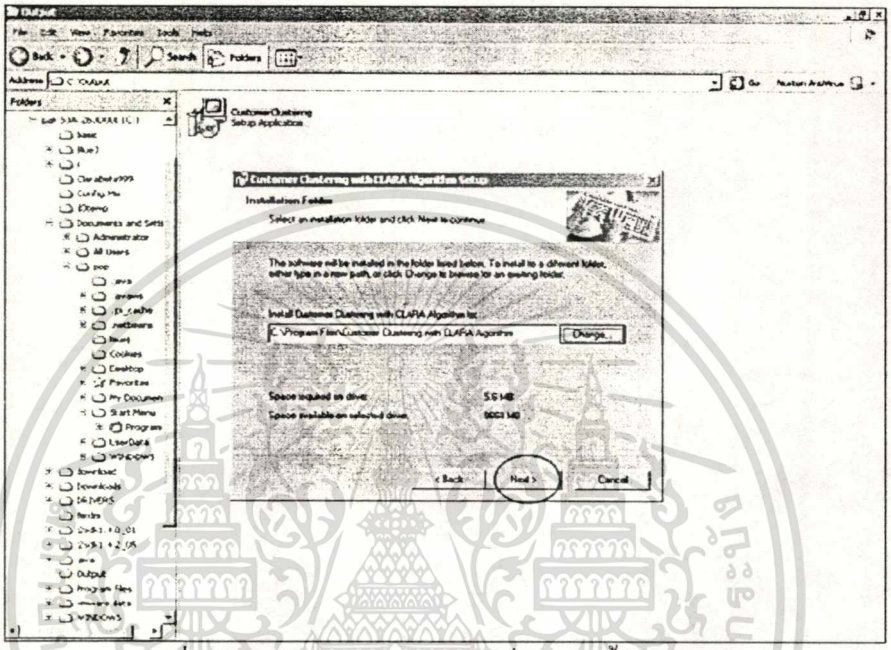
รูปที่ ค.4 แสดงการเปลี่ยน Folder ที่ต้องการจะติดตั้งโปรแกรม



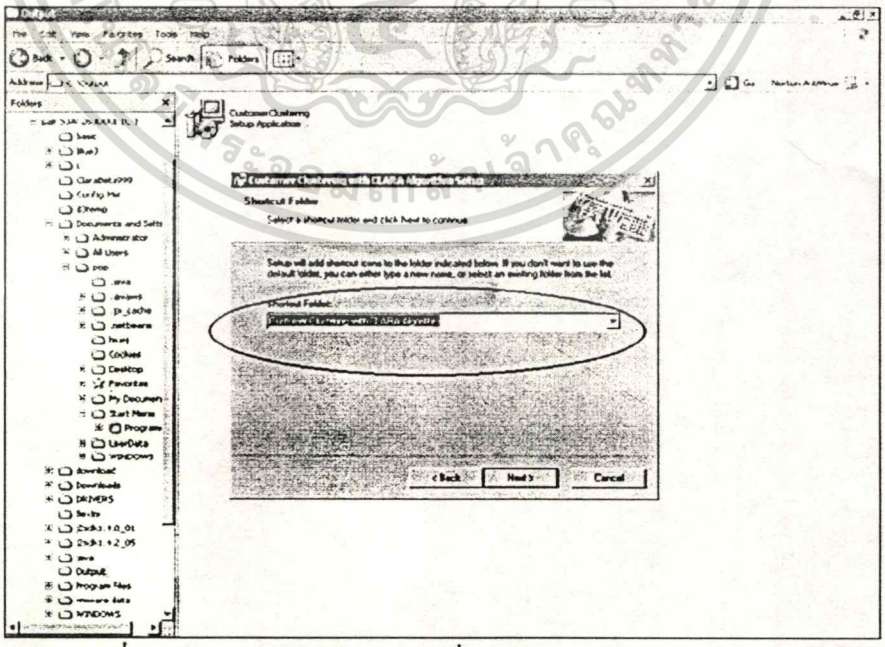
รูปที่ ค.5 แสดงการเลือก Folder ที่ต้องการติดตั้งโปรแกรม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

5. เมื่อเลือก Folder ที่จะติดตั้งโปรแกรมเสร็จเรียบร้อยแล้วให้กดปุ่ม Next เพื่อเข้าสู่ขั้นตอนถัดไป ดังรูปที่ ค.6
6. จากนั้นโปรแกรมจะให้เลือก Folder ที่จะเก็บ shortcut ของโปรแกรม ดังรูปที่ ค.7



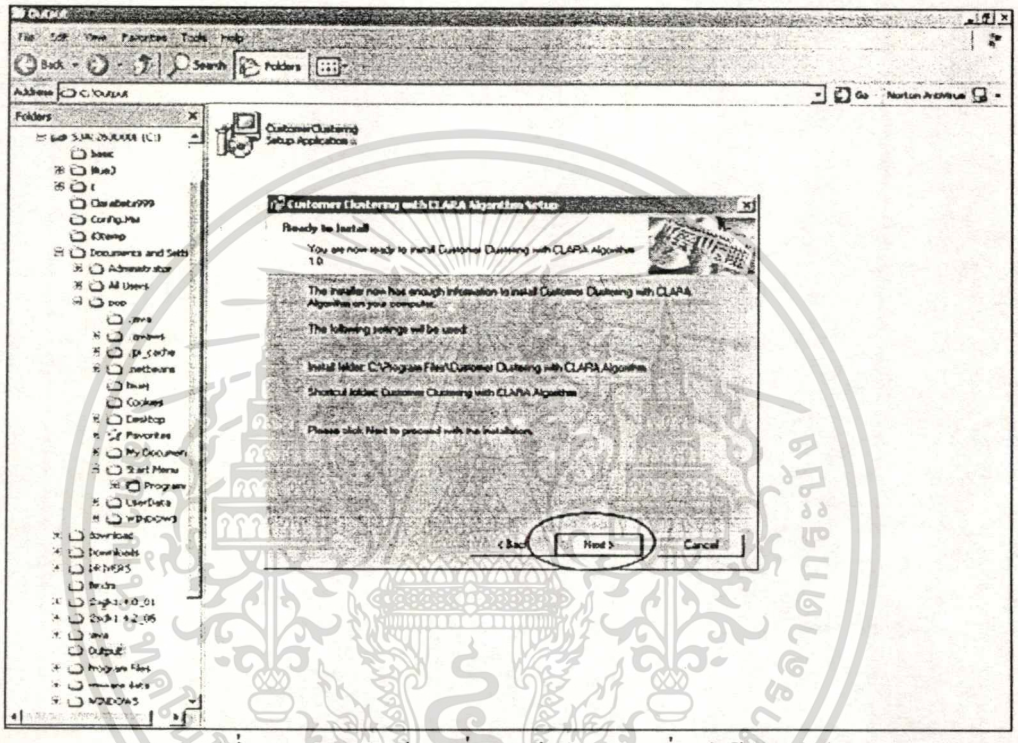
รูปที่ ค.6 แสดงการกดปุ่ม Next เพื่อไปยังขั้นตอนถัดไป



รูปที่ ค.7 แสดงการเลือก Folder ที่จะเก็บ Shortcut ของโปรแกรม

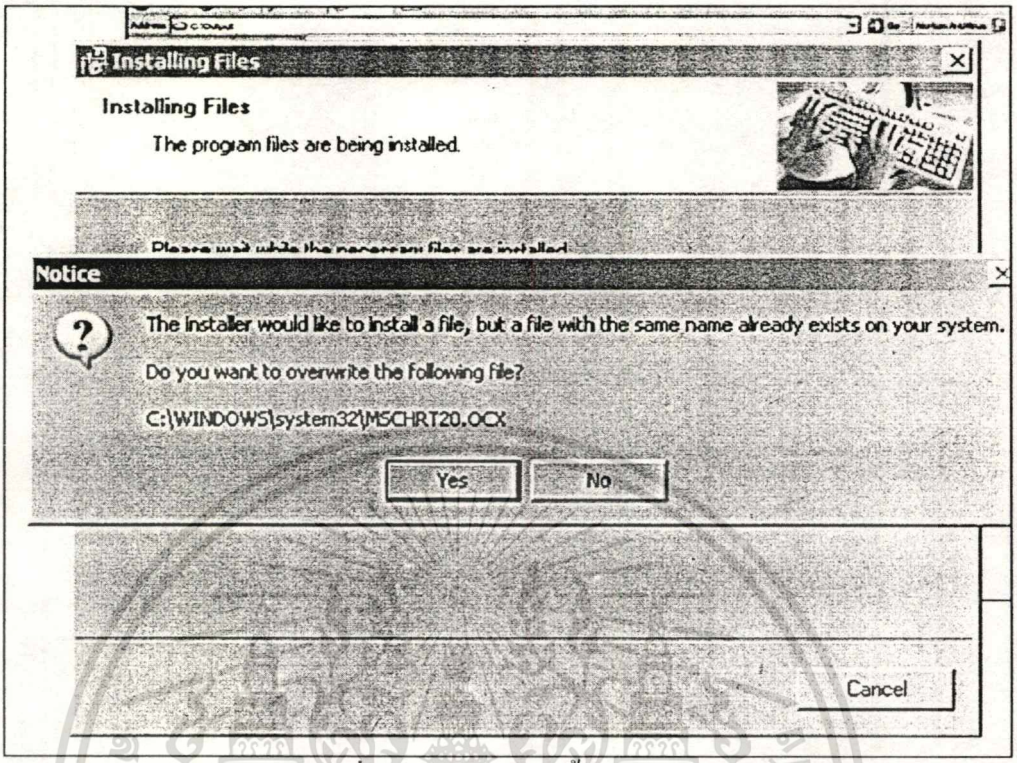
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

7. ขั้นตอนนี้โปรแกรมจะสรุปสิ่งต่างๆที่เราได้เลือกไว้เมื่อสักครู่นี้ขึ้นมาให้ดู ถ้าหากว่าถูกต้องทุกอย่างก็ให้กดปุ่ม Next เพื่อเริ่มสู่การติดตั้งโปรแกรม ดังรูปที่ ค.8
8. ขั้นตอนนี้เป็นการติดตั้งไฟล์ต่างๆของโปรแกรม และเมื่อติดตั้งโปรแกรมเสร็จสมบูรณ์จะผลลัพธ์ดังรูปที่ ค.9 และ ค.10 ตามลำดับ

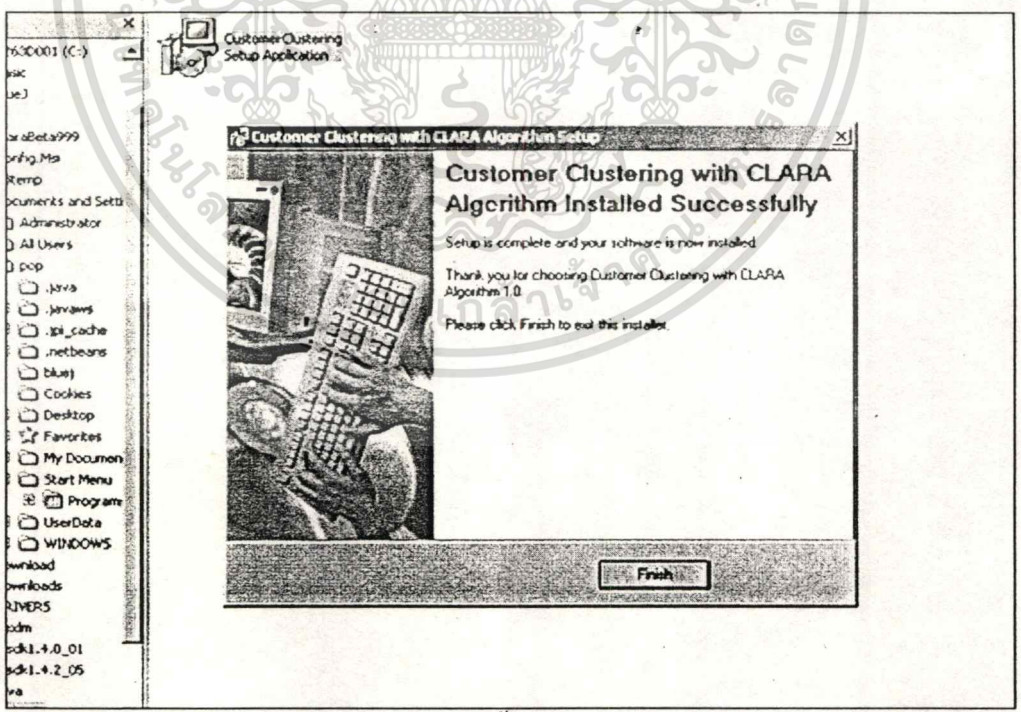


รูปที่ ค.8 แสดงหน้าจอที่สรุปข้อมูลต่างๆที่เราได้เลือกไว้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ ค.9 แสดงการติดตั้งโปรแกรม

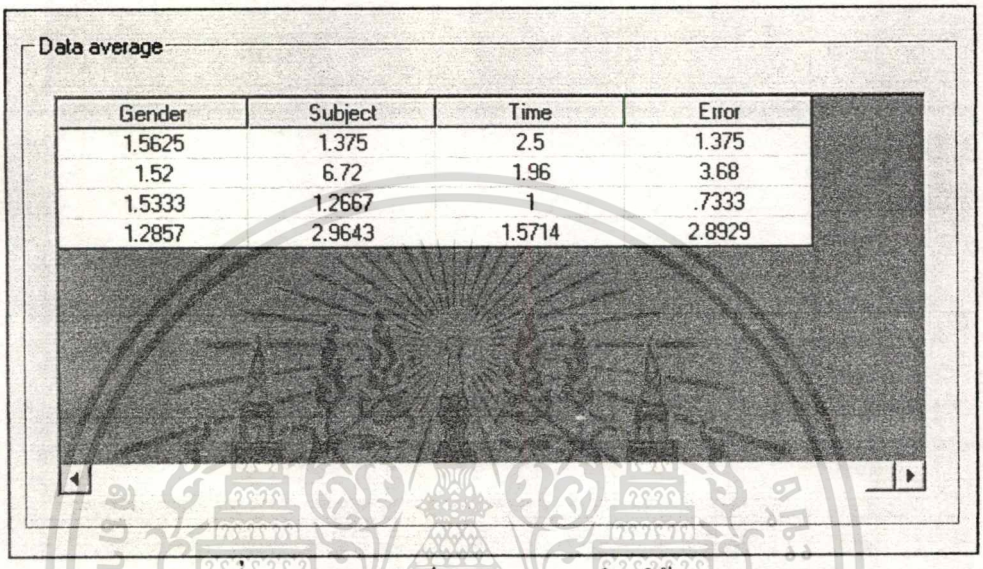


รูปที่ ค.10 แสดงการติดตั้งโปรแกรมที่เสร็จสมบูรณ์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ง. เปรียบเทียบการทำงานของ K-Means Algorithm กับ CLARA Algorithm

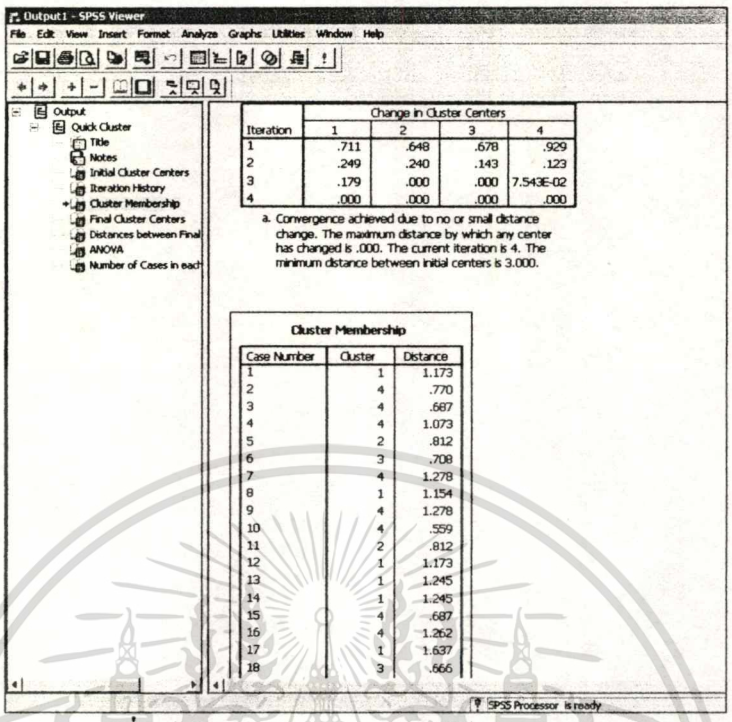
ในส่วนนี้ได้ทำการทดลองเปรียบเทียบประสิทธิภาพการทำงานของ K-Means Algorithm กับ CLARA Algorithm ซึ่งจากการทดลองได้ทำการทดลองนำข้อมูลจำนวน 100 ตัวมาทำการคำนวณโดยใช้ CLARA Algorithm ซึ่งจะได้ผลลัพธ์ ดังรูปที่ ง.1



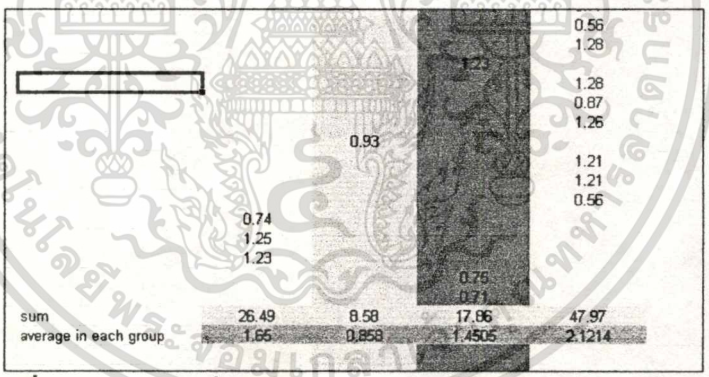
รูปที่ ง.1 แสดงผลลัพธ์จากการคำนวณโดยใช้ CLARA Algorithm

จากนั้นจึงได้นำข้อมูลชุดเดียวกันมาทำคาต้าไมนิ่งโดยใช้ K-Means Algorithm ซึ่งทดสอบโดยการใช้โปรแกรม SPSS เพื่อวิเคราะห์ข้อมูล ซึ่งได้ผลลัพธ์ดังรูปที่ ง.2 และ ง.3

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ ง.2 แสดงผลลัพธ์จากการคำนวณโดยใช้โปรแกรม SPSS



รูปที่ ง.3 แสดงค่าเฉลี่ยของระยะห่างของข้อมูลกับจุดศูนย์กลางของแต่ละกลุ่ม

จากการเปรียบเทียบทั้ง 2 Algorithm ซึ่งจะพบว่าผลลัพธ์ของ CLARA Algorithm มีค่าเฉลี่ยของระยะห่างของข้อมูลกับจุดศูนย์กลางของแต่ละกลุ่มน้อยกว่า K-Means Algorithm เล็กน้อย ซึ่งทั้งนี้อาจจะขึ้นอยู่กับช่วงของข้อมูลในฐานข้อมูลที่เรานำมาคำนวณด้วย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ประวัติผู้เขียน

ชื่อผู้เขียน	นาย นิธิสิทธิ์ สุขแสง
วันเดือนปีเกิด	26 กันยายน 2523
สถานที่เกิด	กรุงเทพมหานคร
ประวัติการศึกษา	สำเร็จการศึกษาระดับมัธยมศึกษาจากโรงเรียนบางกะปี 2541 สำเร็จการศึกษาระดับปริญญาตรีจากคณะวิศวกรรมคอมพิวเตอร์ จากมหาวิทยาลัยเทคโนโลยีมหานคร ปี 2545
ประวัติการทำงาน	อาจารย์พิเศษ สถาบันอินเทอร์เน็ตและการออกแบบ Netdesign

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้