การพัฒนาระบบจัดกลุ่มข้อมูลลูกค้าโดยใช้อัลกอริทึม K-Prototypes

Customer Segmentation Development System using K-Prototypes Algorithm

โดย

นางสาวสุพร เวคีวิทยา

รหัส 45066016

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน

หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

ภาคเรียนที่ 2 ปีการศึกษา 2546

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ชื่อหัวข้อ	การพัฒนาระบบการจัดกลุ่มข้อมูลลูกค้าโดยใช้อัลกอริทึม K-Prototypes
นักศึกษา	นางสาวสุพร เวคีวิทยา
อาจารย์ที่ปรึกษา	ผศ.ดร.วรพจน์ กรีสุระเดช
ระดับการศึกษา	วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2546

บทคัดย่อ

การทำธุรกิจการค้าในปัจจุบันนี้ได้มีการแข่งขันกันอย่างสูง การให้สินเชื่อกับลูกค้าจะเป็นช่องทางหนึ่งในการเพิ่มโอกาสทางธุรกิจให้กับบริษัท ดังนั้นบริษัทจึงต้องมีการนำข้อมูลการชำระเงินของลูกค้าที่มีอยู่ในระบบมาจัดกลุ่มประเภทของลูกค้าและวิเคราะห์เพื่อให้ได้ข้อมูลใหม่ๆ ซึ่งการจัดกลุ่มประเภทของลูกค้าจะสามารถนำไปช่วยในการสนับสนุนการตัดสินใจของผู้บริหาร เพื่อใช้ในการกำหนดวงเงินสินเชื่อ หรือกำหนดระยะเวลาในการชำระเงินให้เหมาะสมกับลูกค้าแต่ละประเภทได้ โดยการจัดกลุ่มข้อมูลเป็นวิธีหนึ่งในการทำดาต้าไมนิ่ง และอัลกอริทึมที่ใช้ในการจัดกลุ่มนั้นมีหลายตัว ในเอกสารประกอบวิชาโครงการพัฒนาระบบงานฉบับนี้จะใช้อัลกอริทึม K-Prototypes ในการจัดกลุ่มซึ่งสามารถใช้ได้กับข้อมูลทั้งประเภทที่เป็นตัวเลขและไม่ใช่ตัวเลข

Title	Customer Segmentation Development System using K-Prototypes Algorithm
Student	Miss Suporn Vakeewittaya
Advisor	Asst.Prof.Dr.Worapoj Kreesuradej
Level of Study	Master of Science in Information Technology
Major	Information Science
Academic Year	2003

ABSTRACT

Nowadays, Investment in business has a high competition. Giving the credit to customers is the way to give the chance for a company. Therefore, the companies have to inspect the data of customer payment in order to analyst and group the type of customer for get new information. Segment of customer is not able to help the executive decision to define credit budget or time of payment which suitable for each type of customer. Segmentation is one method in Data mining.

In this project, we use algorithm "K-prototypes" to segment the data which K-prototypes can use with numeric and categorical data.

กิตติกรรมประกาศ

ในการทำโครงการเรื่องการพัฒนาระบบการจัดกลุ่มข้อมูลลูกค้าโดยใช้อัลกอริทึม K-Prototypes (Customer Segmentation Development System using K-Prototypes Algorithm) สามารถสำเร็จลุล่วงไปด้วยดี ข้าพเจ้าต้องขอขอบพระคุณ ผศ.ดร.วรพจน์ กรีสุระเดช อาจารย์ที่ปรึกษาวิชาโครงการพัฒนาระบบงานที่กรุณาให้คำแนะนำและเป็นที่ปรึกษา อันเป็นประโยชน์ต่อการพัฒนาระบบ รวมทั้งเป็นผู้ตรวจสอบความถูกต้องของโครงการฉบับนี้

นอกจากนี้ข้าพเจ้าต้องขอกราบขอบพระคุณ บิดา มารดา และบุคคลในครอบครัว ที่ได้ให้ความสนับสนุนทางด้านกำลังใจในการเรียนจนการทำโครงการพัฒนาระบบนี้สำเร็จด้วยดี รวมทั้งขอขอบคุณพี่ๆและเพื่อนๆ IS 13.1 ทุกคนที่ให้ความช่วยเหลือในด้านต่างๆ ที่เกี่ยวกับโครงการไว้ ณ ที่นี้

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

หน้า

ตารางที่

สารบัญภาพ

หน้า

ภาพที่

สารบัญภาพ (ต่อ)

หน้า

ภาพที่

สารบัญภาพ (ต่อ)

หน้า

ภาพที่

บทที่ 1
บทนำ

1.1 ความเป็นมาของโครงการ

การดำเนินการทางธุรกิจจะประสบผลสำเร็จ และบรรลุจุดหมายปลายทางที่กำหนดได้นั้น การใช้ทรัพยากรที่มีอยู่ในองค์กรอย่างมีประสิทธิภาพนับว่ามีความสำคัญเป็นอย่างยิ่ง ซึ่งทรัพยากรในองค์กรนั้นประกอบไปด้วยหลายสิ่งหลายอย่าง เช่น คน เงิน วัสดุอุปกรณ์ รวมทั้งเทคโนโลยีและข้อมูลต่างๆที่มีอยู่ในองค์กร

ในการทำธุรกิจนั้น องค์กรจะต้องมีเทคนิคต่างๆมาใช้ในการบริหารเพื่อเป็นการเพิ่มโอกาสทางธุรกิจให้กับตัวองค์กรเอง ซึ่งการให้สินเชื่อกับลูกค้าก็เป็นเทคนิคหนึ่งที่นิยมใช้กันทุกองค์กร และเมื่อลูกค้าทำธุรกิจกับองค์กรไปได้สักระยะหนึ่ง องค์กรสามารถที่จะปรับเปลี่ยนวงเงินสินเชื่อ รวมทั้งการกำหนดระยะเวลาในการชำระเงินของลูกค้าได้ โดยส่วนใหญ่นั้นจะทำการจัดกลุ่มประเภทของลูกค้าจากลักษณะการชำระเงินของลูกค้า แล้วจึงนำผลจากการจัดกลุ่มนี้ไปวิเคราะห์ต่อว่าควรจะเพิ่มหรือลดวงเงินสินเชื่อของลูกค้า แต่ข้อมูลการชำระเงินของลูกค้าจะมีจำนวนมาก ซึ่งถ้าจะใช้คนในการนั่งพิจารณาและจัดกลุ่มจะทำให้เสียเวลาเป็นอย่างมาก ดังนั้นจึงมีแนวความคิดว่าถ้ามีการพัฒนาระบบที่ช่วยในการจัดกลุ่มข้อมูลการชำระเงินลูกค้า จะช่วยให้องค์กรสามารถนำข้อมูลที่ได้จากการจัดกลุ่มไปใช้ในการวิเคราะห์และสนับสนุนการตัดสินใจได้อย่างทันท่วงที

1.2 วัตถุประสงค์ของโครงการ

1) เพื่อศึกษาเทคนิคการทำดาต้าไมนิ่ง (Data Mining)
2) เพื่อศึกษาอัลกอริทึม K-Prototypes ซึ่งเป็นอัลกอริทึมหนึ่งที่ใช้ในการจัดกลุ่มข้อมูล
3) สามารถนำไปเป็นประโยชน์ในการสนับสนุนการวิเคราะห์และแบ่งกลุ่มประเภทของลูกค้า
4) จัดกลุ่มลูกค้าที่มีลักษณะในการชำระเงินใกล้เคียงกันเพื่อนำมาจัดกลุ่มประเภทของลูกค้า
5) นำกลุ่มของลูกค้าที่ได้จากการจัดกลุ่มมาพิจารณาเพื่อให้วงเงินสินเชื่อ รวมทั้งกำหนดระยะเวลาในการชำระเงินของลูกค้า

1.3 ขอบเขตของการดำเนินงาน

1) จัดกลุ่มข้อมูลลูกค้าโดยใช้อัลกอริทึม K-Prototypes ในการจัดกลุ่ม
2) ใช้ได้กับข้อมูลที่ถูกจัดเก็บลงในฐานข้อมูล Microsoft Access 2000 เท่านั้น

1.4 ขั้นตอนการศึกษา

1) กำหนดวัตถุประสงค์ในการจัดกลุ่มข้อมูลลูกค้า
2) ศึกษาขั้นตอนและวิธีการในการทำดาต้าไมนิ่ง
3) ศึกษาอัลกอริทึม K-Prototypes เพื่อใช้ในการจัดกลุ่มข้อมูลลูกค้า
4) ออกแบบและพัฒนาระบบโดยใช้อัลกอริทึม K-Prototypes
5) ตรวจสอบและแก้ไขระบบที่สร้างขึ้นให้สมบูรณ์และถูกต้อง
6) สรุปผลการศึกษา

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1) เข้าใจหลักการและขั้นตอนต่างๆในการทำดาต้าไมนิ่ง รวมทั้งอัลกอริทึม K-Prototypes ที่ใช้ในการจัดกลุ่ม
2) สามารถเป็นแนวทางในการนำดาต้าไมนิ่งไปประยุกต์ใช้กับงานทางธุรกิจ และสามารถเป็นแนวทางให้กับผู้ที่สนใจและศึกษาการทำดาต้าไมนิ่งที่ต้องการเลือกใช้วิธีการจัดกลุ่มข้อมูล
3) สามารถแบ่งข้อมูลลูกค้าออกเป็นกลุ่มได้อย่างรวดเร็ว ซึ่งจะช่วยให้นำไปวิเคราะห์และประยุกต์ใช้กับงานทางด้านธุรกิจได้อย่างทันท่วงที

บทที่ 2
ทฤษฎีที่เกี่ยวข้อง

2.1 ความหมายของดาต้าไมนิ่ง

ในปัจจุบันนี้ข้อมูลมีความสำคัญต่อการดำเนินธุรกิจเป็นอย่างมาก จึงมีความจำเป็นที่จะต้องเก็บข้อมูลที่เกี่ยวข้องกับธุรกิจและนำสารสนเทศมาประยุกต์ เพื่อนำผลที่ได้มาช่วยในการสนับสนุนการตัดสินใจ และการที่มีวิธีการวิเคราะห์ที่ดีและเหมาะสมจะทำให้มีโอกาสเป็นผู้นำทางธุรกิจ ซึ่งส่งผลให้มีการเพิ่มจำนวนฐานข้อมูลในการเก็บข้อมูลมากขึ้นและทำให้เกิดปัญหาตามมา นั่นก็คือการวิเคราะห์ข้อมูลจำนวนมากๆเป็นเรื่องที่ยาก จึงได้มีการคิดค้นเทคโนโลยีที่เรียกว่า "ดาต้าไมนิ่ง" (Data Mining) มาช่วยในการวิเคราะห์ซึ่งจะทำให้ทราบถึงความสัมพันธ์ต่างๆที่ซ่อนอยู่ในฐานข้อมูลที่มีอยู่จำนวนมาก

ดาต้าไมนิ่งเป็นกระบวนการที่ใช้ในการวิเคราะห์ข้อมูล และได้ถูกออกแบบมาเพื่อดึงส่วนที่เป็นข้อมูลที่สำคัญที่เรายังไม่ทราบ และยังช่วยในการสนับสนุนการตัดสินใจในการบริหาร โดยดาต้าไมนิ่งมีข้อดีที่แตกต่างจากเครื่องมอวิเคราะห์อื่นๆคือ ดาต้าไมนิ่งสามารถมองความสัมพันธ์ของข้อมูลในหลายมิติและสามารถใช้กับข้อมูลที่มีขนาดใหญ่ได้

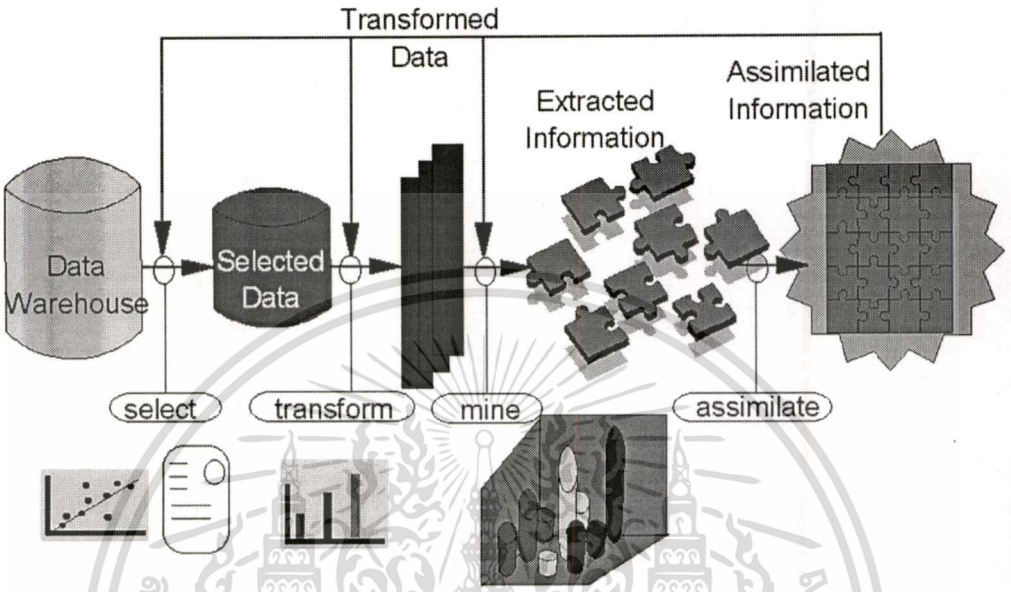
2.2 กระบวนการทำงานของดาต้าไมนิ่ง

ในกระบวนการทำดาต้าไมนิ่งประกอบด้วยขั้นตอนหลักๆ 4 ขั้นตอน คือ การกำหนดจุดประสงค์ทางธุรกิจ การเตรียมข้อมูล การไมนิ่งและการวิเคราะห์ผลที่ได้จากการทำดาต้าไมนิ่งและการนำมาใช้ โดยหลังจากที่ทำการวิเคราะห์ผลแล้วสามารถกลับไปเริ่มทำขั้นตอนใดใหม่อีกครั้งก็ได้ ดังแสดงในรูปที่ 2.1 ส่วนใหญ่มักจะให้ความสำคัญกับการทำไมนิ่ง แต่ในความเป็นจริงแล้วขั้นตอนการเลือกและเตรียมข้อมูลก็เป็นขั้นตอนที่สำคัญเช่นกัน

2.2.1 การกำหนดวัตถุประสงค์ทางธุรกิจ (Business Objective Determination)

การกำหนดวัตถุประสงค์ประสงค์ทางธุรกิจเป็นขั้นตอนแรกของการทำดาต้าไมนิ่ง ขั้นตอนนี้จะเป็นการกำหนดขอบเขตและเป้าหมายของการทำดาต้าไมนิ่ง โดยจะต้องทำความเข้าใจกับปัญหาและความต้องการทางธุรกิจก่อน เนื่องจากปัญหาทางธุรกิจบางปัญหาก็ไม่สามารถทำการ

แก้ไขได้ด้วยการทำดาต้าไมนิ่ง ดังนั้นในขั้นตอนนี้จึงต้องทำการวิเคราะห์ทางธุรกิจรวมทั้งต้องวิเคราะห์ข้อมูลเบื้องต้นว่าเรามีข้อมูลใดอยู่บ้าง เพื่อทำให้การแก้ปัญหาเป็นไปอย่างถูกต้อง



รูปที่ 2.1 แสดงขั้นตอนการทำดาต้าไมนิ่ง

2.2.2 การเตรียมข้อมูล (Data Preparation)

ขั้นตอนการเตรียมข้อมูลนี้เป็นขั้นตอนที่สำคัญมาก เนื่องจากขั้นตอนนี้จะเป็นการจัดเตรียมข้อมูลเพื่อส่งต่อไปยังกระบวนการไมนิ่ง ถ้าเรามีการเตรียมข้อมูลที่ไม่ดีหรือเกิดข้อผิดพลาดจากการเตรียมข้อมูล จะส่งผลให้การทำไมนิ่งนั้นผิดไปจากวัตถุประสงค์ที่ตั้งไว้ ดังนั้นขั้นตอนนี้จึงสำคัญและจำเป็นต้องใช้เวลาในการทำมากที่สุดถึง 60 เปอร์เซนต์ของการทำดาต้าไมนิ่ง ซึ่งในการเตรียมข้อมูลนี้ยังได้ถูกแบ่งออกเป็น 3 ขั้นตอนย่อยดังนี้

<u>ขั้นตอนที่ 1</u> การเลือกข้อมูล (Data Selection)

จุดประสงค์หลักในการเลือกข้อมูลคือ การกำหนดแหล่งข้อมูลต่างๆทั้งในและนอกองค์กร โดยจะทำการระบุลักษณะและเลือกข้อมูลที่ต้องการ ซึ่งการเลือกจะเปลี่ยนแปลงตามวัตถุประสงค์ทางธุรกิจ ในการเลือกข้อมูลนั้นจะต้องไม่เลือกข้อมูลที่ไม่มีสิทธิ์ในการเข้าถึงข้อมูล ข้อมูลที่เป็นความลับซึ่งอาจจะส่งผลให้เกิดปัญหาทางกฎหมาย ดังนั้นการเลือกข้อมูลก็เป็นขั้นตอนที่สำคัญขั้นตอนหนึ่ง

ข้อมูลที่สามารถนำมาใช้ทำดาต้าไมนิ่งนั้นมีได้หลายประเภท เช่น ฐานข้อมูลที่จัดเก็บอยู่ในรูปแบบของตาราง (Relational Database) ฐานข้อมูลที่มีการเก็บรวบรวมข้อมูลจากหลายๆแห่งมาไว้

ในรูปแบบเดียวกันซึ่งเรียกว่า "คลังข้อมูล" (Data Warehouse) เป็นต้น และลักษณะเฉพาะของข้อมูลที่จะนำมาทำดาต้าไมนิ่งจะมีลักษณะดังนี้

1) ข้อมูลมีขนาดใหญ่เกินกว่าที่จะพิจารณาความสัมพันธ์ด้วยตาเปล่า

2) ข้อมูลที่ไม่มีการเปลี่ยนแปลงตลอดช่วงเวลาที่ทำการไมนิ่ง เนื่องจากถ้าข้อมูลมีการเปลี่ยนแปลงตลอดเวลาจะทำให้ผลที่ได้จากการทำไมนิ่งใช้ได้ในช่วงเวลาหนึ่งเท่านั้น ดังนั้นถ้าข้อมูลมีการเปลี่ยนแปลงตลอดเวลาจะต้องแก้ไขปัญหานี้ก่อน โดยอาจจะทำการบันทึกไว้ในฐานข้อมูลหนึ่งและนำฐานข้อมูลนั้นมาทำไมนิ่ง

ขั้นตอนที่ 2 การเตรียมข้อมูลก่อนประมวลผล (Data Preprocessing)

การเตรียมข้อมูลก่อนนำไปประมวลผลจะทำให้ข้อมูลที่ได้มีคุณภาพ และถูกต้องมากขึ้น เนื่องจากข้อมูลที่มาจากขั้นตอนการเลือกข้อมูลนั้นอาจยังมีข้อผิดพลาดอยู่ โดยวิธีในการเตรียมข้อมูลก่อนประมวลผลสามารถทำได้หลายวิธีขึ้นอยู่กับข้อมูลที่เลือกมา ข้อมูลบางประเภทที่เลือกมาอาจจะทำแค่บางวิธี หรืออาจจะต้องทำทุกวิธีเลยก็เป็นไปได้ ซึ่งวิธีเตรียมข้อมูลก่อนการประมวลผลมีดังนี้

1) การทำความสะอาดข้อมูล (Data cleaning)

วิธีการทำความสะอาดข้อมูลจะใช้ในการแก้ไขปัญหาข้อมูลที่ขาดหายไป (Missing data) ปัญหาข้อมูลมีความคลาดเคลื่อน (Noisy data) และปัญหาความไม่สอดคล้องกันของข้อมูล

- Missing data คือการที่ข้อมูลบาง attribute ในบางเรคอร์ดได้ขาดหายไป ซึ่งวิธีแก้ปัญหานี้ก็มีหลายวิธี เช่น ถ้าข้อมูลเกิดขาดหายไปประมาณ 20-30 เปอร์เซนต์และ attribute นั้นไม่ค่อยมีความจำเป็นอาจจะใช้วิธีการตัด attribute นั้นออกไปเลย แต่ถ้าข้อมูลนั้นมีความจำเป็นอาจจะต้องใช้วิธีการเติมค่าที่ขาดหายไป โดยการเติมค่าที่ขาดหายไปนั้นอาจจะเติมค่าทั่วๆไป เช่น ค่า unknown หรืออาจจะเติมด้วยค่าเฉลี่ยของ attribute นั้นก็ได้

- Noisy data คือมีข้อมูลมีความคลาดเคลื่อนไป ความคลาดเคลื่อนที่เกิดขึ้นนี้อาจเกิดจากหลายสาเหตุ เช่น เกิดจากการเก็บข้อมูลผิดพลาด ตัวอย่างเช่นต้องการเก็บข้อมูลอายุพนักงาน 40 ปี แต่ใส่ข้อมูลผิดเป็น 400 ปี กรณีนี้จะส่งผลให้ข้อมูลคลาดเคลื่อนไป โดยที่ค่าคลาดเคลื่อนนี้เรียกว่า Outlier ดังนั้นจึงต้องทำการแก้ไขข้อมูลที่ผิดพลาดเหล่านี้ให้ถูกต้อง โดยใช้วิธีการจัดกลุ่ม การทำ Regression หรือวิธีการ Binning เพื่อทำการตัดค่าที่ผิดพลาดนั้นทิ้งไป

2) การรวมข้อมูล (Data integration)

ในบางครั้งจำเป็นที่จะต้องนำข้อมูลมาจากหลายๆแห่ง ดังนั้นจึงต้องมีการรวมข้อมูลเพื่อให้สะดวกต่อการทำไมนิ่ง แต่การรวมกันของข้อมูลก็ก่อให้เกิดปัญหาได้ โดยปัญหาที่เกิดขึ้นมีหลายรูปแบบ เช่น ข้อมูลเป็น attribute เดียวกันแต่ใช้ชื่อ attribute ต่างกัน ข้อมูลมีชื่อ attribute เดียวกันแต่ตัวข้อมูลต่างกัน หรือข้อมูลเมื่อมารวมกันแล้วทำให้ข้อมูลเกิดความซ้ำซ้อน ซึ่งวิธีแก้ปัญหาเหล่านี้คือ ก่อนที่จะทำการรวมข้อมูลให้ดูที่ meta data ประกอบเพื่อป้องกันมิให้เกิดปัญหาที่กล่าวมา

3) การลดจำนวนข้อมูล (Data reduction)

ข้อมูลที่จะนำมาทำไมนิ่งนั้นส่วนใหญ่จะได้มาจากคลังข้อมูล ซึ่งข้อมูลที่อยู่ในคลังข้อมูลนั้นมีขนาดใหญ่และมีความซับซ้อนมาก ดังนั้นถ้านำข้อมูลจากคลังข้อมูลมาทำดาต้าไมนิ่งทั้งหมดจะทำให้เสียเวลาในการทำไมนิ่งเป็นอย่างมาก จึงจำเป็นที่จะต้องทำการลดจำนวนข้อมูล ซึ่งสามารถทำได้ 2 แบบคือ

3.1) ลดจำนวน attribute (ลดตามแนว column) วิธีการลดจำนวน attribute นั้นสามารถทำได้ 3 วิธี

- วิธีที่ 1 *(step-wise forward selection)* : เริ่มจาก 1 attribute แล้วค่อยๆเพิ่มข้อมูลเข้าไปทีละ attribute ไปเรื่อยจนกว่าค่า error จะเกินกว่าที่จะยอมรับได้
- วิธีที่ 2 *(step-wise backward selection)* : เริ่มจากทุก attribute แล้วค่อยๆตัดออกทีละ attribute จนกระทั่งค่า error จะเกินกว่าที่จะยอมรับได้
- วิธีที่ 3 *(decision-tree induction)* : ใช้ decision tree ในการทำนายว่า attribute ใด ไม่จำเป็นต้องใช้ แล้วจึงตัด attribute นั้นออก

3.2) ลดปริมาณข้อมูล (ลดตามแนว row) วิธีการลดปริมาณข้อมูลจะทำโดยใช้วิธี Sampling ข้อมูลขึ้นมา

ขั้นตอนที่ 3 การแปลงข้อมูล (Data Transformation)

การแปลงข้อมูลมีวัตถุประสงค์ 2 อย่างคือ ทำให้ไมนิ่งมีประสิทธิภาพมากขึ้นและทำให้รูปแบบของข้อมูลสอดคล้องกับโมเดลที่จะนำมาใช้ เนื่องจากข้อมูลที่จะนำมาใช้ทำดาต้าไมนิ่งในบางครั้งอยู่ในรูปแบบที่ไม่เหมาะสมกับอัลกอริทึมที่เลือกใช้ ดังนั้นจึงจำเป็นที่จะต้องทำการแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมกับอัลกอริทึมนั้นๆก่อน โดยวิธีการแปลงข้อมูลมีอยู่หลายวิธีซึ่งขึ้นอยู่กับปัญหาของข้อมูล

1) วิธี Normalization : เป็นวิธีที่แปลงข้อมูลให้อยู่ในช่วงๆหนึ่ง เช่น

Min-max normalization มีสูตรการคำนวณดังนี้

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A$$

v' = ค่าข้อมูลที่ได้หลังจากการแปลง

v = ข้อมูลที่จะนำมาทำการแปลง

$\min_A$ = ค่าต่ำสุดของข้อมูลใน attribute A

$\max_A$ = ค่าสูงสุดของข้อมูลใน attribute A

$new\_\min_A$ = ค่าต่ำสุดของข้อมูลที่ต้องการทำการแปลงข้อมูลของ attribute A

$new\_\max_A$ = ค่าสูงสุดของข้อมูลที่ต้องการทำการแปลงข้อมูลของ attribute A

2) วิธี Discretization : เป็นวิธีที่แปลงข้อมูลที่ต่อเนื่องให้เป็นข้อมูลที่ไม่ต่อเนื่อง เช่น อุณหภูมิเป็นข้อมูลที่ต่อเนื่อง เราอาจจะจัดแบ่งเป็นช่วงๆ คือ ช่วง 0-20 องศาเซลเซียส เป็นช่วงอากาศเย็น ช่วง 21-30 องศาเซลเซียส เป็นช่วงอากาศอุ่น ถ้าอุณหภูมิเป็น 20.1 องศาเซลเซียสจะถูกปัดเป็น 21 และถูกจัดให้อยู่ในกลุ่มอากาศอุ่น ซึ่งความจริงแล้ว 20.1 องศาเซลเซียสไม่ต่างกับ 20 องศาเซลเซียส ควรจัดอยู่ในกลุ่มอากาศเย็นมากกว่า ดังนั้นการแก้ไขปัญหานี้สามารถทำได้โดยการแบ่งช่วงให้ละเอียดมากขึ้นแต่ก็ไม่ควรที่จะละเอียดเกินไป

3) วิธี Generalization : เป็นวิธีที่แปลงข้อมูลโดยมองเป็นภาพรวม ตัวอย่างเช่น จัดกลุ่มถนนเป็นเขต จัดกลุ่มเขตเป็นจังหวัด จัดกลุ่มจังหวัดเป็นประเทศ เป็นต้น

4) วิธี Attribute/Feature construction : เป็นวิธีแปลงข้อมูลโดยการสร้างข้อมูลใหม่จากข้อมูลเดิม เช่น พื้นที่หาจากกว้าง x ยาว

2.2.3 การทำไมนิ่ง (Data Mining Operation)

ขั้นตอนการทำไมนิ่งนี้เป็นการประมวลผลข้อมูลตามอัลกอริทึมที่ได้กำหนดไว้ ในขั้นตอนนี้จะมีความสัมพันธ์กับการวิเคราะห์ข้อมูล โดยการทำไมนิ่งนั้นมีโมเดลอยู่หลายแบบซึ่งการเลือกใช้โมเดลขึ้นอยู่กับวัตถุประสงค์ในการทำไมนิ่ง ซึ่งโมเดลในการทำไมนิ่งมีดังต่อไปนี้

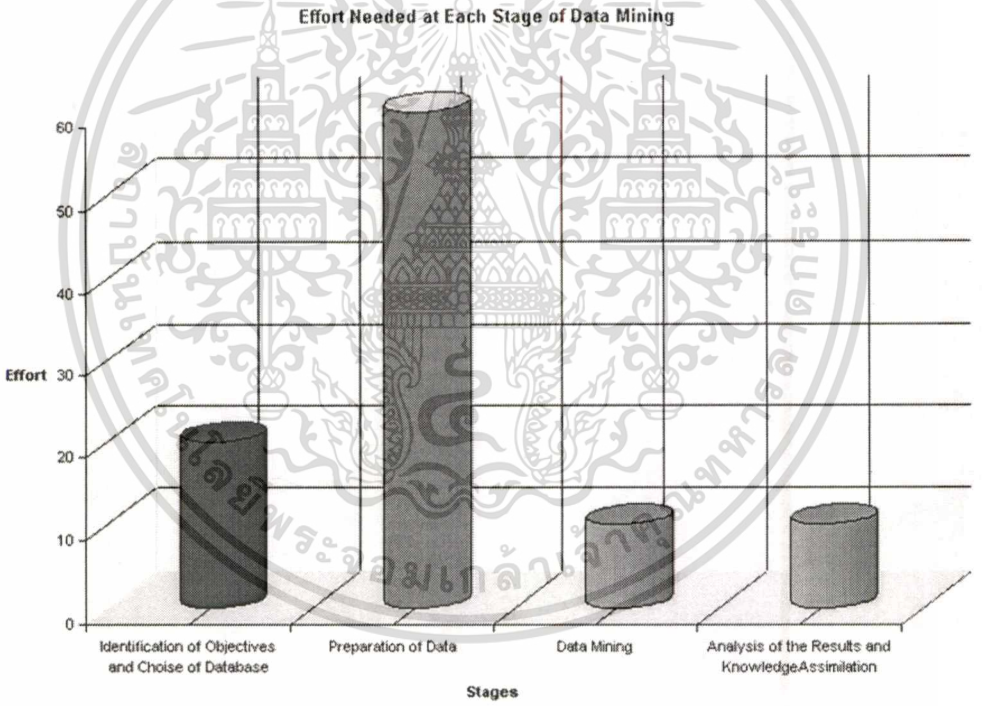
1) Predictive Modeling เป็นโมเดลที่ใช้ในการสร้างแบบจำลองพยากรณ์ โดยจะมีลักษณะที่คล้ายกับการเรียนรู้ของมนุษย์ คือใช้การสังเกตเพื่อที่จะสร้างแบบจำลองของคุณลักษณะที่สำคัญของปรากฏการณ์บางอย่าง โดยข้อมูลที่มีความถูกต้องและสมบูรณ์ จะทำให้แบบจำลองสามารถทำนายผลได้อย่างถูกต้อง ตัวอย่างเช่น ทำนายยอดขายของเดือนถัดไปจากข้อมูลที่มีอยู่

2) Database Segmentation หรือ Database Clustering เป็นโมเดลที่ใช้ในการแบ่งกลุ่มข้อมูล โดยการแบ่งกลุ่มข้อมูลจะแบ่งตามลักษณะที่เหมือนกันของข้อมูล ส่วนใหญ่การแบ่งกลุ่มข้อมูลมัก

ใช้กับข้อมูลลูกค้า หรือกลุ่มตลาดแหมาย ตัวอย่างเช่น เป็นกลุ่มลูกค้าตามรายได้ของลูกค้า หรือแบ่งกลุ่มลูกค้าตามลักษณะการชำระเงินของลูกค้า เป็นต้น

3) Link Analysis เป็นโมเดลที่ใช้วิเคราะห์หาความสัมพันธ์ (Association) ระหว่างข้อมูลว่าข้อมูลแต่ละรายการมีความสัมพันธ์กันอย่างไร ตัวอย่างเช่น ต้องการหาความสัมพันธ์ของสินค้าที่ลูกค้ามักจะซื้อร่วมกัน หรือเมื่อลูกค้าซื้อสินค้าประเภทนี้แล้วจะต้องซื้อสินค้าอีกประเภทหนึ่งต่อเนื่องกัน

4) Deviation Detection เป็นวิธีที่หาค่าที่แตกต่างไปจากค่ามาตรฐาน โดยทั่วไปมักใช้วิธีการทางสถิติ หรืออาศัยการวาดกราฟ แล้วดูการกระจายของข้อมูลว่ามีการกระจายออกไปจากกลุ่มหรือไม่ มักใช้ในการตรวจจับสิ่งผิดปกติต่างๆ เช่น การจับการโกง เป็นต้น



รูปที่ 2.2 แสดงเปอร์เซ็นต์ที่ใช้ในการทำดาต้าไมนิ่งแต่ละขั้นตอน

2.2.4 การวิเคราะห์ผลที่ได้จากทำดาต้าไมนิ่งและการนำความรู้มาใช้ (Analysis of Results and Assimilation of Knowledge)

การวิเคราะห์ผลที่ได้จากการทำไมนิ่งและการนำความรู้ไปใช้ เป็นขั้นตอนสุดท้ายในการทำดาต้าไมนิ่ง นักวิเคราะห์จะต้องนำผลที่ได้จากการไมนิ่งมาตีความหมายและสรุปผล เพื่อนำไปเป็น

สารสนเทศที่ช่วยในการตัดสินใจ ถ้านักวิเคราะห์เห็นว่าผลที่ได้ไม่เป็นไปตามวัตถุประสงค์ที่วางไว้ สามารถย้อนกลับไปแก้ไขในขั้นตอนใดๆได้

2.3 โมเดลการแบ่งกลุ่มข้อมูล (Data Mining Operation : Database Segmentation)

วัตถุประสงค์หลักในการทำ Database Segmentation คือการแบ่งส่วนข้อมูลที่มีลักษณะคล้ายคลึงกันในฐานข้อมูลไว้เป็นกลุ่มเดียวกัน โดยในตอนเริ่มต้นเราจะไม่รู้ว่าจะแบ่งข้อมูลออกเป็นกี่กลุ่ม นั่นคือข้อมูลเหล่านี้จะมีคุณสมบัติหนึ่งที่เหมือนกัน และจะถูกพิจารณาเป็นข้อมูลกลุ่มเดียวกัน โดยคุณสมบัติที่เหมือนกันหมายถึงข้อมูลต่างๆในกลุ่ม (Segment) เดียวกันจะมีลักษณะใกล้เคียงกัน ซึ่งลักษณะที่ใกล้เคียงกันนี้สามารถวัดได้จากความแตกต่างของข้อมูลกับจุดศูนย์กลางกลุ่ม

ลักษณะแอปพลิเคชันโดยทั่วไปที่ใช้การแบ่งกลุ่มข้อมูล

- พวกเครื่องมือที่มีลักษณะ Stand-alone ใช้ในการแบ่งกลุ่มข้อมูล เช่น พวกข้อมูลลูกค้า
- ใช้ในการทำ data preparation ให้กับอัลกอริทึมอื่นๆ เช่นทำเพื่อลด Noisy data
- ใช้ในการทำ Pattern Recognition
- ใช้ในการทำ Image Processing
- ใช้กับเรื่องที่เกี่ยวกับเศรษฐกิจ เช่น การวิจัยตลาด
- และใช้กับเว็บไซต์ต่างๆ เช่น จัดกลุ่มรูปแบบการเข้าถึงเว็บไซต์ที่เหมือนๆกัน

วิธีในการแบ่งกลุ่มข้อมูลนั้นมีหลายวิธี

1) Partitioning algorithms

การทำงานของอัลกอริทึมในลักษณะนี้คือ เริ่มต้นจากการกำหนดกลุ่มที่ต้องการจะแบ่งโดยไม่จำเป็นว่าข้อมูลจะคล้ายกัน จากนั้นจึงนำข้อมูลแต่ละตัวมาลองทดสอบย้ายกลุ่มไปแต่ละกลุ่มและพิจารณาค่าแตกต่างจากจุดศูนย์กลาง แล้วให้ข้อมูลตัวนั้นอยู่ที่กลุ่มที่ให้ค่าแตกต่างจากจุดศูนย์กลางน้อยที่สุด ซึ่งวิธีนี้เป็นวิธีที่นิยมมากที่สุด ในวิธี Partitioning นี้ยังมีอัลกอริทึมให้เลือกใช้อีกหลายอัลกอริทึม เช่น K-Means, K-Modes และ K-Prototypes เป็นต้น อัลกอริทึม K-Means นั้นเหมาะกับข้อมูลประเภทตัวเลข (numeric) อย่างเดียว ส่วน K-Modes เหมาะกับข้อมูลประเภทที่ไม่ใช่ตัวเลข (categorical) ตัวอย่างเช่นเพศ สถานะ การแต่งงาน โดยอัลกอริทึม K-Modes จะแทนค่า means ของกลุ่มข้อมูล (cluster) ด้วยค่าฐานนิยม (modes) และสุดท้ายอัลกอริทึม K-Prototypes ใช้ได้กับข้อมูลทั้งแบบ numeric และข้อมูลแบบ categorical ซึ่งเป็นการรวมอัลกอริทึมระหว่าง K-Means และ K-Modes เข้าด้วยกัน ซึ่งอัลกอริทึม K-Prototypes มักนิยมใช้กับข้อมูลในฐานข้อมูลจริงเพราะในฐานข้อมูลจริง ส่วนใหญ่เป็นข้อมูลทั้งแบบ numeric และแบบ categorical

2) Hierarchy algorithms

เป็นการจัดกลุ่มข้อมูลโดยสร้างเป็นชั้นๆแบบลำดับขั้น ซึ่งวิธีนี้ไม่เป็นที่นิยม เพราะมีค่าใช้จ่ายในการคำนวณสูงและเสียเวลา ส่วนใหญ่จะใช้กับข้อมูลที่ไม่ใช่ตัวเลข เช่น สัญลักษณ์

3) Neural Network

Neural Network เป็นเทคโนโลยีที่มีที่มาจากงานวิจัยด้านปัญญาประดิษฐ์ Artificial Intelligence:AI เพื่อใช้ในการคำนวณค่าฟังก์ชันจากกลุ่มข้อมูลและมีพื้นฐานมาจากสมองของมนุษย์ โดยหลักการทำงานของ Neural Network จะมี 2 ขั้นตอนหลัก คือ ขั้นตอนการเรียนรู้ (Training) และขั้นตอนการนำไปใช้งาน (Deploying)

โดยวิธีการจะทำให้เครื่องเรียนรู้จากตัวอย่างต้นแบบ แล้วฝึก (train) ให้ระบบได้รู้จักที่จะคิดแก้ปัญหาที่กว้างขึ้นได้ ในโครงสร้างของนิวรอลเน็ตจะประกอบด้วยโหนด (node) สำหรับ Input – Output และการประมวลผล กระจายอยู่ในโครงสร้างเป็นชั้น ๆ ได้แก่ input layer , output layer และ hidden layer การประมวลผลของนิวรอลเน็ตจะอาศัยการส่งการทำงานผ่านโหนดต่าง ๆ ใน layer เหล่านี้

2.4 อัลกอริทึม K-Prototypes

K-Protoypes เป็นอัลกอริทึมหนึ่งในวิธี Partitioning ที่สามารถใช้กับข้อมูลที่เป็นทั้งแบบ numeric และแบบ categorical ซึ่งมีพื้นฐานมาจากอัลกอริทึม k-means แต่ต่างกันตรงที่อัลกอริทึม K-Means นั้นสามารถใช้กับข้อมูลแบบ numeric ได้แบบเดียวเท่านั้น

อัลกอริทึม K-Prototypes คล้ายกับอัลกอริทึม K-Means เพราะข้อมูลจะถูกแบ่งกลุ่มตามจุดศูนย์กลางของกลุ่มข้อมูล (k prototypes) แทนที่จะใช้ค่าเฉลี่ยของกลุ่มข้อมูล ดังนั้นเราจึงเรียกอัลกอริทึมนี้ว่า "K-Prototypes" และได้มีการพัฒนาวิธีการปรับเปลี่ยนค่าจุดศูนย์กลางของกลุ่มข้อมูล เพื่อที่จะทำให้ข้อมูลภายในกลุ่มข้อมูลเดียวกันมีความคล้ายกันมากที่สุด ซึ่งการวัดความคล้ายกันของข้อมูลจะได้มาจากแบบ numeric และแบบ categorical

2.4.1 หลักคณิตศาสตร์เบื้องต้นที่ใช้ใน K-Prototypes

ให้ $X = \{X_1, X_2, \ldots, X_n\}$ แทนเซตของข้อมูล n ตัว และ $X_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\}$ แทนข้อมูลที่นำมาจัดกลุ่มโดยที่ m คือค่าจำนวน attribute ของข้อมูล

วัตถุประสงค์ของการจัดกลุ่มข้อมูล X คือการแบ่งแยกข้อมูลใน X ออกเป็น k กลุ่มโดยที่ k เป็นจำนวนเต็มบวก สำหรับ n คือจำนวนในการแบ่งกลุ่มที่เป็นไปได้ซึ่งมีจำนวนมาก วิธีที่ใช้เป็นแนวทางในการแบ่งกลุ่มข้อมูลนั้นใช้หลักการแบ่งข้อมูลที่เรียกว่า "Cost Function"

การคำนวณค่า Cost Function

โดยทั่วไปมักใช้ค่า Cost function ในการวัดค่าความเหมือนกันของข้อมูล ซึ่งทางหนึ่งที่ใช้กำหนด Cost function คือ

$$E = \sum_{l=1}^{k}\sum_{i=1}^{n} y_{il} d(X_i, Q_l) \tag{2.1}$$

โดยที่ $Q_l = [q_{l1}, q_{l2}, \ldots, q_{lm}]$ คือจุดศูนย์กลางของกลุ่มข้อมูล l และ y_{il} คือสมาชิกของ partition matrix Y_{nxl} ส่วน d คือหน่วยที่ใช้วัดความคล้ายกันของข้อมูลซึ่งจะหาได้จากการคำนวณแบบ square Euclidean distance

Y จะมีคุณสมบัติ 2 ข้อคือ $0 \le y_{il} \le 1$ และ $\sum_{l=1}^{k} y_{il} = 1$ ซึ่งถ้า $y_{il} \in \{0,1\}$ จะเรียก Y ว่าเป็น hard partition นอกเหนือจากนี้จะเรียกว่า fuzzy partition ที่ hard partition ค่า $y_{il} = 1$ ซึ่งเป็นการระบุว่าข้อมูล X_i ถูกกำหนดให้อยู่ในกลุ่มข้อมูล l แต่เราจะพิจารณาเพียง hard partition เท่านั้น

จากสมการที่ (2.1) ค่า $E_l = \sum_{i=1}^{n} y_{il} d(X_i, Q_l)$ คือค่า cost ทั้งหมดของข้อมูล X ที่อยู่ในกลุ่มข้อมูล l ตัวอย่างเช่นการกระจายของข้อมูลในกลุ่มข้อมูล l จากจุดศูนย์กลางของกลุ่มข้อมูล (Q_l) ซึ่งค่า E_l จะมีค่าต่ำสุดก็ต่อเมื่อ

$$q_{lj} = \frac{1}{n_l}\sum_{i=1}^{n} y_{il} x_{ij} \tag{2.2}$$

โดยที่ $j = 1, \ldots, m$ และ $n_l = \sum_{i=1}^{n} y_{il}$ เป็นจำนวนข้อมูลในกลุ่มข้อมูล l

ถ้าข้อมูล X มี attribute แบบ categorical แล้วจะสามารถวัดความคล้ายกันได้โดย

$$d(X_i, Q_l) = \sum_{j=1}^{m_r} \left(x_{ij}^r - q_{lj}^r\right)^2 + \gamma_l \sum_{j=1}^{m_c} \delta\left(x_{ij}^c, q_{lj}^c\right) \tag{2.3}$$

โดย $\delta(p,q) = 0$ แล้ว $p = q$

$\delta(p,q)=1$ แล้ว $p \neq q$

x_{ij}^r เป็นค่า numeric attributes ของข้อมูล i และ q_{ij}^r เป็นจุดศูนย์กลางของกลุ่มข้อมูล l

x_{ij}^c เป็นค่า categorical attributes ของข้อมูล i และ q_{ij}^c เป็นจุดศูนย์กลางของกลุ่มข้อมูล l

m_r เป็นจำนวน attribute ของ numeric

m_c เป็นจำนวน attribute ของ categorical

γ_l เป็น weight ของ categorical attribute ของกลุ่มข้อมูล l

ดังนั้นสามารถเขียน E_l ใหม่ได้เป็น

$$E_l = \sum_{i=1}^{n} y_{il} \sum_{j=1}^{m_r} \left(x_{ij}^r - q_{lj}^r\right)^2 + \gamma_l \sum_{i=1}^{n} y_{il} \sum_{j=1}^{m_c} \delta\left(x_{ij}^c, q_{lj}^c\right)$$

$$= E_l^r + E_l^c \tag{2.4}$$

โดย E_l^r เป็นค่า cost ทั้งหมดของ numerical attribute ของข้อมูลที่อยู่ในกลุ่มข้อมูล l ซึ่งจะเป็นค่าต่ำสุด ถ้า q_{ij}^r หามาจากสมการที่ (2.2)

ให้ C_j เป็นเซตที่ประกอบไปด้วยค่าที่เป็นหนึ่งเดียวใน categorical attribute j และ $p\left(c_j \in C_j | l\right)$ เป็นความน่าจะเป็นของค่า c_j ที่จะปรากฏในกลุ่มข้อมูล l ซึ่ง E_l^c ในสมการที่ (2.4) สามารถเขียนใหม่ได้เป็น

$$E_l^c = \gamma_l \sum_{j=1}^{m_c} n_l \left(1 - p\left(q_{lj}^c \in C_j | l\right)\right) \tag{2.5}$$

โดย n_l เป็นจำนวนของข้อมูลในกลุ่มข้อมูล l ซึ่งสามารถหาค่าต่ำสุดของ E_l^c ได้จากบทแทรกที่ 1

บทแทรกที่1: สำหรับกลุ่มข้อมูล l แล้วค่า E_l^c มีค่าต่ำสุดก็ต่อเมื่อ $p\left(q_{lj}^c \in C_j | l\right) \geq p\left(c_j \in C_j | l\right)$ โดย $q_{lj}^c \neq c_j$ สำหรับทุก categorical attributes

ดังนั้นเราสามารถเขียน E ใหม่ได้เป็น

$$E = \sum_{l=1}^{k} \left(E_l^r + E_l^c\right) = \sum_{l=1}^{k} E_l^r + \sum_{l=1}^{k} E_l^c$$

$$= E^r + E^c \tag{2.6}$$

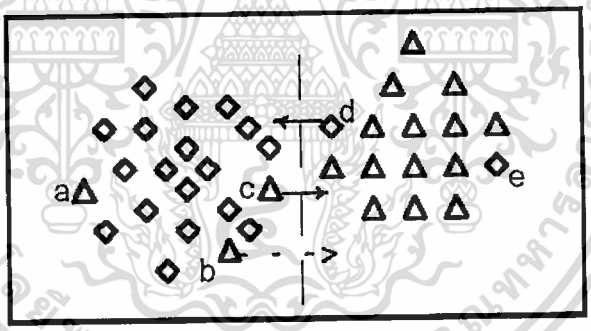
สมการที่ (2.6) เป็น cost function ของกลุ่มข้อมูลทั้ง numeric และ categorical ทั้งค่า E^r และ E^c ไม่เป็นค่าติดลบ ค่าต่ำสุดของ E สามารถหาได้จากค่าต่ำสุดของ E^r และ E^c ซึ่ง

- ค่าต่ำสุดของ E^r หาได้จากสมการที่ (2.2)
- ค่าต่ำสุดของ E^c หาได้จากบทแทรกที่ 1

ดังนั้นสมการที่ (2.2) และบทแทรกที่ 1 จะกำหนดแนวทางในการเลือกจุดศูนย์กลางของกลุ่มข้อมูลเพื่อทำให้ค่า cost function ที่สมการที่ (2.6) มีค่าต่ำสุด

2.4.2 การวัดความเหมือนกัน

การวัดความเหมือนกันของข้อมูลแบบ numeric attributes คือการใช้ square Euclidean distance แต่การวัดความคล้ายกันของข้อมูลแบบ categorical attributes คือการดูจำนวนที่ไม่เข้าคู่กันระหว่างข้อมูลกับจุดศูนย์กลางของกลุ่มข้อมูล ผลกระทบจาก weight γ_l ในการแบ่งกลุ่มข้อมูลสามารถอธิบายได้ดังรูปที่ 2.3



รูปที่ 2.3 ผลกระทบจาก weight γ_l ในการแบ่งกลุ่มข้อมูล

จากรูปที่ 2.3 รูปสามเหลี่ยมและรูปเพชรแทนค่าของ categorical attributes และค่าของ numeric attributes จะแสดงให้เห็นจากตำแหน่งของข้อมูล โดยข้อมูลเหล่านี้จะถูกแบ่งเป็นกลุ่มข้อมูล 2 กลุ่ม

ถ้า $\gamma_l = 0$ หมายถึงว่าการแบ่งข้อมูลจะขึ้นอยู่กับ numeric attribute เท่านั้น เช่นตำแหน่งของข้อมูล ซึ่งผลที่จะได้จากการแบ่งกลุ่มข้อมูลคือจะแบ่งกลุ่มออกเป็น 2 กลุ่มซึ่งแยกโดยเส้นประ

ถ้า $\gamma_l > 0$ ข้อมูล c อาจจะเปลี่ยนไปอยู่ที่กลุ่มข้อมูลด้านขวาเพราะอยู่ใกล้กลุ่มข้อมูลด้านขวา ในทำนองเดียวกันข้อมูล d อาจจะเปลี่ยนไปอยู่ที่กลุ่มข้อมูลด้านซ้าย อย่างไรก็ตามข้อมูล a

และข้อมูล e อาจจะยังคงอยู่ที่กลุ่มข้อมูลเดิมเพราะอยู่ไกลจากกลุ่มข้อมูลที่จะย้ายไป ส่วนข้อมูล b นั้นยังไม่แน่ขึ้นอยู่กับค่า γ_l ถ้า γ_l โน้มเอียงไปทางข้อมูลแบบ categorical ข้อมูล b อาจจะเปลี่ยนไปอยู่กลุ่มข้อมูลด้านขวา นอกเหนือจากนี้จะยังคงอยู่ทางกลุ่มด้านซ้าย

2.4.3 การทำงานของอัลกอริทึม K-Prototypes

ขั้นตอนการทำงานของอัลกอริทึม K-Prototypes

1) เลือก k เริ่มต้น เป็นจำนวนกลุ่มข้อมูลที่ต้องการจัดกลุ่มของข้อมูล X
2) จัดข้อมูลแต่ละตัวที่อยู่ใน X ให้อยู่ในกลุ่มข้อมูลหนึ่ง ซึ่งข้อมูลนั้นอยู่ใกล้กับจุดศูนย์กลางของกลุ่มข้อมูลมากที่สุด ซึ่งหาได้จากสมการที่ (2.3) และทำการปรับเปลี่ยนค่าจุดศูนย์กลางของกลุ่มข้อมูลหลังจากจัดข้อมูลเสร็จแล้ว
3) หลังจากที่ทำการจัดข้อมูล ให้อยู่ในกลุ่มข้อมูลหนึ่งๆแล้ว ให้ทำการทดสอบความคล้ายกันของข้อมูลอีกครั้ง กับจุดศูนย์กลาง ถ้าพบว่าข้อมูลนั้นอยู่ใกล้กับจุดศูนย์กลางของกลุ่มข้อมูลในกลุ่มอื่นมากกว่า ให้ทำการย้ายข้อมูลไปยังกลุ่มข้อมูลนั้น และทำการปรับเปลี่ยนจุดศูนย์กลางของกลุ่มข้อมูลทั้ง 2 กลุ่มใหม่อีกครั้ง
4) ทำซ้ำข้อ 3 จนกระทั่งข้อมูลใน X ทุกตัวไม่เปลี่ยนกลุ่ม

จะเห็นได้ว่าอัลกอริทึม K-Prototypes มีการทำงานหลักอยู่ 3 กระบวนการคือ การเลือกกำหนดค่าจำนวนกลุ่มข้อมูลที่จะจัดกลุ่ม จัดกลุ่มข้อมูลเริ่มต้นและจัดกลุ่มข้อมูลใหม่ ซึ่งกระบวนการแรกจะใช้วิธีการสุ่มค่า k ขึ้นมา ส่วนกระบวนการที่สองเริ่มจากการกำหนดค่าจุดศูนย์กลางของกลุ่มข้อมูลจากนั้นจะทำการจัดข้อมูลให้อยู่ในกลุ่มข้อมูล และทำการปรับเปลี่ยนค่าจุดศูนย์กลางกลุ่มข้อมูลหลังจากการจัดกลุ่มข้อมูลและกระบวนการที่ 3 จะคล้ายกับกระบวนการที่ 2 แต่ต่างตรงที่หลังจากทำการจัดกลุ่มข้อมูลใหม่ จะต้องทำการปรับเปลี่ยนจุดศูนย์กลางกลุ่มข้อมูลของกลุ่มข้อมูลก่อนย้ายและหลังย้าย

ในรูปที่ 2.4 และ 2.5 เป็นกระบวนการบางส่วนที่อยู่ในอัลกอริทึม K-Prototypes ซึ่งประกอบไปด้วยตัวแปรและฟังก์ชันดังต่อไปนี้

X[i] = ข้อมูล

X[i,j] = ค่าของ attribute ของจุดศูนย์กลางกลุ่มข้อมูล

O_prototypes[] = เก็บ numeric attribute ของจุดศูนย์กลางกลุ่มข้อมูล

C_prototypes[] = เก็บ categorical attribute ของจุดศูนย์กลางกลุ่มข้อมูล

O_prototypes[i,j] = ข้อมูลแบบ numeric ที่เป็นสมาชิกของกลุ่มข้อมูล i

C_prototypes[i,j] = ข้อมูลแบบ categorical ที่เป็นสมาชิกของกลุ่มข้อมูล i

Distance() = เป็นฟังก์ชันในการหาค่า square Euclidean distance

Sigma() = เป็นฟังก์ชันในการหาค่า $\delta(\ )$ ในสมการที่ (2.3)

Clustership[] = เก็บกลุ่มข้อมูลที่ข้อมูลนั้นอยู่

ClusterCount[] = เก็บจำนวนข้อมูลที่อยู่ในกลุ่มข้อมูล

SumInCluster[] = ผลรวมค่า numeric ของข้อมูลในกลุ่มข้อมูล และใช้ในการปรับเปลี่ยนค่า numeric attribute

FrequencyInCluster[] = เก็บความถี่ของค่าที่เปลี่ยนแปลงของ categorical attribute ในกลุ่มข้อมูล

HighestFreq() = เป็นฟังก์ชันในการหาค่าของบทแทรกที่ 1 เพื่อปรับเปลี่ยน categorical attribute ของกลุ่มข้อมูล

```
FOR i=1 TO NumberOfObjects
    Mindistance=Distance(X[i],O_prototypes[1])+gamma*Sigma(X[i],C_prototypes[1])
    FOR j=1 TO NumberOfClusters
        distance=Distance(X[i],O_prototypes[j])+gamma*Sigma(X[i],C_prototypes[j])
        IF(distance<Mindistance)
            Mindistance=distance
            cluster=j
        ENDIF
    ENDFOR
    Clustership[i]=cluster
    ClusterCount[cluster]+1
    FOR j=1 TO NumberOfNumericAttributes
        SumInCluster[cluster,j]+X[i,j]
        O_prototypes[cluster,j]=
            SumInCluster[cluster,j]/ClusterCount[cluster]
    ENDFOR
    FOR j=1 TO NumberOfCategoricAttributes
        FrequencyInCluster[cluster,j,X[i,j]]+1
        C_prototypes[cluster,j]=
            HighestFreq(FrequencyInCluster,cluster,j)
    ENDFOR
ENDFOR
```

รูปที่ 2.4 กระบวนการเริ่มต้นจัดกลุ่มข้อมูล

รูปที่ 2.4 แสดงอัลกอริทึมในส่วนของกระบวนการเริ่มต้นจัดกลุ่มข้อมูล โดยมีขั้นตอนการทำงานดังนี้

1) กำหนดค่าต่ำสุดของ cost function

2) คำนวณค่า cost function ของข้อมูล i ที่อยู่ในกลุ่มข้อมูล j

3) เปรียบเทียบค่า cost function ที่คำนวณได้ใหม่กับค่า cost function ที่ต่ำสุด โดยที่ถ้าค่า cost function ที่คำนวณได้ใหม่มีค่าน้อยกว่าค่าต่ำสุดของ cost function ให้ค่าต่ำสุดของ cost function มีค่าเท่ากับค่า cost function ที่คำนวณได้ใหม่ และเก็บค่ากลุ่มข้อมูลที่ค่าต่ำสุดนั้นอยู่

4) กลับไปทำข้อ 2 ซ้ำจนกระทั่งทำครบทุกกลุ่มข้อมูล

5) เก็บกลุ่มข้อมูลที่ข้อมูลนั้นอยู่ซึ่งมีค่า cost function ต่ำสุด

6) เก็บค่าจำนวนข้อมูลในกลุ่มข้อมูลเพิ่มขึ้นหนึ่ง

7) วนลูปเพื่อหาผลรวมของค่า numeric ของข้อมูล i ในกลุ่มข้อมูล cluster และหาค่าต่ำสุดของค่า E^r

8) วนลูปเพื่อหา FrequencyInCluster และหาค่าต่ำสุดของ E^c

กลับไปทำข้อ 1 ซ้ำจนกระทั่งทำครบทุกข้อมูล

```
moves=0
FOR i=1 TO NumberOfObjects
    ...
    (To find the cluster whose protoype is the nearest to object i. Same as Figure 4)
    ...
    IF(Clustership[i]<>cluster)
        moves+1
        oldcluster=Clustership[i]
        ClusterCount[cluster]+1
        ClusterCount[oldcluster]-1
        FOR j=1 TO NumberOfNumericAttributes
            SumInCluster[cluster,j]+X[i,j]
            SumInCluster[oldcluster,j]-X[i,j]
            O_prototypes[cluster,j]=
        SumInCluster[cluster,j]/ClusterCount[cluster]
            O_prototypes[oldcluster,j]=
        SumInCluster[oldcluster,j]/ClusterCount[oldcluster]
        ENDFOR
        FOR j=1 TO NumberOfCategoricalAttributes
            FrequencyInCluster[cluster,j,X[i,j]]+1
            FrequencyInCluster[oldcluster,j,X[i,j]]-1
            C_prototypes[cluster,j]=HighestFreq(cluster,j)
            C_prototypes[oldcluster,j]=HighestFreq(oldcluster,j)
        ENDFOR
    ENDIF
ENDFOR
```

รูปที่ 2.5 กระบวนการจัดกลุ่มข้อมูลใหม่

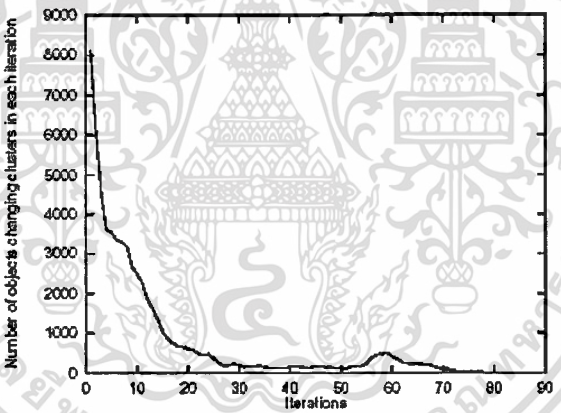
รูปที่ 2.5 แสดงอัลกอริทึมในส่วนของกระบวนการจัดกลุ่มข้อมูลใหม่ โดยมีขั้นตอนการทำงานดังนี้

1) กำหนดค่า moves เริ่มต้นให้เท่ากับศูนย์
2) จัดข้อมูลให้อยู่ในกลุ่มข้อมูล ซึ่งข้อมูลอยู่ใกล้กับจุดศูนย์กลางกลุ่มข้อมูลมากที่สุด (ตามกระบวนการในรูปที่ 2.4)
3) เปรียบเทียบ Clustership[i] ว่าเท่ากับ cluster หรือไม่ถ้าไม่เท่ากัน
   - ให้เพิ่มค่า moves ขึ้น 1

- เพิ่มจำนวนข้อมูลในกลุ่มข้อมูลใหม่ขึ้น 1
- ลดจำนวนข้อมูลในกลุ่มข้อมูลเดิมลง 1
- วนลูปเพื่อหาผลรวมของค่า numeric ของข้อมูล i ในกลุ่มข้อมูล cluster และหาค่าต่ำสุดของค่า E^r ของทั้ง 2 กลุ่มข้อมูล (กลุ่มข้อมูลใหม่และกลุ่มข้อมูลเดิม)
- วนลูปเพื่อหา FrequencyInCluster และหาค่าต่ำสุดของ E^c ของทั้ง 2 กลุ่มข้อมูล (กลุ่มข้อมูลใหม่และกลุ่มข้อมูลเดิม)

4) กลับไปทำข้อ 2 ซ้ำจนกระทั่งครบข้อมูลทุกตัว

อัลกอริทึมจะทำซ้ำจนกว่าข้อมูลจะไม่เปลี่ยนกลุ่ม ในรูปที่ 2.6 แสดงให้เห็นกราฟของอัลกอริทึมนี้ ซึ่งทำการทดสอบกับข้อมูล 75,808 เรคอร์ด จำนวน 20 attribute และแบ่งข้อมูลออกเป็น 64 กลุ่มข้อมูล



รูปที่ 2.6 กราฟการทำงานของอัลกอริทึม K-Prototypes

จากกราฟจะเห็นว่า จำนวนการเปลี่ยนแปลงของข้อมูลจะลดลงอย่างรวดเร็วในช่วงเริ่มต้น และในช่วงท้ายจะมีการเปลี่ยนแปลงเพียงเล็กน้อยก่อนที่จะไม่มีการเปลี่ยนแปลง นั่นคือ ข้อมูลจะมีการเปลี่ยนกลุ่มไปเรื่อยๆจนกระทั่งข้อมูลถูกจัดอยู่ในกลุ่มที่เหมาะสมแล้ว ค่าการเปลี่ยนแปลงของข้อมูลจะเป็นศูนย์

ค่า cost ของอัลกอริทึมนี้มีค่าเท่ากับ $O((t+1)kn)$ โดยที่ค่า n เป็นจำนวนข้อมูลทั้งหมด ส่วนค่า k เป็นจำนวนกลุ่มข้อมูลที่ต้องการจัดกลุ่ม และสุดท้ายค่า t เป็นจำนวนครั้งในทำ

กระบวนการจัดกลุ่มใหม่ซ้ำ โดยทั่วไปค่า k จะต้องน้อยกว่าค่า n มากและค่า t จะไม่เกิน 100 ดังนั้นอัลกอริทึมนี้จะมีประสิทธิภาพอย่างดีกับการจัดกลุ่มข้อมูลขนาดใหญ่

2.4.4 ตัวอย่างการนำข้อมูลมาใช้กับอัลกอริทึม K-Prototypes

ในหัวข้อนี้จะสมมติข้อมูลขึ้นมาเพื่อทำการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม K-Prototypes ในการจัดกลุ่ม ในที่นี้จะสมมติข้อมูลที่มี attribute ทั้งหมด 3 attributes คือรหัสลูกค้า (cus_id), ชื่อบริษัทแม่ของรหัสลูกค้าที่ทำธุรกิจด้วย (company_name) และ วงเงินเครดิตที่กำหนดให้ (creditlimit) โดยข้อมูล cus_id และ company_name จะเป็นข้อมูลประเภท categorical ส่วนข้อมูล creditlimit เป็นข้อมูลประเภท numeric ซึ่งจะแสดงดังตารางที่ 2.1

ตารางที่ 2.1 ประเภทของข้อมูลที่จะนำมาจัดกลุ่ม

ชื่อข้อมูล	ประเภทของข้อมูล
cus_id	char(8)
company_name	char(50)
creditlimit	int

ตารางที่ 2.2 ข้อมูลที่จะนำมาจัดกลุ่ม

cus_id	company_name	credit (หมื่นบาท)
00000101	Greater Pharma	10
00000101	GP Chemical	20
00000101	SDrug	15
00000150	Greater Pharma	15
00000150	SDrug	20
00000216	GP Chemical	10
00000216	SDrug	15

cus_id	company_name	credit (หมื่นบาท)
00000354	Greater Pharma	10
00000460	GP Chemical	20

ขั้นตอนในการจัดกลุ่มข้อมูลในตารางที่ 2.2 มีดังนี้

1) กำหนดกลุ่มข้อมูล ในที่นี้จะจัดกลุ่มข้อมูลออกเป็น 3 กลุ่ม
2) จัดข้อมูลให้อยู่ในแต่ละกลุ่ม จะได้ดังตารางที่ 2.3, 2.4 และ 2.5

ตารางที่ 2.3 ข้อมูลกลุ่มที่ 1

cus_id	company_name	credit (หมื่นบาท)
00000101	**Greater Pharma**	**10**
00000150	SDrug	20
00000354	Greater Pharma	10

ตารางที่ 2.4 ข้อมูลกลุ่มที่ 2

cus_id	company_name	credit (หมื่นบาท)
00000101	**GP Chemical**	**20**
00000216	GP Chemical	10
00000460	GP Chemical	20

ตารางที่ 2.5 ข้อมูลกลุ่มที่ 3

cus_id	company_name	credit (หมื่นบาท)
00000101	SDrug	15
00000150	Greater Pharma	15
00000216	**SDrug**	**15**

3) กำหนดจุดศูนย์กลางกลุ่มข้อมูลในแต่ละกลุ่ม ซึ่งกำหนดได้เป็นดังนี้ (ที่ตาราง 2.3, 2.4, 2.5 บรรทัดที่เป็นตัวหนาหมายถึงจุดศูนย์กลางของกลุ่มข้อมูล)

   กลุ่มที่ 1 คือ cus_id = 00000101, company_name = Greater Pharma

   กลุ่มที่ 2 คือ cus_id = 00000101, company_name = GP Chemical

   กลุ่มที่ 3 คือ cus_id = 00000216, company_name = SDrug

4) นำข้อมูลแต่ละตัวมาหาค่า distance ในแต่ละกลุ่มข้อมูล ในที่นี้จะยกตัวอย่างให้เห็นโดยสมมติว่าจะทำการจัดกลุ่มข้อมูลจากตารางที่ 2.4 โดยเลือกข้อมูลที่ cus_id = 00000216 และ company_name = GP Chemical ดังนั้นจะต้องทำการหาค่า distance จากตัวข้อมูลกับจุดศูนย์กลางของกลุ่มข้อมูลในแต่ละกลุ่ม เพื่อเลือกว่าควรจะจัดข้อมูลนี้ไว้ที่กลุ่มใดมากที่สุด โดยจะกำหนดให้

   - ค่า weight ของ cus_id เท่ากับ 0.5
   - ค่า weight ของ company_name เท่ากับ 0.3

เริ่มพิจารณาทีละกลุ่ม

กลุ่มที่ 1 หาค่า distance ของข้อมูล

$$\begin{bmatrix} 00000101 \\ Makro \\ 10 \end{bmatrix} \text{ กับ } \begin{bmatrix} 00000216 \\ Carrefour \\ 10 \end{bmatrix}$$

ซึ่งสามารถหาค่าได้จากสมการที่ (2.3) จะได้เป็น

$$\text{Distance} = (10-10)^2 + 0.5(1) + 0.3(1)$$

$$= 0 + 0.5 + 0.3$$

$= 0.8$

<u>กลุ่มที่ 2</u> หาค่า distance ของข้อมูล

$$\begin{bmatrix} 00000101 \\ Carrefour \\ 20 \end{bmatrix} \text{ กับ } \begin{bmatrix} 00000216 \\ Carrefour \\ 10 \end{bmatrix}$$

Distance $= (20-10)^2 + 0.5(1) + 0.3(0)$

$= 100 + 0.5 + 0$

$= 100.5$

<u>กลุ่มที่ 3</u> หาค่า distance ของข้อมูล

$$\begin{bmatrix} 00000216 \\ Lotus \\ 15 \end{bmatrix} \text{ กับ } \begin{bmatrix} 00000216 \\ Carrefour \\ 10 \end{bmatrix}$$

Distance $= (15-10)^2 + 0.5(0) + 0.3(1)$

$= 25 + 0 + 0.3$

$= 25.3$

จากการคำนวณค่า distance ในแต่ละกลุ่มจะเห็นว่าถ้าจัดข้อมูลไว้ที่กลุ่มที่ 1 จะหาค่า distance ได้ค่าต่ำที่สุด ดังนั้นจึงจัดข้อมูลไว้ที่กลุ่มที่ 1 ซึ่งข้อมูลที่จัดได้จะแสดงในตารางที่ 2.6

ตารางที่ 2.6 ข้อมูลกลุ่มที่ 1 ที่ได้หลังจากการคำนวณ

cus_id	company_name	credit (หมื่นบาท)
00000101	Greater Pharma	10
00000150	SDrug	20
00000216	GP Chemical	10
00000354	Greater Pharma	10

ตารางที่ 2.7 ข้อมูลกลุ่มที่ 2 ที่ได้หลังจากการคำนวณ

cus_id	company_name	credit (หมื่นบาท)
00000101	GP Chemical	20
00000460	GP Chemical	20

5) ทำการคำนวณหาค่าจุดศูนย์กลางของกลุ่มที่ 1 และกลุ่มที่ 2 ใหม่ โดยที่ถ้าเป็นข้อมูลประเภท Numeric ให้ใช้วิธีการหาค่าเฉลี่ย และถ้าเป็นข้อมูลประเภท Categorical จะใช้วิธีฐานนิยม ดังนั้นในกลุ่มที่ 1 และกลุ่มที่ 2 จะหาจุดศูนย์กลางกลุ่มข้อมูลใหม่ได้เป็น

<u>กลุ่มที่ 1</u> :

- cus_id เลือกตัวไหนก็ได้เพราะมีค่าฐานนิยมเป็น 1 ทั้งหมด
- company_name เลือก Greater Pharma เพราะมีค่าฐานนิยมมากที่สุด คือเท่ากับ 2
- creditlimit ใช้วิธีหาค่าเฉลี่ยจะได้เป็น $\frac{10+20+10+10}{4} = 10.25$

$\therefore$ จุดศูนย์กลางกลุ่มที่ 1 = $\begin{bmatrix} 00000101 \\ Makro \\ 10.25 \end{bmatrix}$

<u>กลุ่มที่ 2</u> :

- cus_id เลือกตัวไหนก็ได้เพราะมีค่าฐานนิยมเป็น 1 ทั้ง 2 ตัว
- company_name = GP Chemical
- creditlimit ใช้วิธีหาค่าเฉลี่ยได้เป็น $\frac{20+20}{2} = 20$

$\therefore$ จุดศูนย์กลางกลุ่มที่ 2 = $\begin{bmatrix} 00000101 \\ Carrefour \\ 20 \end{bmatrix}$

บทที่ 3

การออกแบบระบบ

3.1 ระบบงาน

ในหัวข้อระบบงานนี้จะกล่าวถึงการทำงานของระบบจัดกลุ่มข้อมูลลูกค้าที่จะพัฒนาขึ้น ซึ่งประกอบด้วยส่วนหลักๆ 3 ส่วน คือ ส่วนนำข้อมูลเข้า (Input) ส่วนวิเคราะห์และประมวลผล (Process) และสุดท้ายคือส่วนแสดงผล (Output)

3.1.1 ส่วนนำข้อมูลเข้า

ส่วนนำข้อมูลเข้าจะเป็นขั้นตอนในการนำข้อมูลเข้ามาเพื่อใช้ในการจัดกลุ่ม ซึ่งอยู่ในขั้นตอนการประมวลผล โดยข้อมูลที่จะต้องนำเขามีดังนี้

1) ไฟล์ฐานข้อมูล ที่ประกอบด้วยข้อมูลลูกค้าที่ต้องการจัดกลุ่ม โดยฐานข้อมูลที่เลือกเข้ามานั้นจะต้องเป็นฐานข้อมูล Microsoft Access เท่านั้น
2) ตารางข้อมูล โดยเลือกจากฐานข้อมูลที่เลือกไว้ในข้อ 1
3) ฟิลด์ข้อมูล โดยเลือกจากตารางข้อมูลที่เลือกไว้ในข้อ 2
4) จำนวนกลุ่มที่ต้องการจัดแบ่ง

3.1.2 ส่วนวิเคราะห์และประมวลผล

ขั้นตอนนี้จะเป็นการประมวลผลข้อมูลที่นำเข้ามาจากส่วนนำข้อมูลเข้าโดยการประมวลผลจะแบ่งออกเป็น 2 ส่วนคือส่วนการเตรียมข้อมูลและส่วนการทำไมนิ่ง

<u>ส่วนการเตรียมข้อมูล</u> เป็นขั้นตอนในการเตรียมข้อมูลเพื่อจะนำไปใช้ต่อในขั้นตอนการทำไมนิ่ง โดยขั้นตอนการทำงานมีดังนี้

1) เลือกฟิลด์ข้อมูลที่ต้องการทำการจัดกลุ่ม
2) ทำการ Clean ข้อมูล โดยจะตรวจสอบหาค่าที่ขาดหายไป (Missing value) ถ้าตรวจพบจะทำการแก้ไขให้เรียบร้อย

<u>ส่วนการทำไมนิ่ง</u> การจัดกลุ่มข้อมูลตามอัลกอริทึม K-Prototypes โดยขั้นตอนการประมวลผลมีขั้นตอนดังนี้

1) Random จุดศูนย์กลางกลุ่มข้อมูลขึ้นมาตามจำนวนกลุ่มที่ต้องการจะแบ่ง เช่น ต้องการจัดกลุ่มข้อมูลออกเป็น 3 กลุ่ม จะทำการ Random ข้อมูลขึ้นมา 3 เรคอร์ดเพื่อเป็นจุดศูนย์กลางกลุ่มข้อมูล
2) กำหนดค่า distance ที่ต่ำสุด (mindistance) โดยคำนวณจากการเปรียบเทียบกับจุดศูนย์กลางข้อมูลกลุ่มที่ 1
3) คำนวณหาค่า distance กับจุดศูนย์กลางกลุ่มข้อมูลกลุ่มถัดไป
4) เปรียบเทียบว่ามีค่าน้อยกว่าค่า mindistance หรือไม่
    - ถ้ามีค่าน้อยกว่าให้ค่า mindistance = ค่า distance และเก็บเลขที่กลุ่มที่ให้ค่า distance ที่น้อยกว่า
5) ตรวจสอบว่าเปรียบเทียบกับจุดศูนย์กลางกลุ่มข้อมูลครบทุกกลุ่มหรือยัง
    - ถ้ายังเปรียบเทียบไม่ครบ ให้กลับไปทำข้อ 3 ใหม่
6) ตรวจสอบว่าข้อมูลจะต้องย้ายกลุ่มหรือไม่
    - ถ้าต้องย้ายกลุ่ม ให้คำนวณค่าจุดศูนย์กลางกลุ่มข้อมูลกลุ่มเดิม และกลุ่มใหม่
7) ตรวจสอบว่าทำข้อมูลครบทุกตัวหรือยัง
    - ถ้ายังทำไม่ครบข้อมูลทุกตัว ให้กลับไปทำข้อ 2-6 ใหม่
8) วนกลับไปทำข้อ 2 ใหม่

3.1.3 ส่วนแสดงผล

หลังจากทำการจัดกลุ่มข้อมูลแล้วจะทำการแสดงผลข้อมูลทางจอภาพ โดยข้อมูลที่จะทำการแสดงผลจะมีดังนี้

1) จุดศูนย์กลางกลุ่มข้อมูลของทุกกลุ่ม
2) ข้อมูลหลังจากการแปลงข้อมูลทุกตัวที่นำมาจัดกลุ่ม รวมทั้งบอกว่าข้อมูลตัวนี้อยู่ในกลุ่มใด
3) จำนวนข้อมูลในแต่ละกลุ่ม

3.2 ขั้นตอนการดำเนินงาน

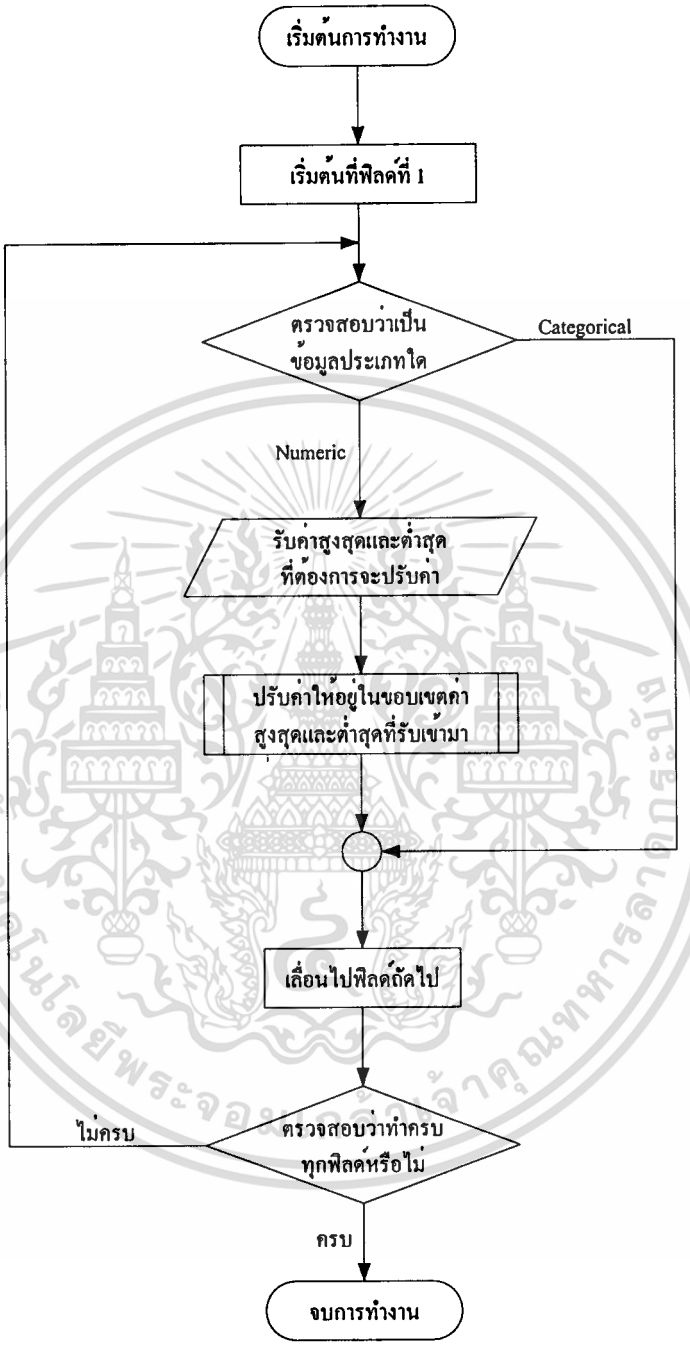
ในการทำงานของระบบมีขั้นตอนตามลำดับดังต่อไปนี้

1. เลือกไฟล์ฐานข้อมูล (Microsoft Access)
2. เลือกตารางข้อมูลและฟิลด์ข้อมูลที่ต้องการจัดกลุ่ม
3. เตรียมข้อมูล เช่น ทำการ Clean ข้อมูล ทำการ Normalize ข้อมูล
4. ใส่ค่า weight ให้กับข้อมูล Categorical (ถ้ามี)

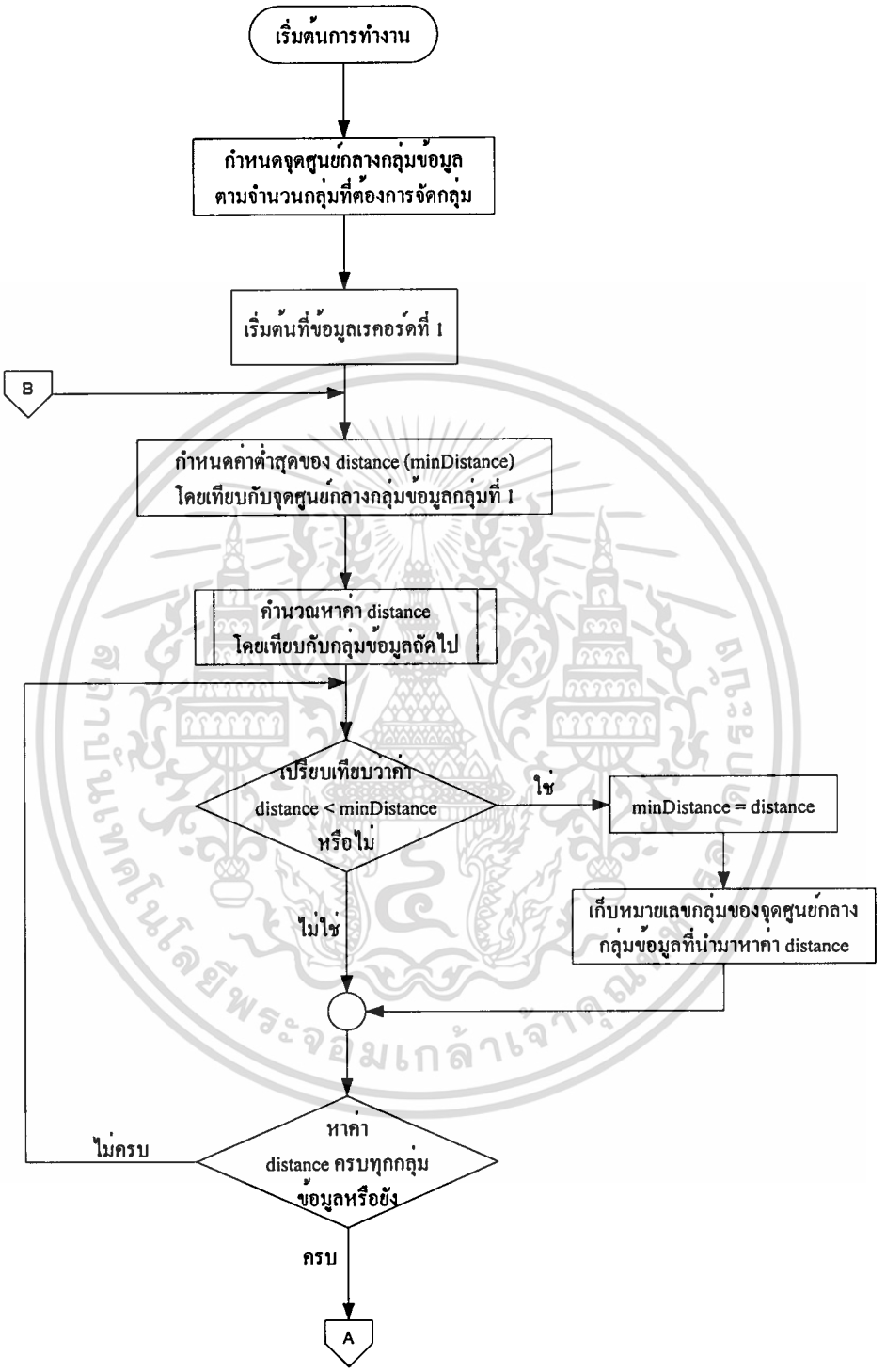
5. ใส่จำนวนกลุ่มที่ต้องการจัดกลุ่ม
6. จัดกลุ่มข้อมูลโดยใช้อัลกอริทึม K-Prototypes
7. แสดงผลการจัดกลุ่ม
8. บันทึกผลการจัดกลุ่ม (ถ้าต้องการทำการบันทึก)
9. จบการทำงาน

แสดงด้วย System Flow Diagram ดังนี้

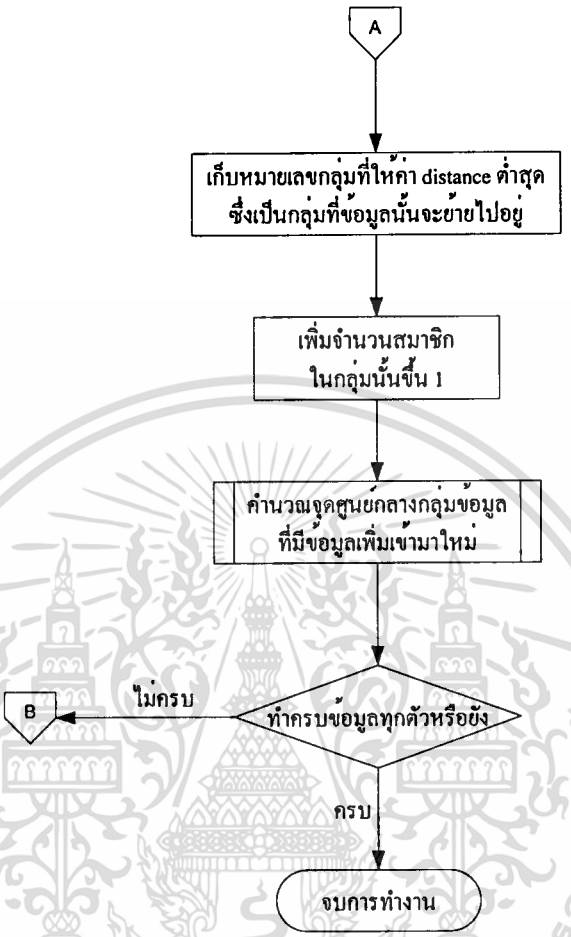
รูปที่ 3.1 ผังงานแสดงขั้นตอนการทำงานหลักของระบบ



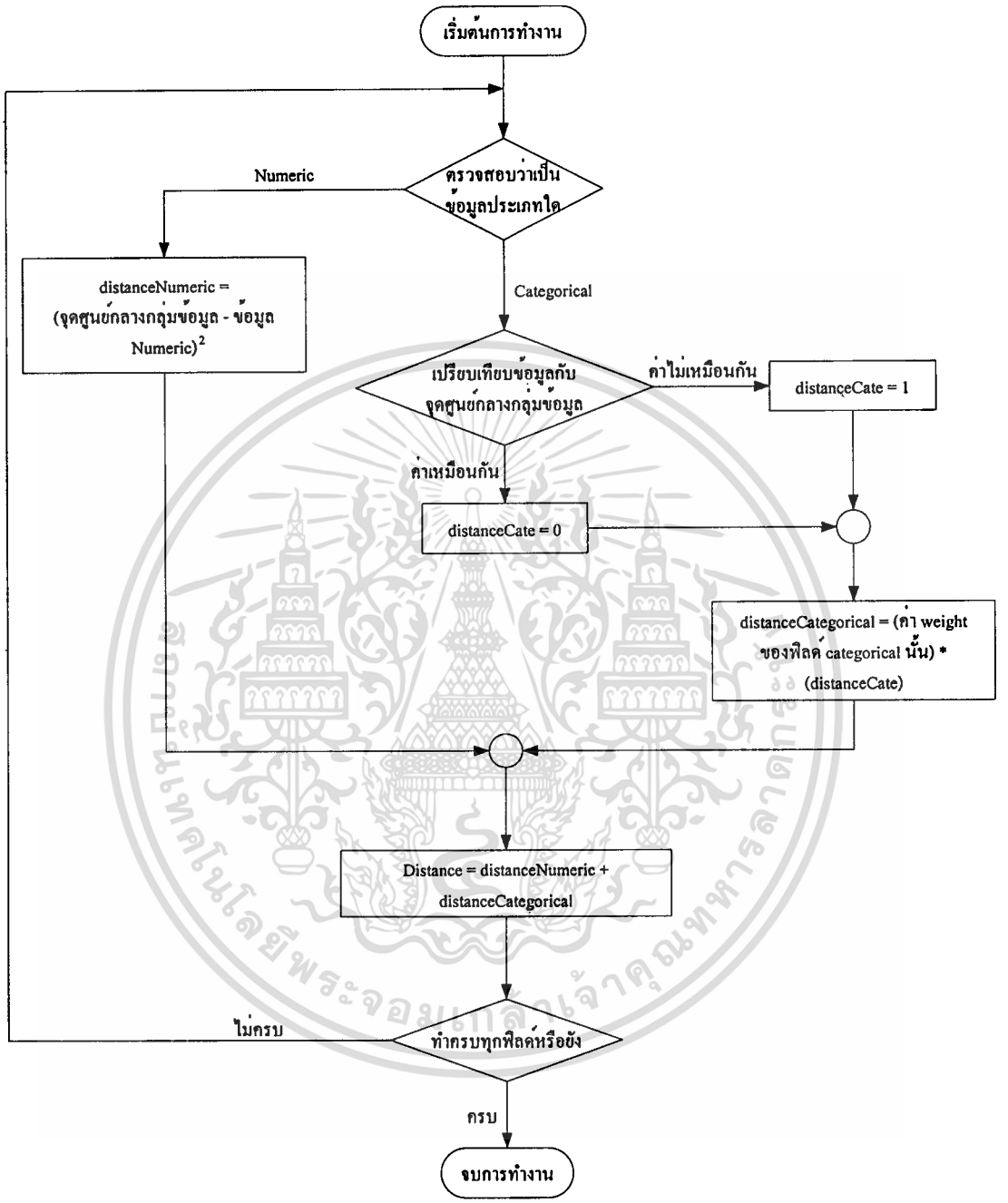
รูปที่ 3.2 ผังงานย่อยแสดงขั้นตอนการแปลงข้อมูลให้อยู่ในรูปที่เหมาะสมกับการจัดกลุ่ม



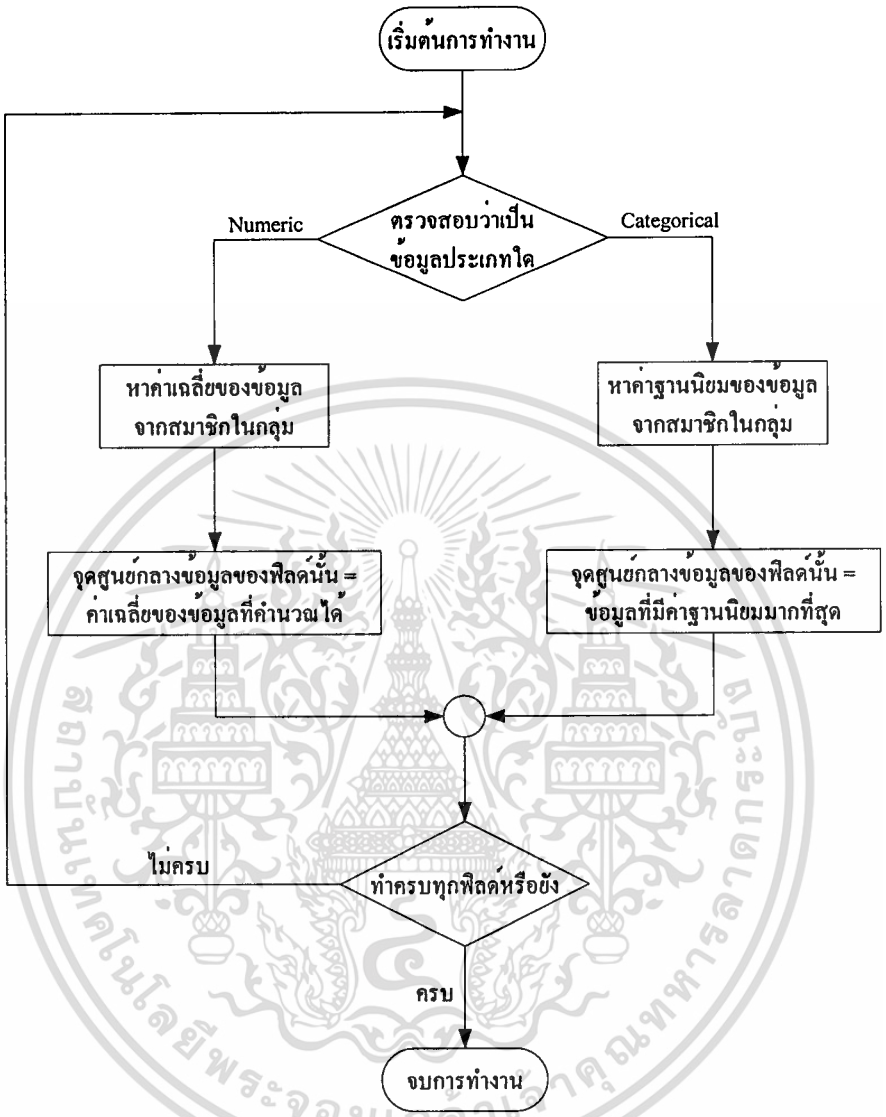
รูปที่ 3.3 ผังงานย่อยแสดงขั้นตอนการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม K-Prototypes



รูปที่ 3.3 (ต่อ) ผังงานย่อยแสดงขั้นตอนการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม K-Prototypes



รูปที่ 3.4 ผังงานย่อยแสดงขั้นตอนการคำนวณค่า Distance ของข้อมูลในแต่ละเรคอร์ด



รูปที่ 3.5 ผังงานย่อยแสดงขั้นตอนการคำนวณจุดศูนย์กลางกลุ่มข้อมูลใหม่

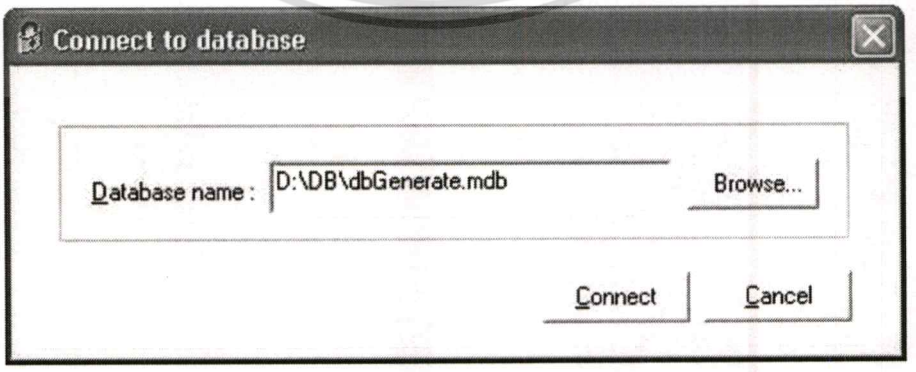
บทที่ 4
การประยุกต์ใช้ดาต้าไมนิ่งเพื่อทำการจัดกลุ่มข้อมูลลูกค้า

4.1 กำหนดวัตถุประสงค์

ในบริษัทต่างๆนั้นมักจะมีการให้สินเชื่อกับลูกค้า และหลังจากที่ทำการให้สินเชื่อไปแล้ว บางครั้งจำเป็นที่จะต้องมีการปรับเปลี่ยน เช่น เพิ่มช่วงเวลาในการชำระเงินของลูกค้า หรือเพิ่มส่วนลดให้กับลูกค้า ซึ่งปัญหาที่มักเกิดขึ้นในการปรับเปลี่ยนสินเชื่อลูกค้าคือ ข้อมูลในการนำมาพิจารณานั้นมีจำนวนมาก ทำให้ยากต่อการวิเคราะห์ข้อมูลทั้งหมดให้เสร็จภายในเวลาอันรวดเร็ว ดังนั้นจึงมีความคิดที่จะนำดาต้าไมนิ่งมาประยุกต์ใช้กับการแก้ปัญหาในครั้งนี้ โดยมีวัตถุประสงค์เพื่อจัดกลุ่มข้อมูลลูกค้าออกเป็นประเภทต่างๆ ซึ่งการแบ่งกลุ่มข้อมูลลูกค้าจะแบ่งโดยใช้ข้อมูลการชำระเงินของลูกค้าว่ามีลักษณะการชำระเงินเป็นอย่างไร เช่น ชำระตามกำหนดหรือไม่ ถ้าไม่ตามกำหนดจะดูว่าชำระล่าช้าไปกี่วัน จากนั้นข้อมูลที่ได้จากการจัดกลุ่มนี้เพื่อไปวิเคราะห์ว่าควรจะปรับเปลี่ยนสินเชื่อของลูกค้าอย่างไร และควรจะเพิ่มส่วนลดให้กับลูกค้าหรือไม่ เพื่อช่วยสนับสนุนการตัดสินใจของผู้บริหาร

4.2 การเตรียมข้อมูล

ในขั้นตอนนี้จะทำการจัดเตรียมข้อมูลเพื่อส่งต่อไปยังกระบวนการถัดไป คือการทำไมนิ่งก่อนอื่นนั้นจะต้องทำการติดต่อกับฐานข้อมูล (ดังรูปที่ 4.1)



รูปที่ 4.1 หน้าจอแสดงการเลือกฐานข้อมูล

ขั้นตอนที่ 1 การเลือกข้อมูล (Data Selection)

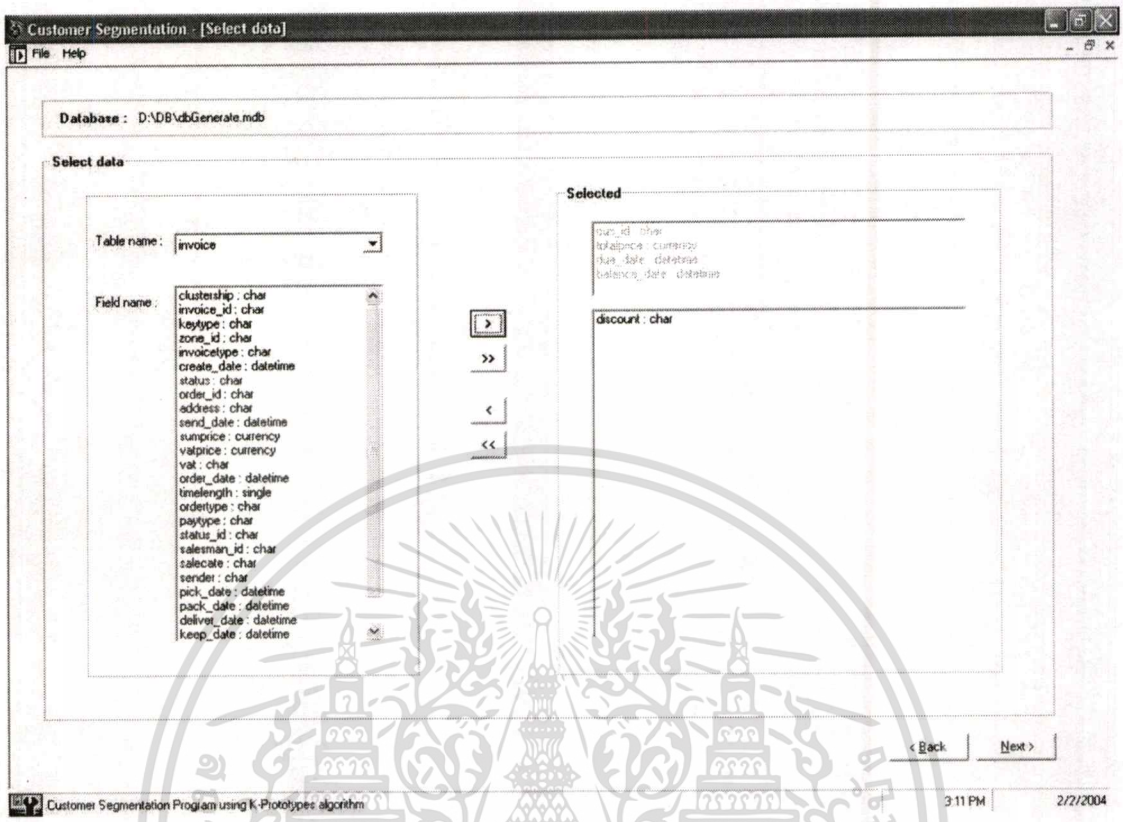
ข้อมูลที่นำมาใช้เป็นข้อมูลที่นำมาจากฐานข้อมูลของบริษัทแห่งหนึ่ง โดยทำการจัดเก็บในฐานข้อมูล Microsoft Access 2000 ข้อมูลที่จะนำมาจัดกลุ่มนี้มาจากตารางที่เก็บข้อมูลใบ Invoice มีชื่อตารางว่า "invoice" ซึ่งได้ทำการคัดเลือกเฉพาะฟิลด์ที่สนใจในตารางดังนี้

ตารางที่ 4.1 ตารางข้อมูลใบ invoice ที่ทำการเลือกมาจัดกลุ่ม

ชื่อข้อมูล	ประเภทของข้อมูล
รหัสลูกค้า (cus_id)	Text
ยอดรวมเงิน (totalprice)	Currency
วันกำหนดชำระเงิน (due_date)	Date/Time
วันที่ชำระเงินครบ (balance_date)	Date/Time
ส่วนลดในการซื้อสินค้า (discount)	Text

เนื่องจากโปรแกรมถูกพัฒนาขึ้น เพื่อรองรับการจัดกลุ่มข้อมูลจากลักษณะการชำระเงินของลูกค้า ดังนั้นการใช้งานโปรแกรมในฐานข้อมูลจะต้องประกอบไปด้วยฟิลด์ 4 ฟิลด์ดังนี้ คือ cus_id, totalprice, due_date และ balance_date จะสังเกตได้จากรูปที่ 4.2 ที่เฟรม Selected จะมี list box 2 อัน โดย list box ด้านบนจะเป็นฟิลด์ที่ถูกเลือกเป็น default ไว้ และ list box ด้านล่างจะเป็น list ที่แสดงฟิลด์อื่นๆที่ผู้ใช้ต้องการเลือกเพิ่มเข้ามาเพื่อใช้ในการจัดกลุ่ม

จากวัตถุประสงค์ที่ต้องการดูว่าควรจะเพิ่มส่วนลดให้กับลูกค้าด้วยหรือไม่ ในที่นี้การวิเคราะห์จึงจำเป็นจะต้องเลือกฟิลด์ส่วนลดในการซื้อสินค้าเพื่อมาทำการจัดกลุ่มด้วย

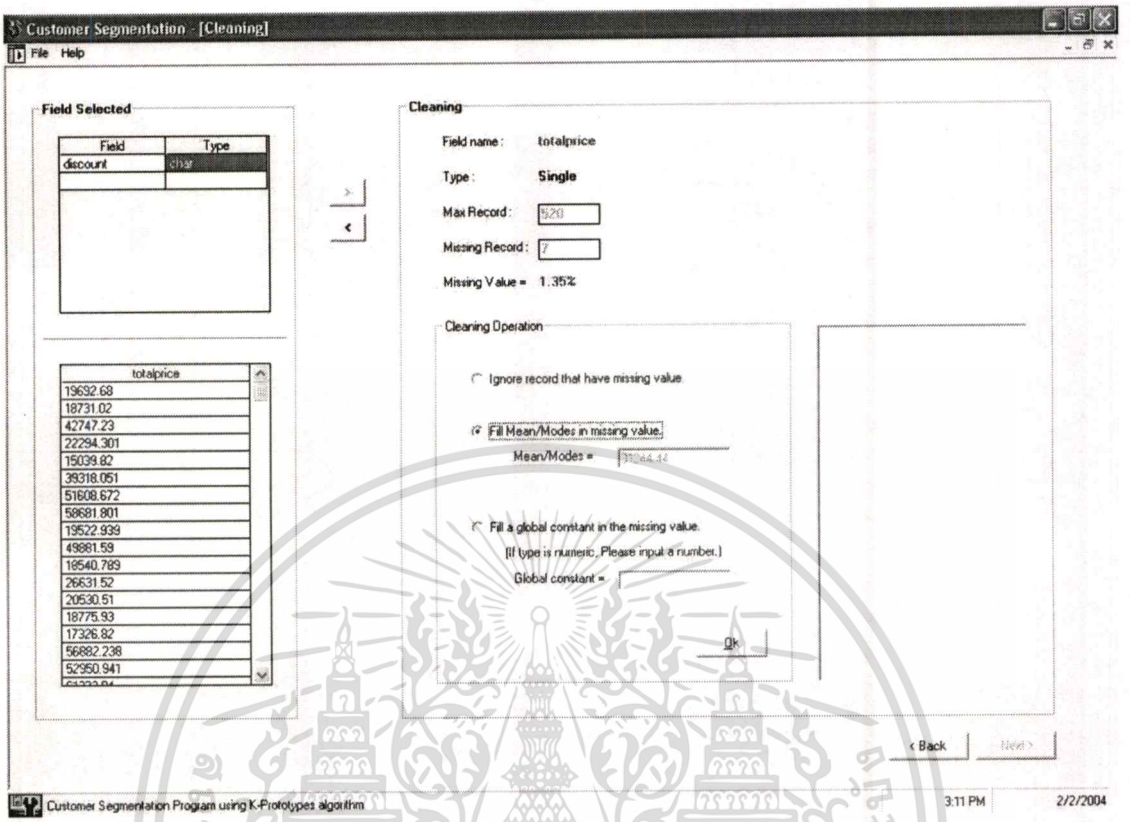


รูปที่ 4.2 หน้าจอแสดงการเลือกฟิลด์ข้อมูลที่ต้องการจัดกลุ่ม

ขั้นตอนที่ 2 การเตรียมข้อมูลก่อนประมวลผล (Data Preprocessing)

ในขั้นตอนนี้จะทำการ clean ข้อมูล โดยจะทำการกำจัดข้อมูลที่ขาดหายไป (Missing data) ดังรูปที่ 4.3 ซึ่งโปรแกรมที่พัฒนาขึ้นนี้มีวิธีกำจัดข้อมูลที่ขาดหายไป 3 วิธีคือ

1) ลบเรคอร์ดที่มีค่า Missing
2) เติมค่าเฉลี่ย (Mean) ในกรณีที่เป็นข้อมูลประเภท Numeric หรือเติมค่าฐานนิยม (Modes) ในกรณีที่เป็นข้อมูลประเภท Categorical แทนข้อมูลที่ขาดหายไป
3) เติมค่าทั่วๆไป โดยผู้ใช้งานโปรแกรมเป็นผู้ใส่ค่าลงไป



รูปที่ 4.3 หน้าจอแสดงการ Clean ข้อมูล

ขั้นตอนที่ 3 การแปลงข้อมูล (Data Transformation)

ขั้นตอนนี้ในโปรแกรมที่พัฒนาขึ้นจะทำการแปลงข้อมูลโดยแบ่งออกเป็น 2 ส่วนคือ

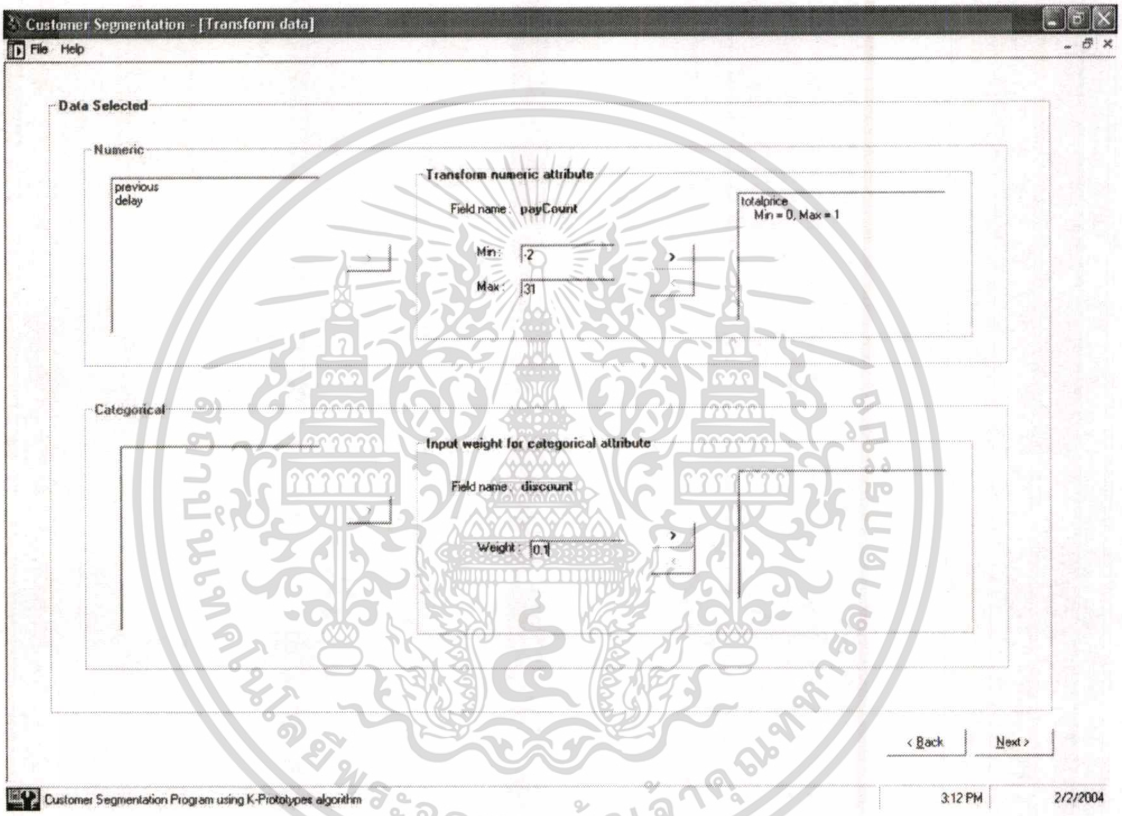
- ส่วนที่ 1 เป็นส่วนที่ตัวโปรแกรมจะทำการแปลงข้อมูลให้โดยอัตโนมัติ นั่นคือโปรแกรมจะทำการนำฟิลด์ข้อมูล “totalprice” ที่เป็นของลูกค้ารายเดียวกัน (มีรหัสลูกค้าเดียวกัน) มาหาค่าเฉลี่ยและนำฟิลด์ข้อมูล “due_date” กับ “balance_date” มาคำนวณและแปลงข้อมูลออกเป็น 3 ฟิลด์ดังนี้

  1) “payCount” เป็นฟิลด์ข้อมูลที่เก็บจำนวนวันว่าลูกค้าชำระเงินก่อนกำหนดหรือล่าช้าเป็นจำนวนกี่วัน ตัวอย่างเช่น วันกำหนดชำระเงินเป็นวันที่ 15/10/2546 และวันที่ชำระเงินครบเป็นวันที่ 22/10/2546 จะคำนวณค่า payCount ออกมาได้เป็น 3 (ชำระล่าช้า 3 วัน) แต่ถ้าวันที่ชำระเงินครบเป็นวันที่ 8/10/2546 ค่า payCount ที่คำนวณได้จะเป็น -2 (ชำระก่อนกำหนด 2 วัน)

  2) “previous” เป็นฟิลด์ข้อมูลที่เก็บเปอร์เซ็นต์ที่ลูกค้าชำระเงินก่อนกำหนด

  3) “delay” เป็นฟิลด์ข้อมูลที่เก็บเปอร์เซ็นต์ที่ลูกค้าชำระเงินช้ากว่ากำหนด

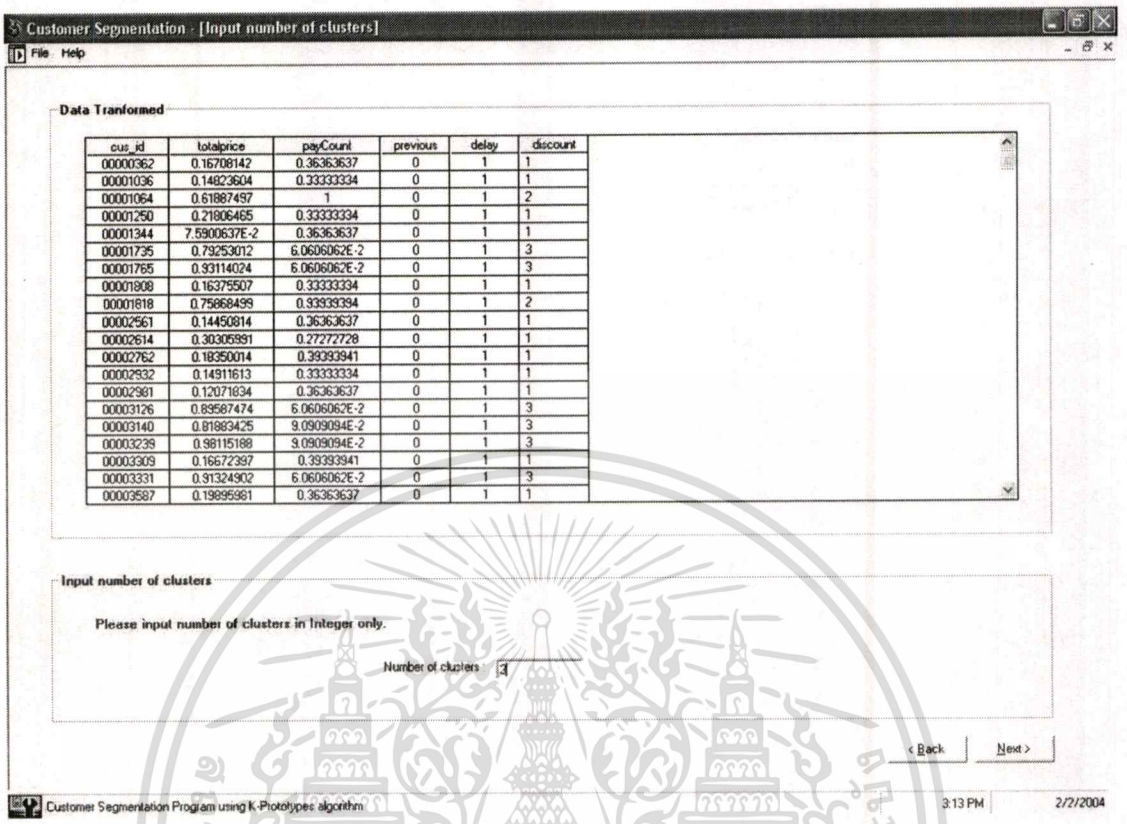
- ส่วนที่ 2 เป็นส่วนที่ทำการแปลงข้อมูลโดยผู้ใช้งานโปรแกรม ซึ่งจะทำการแปลงข้อมูลได้เฉพาะข้อมูลที่เป็นประเภท Numeric เท่านั้น ดังรูปที่ 4.4 โดยผู้ใช้สามารถแปลงให้อยู่ในช่วงที่ผู้ใช้งานต้องการ เพื่อความเหมาะสมในการทำดาต้าไมนิ่ง การแปลงข้อมูลประเภทตัวเลขนี้ได้ใช้วิธี Normalization โดยใช้สูตร Min-Max Normalization ในการคำนวณ



รูปที่ 4.4 หน้าจอแสดงการแปลงข้อมูล

4.3 การทำไมนิ่ง

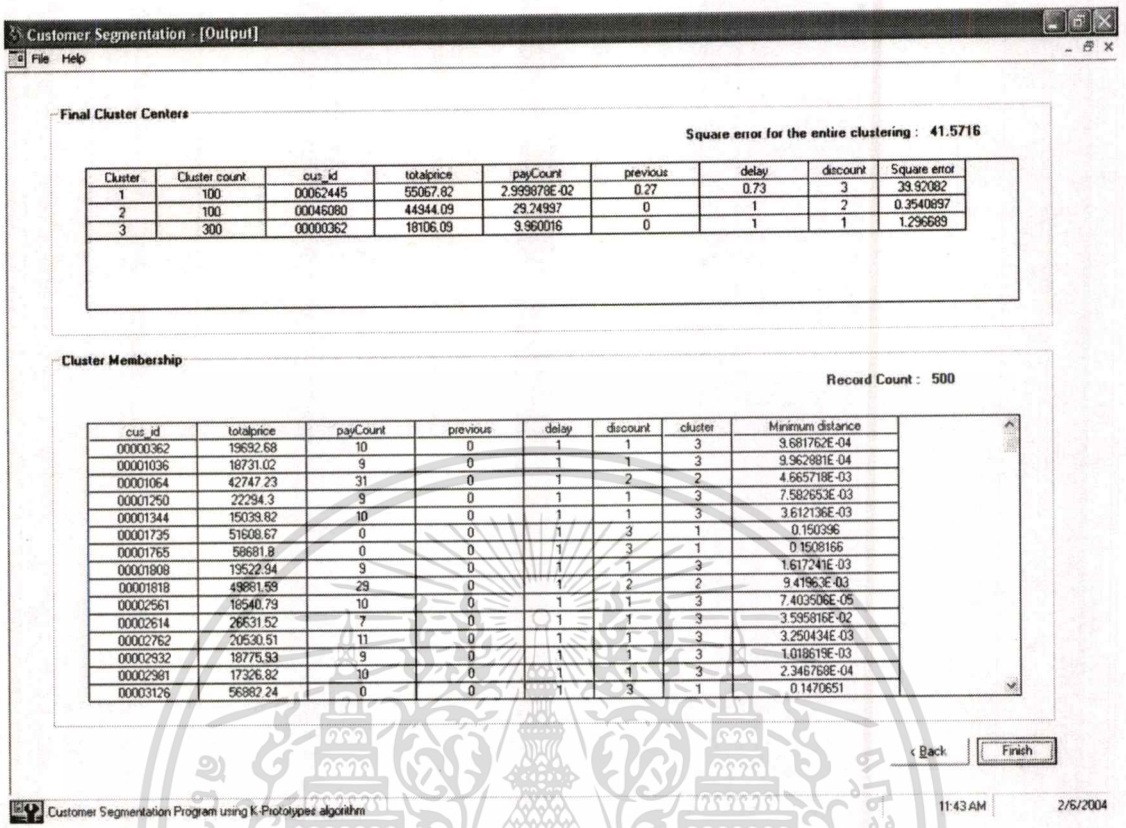
ขั้นตอนการทำไมนิ่งเป็นขั้นตอนที่นำเอาข้อมูลที่ผ่านการเตรียมข้อมูลทั้ง 3 ขั้นตอนย่อยที่ได้กล่าวมาแล้วในหัวข้อที่ 4.2 โดยข้อมูลที่เป็นประเภท Numeric จะถูกทำการแปลงค่าของข้อมูลให้อยู่ในช่วงที่เหมาะสมกับการทำไมนิ่งและข้อมูลประเภท Categorical จะต้องทำการใส่ค่า weight ซึ่งได้ทำการใส่ค่าตั้งแต่ขั้นตอนการแปลงข้อมูลแล้ว ดังรูปที่ 4.4 มาผ่านอัลกอริทึม K-Prototypes ในขั้นตอนการทำไมนิ่งนี้จะต้องมีการใส่ค่าจำนวนกลุ่มข้อมูลที่ต้องการจะแบ่ง (ค่า k)



รูปที่ 4.5 หน้าจอแสดงข้อมูลที่ผ่านการแปลงและให้ใส่ค่าจำนวนกลุ่มที่ต้องการแบ่ง

จากรูปที่ 4.5 ที่เฟรมด้านบนจะแสดงข้อมูลที่ผ่านการแปลงข้อมูลแล้ว ส่วนเฟรมด้านล่างจะเป็นส่วนที่ให้ผู้ใช้งานทำการใส่ค่าจำนวนกลุ่มที่ต้องการจะจัดกลุ่ม

หลังจากที่ใส่ค่าจำนวนกลุ่มที่ต้องการจะจัดกลุ่มแล้วนั้น โปรแกรมจะทำการจัดกลุ่มข้อมูลโดยผ่านอัลกอริทึม K-Prototypes ซึ่งจะแสดงผลออกมาดังรูปที่ 4.6 ซึ่งหน้าจอนี้จะแบ่งออกเป็น 2 ส่วนส่วนที่ 1 จะแสดงจุดศูนย์กลางกลุ่มข้อมูลที่ได้จากการจัดกลุ่มและจำนวนข้อมูลในแต่ละกลุ่ม รวมทั้งยังค่า Square Error ของแต่ละกลุ่มข้อมูลและบอกผลรวมของค่า Square Error ของทุกกลุ่มข้อมูลที่มุมบนด้านขวา ในส่วนที่ 2 จะแสดงข้อมูลทั้งหมดรวมและบอกว่าข้อมูลแต่ละเรคอร์ดนั้นจัดอยู่ในกลุ่มใด รวมทั้งยังบอกค่า Minimum distance ที่ห่างจากจุดศูนย์กลางกลุ่มข้อมูล



รูปที่ 4.6 หน้าจอแสดงผลการจัดกลุ่มข้อมูล

4.4 การวิเคราะห์ผลที่ได้จากการทำดาต้าไมนิ่งและการนำความรู้มาใช้

หลังจากทำการไมนิ่งเสร็จแล้วจะได้ผลลัพธ์ ซึ่งจะต้องนำไปใช้วิเคราะห์ต่อ แต่ก่อนที่เราจะวิเคราะห์ข้อมูลนั้น จะต้องทราบความหมายของตัวแปรประเภท Categorical ก่อน ตัวอย่างเช่น ฟิลด์ส่วนลดในการซื้อสินค้ามี 3 ค่า โดยแต่ละค่ามีความหมายดังนี้

- 1 = ได้รับส่วนลด 0-10%
- 2 = ได้รับส่วนลด 10-20%
- 3 = ได้รับส่วนลด 25-30%

จากรูปที่ 4.6 ผลจากการจัดกลุ่มข้อมูลออกเป็น 3 กลุ่มจะเห็นว่า

<u>กลุ่มที่ 1</u> มีจำนวนข้อมูลทั้งหมดในกลุ่มนี้เท่ากับ 100 เรคอร์ด โดยลักษณะโดยทั่วไปของกลุ่มนี้คือมียอดรวมในการซื้อสินค้า (totalprice) ประมาณ 55,067.82 บาท มีการชำระเงินครบก่อนกำหนดประมาณ 0.03 วัน โดยคิดเป็นเปอร์เซ็นต์ที่ลูกค้าชำระก่อนกำหนดเป็น 27% และชำระเงินล่าช้ากว่า

กำหนดเป็น 73% ประเภทของส่วนลดในกลุ่มนี้คือได้รับส่วนลดในประเภทที่ 3 นั่นคือได้รับส่วนลด 25-30%

กลุ่มที่ 2 มีจำนวนข้อมูลทั้งหมดในกลุ่มนี้เท่ากับ 100 เรคอร์ด โดยลักษณะโดยทั่วไปของกลุ่มนี้คือมียอดรวมในการซื้อสินค้า (totalprice) ประมาณ 44,944.09 บาท มีการชำระเงินครบล่าช้ากว่าที่กำหนดไว้ประมาณ 29.25 วัน โดยคิดเป็นเปอร์เซ็นต์ที่มีลูกค้าชำระก่อนกำหนดเป็น 0% และมีลูกค้าชำระเงินล่าช้ากว่ากำหนด 100% ประเภทของส่วนลดในกลุ่มนี้คือได้รับส่วนลดในประเภทที่ 2 นั่นคือได้รับส่วนลด 10-20%

กลุ่มที่ 3 มีจำนวนข้อมูลทั้งหมดในกลุ่มนี้เท่ากับ 300 เรคอร์ด โดยลักษณะโดยทั่วไปของกลุ่มนี้คือมียอดรวมในการซื้อสินค้า (totalprice) ประมาณ 18,106.09 บาท มีการชำระเงินครบล่าช้ากว่าที่กำหนดไว้ประมาณ 9.96 วัน โดยคิดเป็นเปอร์เซ็นต์ที่มีลูกค้าชำระก่อนกำหนดเป็น 0% และมีลูกค้าชำระเงินล่าช้ากว่ากำหนด 100% ประเภทของส่วนลดในกลุ่มนี้คือได้รับส่วนลดในประเภทที่ 1 นั่นคือได้รับส่วนลด 0-10%

จากผลการจัดกลุ่มทั้ง 3 กลุ่มข้างต้น จะได้ว่ากลุ่มที่ 1 และ 2 มียอดการสั่งซื้อที่สูง ต่างกันตรงที่กลุ่มที่ 1 ลักษณะการชำระเงินที่ตรงเวลาแต่กลุ่ม 2 มีการชำระเงินที่ล่าช้ากว่ากำหนดมาก ส่วนกลุ่มที่ 3 ซึ่งเป็นกลุ่มที่มีข้อมูลในกลุ่มจำนวนมากที่สุด มียอดการสั่งซื้อสินค้าไม่สูงมากเพียงแต่ชำระเงินล่าช้ากว่ากำหนดเพียงเล็กน้อย ดังนั้นจึงสามารถจัดลูกค้าออกเป็น 3 ประเภทดังนี้

- ประเภทที่ 1 (ข้อมูลจัดอยู่ในกลุ่มที่ 1) เป็นกลุ่มลูกค้าที่มียอดการสั่งซื้อไม่มาก และมีการชำระเงินที่ล่าช้ากว่ากำหนดไม่มาก โดยจะได้รับส่วนลดการซื้อสินค้าอยู่ในช่วง 0-10%
- ประเภทที่ 2 (ข้อมูลจัดอยู่ในกลุ่มที่ 2) เป็นกลุ่มลูกค้าที่มียอดการสั่งซื้อสูง แต่มีการชำระเงินที่ล่าช้ากว่ากำหนดมาก ซึ่งจะได้รับส่วนลดการซื้อสินค้าอยู่ในช่วง 10-20%
- ประเภทที่ 3 (ข้อมูลจัดอยู่ในกลุ่มที่ 3) เป็นกลุ่มลูกค้าที่มียอดการสั่งซื้อสูง และมีการชำระเงินที่ตรงต่อเวลา โดยได้รับส่วนลดการซื้อสินค้าอยู่ในช่วง 25-30%

ผลที่ได้จากการจัดกลุ่มนั้น บอกว่าลูกค้าส่วนใหญ่มีลักษณะการซื้อสินค้าและการชำระเงินเป็นดังกลุ่มที่ 3 โดยจะสังเกตได้จากจำนวนข้อมูลที่ทำการจัดกลุ่มออกมา ซึ่งผลการจัดกลุ่มจะช่วยให้เราสามารถวิเคราะห์และแบ่งประเภทของลูกค้าออกเป็นกลุ่มได้อย่างรวดเร็ว สามารถนำผลที่ได้ไปช่วยสนับสนุนการตัดสินใจของผู้บริหารได้ ซึ่งตรงกับวัตถุประสงค์ที่ได้ตั้งไว้ในข้างต้น

บทที่ 5
สรุปผลการศึกษาและข้อเสนอแนะ

โครงการพัฒนาระบบนี้เป็นโครงการที่จัดทำขึ้น เพื่อศึกษาและนำเสนอให้เห็นถึงการนำดาต้าไมนิ่งมาประยุกต์ใช้กับธุรกิจ โดยใช้วิธีการจัดกลุ่มข้อมูลซึ่งเป็นวิธีการทำไมนิ่งอย่างหนึ่ง ซึ่งนำมาใช้แบ่งกลุ่มประเภทของลูกค้า เพื่อนำผลลัพธ์ที่ได้ไปใช้ในการปรับเปลี่ยนสินเชื่อหรือเพิ่มส่วนลดให้แก่ลูกค้าได้

5.1 สรุปผลการดำเนินงาน

การทำดาต้าไมนิ่งเป็นกระบวนการในการค้นหาข้อมูลที่ยังไม่เคยเห็น และเป็นประโยชน์ออกมาจากฐานข้อมูลที่มีอยู่ เพื่อนำข้อมูลนั้นไปช่วยในการสนับสนุนการตัดสินใจและยังสามารถนำไปประยุกต์ใช้กับงานทางด้านธุรกิจ โดยการนำดาต้าไมนิ่งมาประยุกต์ใช้นั้นก่อนอื่นจะต้องทำความเข้าใจกับปัญหาที่เกิดขึ้น จากนั้นจะต้องทำการตั้งวัตถุประสงค์ของการทำดาต้าไมนิ่ง โดยถ้าทำการตั้งวัตถุประสงค์ไว้ไม่ดี หลังจากที่ทำดาต้าไมนิ่งแล้วผลลัพธ์ที่ได้จะไม่เป็นประโยชน์ต่อการแก้ปัญหา เพราะอาจเกิดจากการตั้งวัตถุประสงค์ผิดพลาดทำให้การเลือกวิธีการทำไมนิ่งผิดไปด้วย ในทางกลับกัน ถ้าทำการกำหนดวัตถุประสงค์ได้ถูกต้องก็จะนำไปสู่ขั้นตอนการทำไมนิ่งที่ถูกต้องและสามารถนำผลลัพธ์ที่ได้ไปใช้แก้ปัญหาได้เป็นอย่างดี

โครงการพัฒนาระบบงานนี้ได้ใช้เทคนิคการจัดกลุ่มข้อมูล (Data Segmentation) และเลือกใช้อัลกอริทึม K-Prototypes เป็นอัลกอริทึมในการทำไมนิ่ง โดยผลที่ได้จากการศึกษาและพัฒนาระบบงานในครั้งนี้พบว่า ผลลัพธ์ที่ได้จากการจัดกลุ่มสามารถช่วยระบุได้ว่าลูกค้าแต่ละรายจัดว่าเป็นลูกค้าประเภทใดบ้าง เช่น ลูกค้าที่มียอดการซื้อสินค้าสูงและชำระเงินตรงตามเวลาที่กำหนดจะถูกจัดเป็นลูกค้าชั้นดี และได้นำผลการจัดกลุ่มนี้ไปช่วยสนับสนุนการตัดสินใจว่าจะปรับเพิ่มระยะเวลากำหนดในการชำระเงิน หรืออาจจะเพิ่มส่วนลดให้กับกลุ่มลูกค้าชั้นดีนี้ได้ ซึ่งระบบงานที่พัฒนาขึ้นนี้ทำให้ประหยัดเวลาในการวิเคราะห์ข้อมูลได้เป็นอย่างมาก เนื่องจากข้อมูลที่มีอยู่จำนวนมากนั้นได้ถูกจัดออกมาเป็นกลุ่ม ทำให้ผู้บริหารวิเคราะห์ข้อมูลน้อยลง ทั้งนี้ข้อมูลที่เป็นผลลัพธ์จากการทำไมนิ่งจะเป็นประโยชน์ได้จะต้องอาศัยกระบวนการต่างๆทุกขั้นตอน ดังที่กล่าวมาแล้ว คือการกำหนดวัตถุประสงค์ การเตรียมข้อมูล การทำไมนิ่ง และสุดท้ายคือการวิเคราะห์

ข้อมูลและนำไปใช้ โดยถ้ามีขั้นตอนใดผิดพลาดไปอาจทำให้ผลลัพธ์ที่ออกมาไม่เป็นประโยชน์ต่อการแก้ปัญหาทางธุรกิจ

5.2 ข้อเสนอแนะ

1) โครงการพัฒนาระบบงานนี้สามารถใช้กับธุรกิจใดๆก็ได้ โดยธุรกิจนั้นจะต้องมีการเก็บข้อมูลใบ invoice ซึ่งในตารางข้อมูลใบ invoice จะต้องประกอบด้วยฟิลด์ 4 ฟิลด์ดังนี้ คือ cus_id, totalprice, due_date และ balance_date โดยระบบงานนี้ยังสามารถเลือกข้อมูลฟิลด์อื่นที่นอกเหนือจากฟิลด์ที่กล่าวมาได้อีก
2) ระบบงานนี้จะทำการจัดกลุ่มข้อมูลเพื่อนำผลที่ได้ไปปรับเปลี่ยนสินเชื่อ หรือสร้างกลยุทธ์อื่นๆให้กับลูกค้า โดยกลุ่มลูกค้านี้เป็นลูกค้าที่มีการซื้อสินค้ากับทางบริษัทแล้ว จึงไม่สามารถนำผลจากการทำไมนิ่งของระบบไปใช้ในการกำหนดช่วงระยะเวลากำหนดชำระเงินหรือกำหนดส่วนลดให้กับลูกค้าใหม่ที่ยังไม่เคยทำธุรกิจกับทางบริษัทมาก่อน

บรรณานุกรม

Cabena, I. et al. 1997. **Discovering Data Mining from Concept to Implementation.** New Jersy: Prentice Hall.

Han, J. and Kamber, M. 2001. **Data Mining: concept and techniques.** United States of America: Academic Press.

Jagoda, C. and Frank, C. **Data Mining in a Scientific Environment.** [Online]. Available: http://www.csu.edu.au/special/auugwww96/proceedings/crawford/crawford.html

Kira Tarapanoff, et al. 2002. **Intelligence obtained by applying data mining to a database of French theses on the subject of Brazil.** [Online]. Available: http://informationr.net/ir/7-1/paper117.html

Zhexue Huang. **Clustering large data sets with mixed numeric and categorical values.** [Online]. Available: http://www.act.cmis.csiro.au/gjw/papers/apkdd.pdf

ภาคผนวก

ภาคผนวก ก

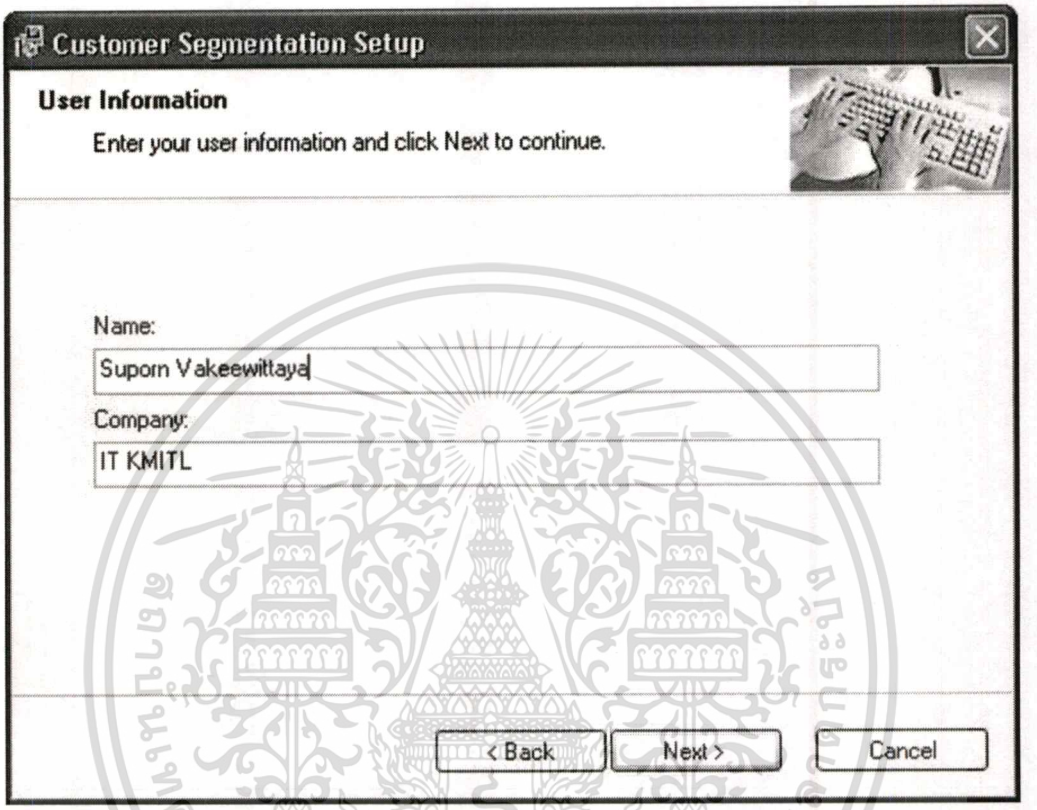
การติดตั้งโปรแกรม

1. เริ่มการติดตั้งโปรแกรม โดยทำการดับเบิ้ลคลิกที่ไอคอน 

จาก Floppy disk (A:) จะปรากฏหน้าจอดังรูป ก.1

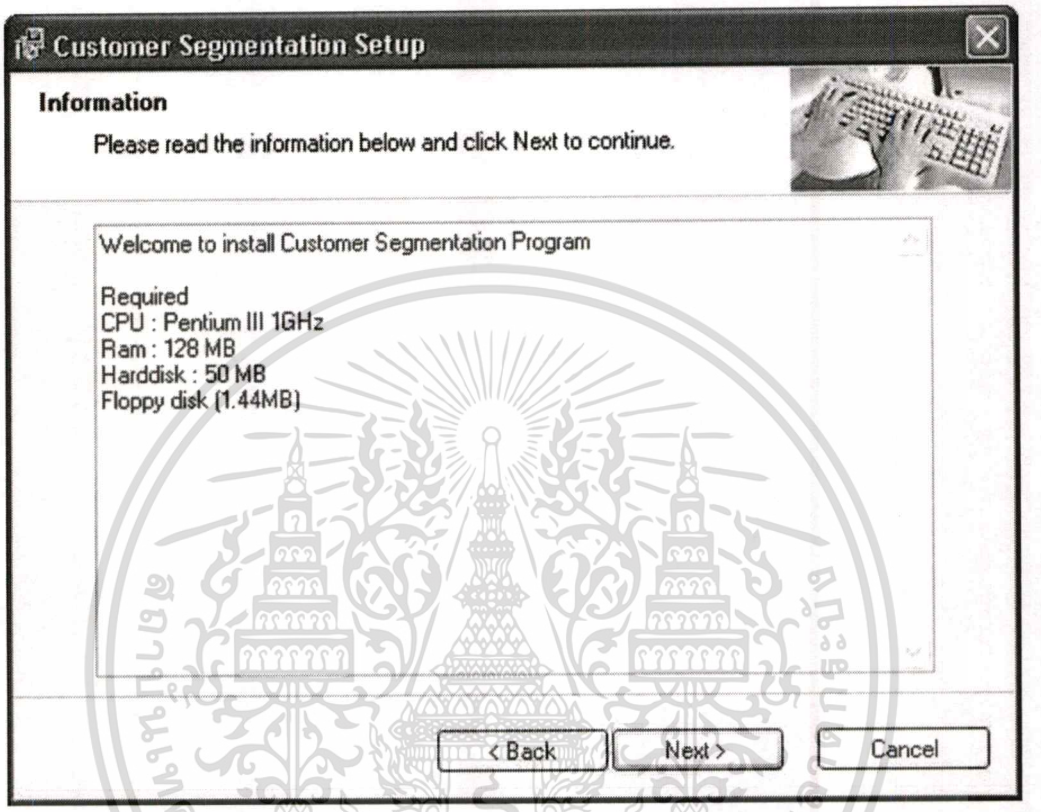
รูปที่ ก.1 หน้าจอแรกเมื่อเข้าสู่การติดตั้งโปรแกรม

2. จะปรากฏหน้าจอดังรูป ก.2 โดยหน้าจอนี้จะแสดงชื่อของตัวเครื่อง จากนั้นคลิก Next > เพื่อไปยังหน้าถัดไป



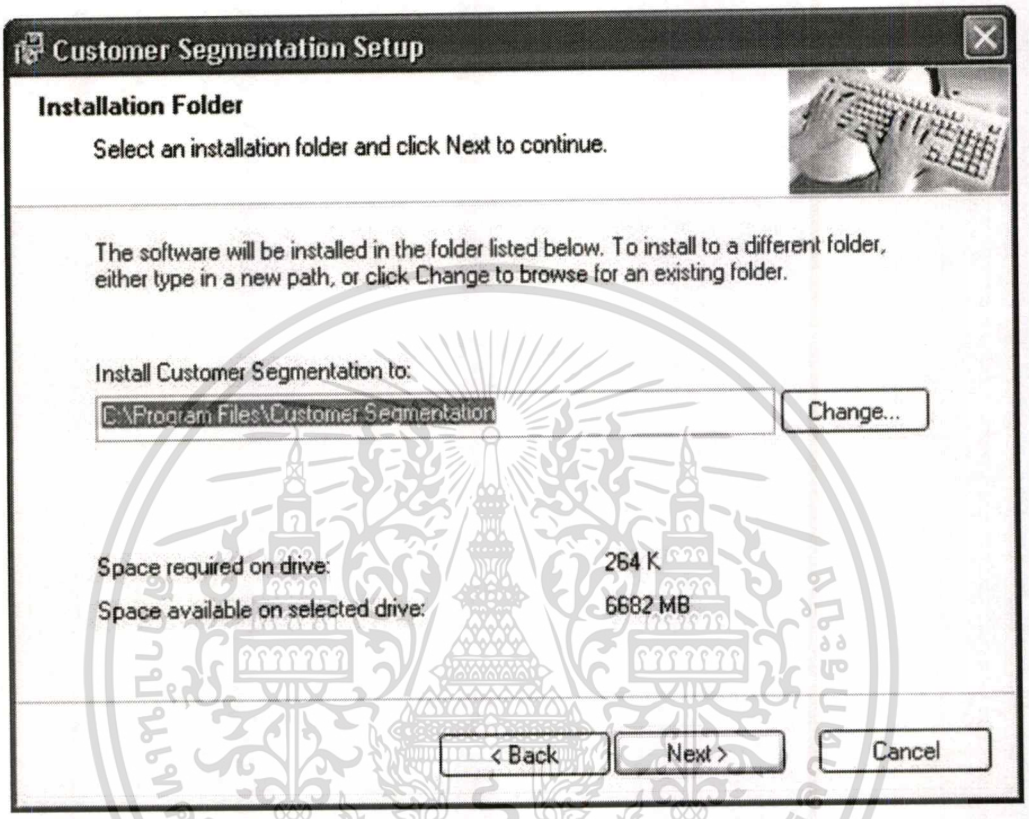
รูปที่ ก.2 หน้าจอที่ 2 ของการติดตั้งโปรแกรม

3. จะปรากฏหน้าจอดังรูป ก.3 ซึ่งหน้าจอนี้จะแสดงข้อมูลความต้องการของ Hardware จากนั้นให้ทำการคลิก Next > เพื่อไปยังหน้าถัดไป



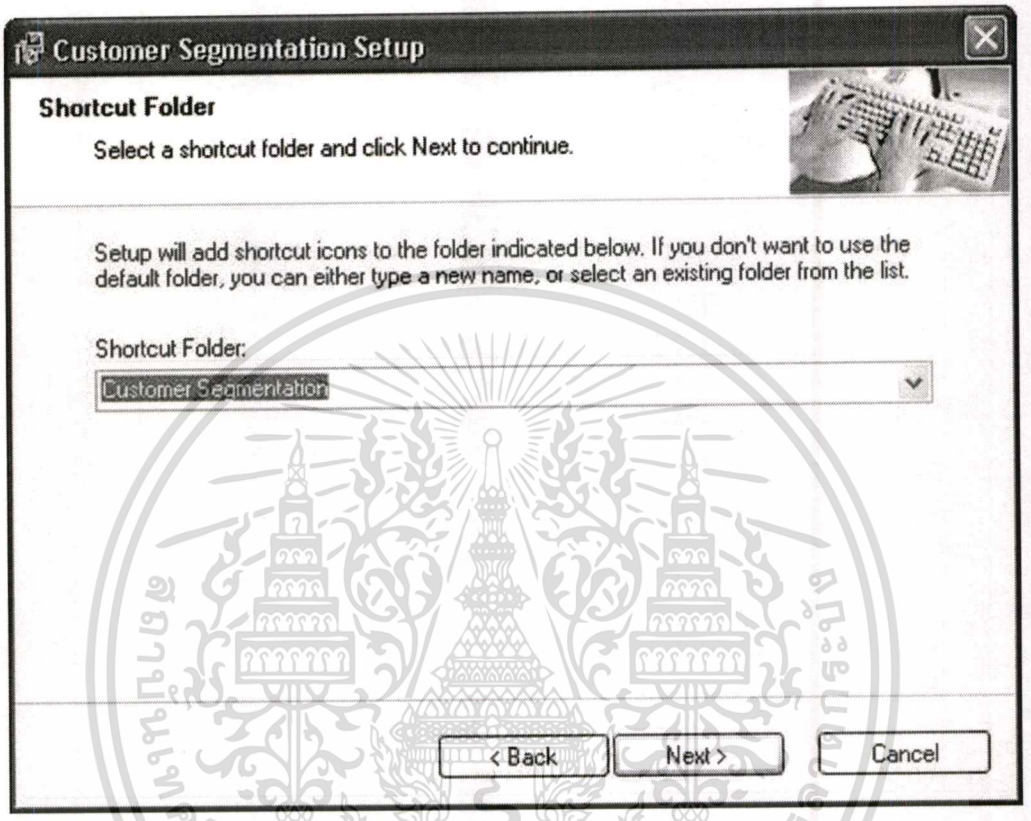
รูปที่ ก.3 หน้าจอที่ 3 ของการติดตั้งโปรแกรม

4. จะปรากฏหน้าจอดังรูป ก.4 ซึ่งหน้าจอนี้จะให้เลือก Path ที่ต้องการจะลงโปรแกรม จากนั้นคลิก Next > เพื่อไปยังหน้าถัดไป



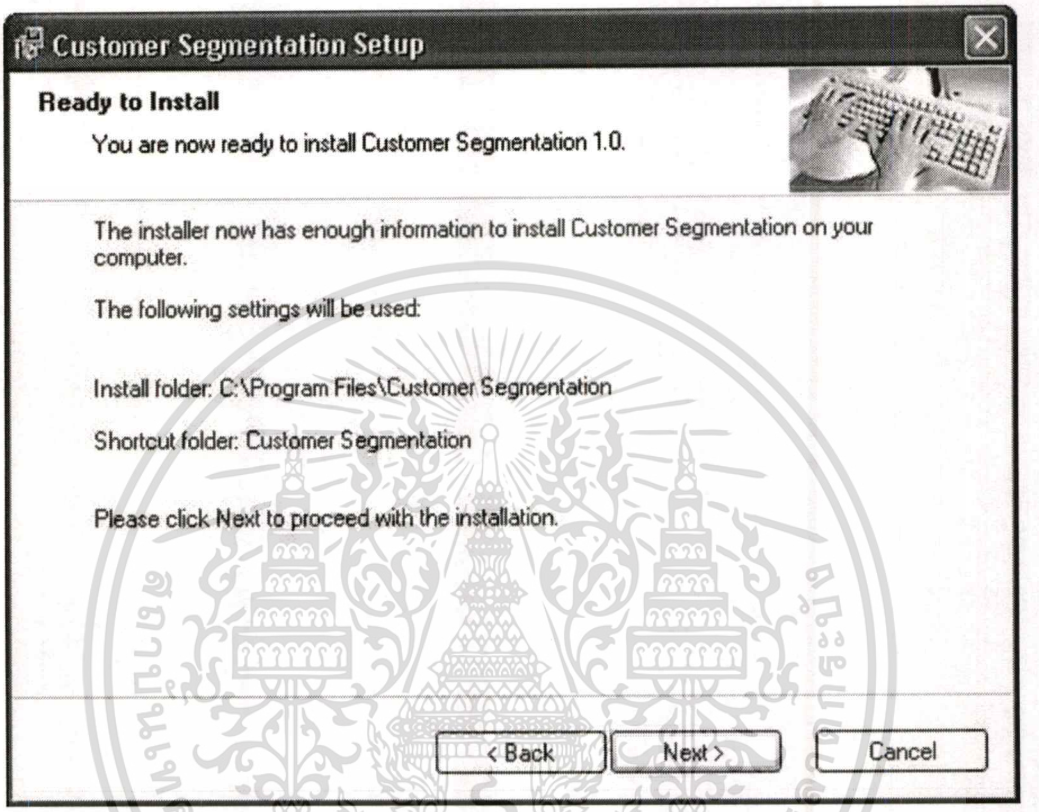
รูปที่ ก.4 หน้าจอที่ 4 ของการติดตั้งโปรแกรม

5. จะปรากฏหน้าจอดังรูป ก.5 ซึ่งหน้าจอนี้จะให้ทำการตั้งชื่อโฟลเดอร์ของโปรแกรม จากนั้นคลิก Next > เพื่อไปยังหน้าจอถัดไป



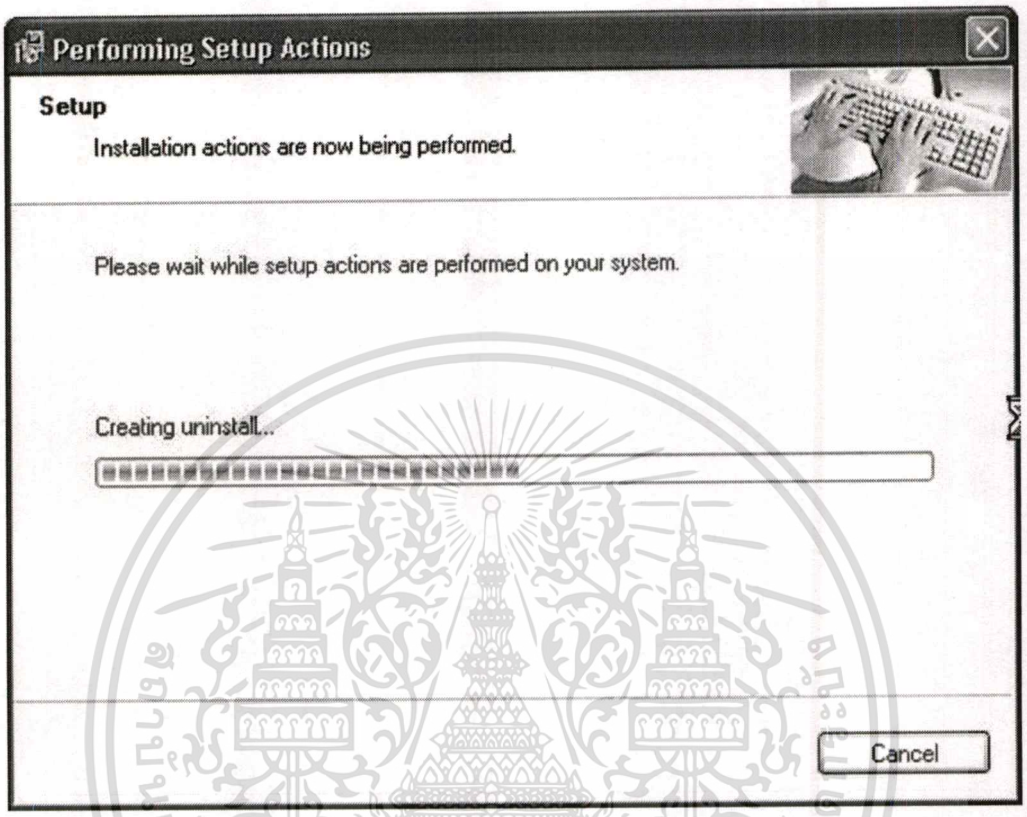
รูปที่ ก.5 หน้าจอที่ 5 ของการติดตั้งโปรแกรม

6. หน้าจอรูป ก.6 จะบอกรายละเอียดว่าตัวโปรแกรมจะทำการ Install ไว้ที่ใด จากนั้นคลิก Next > เพื่อลงโปรแกรม



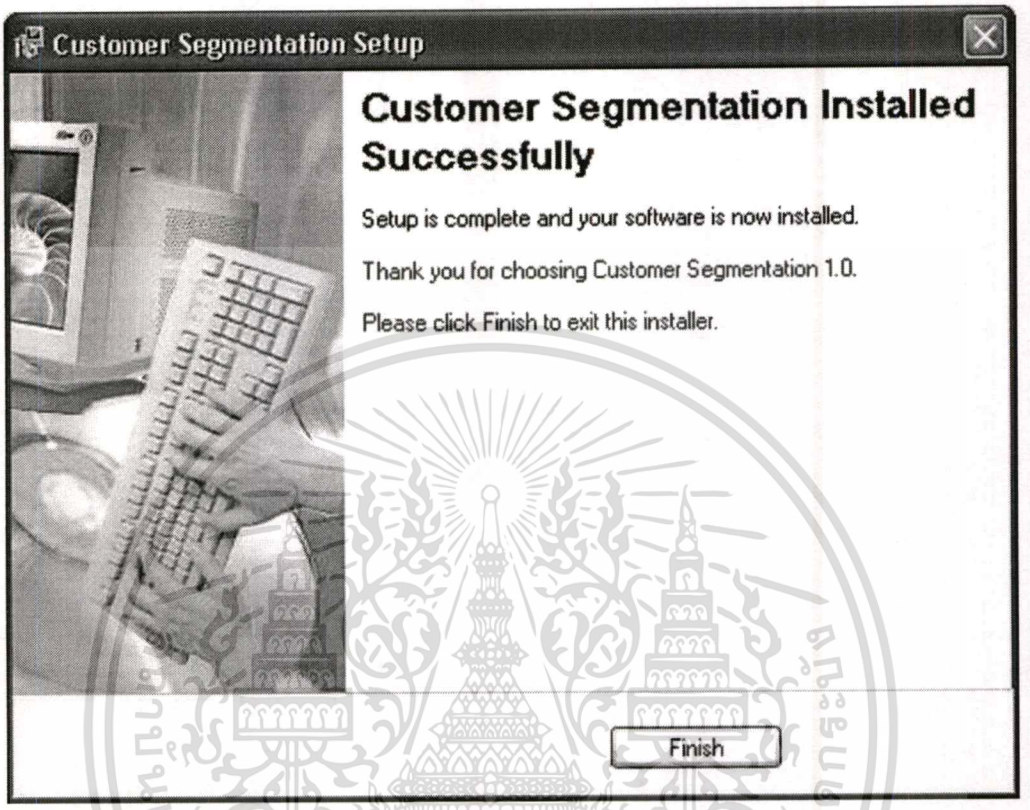
รูปที่ ก.6 หน้าจอที่ 6 ของการติดตั้งโปรแกรม

7. จะปรากฏหน้าจอดังรูป ก.7 เป็นหน้าจอที่แสดงว่าโปรแกรมกำลังถูกติดตั้ง



รูปที่ ก.7 หน้าจอที่ 7 ของการติดตั้งโปรแกรม

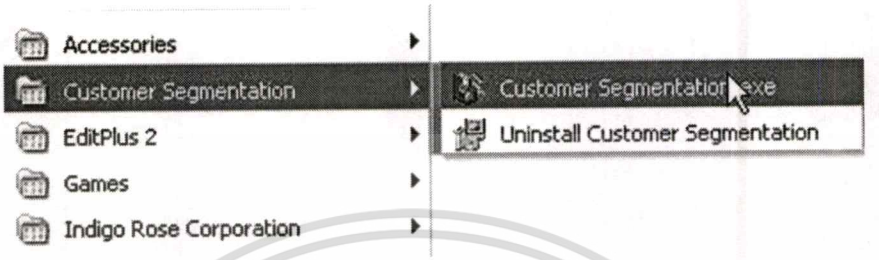
8. คลิก Finish เพื่อสิ้นสุดการลงโปรแกรม



รูปที่ ก.8 หน้าจอสุดท้ายของการติดตั้งโปรแกรม

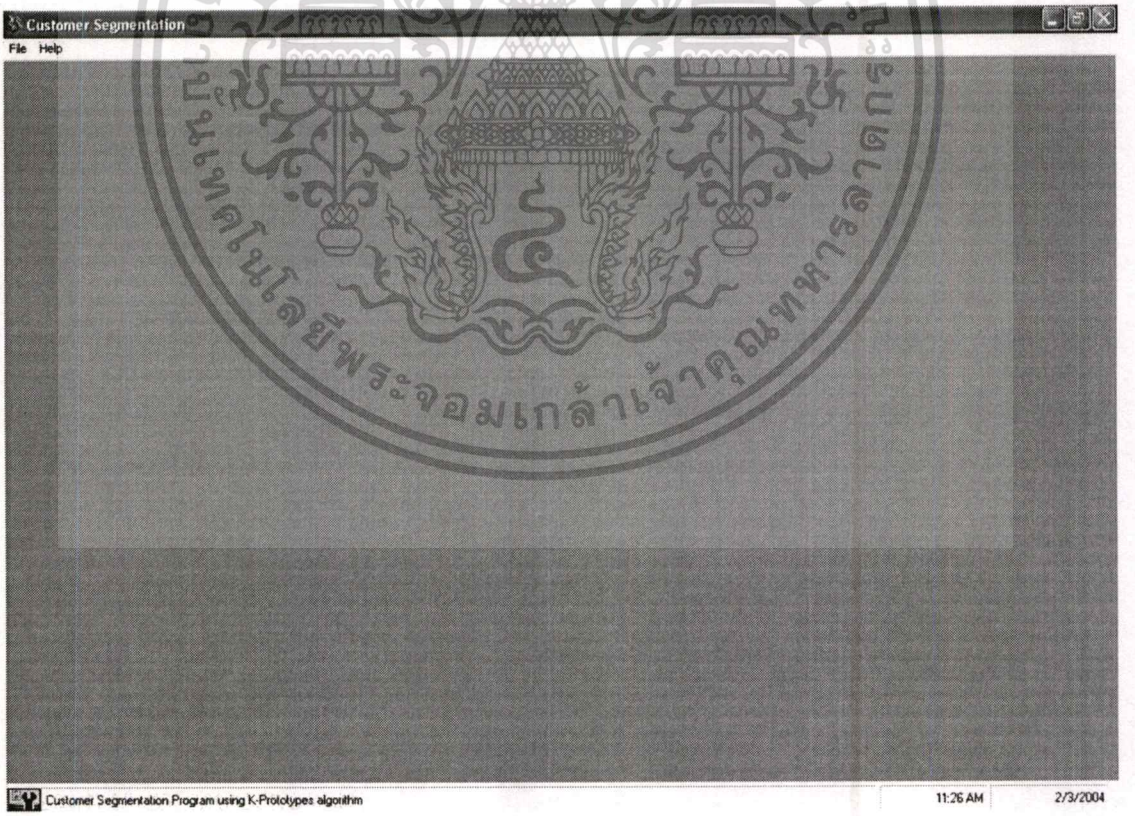
การใช้งานโปรแกรม

1. เริ่มต้นการใช้งานโปรแกรม Customer Segmentation โดยการเลือกที่ Start > All Programs > Customer Segmentation > Customer Segmentation.exe ดังรูปที่ ก.9



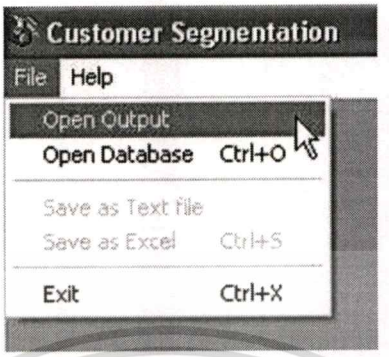
รูปที่ ก.9 หน้าจอการเริ่มต้นใช้งานโปรแกรม

จากนั้นจะปรากฏหน้าจอดังรูป ก.10



รูปที่ ก.10 หน้าจอเริ่มต้นของโปรแกรม Customer Segmentation

2. เลือกเมนู File จะปรากฏเมนูย่อยดังรูป ก.11



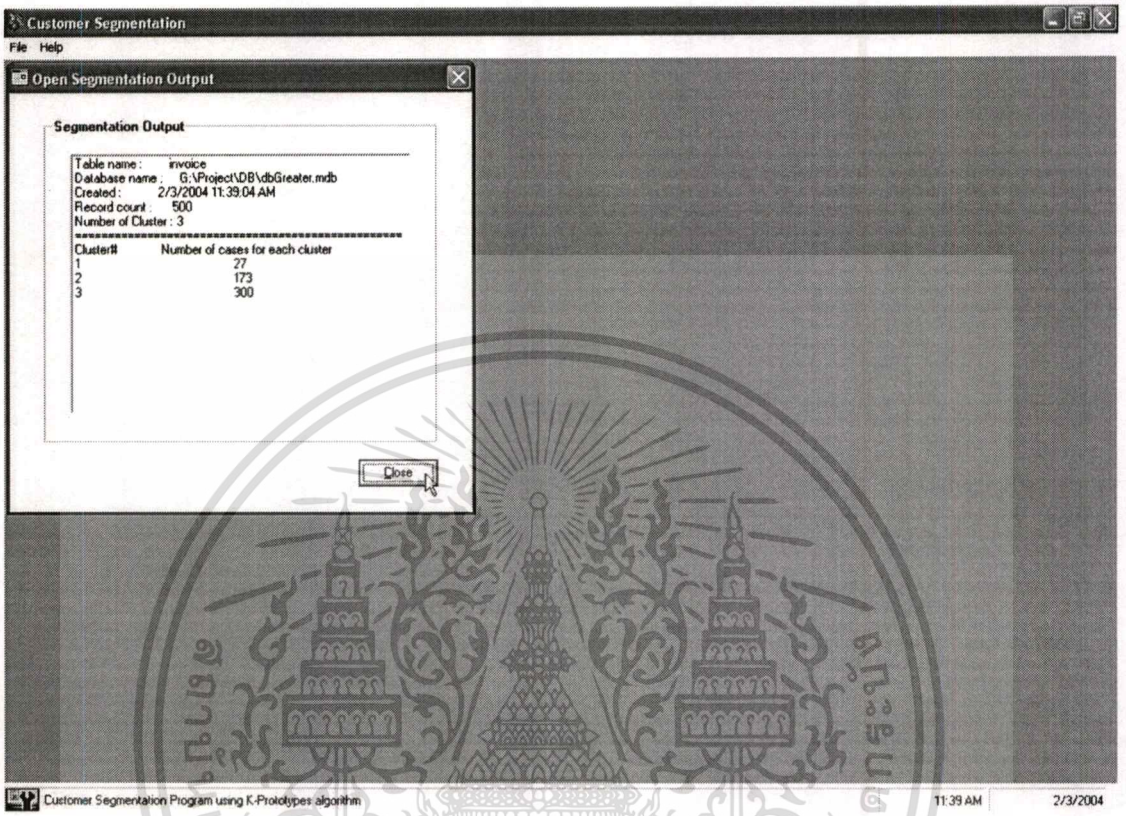
รูปที่ ก.11 หน้าจอแสดงผลเมนู File

- New Project – ทำหน้าที่เริ่มทำการจัดกลุ่มใหม่ตั้งแต่เริ่มต้น
- Open Output – ทำหน้าที่เปิดไฟล์ผลลัพธ์ที่เคยทำการบันทึกไว้แล้ว
- Open Database – ทำหน้าที่เปิดไฟล์ฐานข้อมูลที่ต้องการนำมาจัดกลุ่มข้อมูล
- Save as Text file – ทำหน้าที่บันทึกผลการจัดกลุ่มข้อมูลเป็นไฟล์ประเภท Text
- Save as Excel – ทำหน้าที่บันทึกผลการจัดกลุ่มข้อมูลเป็นไฟล์ประเภท Excel
- Exit – ทำการออกจากโปรแกรม

2.1) ถ้าเลือกที่ File > New Project จะปรากฏหน้าจอดังรูป ก.12

รูปที่ ก.12 หน้าจอแสดงผลเมื่อเลือกเมนู File > New Project

2.2) ถ้าเลือกที่ File > Open Output จะปรากฏหน้าจอดังรูปที่ ก.13



รูปที่ ก.13 หน้าจอแสดงผลเมื่อเลือกเมนู Open Output

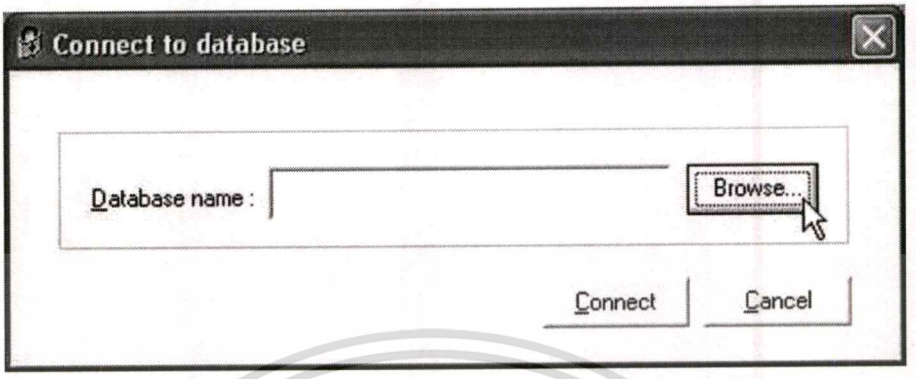
หน้าจอนี้จะแสดงรายละเอียดโดยสรุปของข้อมูลที่เคยทำการจัดกลุ่มมาแล้ว โดยจะบอก

- ชื่อตารางที่ทำการจัดกลุ่ม
- ชื่อฐานข้อมูลรวมทั้ง Path ของฐานข้อมูลนั้น
- วันที่ทำการจัดกลุ่ม จำนวนข้อมูลทั้งหมดที่ทำการจัดกลุ่ม
- จำนวนกลุ่มที่ทำการจัดกลุ่ม
- และสุดท้ายจะบอกว่าแต่ละกลุ่มข้อมูลมีจำนวนข้อมูลเท่าไร

โดยถ้าเลือกที่ปุ่ม 

จะทำการปิดหน้าต่างแสดงผล Output

2.3) ถ้าเลือกที่ File > Open Database จะปรากฏหน้าจอดังรูป ก.14



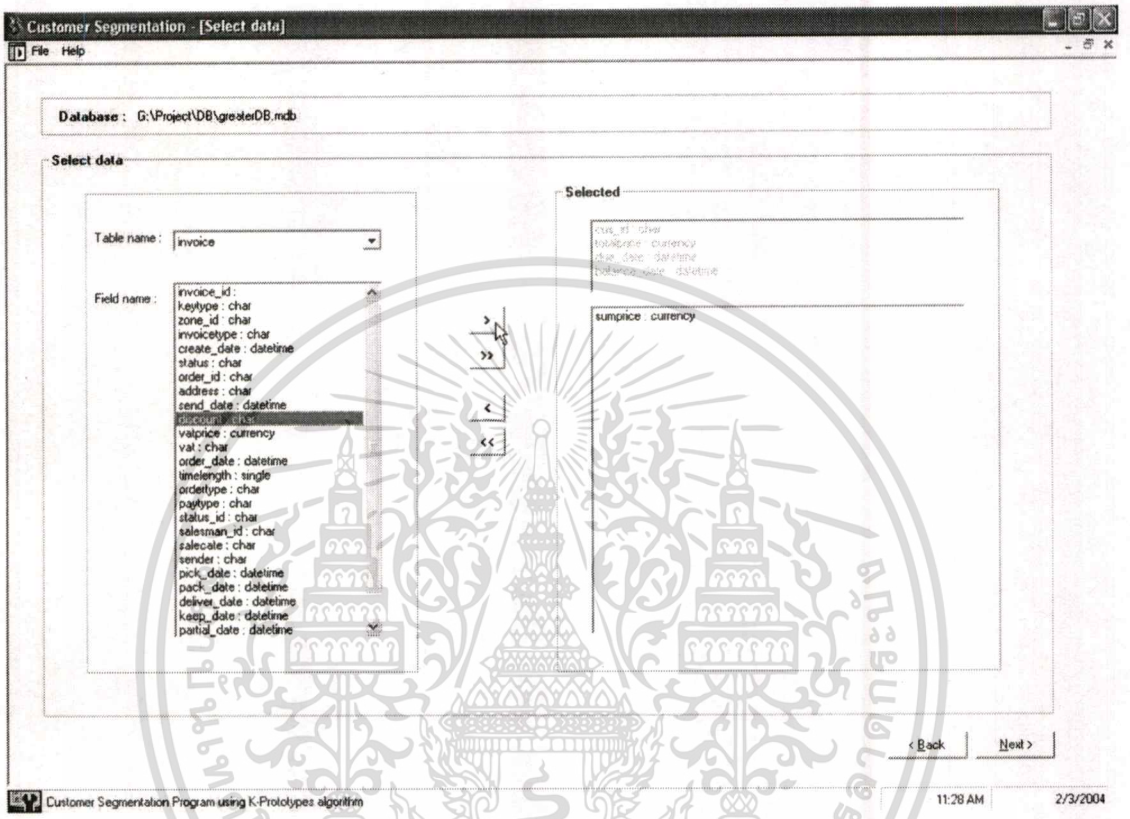
รูปที่ ก.14 หน้าจอแสดงผลที่ใช้ในการติดต่อฐานข้อมูล

2.3.1) เลือกฐานข้อมูล

<u>ขั้นที่ 1</u> คลิกที่ปุ่ม Browse... เพื่อทำการเลือกฐานข้อมูล

<u>ขั้นที่ 2</u> หลังจากที่ทำการเลือกฐานข้อมูลเรียบร้อยแล้วให้ทำการคลิกที่ปุ่ม Connect เพื่อทำการติดต่อกับฐานข้อมูล แต่ถ้าต้องการยกเลิกการเลือกฐานข้อมูลให้คลิกที่ปุ่ม Cancel

2.3.2) หลังจากที่คลิกที่ปุ่ม Connect จากขั้นตอนที่ 2.2.1 แล้วจะปรากฏหน้าจอดังรูป ก.15 เพื่อให้ทำการเลือกตารางข้อมูลและฟิลด์ข้อมูล



รูปที่ ก.15 หน้าจอแสดงผลการเลือกข้อมูล

ขั้นที่ 1 เลือกที่ invoice (Combo box) เพื่อทำการเลือกตารางข้อมูลจากไฟล์ฐานข้อมูล

ขั้นที่ 2 เลือกฟิลด์ข้อมูลจาก List box ในเฟรมด้านซ้าย แล้วคลิกที่ปุ่มดังนี้

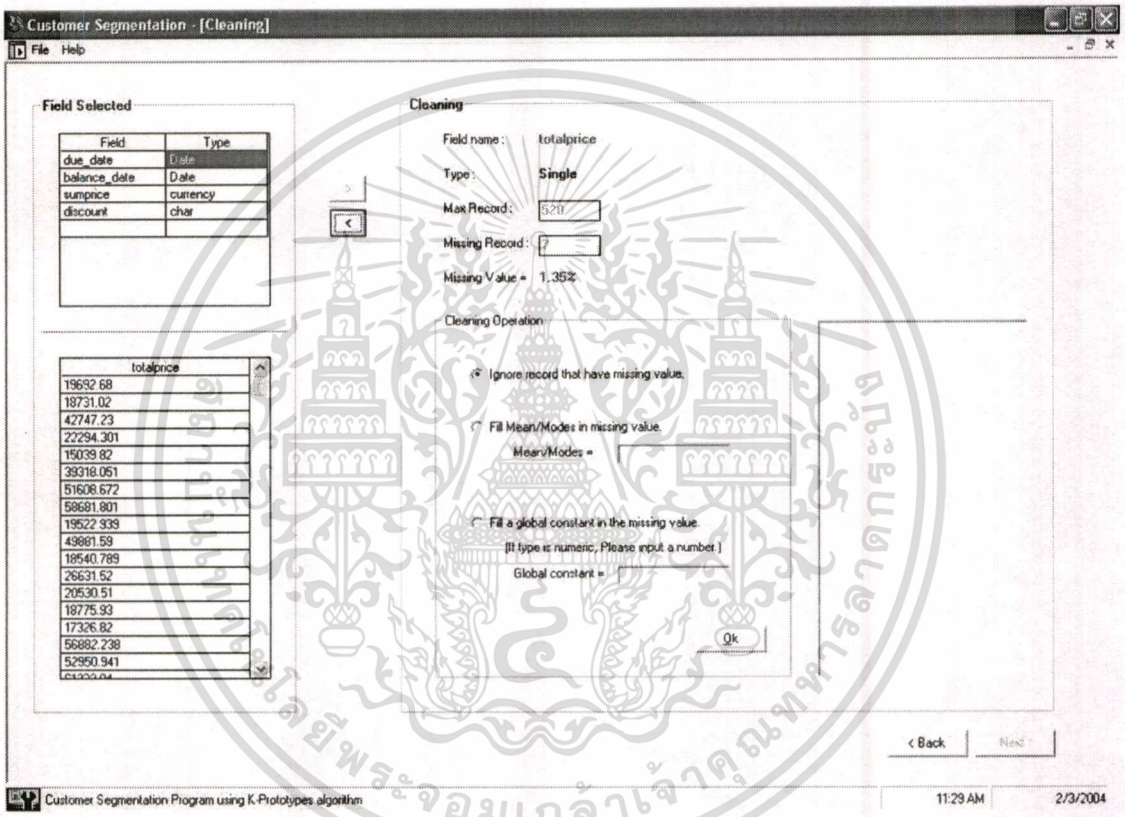
- > เพื่อทำการเลือกฟิลด์ข้อมูลที่ต้องการ
- >> เพื่อทำการเลือกฟิลด์ข้อมูลทั้งหมด

แต่ถ้าต้องการยกเลิกการเลือกข้อมูล ให้คลิกที่ปุ่มดังนี้

- < เพื่อทำการย้ายฟิลด์ข้อมูลกลับ
- << เพื่อทำการย้ายฟิลด์ข้อมูลกลับทั้งหมด

ขั้นที่ 3 คลิกที่ปุ่ม Next > เพื่อทำขั้นตอนถัดไป แต่ถ้าต้องการเลือกฐานข้อมูลอื่นให้คลิกที่ปุ่ม < Back จะกลับไปยังหน้าจอเลือกฐานข้อมูล

2.3.3) หลังจากที่ทำการเลือกข้อมูลมาแล้ว จะทำการ Clean ข้อมูลซึ่งจะปรากฏหน้าจอดังรูปที่ ก. 16



รูปที่ ก.16 หน้าจอแสดงผลการ Clean ข้อมูล

ขั้นที่ 1 เลือกฟิลด์ข้อมูลที่ List ด้านบนซ้าย แล้วคลิกที่ปุ่ม > จะแสดงข้อมูลของฟิลด์ที่เลือกนั้นใน List ด้านล่างซ้าย และที่เฟรมด้านขวาจะบอกชื่อฟิลด์ ประเภทของฟิลด์ จำนวนข้อมูลทั้งหมดในฟิลด์นั้น จำนวนข้อมูลที่ขาดหายไป และสุดท้ายจะบอกเป็นเปอร์เซ็นต์ที่ข้อมูลขาดหายไปในฟิลด์นั้น

ขั้นที่ 2 ทำการเลือกวิธีในการจัดการกับข้อมูลที่ขาดหายไปซึ่งมีให้เลือก 3 วิธีคือ

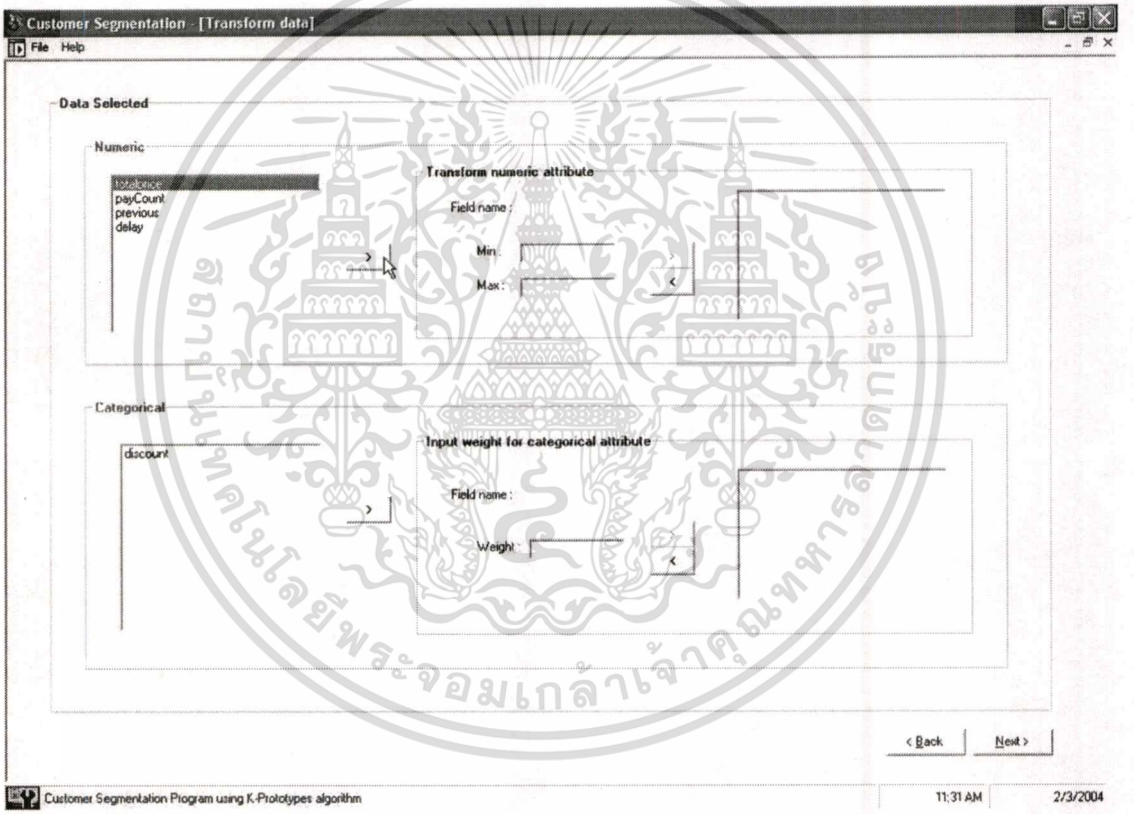
- ลบเรคอร์ดนั้นออกจากฐานข้อมูล

- เติมค่าเฉลี่ยในกรณีที่เป็นข้อมูลประเภท Numeric และเติมค่าฐานนิยมในกรณีที่เป็นข้อมูลประเภท Categorical
- เติมค่าทั่วไปโดยผู้ใช้โปรแกรมจะเป็นผู้ใส่ค่าลงไปเอง

<u>ขั้นที่ 3</u> คลิกที่ปุ่ม Ok และวนกลับไปทำที่ขั้นที่ 1-3 ใหม่จนครบทุกฟิลด์

<u>ขั้นที่ 4</u> คลิกที่ปุ่ม Next > เพื่อไปยังขั้นตอนการแปลงข้อมูล

2.3.4) เมื่อทำการ Clean ข้อมูลเสร็จแล้วจะปรากฏหน้าจอการแปลงข้อมูล ดังรูปที่ ก.17



รูปที่ ก.17 หน้าจอแสดงผลการแปลงข้อมูล

<u>ขั้นที่ 1</u> ทำการแปลงข้อมูลประเภท Numeric ให้อยู่ในช่วงที่ต้องการ โดยทำการเลือกฟิลด์ข้อมูลประเภท Numeric และทำการคลิกที่ปุ่ม > จะปรากฏชื่อฟิลด์ข้อมูล ค่าต่ำสุดและค่าสูงสุดเดิมที่อยู่ในฐานข้อมูล

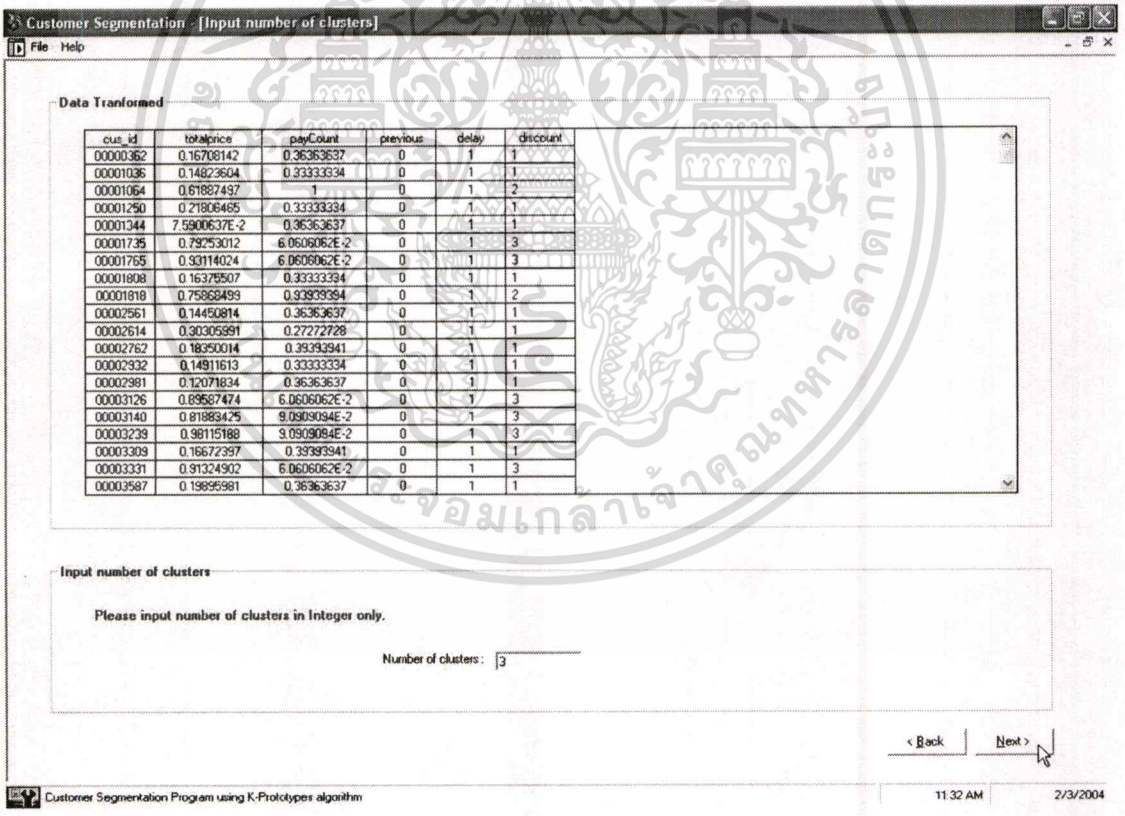
ขั้นที่ 2 ใส่ค่าที่ต้องการแปลงให้อยู่ในช่วง โดยใส่ค่าต่ำสุดที่กล่องข้อความ Min : และใส่ค่าสูงสุดที่กล่องข้อความ Max : จากนั้นคลิกที่ปุ่ม > เมื่อทำการแปลงข้อมูลไปแล้วแต่ต้องการจะแก้ไขข้อมูลช่วงอีกครั้งให้ทำการคลิกที่ปุ่ม <

ขั้นที่ 3 ทำข้อ 1-2 ซ้ำจนครบทุกข้อมูลที่อยู่ใน List ข้อมูลประเภท Numeric

ขั้นที่ 4 ขั้นตอนนี้จะทำการกำหนดค่า weight ให้กับข้อมูลประเภท Categorical ซึ่งการใช้งานปุ่มจะเหมือนกับวิธีการแปลงข้อมูลของข้อมูลประเภท Numeric

ขั้นที่ 5 คลิกที่ปุ่ม Next > เพื่อไปยังขั้นตอนการใส่ค่าจำนวนกลุ่มข้อมูลที่ต้องการจัดกลุ่ม

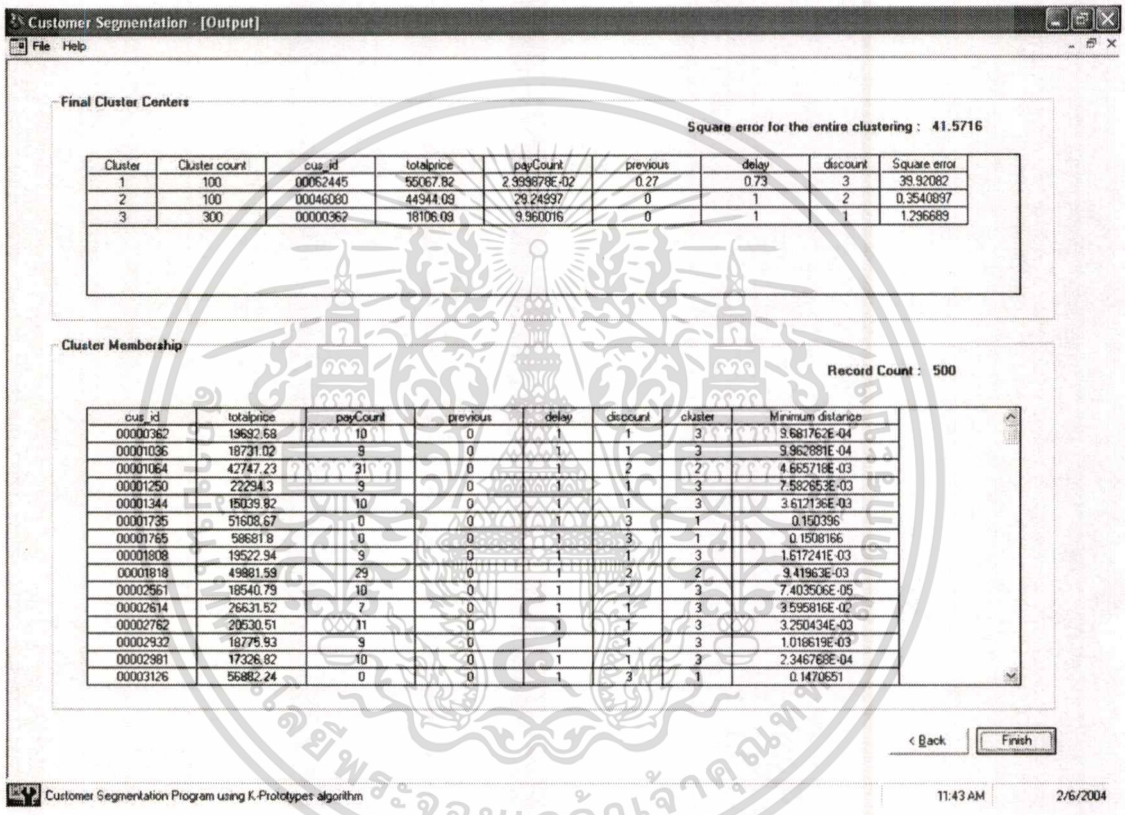
2.3.5) หลังจากที่ทำการเตรียมข้อมูลจากข้อที่ 2.2.2 ถึงข้อที่ 2.2.4 แล้วขั้นตอนนี้จะทำการใส่ค่าจำนวนกลุ่มข้อมูลที่ต้องการจัดกลุ่ม ซึ่งจะปรากฏหน้าจอดังรูป ก 18



รูปที่ ก.18 หน้าจอแสดงผลการใส่ค่าจำนวนกลุ่มข้อมูล

เฟรมด้านบนจะแสดงข้อมูลหลังจากทำการแปลงข้อมูลแล้ว ส่วนเฟรมด้านล่างจะให้ผู้ใช้โปรแกรมใส่ค่าจำนวนกลุ่มข้อมูลที่ต้องการจัดกลุ่ม และจากนั้นให้ทำการคลิกที่ปุ่ม Next >

2.3.6) เมื่อใส่จำนวนกลุ่มแล้ว โปรแกรมจะทำการจัดกลุ่มให้ตามจำนวนกลุ่มที่ต้องการและจะปรากฏหน้าจอดังรูป ก.19

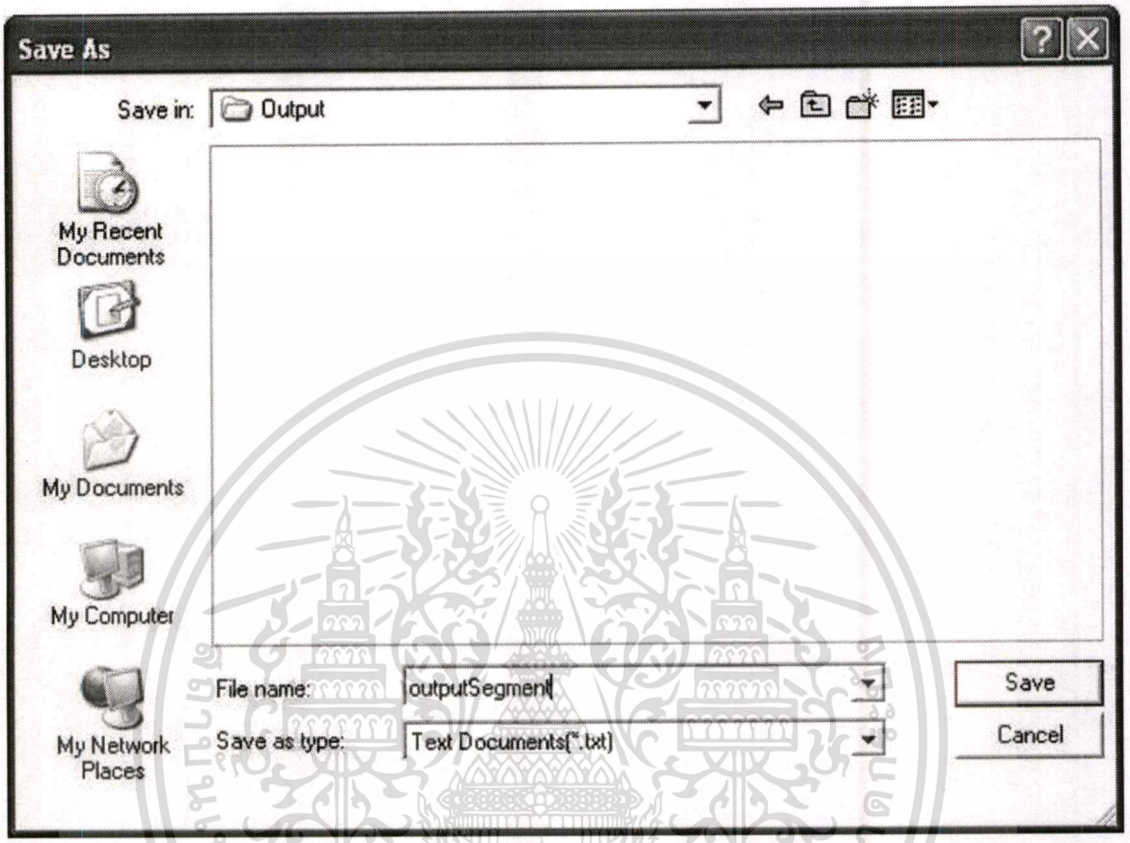


รูปที่ ก.19 หน้าจอแสดงผลการจัดกลุ่มข้อมูล

ที่เฟรมด้านบนจะแสดงจุดศูนย์กลางกลุ่มข้อมูลในแต่ละกลุ่มข้อมูล ส่วนเฟรมด้านล่างจะแสดงข้อมูลทั้งหมด รวมทั้งบอกกลุ่มที่ข้อมูลนั้นอยู่และบอกระยะห่างจากจุดศูนย์กลางกลุ่มข้อมูลนั้น (Minimum distance)

- ถ้าต้องการจบโปรแกรมให้คลิกที่ปุ่ม Finish
- ถ้าต้องการกลับไปทำขั้นตอนอื่นๆในก่อนหน้านี้ให้คลิกที่ปุ่ม < Back

2.4) ถ้าเลือกที่ File > Save as Text file จะปรากฏหน้าจอดังต่อไปนี้

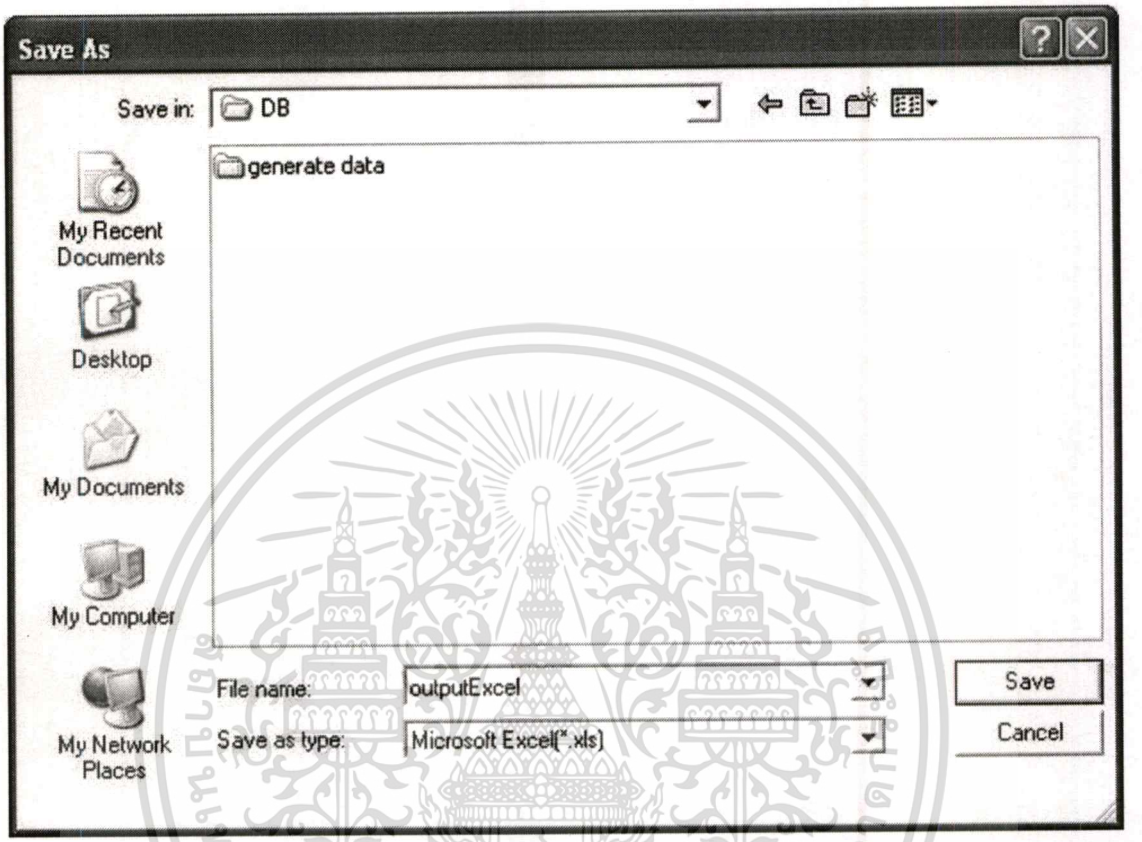


รูปที่ ก.20 หน้าจอแสดงผลเมื่อเลือกเมนู File > Save as Text file

ใส่ชื่อไฟล์ข้อมูลที่ต้องการบันทึกในกล่องข้อความ File Name จากนั้นให้คลิกที่ปุ่ม Save เพื่อทำการบันทึกผลลัพธ์

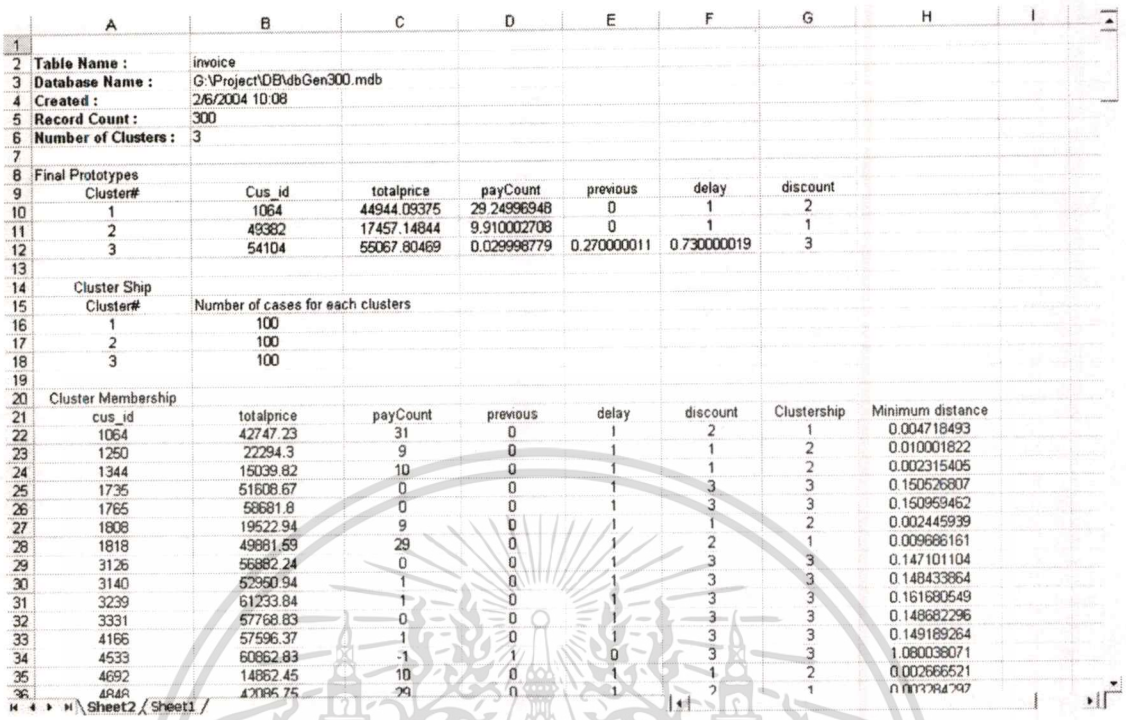
หลังจากทำการบันทึกไฟล์ข้อมูลเป็นไฟล์ประเภท Text แล้ว สามารถเปิดดูผลลัพธ์ได้อีกโดยการเปิดดูที่เมนู Open Output หรือสามารถเปิดดูได้จากโปรแกรม Notepad

2.5) ถ้าเลือกที่ File > Save as Excel จะปรากฏหน้าจอดังต่อไปนี้



รูปที่ ก.21 หน้าจอแสดงผลเมื่อเลือกเมนู File > Save as Excel

หลังจากที่ทำการบันทึกผลลัพธ์เป็นไฟล์ Excel แล้วสามารถเปิดดูผลลัพธ์ได้จากโปรแกรม Excel จากรูปเป็นตัวอย่างของ Output ที่ Save เป็นไฟล์ Excel

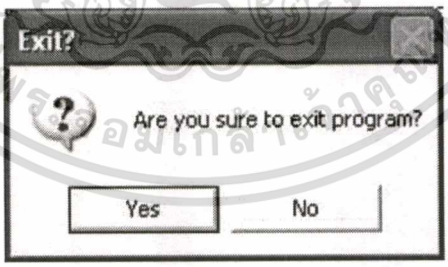


	A	B	C	D	E	F	G	H	I
1									
2	**Table Name :**	invoice							
3	**Database Name :**	G:\Project\DB\dbGen300.mdb							
4	**Created :**	2/6/2004 10:08							
5	**Record Count :**	300							
6	**Number of Clusters :**	3							
7									
8	Final Prototypes								
9	Cluster#	Cus_id	totalprice	payCount	previous	delay	discount		
10	1	1064	44944.09375	29.24996948	0	1	2		
11	2	49382	17457.14844	9.910002708	0	1	1		
12	3	54104	55067.80469	0.029998779	0.270000011	0.730000019	3		
13									
14	Cluster Ship								
15	Cluster#	Number of cases for each clusters							
16	1	100							
17	2	100							
18	3	100							
19									
20	Cluster Membership								
21	cus_id	totalprice	payCount	previous	delay	discount	Clustership	Minimum distance	
22	1064	42747.23	31	0	1	2	1	0.004718493	
23	1250	22294.3	9	0	1	1	2	0.010001822	
24	1344	15039.82	10	0	1	1	2	0.002315405	
25	1735	51608.67	0	0	1	3	3	0.150526807	
26	1765	58681.8	0	0	1	3	3	0.150959462	
27	1808	19522.94	9	0	1	1	2	0.002445939	
28	1818	49881.59	29	0	1	2	1	0.009686161	
29	3126	56882.24	0	0	1	3	3	0.147101104	
30	3140	52950.94	1	0	1	3	3	0.148433864	
31	3239	61233.84	1	0	1	3	3	0.161680549	
32	3331	57768.83	0	0	1	3	3	0.148682296	
33	4166	57596.37	1	0	1	3	3	0.149189264	
34	4533	60862.83	-1	1	0	3	3	1.080038071	
35	4692	14862.45	10	0	1	1	2	0.002666521	
36	4848	42086.75	29	0	1	2	1	0.003284297	

Sheet2 / Sheet1 /

รูปที่ ก.22 หน้าจอแสดงผลการบันทึกข้อมูลในไฟล์ประเภท Excel

2.6) ถ้าเลือกที่ File > Exit จะปรากฏหน้าจอดังต่อไปนี้ เพื่อทำการยืนยันว่าต้องการออกจากโปรแกรมหรือไม่



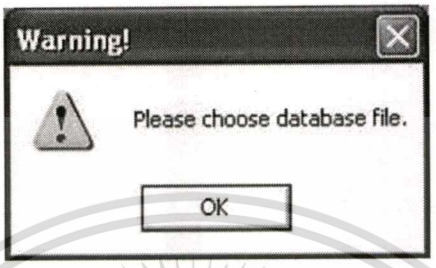
รูปที่ ก.23 หน้าจอแสดงผลเมื่อต้องการออกจากโปรแกรม

- ถ้าต้องการออกจากโปรแกรมให้คลิกที่ปุ่ม Yes
- ถ้าไม่ต้องการออกจากโปรแกรมให้คลิกที่ปุ่ม No

3. Message Box ต่างๆที่มีในโปรแกรมมีดังนี้

3.1) หน้าจอการเลือกฐานข้อมูล

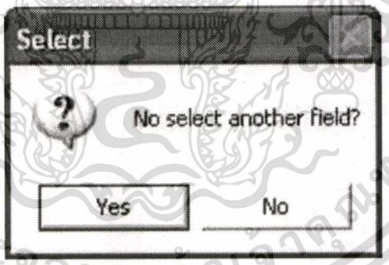
- แสดงเมื่อไม่ได้ทำการเลือกฐานข้อมูล แล้วคลิกที่ปุ่ม Connect



รูปที่ ก.24 ข้อความเตือนเมื่อไม่ได้ทำการเลือกฐานข้อมูล

3.2) หน้าจอการเลือกตารางข้อมูลและฟิลด์ข้อมูล

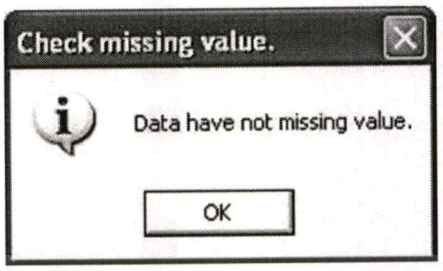
3.2.1) แสดงเมื่อผู้ใช้โปรแกรมไม่ได้ทำการเลือกฟิลด์ข้อมูล โดยจะโปรแกรมจะถามว่าต้องการจะไม่เลือกฟิลด์ข้อมูลใช่หรือไม่



รูปที่ ก.25 หน้าจอถามเมื่อไม่ได้ทำการเลือกฟิลด์ข้อมูล

- คลิกที่ปุ่ม Yes ถ้าต้องการยืนยันว่าจะไม่เลือกฟิลด์ข้อมูลอื่นๆ
- คลิกที่ปุ่ม No ถ้าต้องการเลือกฟิลด์ข้อมูลอื่นเพิ่ม

3.2.2) แสดงเมื่อข้อมูลที่เลือกมาจากฐานข้อมูล ไม่พบค่าที่ขาดหายไป (Missing value)



รูปที่ ก.26 ข้อความเตือนเมื่อไม่พบ Missing Value

3.3) หน้าจอการ Clean ข้อมูล

- แสดงเมื่อผู้ใช้โปรแกรมเลือกวิธีการจัดการกับข้อมูลที่ขาดหายไป โดยใช้วิธีเติมค่าทั่วไป ซึ่งจะปรากฏหน้าต่างนี้เพื่อแจ้งเตือนให้ผู้ใช้ใส่ค่าทั่วไปก่อน

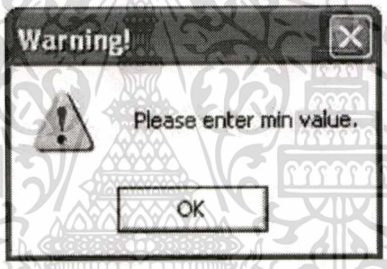
รูปที่ ก.27 ข้อความเตือนเมื่อไม่ได้ใส่ค่าทั่วไป

3.4) หน้าจอการแปลงข้อมูลประเภท Numeric และกำหนดค่า weight ให้กับข้อมูลประเภท Categorical

3.4.1) แสดงเมื่อทำการแปลงข้อมูลประเภท Numeric อยู่แล้วมีการเลือกฟิลด์ข้อมูลประเภท Numeric ตัวอื่นเข้ามาอีก ซึ่งจะปรากฏหน้าต่างนี้เพื่อแจ้งเตือนว่าไม่ให้ทำการเลือกข้อมูลอื่นอีกในขณะที่ทำการแปลงค่าข้อมูลอยู่

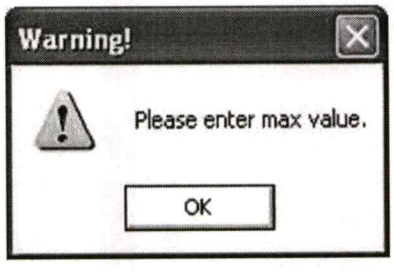
รูปที่ ก.28 ข้อความเตือนเมื่อมีการเลือกข้อมูลซ้ำ

3.4.2) แสดงเมื่อทำการแปลงข้อมูลประเภท Numeric และไม่ได้ใส่ค่าต่ำสุดของข้อมูลที่ต้องการจะแปลงหรือปล่อยให้เป็นค่าว่างนั่นเอง โดยโปรแกรมจะทำการแจ้งเตือนให้ใส่ค่าต่ำสุดของข้อมูล



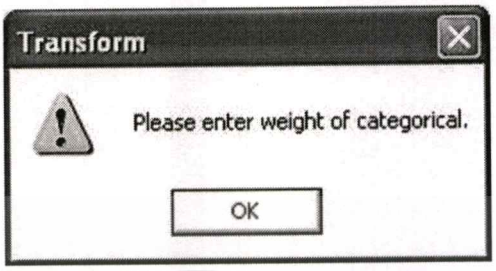
รูปที่ ก.29 ข้อความเตือนเมื่อไม่ได้ใส่ค่าต่ำสุดที่ต้องการแปลง

3.4.3) แสดงเมื่อทำการแปลงข้อมูลประเภท Numeric และไม่ได้ใส่ค่าสูงสุดของข้อมูลที่ต้องการจะแปลงหรือปล่อยให้เป็นค่าว่างนั่นเอง โดยโปรแกรมจะทำการแจ้งเตือนให้ใส่ค่าสูงสุดของข้อมูล



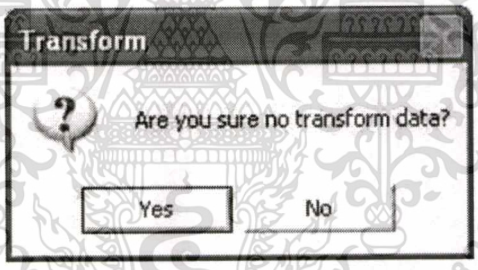
รูปที่ ก.30 ข้อความเตือนเมื่อไม่ได้ใส่ค่าสูงสุดที่ต้องการแปลง

3.4.4) แสดงเมื่อผู้ใช้โปรแกรมไม่ได้ทำการใส่ค่า weight ให้กับตัวแปรประเภท Categorical



รูปที่ ก.31 ข้อความเตือนเมื่อไม่ได้ใส่ค่า weight ให้กับข้อมูลประเภท Categorical

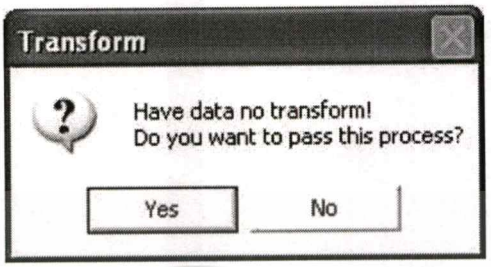
3.4.5) แสดงเมื่อข้อมูลประเภท Numeric ไม่ได้ทำการแปลงข้อมูล โดยโปรแกรมจะถามเพื่อยืนยันว่าไม่ต้องการแปลงข้อมูลจริง



รูปที่ ก.32 หน้าจอถามว่าต้องการแปลงข้อมูลหรือไม่

- คลิกที่ปุ่ม Yes เพื่อทำการยืนยันว่าไม่ต้องการแปลงข้อมูล
- คลิกที่ปุ่ม No เพื่อทำการกลับไปแปลงข้อมูล

3.4.6) แสดงเมื่อข้อมูลประเภท Numeric บางตัวไม่ได้ทำการแปลงข้อมูล โดยโปรแกรมจะถามเพื่อยืนยันว่าไม่ต้องการแปลงข้อมูลที่เหลืออยู่

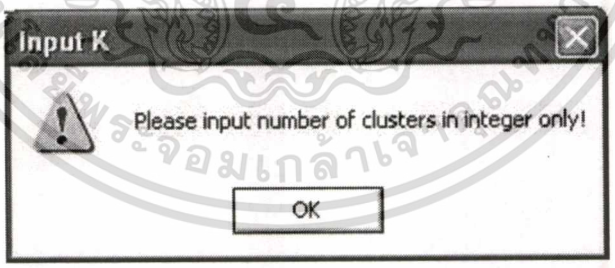


รูปที่ ก.33 หน้าจอถามว่าต้องการแปลงข้อมูลที่เหลือหรือไม่

- คลิกที่ปุ่ม Yes เพื่อทำการยืนยันว่าไม่ต้องการแปลงข้อมูลที่เหลืออยู่และข้ามกระบวนการแปลงข้อมูลนี้ไป
- คลิกที่ปุ่ม No เพื่อทำการกลับไปแปลงข้อมูลที่เหลืออยู่

3.5) หน้าจอการใส่ค่าจำนวนกลุ่มที่ต้องการจัดกลุ่ม

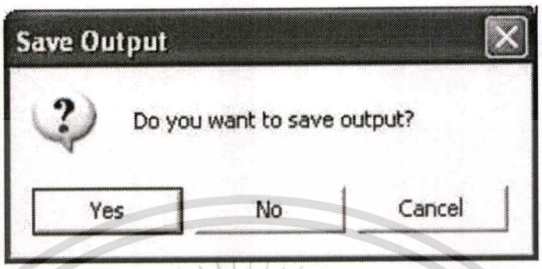
- แสดงเมื่อผู้ใช้โปรแกรมใส่ค่าจำนวนกลุ่มเป็นเลขทศนิยม หรือไม่ใช่ตัวเลข



รูปที่ ก.34 ข้อความเตือนเมื่อใส่ค่าจำนวนกลุ่มที่ไม่ใช่เลขประเภท Integer

3.6) หน้าจอแสดงผลการจัดกลุ่ม

- แสดงเมื่อผู้ใช้คลิกที่ปุ่ม Finish โดยโปรแกรมจะปรากฏหน้าจอนี้เพื่อถามว่าต้องการบันทึกข้อมูลผลลัพธ์ที่ได้จากการจัดกลุ่มหรือไม่



รูปที่ ก.35 หน้าจอถามว่าต้องการบันทึกผลลัพธ์หรือไม่

หลังจากที่ทำ Message Box นี้เสร็จแล้ว โปรแกรมจะกลับไปเริ่มต้นที่หน้าเมนูหลักอีกครั้ง

ประวัติผู้เขียน

ชื่อผู้เขียน	นางสาวสุพร เวคีวิทยา
วันเดือนปีเกิด	29 ตุลาคม 2524
สถานที่เกิด	กรุงเทพมหานคร
มัธยมศึกษา	โรงเรียนสายน้ำผึ้ง
ปริญญาตรี	สาขาคณิตศาสตร์ประยุกต์ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ปีที่สำเร็จการศึกษา	ปีการศึกษา 2544