การพัฒนาระบบสำหรับจัดกลุ่มลูกค้าเพื่อทำรายการส่งเสริมการขาย ผ่านทางเว็บไซต์ โดยใช้อัลกอริทึม k-prototypes

The Development of the System for Customer Clustering for Website Promotion Using k-prototypes Algorithm

โดย

นางสาว นภาศรี รุ่งธีรวัฒนานนท์

อาจารย์ที่ปรึกษา

ผศ.ดร.วรพจน์ กรีสุระเดช

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน
หลักสูตรวิทยาศาตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
ภาคเรียนที่ 2 ปีการศึกษา 2546
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ชื่อหัวข้อ	การพัฒนาระบบสำหรับจัดกลุ่มลูกค้าเพื่อทำรายการส่งเสริมการขายผ่านทางเว็บไซต์ โดยใช้อัลกอริทึม k-prototypes
นักศึกษา	นางสาวนภาศรี รุ่งธีรวัฒนานนท์
อาจารย์ที่ปรึกษา	ผศ.ดร.วรพจน์ กรีสุระเดช
ระดับการศึกษา	วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2546

บทคัดย่อ

ปัจจุบันการทำธุรกิจผ่านทางอินเตอร์เน็ตมีการแข่งขันกันสูง ทำให้แต่ละเว็บไซต์ต้องหากลยุทธ์เพื่อแย่งชิงลูกค้า กลยุทธ์หนึ่งคือการทำความเข้าใจในตัวลูกค้าหรือกลุ่มของลูกค้า เพื่อวางแผนกลยุทธ์ในการทำธุรกิจให้ตรงกับความต้องการของลูกค้า จุดประสงค์ในการพัฒนาโครงการคือ การจัดกลุ่มลูกค้าที่มีลักษณะใกล้เคียงกัน เพื่อนำเสนอรายการส่งเสริมการขายให้ตรงตามกลุ่มเป้าหมาย โดยใช้เทคนิคการแบ่งกลุ่มข้อมูลด้วยอัลกอริทึม k-prototypes ผลลัพธ์ที่ได้จะทำให้เข้าใจความต้องการของลูกค้ามากขึ้น และสามารถกำหนดเป้าหมายทางการตลาดได้ชัดเจนมากขึ้น

Title	The Development of the System for Customer Clustering for Website Promotion Using k-prototypes Algorithm
Student	Miss Napasri Roongtrirawattananont
Advisor	Asst. Prof. Dr.Worapoj Kreesuradej
Level of Study	Master of Science in Information Technology
Major	Information Science
Academic Year	2003

ABSTRACT

Nowadays, there are varied kinds of doing business via internet. Because there is intense competition, each of the websites has to develop their strategies for attracting customers. One of the strategies is to realize and understand customers and plan strategies for doing business that meets customers' requirements. The purpose of developing the project is clustering customers into similar groups and offer appropriate promotion to customers. K-Prototypes algorithm technique is used for categorizing, which its results can give us not only more understanding of customers' requirements but also an opportunity to set out market aims more distinct.

กิตติกรรมประกาศ

สำหรับการพัฒนาระบบงานครั้งนี้ ข้าพเจ้าขอขอบพระคุณ ผศ.ดร.วรพจน์ กรีสุระเดช อาจารย์ที่ปรึกษาวิชาโครงการพัฒนาระบบงานเป็นอย่างสูง ที่ได้กรุณาให้คำแนะนำและคำปรึกษา ทำให้การพัฒนาระบบสำเร็จลุล่วงด้วยดี

นอกจากนี้ข้าพเจ้าขอขอบพระคุณคุณพ่อคุณแม่ที่เป็นกำลังใจในการเรียนมาโดยตลอด และขอขอบคุณนายสราวุธิ ราษฎร์นิยม และเพื่อนๆคณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยกรุงเทพ รุ่น 12 ที่ให้คำแนะนำติชมอันเป็นประโยชน์ในการพัฒนางานชิ้นนี้

ข้าพเจ้าหวังเป็นอย่างยิ่งว่า โครงการพัฒนาระบบงานชิ้นนี้ จะเป็นประโยชน์และให้ความรู้แก่ผู้สนใจและบุคคลทั่วไป หากมีข้อบกพร่องประการใด ข้าพเจ้าขอน้อมรับไว้เพื่อพัฒนาระบบให้ดีขึ้นในโอกาสต่อไป

นางสาวนภาศรี รุ่งธีรวัฒนานนท์

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

ภาพที่

สารบัญภาพ (ต่อ)

หน้า

ภาพที่

สารบัญภาพ (ต่อ)

หน้า

ภาพที่

บทที่ 1
บทนำ

1.1 หลักการและเหตุผล

ในการทำธุรกิจ E-Commerce นั้นมีหลากหลายรูปแบบและมีการแข่งขันกันสูง ทำให้แต่ละเว็บไซต์ต้องหากลยุทธ์ เพื่อให้ลูกค้าที่เข้ามาเยี่ยมชมเว็บไซต์กลับมาใช้บริการเว็บไซต์อีก กลยุทธ์หนึ่งที่มีความสำคัญต่อความสำเร็จของธุรกิจคือ การทำความเข้าใจในตัวลูกค้า หรือกลุ่มของลูกค้า เพราะยิ่งรู้ข้อมูลของลูกค้ามากเท่าไร ก็จะยิ่งเข้าใจลูกค้ามากขึ้น โอกาสที่จะทำธุรกิจให้ตรงกับความต้องการของตลาดก็จะมากขึ้นไปด้วย ดังนั้นหลายๆเว็บไซต์จึงพยายามขอข้อมูลที่จะเป็นประโยชน์ของลูกค้าไว้ เพื่อเป็นช่องทางในการติดต่อกลับภายหลัง ซึ่งข้อมูลดังกล่าวเราสามารถค้นหาความรู้ได้ ด้วยเทคนิคต่างๆของดาต้าไมนิ่ง ที่สามารถวิเคราะห์ และกลั่นกรองข้อมูลที่จำนวนมากเหล่านี้ เพื่อให้ได้ข้อมูลที่มีประโยชน์และไม่เคยรู้มาก่อนล่วงหน้า

โครงการพัฒนาระบบงานนี้ จึงได้การนำเอาดาต้าไมนิ่งมาช่วยในการแบ่งกลุ่มลูกค้า เพราะการแบ่งลูกค้าเป็นจุดเริ่มต้นที่สำคัญในการทำความเข้าใจลูกค้าแต่ละกลุ่ม โดยนำเอาข้อมูลที่มีอยู่มาผ่านกระบวนการให้กลายเป็นความรู้ ซึ่งจะช่วยให้ผู้บริหารสามารถวางแผนการดำเนินธุรกิจ ช่วยลดต้นทุน และเพิ่มกิจกรรมทางการตลาดที่หลากหลายมากขึ้น และมีโอกาสที่จะลดเวลาในการสร้างรายการส่งเสริมการขายใหม่ๆ ทำให้ลูกค้าเกิดความพึงพอใจ ซึ่งเป็นการรักษาลูกค้าไว้และเพิ่มลูกค้าใหม่ที่ได้มาจากการเรียนรู้ กลยุทธ์ดังกล่าวจะช่วยสร้างผลกำไร และนำไปสู่ช่องทางที่จะทำให้ธุรกิจเติบโตมากยิ่งขึ้น

1.2 วัตถุประสงค์ในการพัฒนาระบบ

การศึกษาโครงการพัฒนาระบบงานนี้มีวัตถุประสงค์เพื่อศึกษาเทคนิคของดาต้าไมนิ่งมาใช้ในการแบ่งกลุ่มข้อมูล เพื่อจัดกลุ่มลูกค้าที่มีลักษณะใกล้เคียงกัน จะช่วยให้ผู้บริหารสามารถวางแผนกลยุทธ์และกำหนดเป้าหมายทางการตลาดในการนำเสนอรายการส่งเสริมการขายหรือสิทธิพิเศษและบริการต่างๆ ให้ตรงตามกลุ่มเป้าหมายได้ชัดเจนมากขึ้น

1.3 ขอบเขตของการพัฒนาระบบ

โครงการพัฒนาระบบงานนี้ ได้กำหนดขอบเขตของการศึกษาไว้ ดังนี้

1. ศึกษาถึงการนำเอาเทคนิคดาต้าไมนิ่งมาประยุกต์ใช้ โดยอาศัยหลักการของ Database Segmentation (Clustering Analysis) ด้วยอัลกอริทึม k-prototypes ในการแบ่งกลุ่มข้อมูลลูกค้า
2. ข้อมูลที่นำมาใช้ในการวิเคราะห์นั้น เป็นข้อมูลของลูกค้าที่ได้สมัครเป็นสมาชิกกับเว็บไซต์ โดยจะนำข้อมูลมาผ่านขั้นตอนต่างๆของดาต้าไมนิ่ง แล้วจัดกลุ่มลูกค้าที่มีลักษณะใกล้เคียงกัน

1.4 ขั้นตอนและวิธีการดำเนินงาน

ในการศึกษาโครงการพัฒนาระบบงานนี้ เพื่อให้ศึกษาคลอบคลุมวัตถุประสงค์และขอบเขตในการพัฒนาระบบ จึงได้กำหนดขั้นตอนในการศึกษาไว้ดังนี้

1. กำหนดวัตถุประสงค์ในการแบ่งกลุ่มลูกค้า
2. ศึกษาแนวคิด ขั้นตอน และกระบวนการในการทำดาต้าไมนิ่งโดยเลือกใช้อัลกอริทึม k-prototypes
3. เก็บรวบรวม และศึกษาข้อมูลที่จะนำมาใช้ในการแบ่งกลุ่ม โดยเลือกใช้ข้อมูลของลูกค้าที่ได้สมัครเป็นสมาชิกกับเว็บไซต์
4. เลือกข้อมูล และเตรียมข้อมูลให้อยู่ในรูปแบบที่เหมาะสมกับอัลกอริทึม
5. ออกแบบและพัฒนาระบบงานเพื่อวิเคราะห์ข้อมูล

1.5 ประโยชน์ที่คาดว่าจะได้รับ

จากการศึกษาและพัฒนาระบบงานเพื่อแบ่งกลุ่มลูกค้าที่มีลักษณะใกล้เคียงกัน คาดว่าจะให้ประโยชน์และเป็นแนวทางในการวางแผนทางธุรกิจ ดังนี้

1. ช่วยทำให้เข้าใจความต้องการของลูกค้ามากขึ้น เช่น เข้าใจเอกลักษณ์ของลูกค้าที่จะทำกำไรให้กับเว็บไซต์
2. สามารถกำหนดเป้าหมายทางการตลาดได้ชัดเจนมากขึ้น เช่น สร้างและเสนอรายการส่งเสริมการขายที่ตรงกับความต้องการของลูกค้าในเวลาที่เหมาะสม
3. เมื่อมีรายการส่งเสริมการขายที่ตรงใจลูกค้ากลุ่มเป้าหมายแล้ว จะส่งผลให้ลูกค้ากลุ่มดังกล่าวกลับมาเยี่ยมชมเว็บไซต์อีก (Customer loyalty)

บทที่ 2
ดาต้าไมนิ่งและทฤษฎีที่เกี่ยวข้อง

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

2.1 ดาต้าไมนิ่ง (Data Mining)

ดาต้าไมนิ่ง (Data Mining) หรือการค้นหาความรู้จากฐานข้อมูลขนาดใหญ่ (Knowledge Discovery from very large Database: KDD) คือขั้นตอนหรือกระบวนการทำงานที่กลั่นกรองข้อมูล (Extract Data) จากฐานข้อมูลขนาดใหญ่ โดยดึงข้อมูลที่ซ่อนเร้นอยู่หรือส่วนที่เป็นนัยของข้อมูลที่เราไม่ทราบมาก่อน เพื่อให้ได้สารสนเทศที่มีประโยชน์ (Useful Information) แล้วนำสารสนเทศนั้นมาเป็นฐานความรู้เพื่อช่วยในการบริหารงานและสนับสนุนการตัดสินใจทางด้านธุรกิจ

สารสนเทศที่ได้จากการทำดาต้าไมนิ่งนั้น จะต้องมีลักษณะดังนี้ คือ

- ข้อมูลที่ไม่รู้มาก่อน(Unknown) คือข้อมูลที่ผู้ใช้งานไม่รู้มาก่อนล่วงหน้าและยังไม่ชัดเจน ไม่สามารถตั้งสมมติฐานล่วงหน้าว่ามีแบบแผนอย่างไร
- ข้อมูลที่มีความถูกต้อง (Valid) คือข้อมูลที่ค้นพบใหม่นั้น จะต้องมีความถูกต้อง มีเหตุผล
- ข้อมูลที่สามารถนำไปใช้ในทางปฏิบัติได้ (Actionable) คือข้อมูลจะต้องถูกแปลงออกมา และสามารถนำไปใช้ประโยชน์ได้เพื่อสนับสนุนการตัดสินใจ ทำให้เกิดความได้เปรียบในเชิงธุรกิจ

2.2 กระบวนการทำงานของดาต้าไมนิ่ง (Data Mining Process)

กระบวนการทำงานของดาต้าไมนิ่งเป็นกระบวนการในการสร้างแบบจำลอง (Model) ของกลุ่มข้อมูล เพื่อให้เกิดความเข้าใจในแนวโน้ม รูปแบบ ความเกี่ยวข้องสัมพันธ์กันของกลุ่มข้อมูล และสามารถทำนายลักษณะของข้อมูลนั้นๆได้ ซึ่งกระบวนการทำงานของดาต้าไมนิ่งนั้นมี 5 ขั้นตอน คือ

1. การกำหนดวัตถุประสงค์ทางธุรกิจ (Business Objective Determination)
2. การเตรียมข้อมูล (Data Preparation)
3. การทำดาต้าไมนิ่ง (Data Mining)
4. การวิเคราะห์ผลลัพธ์ (Analysis of Result)
5. การนำความรู้ที่ได้ไปใช้ (Assimilation of Knowledge)

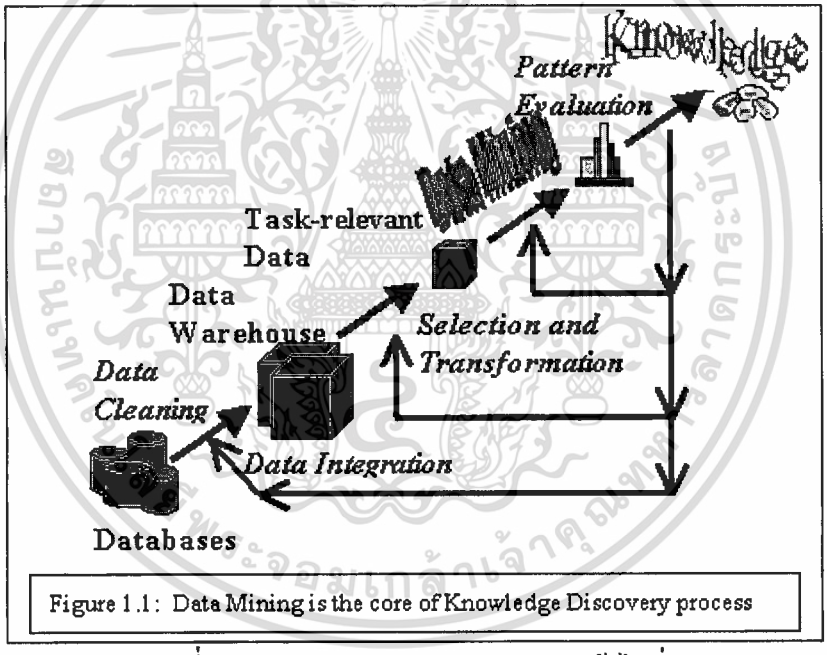


Figure 1.1: Data Mining is the core of Knowledge Discovery process

รูปที่ 2.1 แสดงกระบวนการทำงานของดาต้าไมนิ่ง

ขั้นตอนที่ 1 การกำหนดวัตถุประสงค์ทางธุรกิจ (Business Objective Determination)

การกำหนดวัตถุประสงค์ทางธุรกิจนั้นเป็นขั้นตอนสำคัญของกระบวนการทำดาต้าไมนิ่ง ซึ่งจะต้องกำหนดวัตถุประสงค์ให้ชัดเจน ต้องเข้าใจถึงปัญหาและความต้องการทางธุรกิจ เพราะวัตถุประสงค์จะเป็นตัวกำหนดทิศทางในการทำดาต้าไมนิ่ง หากวัตถุประสงค์ที่กำหนดไม่ชัดเจนหรือคลุมเครือ เมื่อทำดาต้าไมนิ่งเสร็จแล้วอาจไม่สามารถนำไปใช้ได้จริง ต้องกลับไปเริ่มต้นใน

ขั้นตอนแรกใหม่อีกครั้ง นอกจากนี้ยังต้องคำนึงถึงความเป็นไปได้ในการทำ และมีความจำเป็นหรือไม่ เพราะบางครั้งการทำดาต้าไมนิ่งอาจไม่ได้ช่วยแก้ปัญหาที่เกิดขึ้นก็เป็นได้

ขั้นตอนที่ 2 การเตรียมข้อมูล (Data Preparation)

ข้อมูลที่สามารถทำดาต้าไมนิ่งได้นั้นมีหลายรูปแบบ เช่น Relational Database, Data Warehouses, Transaction Database และฐานข้อมูลที่จัดเก็บในรูปแบบอื่นๆ เช่น Object-Oriented Databases, Text Databases, Multimedia Databases และข้อมูลในรูปของเว็บไซต์ จะเห็นได้ว่าข้อมูลที่สามารถนำมาใช้นั้นมีความหลากหลายมาก ดังนั้นขั้นตอนการเตรียมข้อมูล จึงถือเป็นหัวใจของกระบวนการทำดาต้าไมนิ่งเพราะหากไม่ทำขั้นตอนนี้อย่างมีประสิทธิภาพแล้ว จะทำให้ผลลัพธ์ที่ได้ในขั้นตอนสุดท้ายไม่มีความถูกต้อง และไม่สามารถนำไปใช้ได้จริง ขั้นตอนที่สำคัญนี้จะใช้เวลามากถึงประมาณ 60 เปอร์เซ็นต์ของเวลาทั้งหมดในการทำดาต้าไมนิ่ง ซึ่งมากกว่าขั้นตอนอื่นๆ เนื่องจากข้อมูลต่างๆที่นำมาใช้นั้นอาจรวบรวมมาจากหลายแหล่ง และอาจมีความแตกต่างกัน ข้อมูลที่ได้จากขั้นตอนนี้จะต้องถูกต้อง มีคุณภาพ เพื่อให้สามารถนำไปใช้ในขั้นตอนการทำ Data Mining Operation ในขั้นตอนถัดไป

ขั้นตอนในการเตรียมข้อมูลแบ่งออกเป็น 3 ขั้นตอน ดังนี้

1. การคัดเลือกข้อมูล (Data Selection)

การคัดเลือกข้อมูลจะเป็นการระบุและเลือกข้อมูลที่ต้องการจากแหล่งข้อมูลต่างๆทั้งภายในและภายนอกองค์กร การคัดเลือกข้อมูลที่มีอยู่จำนวนมากนั้นจะต้องคำนึงถึงวัตถุประสงค์ทางธุรกิจที่กำหนดไว้ ข้อมูลที่ถูกคัดเลือกจะต้องเป็นตัวแทนที่ดีของข้อมูลทั้งหมด เป็นข้อมูลที่ไม่มีการเปลี่ยนแปลงตลอดช่วงเวลาที่ทำการไมนิ่ง เพื่อให้ได้ผลลัพธ์ที่มีความถูกต้อง นอกจากนี้ยังต้องคำนึงถึงการได้มาของข้อมูลว่ามีสิทธิ์ในการเข้าถึงข้อมูลหรือไม่ เช่นในกรณีเป็นข้อมูลจากภายนอกองค์กร

การคัดเลือกข้อมูลนั้นจะต้องพิจารณาถึงชนิดของข้อมูล ซึ่งมี 2 ลักษณะได้แก่

1) Quantitative Data (ข้อมูลที่เป็นตัวเลข) แบ่งเป็น
   - Discrete คือค่าที่เป็นตัวเลขจำนวนเต็ม (Integer) เช่น 1, 2, 10, 25, 99 เป็นต้น
   - Continuous คือค่าที่เป็นเลขจำนวนจริง (Real Number) เช่น 0.25, 8.7, 22.7, 120.1 เป็นต้น
2) Categorical Data (ข้อมูลที่เป็นกลุ่ม) แบ่งเป็น
   - Nominal คือข้อมูลที่ลำดับของข้อมูลไม่มีความสำคัญ เช่น เพศ (ชาย, หญิง) เป็นต้น

- Ordinal คือข้อมูลที่ลำดับของข้อมูลมีความสำคัญ เช่น ระดับการศึกษา (ต่ำกว่าปริญญาตรี; ระดับปริญญาตรี, สูงกว่าปริญญาตรี) เป็นต้น

นอกจากนี้ ยังต้องพิจารณาถึง

1) ระดับของข้อมูลที่พิจารณา

ซึ่งขึ้นกับวัตถุประสงค์ที่กำหนดไว้ว่าจะนำข้อมูลในระดับรายการ (Item) หรือนำข้อมูลที่สรุปแล้วมาใช้

2) ลักษณะการจัดเก็บของข้อมูล

จากข้อมูลที่ดึงจากหลายแหล่งนั้นอาจใช้ภาษาคอมพิวเตอร์และระบบปฏิบัติการแตกต่างกัน ทำให้มีผลกระทบในการนำข้อมูลมาวิเคราะห์ เช่นข้อมูลอาจถูกเก็บในลักษณะ ASCII Code หรือ EBCDIC Code แต่ระบบดาต้าไมนิ่งที่จะนำมาใช้รองรับข้อมูลที่เก็บในลักษณะ Floating Point เท่านั้น

3) ความแตกต่างกันของข้อมูลแต่ละแหล่ง

การที่นำข้อมูลจากหลายแหล่งนั้น แต่ละแหล่งอาจมีรูปแบบ ความหมายและลักษณะการจัดเก็บข้อมูลที่แตกต่างกัน

4) ข้อมูลที่เป็นข้อความ

ข้อมูลที่เก็บในรูปแบบข้อความ (Text) อาจจะถูกจัดเก็บแตกต่างกัน ทั้งที่ข้อมูลนั้นมีความหมายเหมือนกัน เช่นข้อมูลระบบโทรศัพท์ที่ใช้ DTAC กับ Dtac หรือ YES กับ yes เป็นต้น ซึ่งระบบดาต้าไมนิ่งจะมองข้อมูลเหล่านี้ว่ามีความแตกต่างกัน ไม่ใช่ข้อมูลเดียวกัน

2. การเตรียมข้อมูลก่อนการประมวลผล (Data Preprocessing)

ข้อมูลที่จะนำมาทำดาต้าไมนิ่งนั้น จะต้องผ่านกระบวนการเตรียมข้อมูลก่อนเพื่อให้ได้ข้อมูลที่มีคุณภาพ เนื่องจากว่าข้อมูลที่เรามีอยู่นั้นอาจจะเป็นข้อมูลที่ยังไม่สมบูรณ์ ข้อมูลไม่ครบ หรือข้อมูลกระจัดกระจาย เป็นต้น เพราะข้อมูลที่มีคุณภาพนั้น จะทำให้ผลลัพธ์ที่ได้จากการทำดาต้าไมนิ่งนั้นถูกต้อง แม่นยำ และเชื่อถือได้ ซึ่งกระบวนการเตรียมข้อมูลมีขั้นตอนดังนี้

2.1 การทำข้อมูลให้สะอาด (Data cleaning)

เนื่องจากข้อมูลที่จะนำมาวิเคราะห์นั้นอาจยังไม่มีความสมบูรณ์ ไม่ครบถ้วน ดังนั้นการทำ Data cleaning จะเป็นการระบุลักษณะและเลือกข้อมูลที่ต้องการและนำข้อมูลที่ไม่ต้องการออกไป อาจจะเป็นข้อมูลที่ไม่สามารถแยกความแตกต่างได้ หรือข้อมูลที่ไม่มีความสัมพันธ์กันเลย เช่น ชื่อ ที่อยู่ หรือข้อมูลบางส่วนอาจไม่มีความสำคัญ การทำข้อมูลให้สะอาดนั้นจะแก้ไขข้อมูลที่ไม่สมบูรณ์ ซึ่งอาจส่งผลกระทบต่อการวิเคราะห์ในขั้นตอนถัดไป เราพิจารณาได้ 2 ประเด็นคือ

- Missing Value คือ ค่าของข้อมูลบางค่าหายไปซึ่งอาจเกิดจากความผิดพลาดของอุปกรณ์ในการบันทึกข้อมูล หรือความผิดพลาดที่เกิดจากมนุษย์ ถ้าข้อมูลที่ขาดหายไปมีจำนวนมากและเป็นข้อมูลที่ไม่สำคัญมากนัก อาจแก้ไขโดยไม่นำข้อมูลชุดนั้นมาพิจารณา แต่ถ้าข้อมูลที่ขาดหายไปมีจำนวนน้อยและมีความสำคัญก็อาจจะแก้ไขโดยใช้วิธีการต่างๆ เช่น การแทนที่ค่าว่างด้วยค่าเฉลี่ย หรือใช้วิธีการทำนายค่า เป็นต้น

- Noisy Data คือ ค่าของข้อมูลที่มีลักษณะแตกต่างจากค่าของข้อมูลที่คาดการณ์ไว้ ซึ่งอาจเกิดขึ้นจากความผิดพลาดในการบันทึกข้อมูล ข้อมูลที่มีความคลาดเคลื่อนไปนี้ หากต้องการนำไปใช้จะต้องมีการแก้ไขก่อน หรือตัดค่านั้นทิ้ง โดยใช้เทคนิคต่างๆเช่น การแบ่งข้อมูลออกเป็นกลุ่มๆ (Cluster Analysis) แล้วพิจารณาว่าข้อมูลตัวไหนอยู่นอกกลุ่มจะเป็น Noisy หรือเทคนิค Binning Method ที่ทำการเรียงลำดับข้อมูล แล้วแบ่งข้อมูลออกเป็นกลุ่มๆเท่ากัน แล้วหาค่าเฉลี่ยของข้อมูลในแต่ละกลุ่ม สุดท้ายจะนำค่าเฉลี่ยนั้นมาแทนที่ข้อมูลทุกตัวในกลุ่ม เป็นต้น วิธีการดังกล่าวเพื่อให้ได้ข้อมูลที่มีความถูกต้องมากขึ้น

2.2 การรวมข้อมูลจากแหล่งต่างๆ (Data integration)

การรวมข้อมูลจากหลายแหล่งมาไว้ด้วยกันนั้น อาจทำให้เกิดปัญหาขึ้นเช่น ข้อมูลข้อมูลเดียวกันแต่อยู่คนละฐานข้อมูลทำให้การตั้งชื่อ attribute แตกต่างกันด้วยเช่น CustomerID กับ Customer-ID หรือหน่วยวัดต่างๆที่เก็บค่าๆเดียวกัน แต่เก็บกันคนละหน่วย เช่น เซนติเมตร กับ นิ้ว เป็นต้น ทำให้ข้อมูลมีมากเกินจำเป็น และเมื่อเรานำข้อมูลจำนวนมากมาวิเคราะห์ จึงยากที่จะทราบว่าข้อมูลชุดใดมีความสัมพันธ์กันบ้าง ดังนั้นอาจใช้เทคนิค Correlational analysis มาช่วยตรวจสอบถึงความสัมพันธ์ของข้อมูลต่างๆ เพื่อลดความซ้ำซ้อน

2.3 การลดขนาดของข้อมูล (Data reduction)

ในคลังข้อมูลมีการเก็บข้อมูลจำนวนมาก การวิเคราะห์ข้อมูลนั้นจะต้องใช้เวลามาก ดังนั้นจึงต้องมีการลดขนาดของข้อมูลที่มีปริมาณมากให้เล็กลง แต่ยังคงเป็นข้อมูลที่เป็นตัวแทนของข้อมูลทั้งหมด คุณสมบัติของข้อมูลไม่เปลี่ยนแปลงที่จะทำให้ได้ผลลัพธ์เดิมเหมือนกับการวิเคราะห์จากข้อมูลทั้งหมด เพื่อช่วยประหยัดเวลาในการวิเคราะห์ข้อมูล

3. การแปลงข้อมูล (Data Transformation)

การแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมนั้นจะทำให้ได้ข้อมูลที่สอดคล้องกับอัลกอริทึมหรือโมเดลที่เราจะใช้ในขั้นตอน Data Mining Operation เช่นการแปลงข้อมูลให้อยู่ในช่วงที่เรากำหนด เช่นแปลงให้ข้อมูลอยู่ในช่วง 0 ถึง 1 หรือการแปลงข้อมูลที่ต่อเนื่องให้เป็นข้อมูลที่ไม่ต่อเนื่องเช่น ข้อมูลรายได้ 8,000 12,000 40,000 100,000 บาท แปลงเป็นต่ำกว่า 10,000 บาทคือรายได้ต่ำ ระหว่าง 10,001 ถึง 50,000 บาท คือรายได้ปานกลาง และตั้งแต่ 100,000 บาทขึ้น

ไปคือรายได้สูง เป็นต้น หรือการแปลงตัวแปรแบบ Categorical ให้เป็น Numeric เช่น AIS = 001, DTAC = 010, ORANGE = 011 เป็นต้น

ขั้นตอนที่ 3 การทำดาต้าไมนิ่ง (Data Mining)

ขั้นตอนนี้เป็นขั้นตอนการทำไมนิ่ง โดยเลือกใช้โอเปอร์เรชั่นให้ตรงกับวัตถุประสงค์ที่กำหนดไว้ในขั้นตอนแรก และนำข้อมูลที่เตรียมไว้มาใช้กับอัลกอริทึมที่เหมาะสม การพิจารณาว่าเหมาะสมหรือไม่นั้น ควรพิจารณาว่าอัลกอริทึมที่จะเลือกใช้นั้นสามารถให้ผลลัพธ์ตรงตามจุดประสงค์หรือไม่ มีความสามารถรองรับชนิดของข้อมูล มีความยืดหยุ่นในการรองรับขนาดของข้อมูล และมีประสิทธิภาพเมื่อนำมาประยุกต์ใช้มากน้อยเพียงใด หลังจากเลือกอัลกอริทึมที่เหมาะสมเรียบร้อยแล้ว ก็ทำการวิเคราะห์โดยประมวลข้อมูลกับอัลกอริทึม สุดท้ายก็จะได้ผลการวิเคราะห์ แต่ในทางปฏิบัติแล้วการวิเคราะห์เพียงครั้งเดียวอาจยังไม่ให้ผลลัพธ์ที่สามารถนำไปใช้ได้ ดังนั้นอาจต้องย้อนกลับไปทำขั้นตอนก่อนหน้าใหม่ หรือนำผลการวิเคราะห์ที่ได้กลับไปผ่านอัลกอริทึมอีก เพื่อให้ได้ผลลัพธ์ที่ถูกต้องและสามารถนำไปใช้ได้จริง ซึ่งขึ้นกับโอเปอร์เรชั่นที่เลือกใช้

โอเปอร์เรชั่นของดาต้าไมนิ่ง (Data Mining Operations) แบ่งเป็น 4 เทคนิค ดังนี้

1. Predictive Modeling

Predictive Model เป็นเทคนิคที่ใช้ในการสร้างแบบจำลองพยากรณ์ เพื่อวิเคราะห์ข้อมูลที่มีอยู่เพื่อทำนายแนวโน้มของข้อมูลที่คาดว่าจะเกิดขึ้นในอนาคต ซึ่งการสร้างโมเดลนั้นประกอบด้วย 2 ขั้นตอนคือ ขั้นตอนแรกคือการ Training เป็นขั้นตอนในการสร้างโมเดลโดยการเรียนรู้จากข้อมูลเดิมที่มีอยู่ จากนั้นขั้นตอนที่ 2 คือ Testing เป็นขั้นตอนการประมาณความถูกต้องโดยอาศัยข้อมูลที่ไม่เคยใช้ในการสร้างโมเดลมาทดสอบ ซึ่งผลลัพธ์ที่ได้นั้นจะถูกนำมาเปรียบเทียบกับผลลัพธ์ที่หามาได้จากโมเดลเพื่อทดสอบความถูกต้อง และ Predictive Modeling แบ่งออกเป็น 2 แบบ คือ

- Classification เป็นกระบวนการสร้างโมเดลจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดตามลักษณะที่กำหนดไว้ล่วงหน้า เช่น จัดกลุ่มนักเรียนว่า ดีมาก ดี ปานกลาง ไม่ดี โดยพิจารณาจากประวัติและผลการเรียน หรือแบ่งประเภทของลูกค้าว่าเชื่อถือได้หรือไม่ โดยพิจารณาข้อมูลที่มีอยู่
- Value Prediction (Forecasting) เป็นการทำนายค่าที่ต้องการจากข้อมูลที่มีอยู่ เช่น หายอดขายของเดือนถัดไปจากข้อมูลที่มีอยู่ หรือทำนายโรคจากอาการของคนไข้ในอดีต เป็นต้น

2. Database Segmentation (Clustering Analysis)

Database Segmentation เป็นเทคนิคในการแบ่งกลุ่มข้อมูล เพื่อวิเคราะห์หาความเหมือนหรือแตกต่างกันของข้อมูลในแต่ละกลุ่ม โดยไม่สามารถรู้ล่วงหน้าว่าจะแบ่งข้อมูลออกเป็นกี่กลุ่ม และมีลักษณะอย่างไรบ้าง

3. Link Analysis

Link Analysis เป็นเทคนิคที่ค้นหาความสัมพันธ์ (Associations) ระหว่างเรคอร์ดเดียวหรือ กลุ่มของเรคอร์ดในฐานข้อมูลว่าแต่ละรายการมีความสัมพันธ์กันหรือไม่ อย่างไร เช่นการวิเคราะห์เพื่อหาความสัมพันธ์ที่ซ่อนอยู่ระหว่างสินค้า

4. Deviation Detection

Deviation Detection เป็นเทคนิคในการหาค่าที่แตกต่างไปจากมาตรฐาน หรือค่าที่คาดคิดไว้ว่าต่างไปมากน้อยเพียงใด โดยทั่วไปมักใช้วิธีการทางสถิติ หรือการแสดงให้เห็นภาพ (Visualization) ซึ่งเทคนิคนี้ใช้ในการตรวจสอบ ลายเซ็นปลอม หรือบัตรเครดิตปลอม รวมทั้งการตรวจหาจุดบกพร่องของชิ้นงานในโรงงานอุตสาหกรรม

ขั้นตอนที่ 4 การวิเคราะห์ผลลัพธ์ (Analysis of Result)

เป็นขั้นตอนที่จะทำความเข้าใจกับแบบจำลองที่ได้สร้างขึ้น โดยจะทำการแปลความหมายและประเมินผลที่ได้จากกระบวนการทำดาต้าไมนิ่ง ผลลัพธ์ที่ได้นั้นมีการค้นพบรูปแบบ (Patterns) อะไรที่น่าสนใจหรือไม่ ซึ่งรูปแบบที่น่าสนใจนั้นจะต้องมีคุณสมบัติดังนี้คือ เป็นรูปแบบที่คนสามารถเข้าใจได้ง่าย มีความถูกต้อง น่าเชื่อถือ และก่อให้เกิดประโยชน์ ทั้งนี้อาจต้องใช้ความรู้ทางด้านการวิเคราะห์ข้อมูลและการวิเคราะห์ทางธุรกิจร่วมด้วย แต่หากผลลัพธ์ที่ได้ไม่บรรลุตามวัตถุประสงค์ที่ตั้งไว้ ก็ต้องพิจารณาว่าผิดพลาดที่ขั้นตอนใด อาจต้องย้อนกลับไปแก้ไขขั้นตอนนั้นๆ เพื่อให้ได้ผลลัพธ์ที่สามารถนำไปใช้ได้

ขั้นตอนที่ 5 การนำความรู้ที่ได้ไปใช้ (Assimilation of Knowledge)

ขั้นตอนสุดท้ายคือการนำความรู้หรือสารสนเทศที่ได้ไปใช้ตามจุดประสงค์ที่ได้กำหนดไว้ในขั้นตอนแรก เมื่อนำสารสนเทศใหม่ที่เกิดขึ้นมาใช้กับการดำเนินการหรือการตัดสินใจทางธุรกิจจะก่อให้เกิดแนวทางใหม่ๆในการดำเนินธุรกิจ และแนวทางที่จะนำสารสนเทศใหม่ไปใช้ให้เกิดประโยชน์

Market Management	Risk Management	Fraud Management
■ *Target Marketing* ■ *Customer Relationship Management* ■ *Market Basket Analysis* ■ *Cross Selling* ■ *Marketing Segmentation*	■ *Forecasting* ■ *Customer Relation* ■ *Improved Underwriting* ■ *Quality Control* ■ *Competitive Analysis*	■ *Fraud Detection*

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

รูปที่ 2.2 แสดงการนำดาต้าไมนิ่งไปประยุกต์ใช้ในธุรกิจต่างๆ

จากที่กล่าวมาข้างต้น จะเห็นว่ากระบวนการทำงานของดาต้าไมนิ่งนั้นมีหลายขั้นตอน และมีเทคนิคมากมาย สำหรับโครงการพัฒนาระบบงานที่ได้ศึกษานี้ จะนำเสนอเทคนิคของ Database Segmentation (Clustering Analysis) ด้วยอัลกอริทึม k-prototypes เพื่อแบ่งกลุ่มลูกค้า โดยรายละเอียดจะกล่าวถึงในบทถัดไป

บทที่ 3
เนื้อหาและหลักการของอัลกอริทึม k-prototypes

เทคนิคหนึ่งในการวิเคราะห์ข้อมูลของดาต้าไมนิ่งนั้น คือ เทคนิคการทำ Database Segmentation โดยมีจุดประสงค์เพื่อแบ่งกลุ่มข้อมูลให้กลายเป็นกลุ่มย่อยๆ โดยที่แต่ละกลุ่มจะมีรูปแบบ (Pattern) ที่มีลักษณะคล้ายคลึงกันหรือเหมือนกันภายใน Database Segmentation Methods นั้นก็ยังสามารถแบ่งได้อีกหลายประเภท สำหรับอัลกอริทึมที่จะศึกษานั้นจัดอยู่ใน Partitioning methods การทำงานนั้นจะแบ่งจำนวนข้อมูล n ออกเป็น Cluster ต่างๆ จำนวน k Cluster แล้วกำหนดค่า k หรือจำนวน Cluster ว่าต้องการจัดกลุ่มเป็นกี่กลุ่ม แล้วนำเอาข้อมูลแต่ละตัวไปเปรียบเทียบกับจุดศูนย์กลางของกลุ่มต่างๆที่สร้างขึ้นทุกกลุ่ม จนกระทั่งได้ค่าที่เหมาะสมที่มีค่าเป็นศูนย์หรือเบนเข้าหาค่าที่ตั้งไว้

3.1 อัลกอริทึม *K-Means*

อัลกอริทึมที่ได้รับความนิยมสำหรับเทคนิค Partitioning methods นี้คือ *k-Means* Algorithm โดยจะค่าหาความเหมือนหรือต่างของข้อมูลแต่ละตัว จากการวัดระยะห่าง (Square-Error) ซึ่งเป็นการคำนวณแบบ Euclidean distance method กับจุดศูนย์กลาง (Centriod) ของ Cluster แล้วทำแบบวนรอบไปเรื่อยๆ (Iterative) จนกระทั่งค่า Square-Error ไม่เปลี่ยนแปลงหรือเบนเข้าหาค่าค่าหนึ่งที่กำหนดไว้

ประสิทธิภาพของอัลกอริทึมนี้คือ $O(tkn)$

โดย n คือ จำนวนของ object

k คือ จำนวนของ Cluster

t คือ จำนวนรอบที่วนของขั้นตอน reallocate

ข้อดีของอัลกอริทึม *k-Means* นี้คือ ง่าย ทำงานได้เร็วแม้จะมีข้อมูลขนาดใหญ่ แต่ก็มีข้อจำกัดคือ ทำงานได้กับข้อมูลประเภท Numeric เท่านั้น หากไม่ใช่ข้อมูลประเภท Numeric ก็จะต้องแปลง (map) ให้เป็นตัวเลขก่อน เช่น เพศชายแปลงเป็น 0 เพศหญิงแปลงเป็น 1 เป็นต้น แต่ลักษณะทั่วไปของข้อมูลที่จะทำดาต้าไมนิ่งนั้นมักมีทั้งข้อมูลที่เป็น Numeric และ Categorical ซึ่งการแปลงข้อมูล Categorical ให้เป็นข้อมูล Numeric นั้นบางครั้งก็อาจทำให้ความหมายคลาดเคลื่อนไปได้

จึงมีการพัฒนา Database Segmentation Algorithm เพื่อแก้ปัญหานี้ขึ้น โดยอัลกอริทึมนี้อยู่บนพื้นฐานของ k-means แต่ว่าได้ลดข้อจำกัดในเรื่องข้อมูล Numeric ลง โดยที่ยังคงรักษาประสิทธิภาพไว้เช่นเดิม อัลกอริทึมดังกล่าวคือ *k-prototypes* Algorithm ซึ่งจะสามารถทำงานได้ทั้งข้อมูลที่เป็น Numeric และ Categorical

3.2 หลักการทางคณิตศาสตร์เบื้องต้น

กำหนดให้เซตของ object n คือ $X = \{X_1, X_2, \ldots, X_n\}$ และให้ $X_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\}$ แทน attribute ของ object โดยที่มีจำนวน m attributes การแบ่งกลุ่ม X นั้นจะทำการแบ่งข้อมูลแต่ละตัวใน X ให้เป็นกลุ่มย่อยๆ k กลุ่ม ดังนั้น k จะเป็นได้เฉพาะจำนวนเต็มบวกเท่านั้น คือ $k = \{1,2,\ldots,n\}$ โดยที่ n คือจำนวนกลุ่มที่เป็นไปได้ในการแบ่งข้อมูล แต่ n มีจำนวนมากเกินไปจนยากที่จะแบ่งข้อมูลเป็นกลุ่มให้เหมาะสม ดังนั้นจึงมีแนวทางในการค้นหากลุ่มที่เหมาะสมกับข้อมูล นั่นคือการใช้หลักการ Cost function

- **การหาค่า Cost Function**

เราจะใช้ Cost function ในการแบ่งข้อมูลเป็นกลุ่มๆ ซึ่งค่า Cost function คือ

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

โดยที่

$Q_l = [q_{l1}, q_{l2}, \ldots, q_{lm}]$ แทนจุดศูนย์กลางของกลุ่มข้อมูล l

y_{il} คือ ค่าของ y ที่เมทริกซ์ nxl

d คือ การวัดค่าความเหมือนกันของข้อมูลตามการคำนวณแบบ square Euclidean distance

คุณสมบัติของ Y 2 ประการ คือ

1. $0 \le y_{il} \le 1$
2. $\sum_{l=1}^{k} y_{il} = 1$

ถ้า $y_{il} \in \{0,1\}$ จะเรียก Y ว่าเป็น hard partition ถ้า $y_{il} = 1$ แสดงว่าobject X_i ถูกกำหนดให้อยู่ในกลุ่มข้อมูล l

จากสมการที่ (1) นั้นจะได้ว่า $E_l = \sum_{i=1}^{n} y_{il} d(X_i, Q_l)$ ซึ่งก็การคำนวณคือค่า total cost ของข้อมูลเพื่อกำหนดให้ข้อมูล X ไปอยู่ในกลุ่มข้อมูล l เช่น การกระจาย object ในกลุ่มข้อมูล l จากจุดศูนย์กลางของกลุ่มข้อมูล Q_l

ค่า E_l จะมีค่าต่ำสุดเมื่อ

$$q_{lj} = \frac{1}{n_l}\sum_{i=1}^{n} y_{il} x_{ij} \tag{2}$$

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

เมื่อ X เป็นข้อมูลประเภท categorical แล้วจะใช้วิธีวัดความเหมือนของข้อมูลได้โดย

$$d(X_i, Q_l) = \sum_{j=1}^{m_r}\left(x_{ij}^r - q_{lj}^r\right)^2 + \gamma_l \sum_{j=1}^{m_c}\delta\left(x_{ij}^c, q_{lj}^c\right) \tag{3}$$

โดยที่ $\delta(p,q) = 0$ เมื่อ $p = q$

$\delta(p,q) = 1$ เมื่อ $p \neq q$

x_{ij}^r คือ ค่าของ attributed ของ object i ที่เป็นข้อมูลประเภท numeric

q_{lj}^r คือ จุดศูนย์กลางของกลุ่มข้อมูล l ที่เป็นข้อมูลประเภท numeric

x_{ij}^c คือ ค่าของ attributed ของ object i ที่เป็นข้อมูลประเภท categorical

q_{lj}^c คือ จุดศูนย์กลางของกลุ่มข้อมูล l ที่เป็นข้อมูลประเภท categorical

m_r คือ จำนวน attribute ของข้อมูลประเภท numeric

m_c คือ จำนวน attribute ของข้อมูลประเภท categorical

γ_l คือ weight ของ categorical attribute สำหรับกลุ่มข้อมูล l

จากสมการที่ (3)สามารถเขียน E_l ใหม่ได้เป็น

$$\begin{aligned} E_l &= \sum_{i=1}^{n} y_{il}\sum_{j=1}^{m_r}\left(x_{ij}^r - q_{lj}^r\right)^2 + \gamma_l \sum_{i=1}^{n} y_{il}\sum_{j=1}^{m_c}\delta\left(x_{ij}^c, q_{lj}^c\right) \\ &= E_l^r + E_l^c \end{aligned} \tag{4}$$

โดยที่ E_l^r คือค่า total cost ของ numeric attribute ของ object ทั้งหมดในกลุ่มข้อมูล l และ E_l^r จะมีค่าต่ำสุด เมื่อคำนวณ q_{ij}^r จากสมการที่ (2)

กำหนดให้ C_j เป็นเซตของค่าที่มีลักษณะเฉพาะเพียงอย่างเดียว (unique value) ทั้งหมดใน categorical attribute j และ $p\left(c_j \in C_j \middle| l\right)$ คือความน่าจะเป็นของค่า c_j ที่จะเกิดขึ้นในกลุ่มข้อมูล l

ดังนั้นจะเขียน E_l^c ได้เป็น

$$E_l^c = \gamma_l \sum_{j=1}^{m_c} n_l\left(1 - p\left(q_{lj}^c \in C_j \middle| l\right)\right) \tag{5}$$

โดยที่ n_l เป็นจำนวนของ object ในกลุ่มข้อมูล l ซึ่งหาค่าต่ำสุดของ E_l^c หาได้จากบทแทรกที่ 1

บทแทรกที่ 1 : สำหรับกลุ่มข้อมูล l แล้ว E_l^c จะมีค่าต่ำสุดเมื่อ $p\left(q_{ij}^c \in C_j | l\right) \geq p\left(c_j \in C_j | l\right)$ โดยที่ $q_{ij}^c \neq c_j$ สำหรับทุก attributes ของข้อมูลประเภท categorical

ท้ายที่สุดจะสามารถเขียน E ได้เป็น

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

$$= E^r + E^c \qquad (6)$$

สมการที่ (6) เป็นการคำนวณ cost function เพื่อแบ่งกลุ่มข้อมูลที่เป็น numeric และ categorical โดยที่ทั้งค่า E^r และ E^c จะต้องไม่เป็นค่าลบ ค่าต่ำสุดของ E หาได้จากค่าต่ำสุดของ E^r (total cost ของ numeric attribute ของข้อมูลทั้งหมด) และ E^c (total cost ของ categorical attribute ของข้อมูลทั้งหมด) ซึ่ง

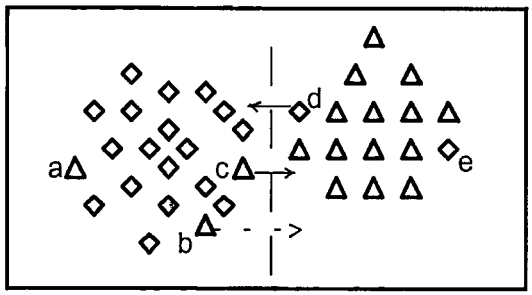
- E^r หาค่าต่ำสุดของข้อมูลประเภท numeric ได้โดยการคำนวณค่าของจุดศูนย์กลาง k กลุ่ม จากสมการที่ (2)
- E^c หาค่าต่ำสุดของข้อมูลประเภท categorical ได้โดยการคำนวณค่าของจุดศูนย์กลาง k กลุ่ม จากบทแทรกที่ 1

ดังนั้นสมการที่ (2) และบทแทรกที่ 1 จะเป็นค่าที่ใช้เลือกจุดศูนย์กลางของกลุ่มข้อมูล โดยดูจากค่า cost function ที่มีค่าน้อยที่สุดในสมการที่ (6)

หลักการทางคณิตศาสตร์ดังกล่าวจะเป็นพื้นฐานของ *k-prototypes* Algorithm

3.3 การวัดค่าความเหมือนของข้อมูล (Similarity Measure)

Similarity Measure ของข้อมูลที่เป็น Numeric คือค่า square Euclidean distance ส่วน Similarity Measure ของข้อมูลที่เป็น Categorical คือจำนวนครั้งที่ไม่ตรงกันระหว่าง object กับค่า cluster prototypes สำหรับค่า weight γ_l นั้น อธิบายได้ด้วยรูปที่ 1 โดยกำหนดให้เครื่องหมายสามเหลี่ยมแทนข้อมูลประเภท Categorical และเครื่องหมายข้าวหลามตัดแทนข้อมูลประเภท Numeric ซึ่งข้อมูลจะถูกแบ่งออกเป็น 2 clusters



รูปที่ 3.1 แสดงอิทธิพลของ weight γ_l ที่มีผลต่อการ clustering

ถ้า weight $\gamma_l = 0$ การ clustering จะขึ้นกับข้อมูลประเภท Numeric เท่านั้น นั่นคือจะแบ่งกลุ่มของ object ออกเป็น 2 clusters โดยมีเส้นขีดแบ่งระหว่าง cluster

ถ้า weight $\gamma_l > 0$ แล้ว object c อาจจะเปลี่ยนไปยังด้านขวาของ cluster เนื่องจากใกล้กับ cluster ที่เป็น categorical (เป็นสามเหลี่ยมเหมือนกัน) ในทางเดียวกัน object d ก็อาจจะเปลี่ยนไปยังด้านซ้ายของ cluster เนื่องจากใกล้กับ cluster ที่เป็น Numeric (เป็นข้าวหลามตัดเหมือนกัน) แต่สำหรับ object a จะยังคงอยู่ใน cluster ด้านซ้ายเพราะ object a อยู่ไกลจาก cluster ด้านขวา ถึงแม้ว่ามันจะเป็นข้อมูล categorical ก็ตามซึ่งเหมือนกับ object e จะยังคงอยู่ใน cluster ด้านขวาต่อไป สำหรับ object b นั้นจะไม่แน่นอนขึ้นอยู่กับ weight γ_l ว่าเอนเอียงไปทางข้อมูลประเภท Numeric หรือ Categorical ถ้า weight เอียงไปทางข้อมูลประเภท Categorical แล้ว object b ก็อาจจะเปลี่ยนไปยังด้านขวา แต่ถ้าเอียงไปทางข้อมูลประเภท Numeric ก็อาจจะยังคงอยู่ทางด้านซ้าย

3.4 การทำงานของอัลกอริทึม *k-prototypes*

หลักการทำงานของ *k-prototypes* Algorithm มีดังนี้คือ

1. เลือก k ที่จะเป็นจุดศูนย์กลางจากกลุ่มข้อมูล X ในแต่ละกลุ่ม
2. ให้ object แต่ละตัวในกลุ่มข้อมูล X ไปเปรียบเทียบกับจุดศูนย์กลางของแต่ละกลุ่ม เพื่อหาว่าใกล้กับกลุ่มใดมากที่สุด โดยเป็นไปตามสมการที่ (3) เมื่อกำหนดให้ object อยู่ในกลุ่มใดกลุ่มหนึ่งแล้ว ก็ให้ update จุดศูนย์กลางของแต่ละกลุ่ม
3. หลังจากที่ object ทั้งหมดถูกกำหนดให้อยู่ในกลุ่มต่างๆแล้ว ให้ทำการทดสอบความเหมือนกันของข้อมูลอีกครั้งในแต่ละกลุ่ม ถ้าพบว่ามี object ใดอยู่ใกล้กับกลุ่มอื่นมากกว่าก็ให้ย้าย object นั้นไป และ update จุดศูนย์กลางของกลุ่มเดิมและกลุ่มใหม่
4. ทำขั้นตอนที่ 3 ซ้ำจนกระทั่งไม่มี object ใดต้องเปลี่ยนกลุ่มอีก จากนั้นก็ให้ทดสอบข้อมูลทั้งหมดอีกครั้ง

จากอัลกอริทึมนี้จะประกอบไปด้วย 3 กระบวนการ คือ

1. *initial prototypes selection* จะสุ่มเลือก *k* เพื่อเป็นจุดศูนย์กลาง
2. *initial allocation* เริ่มต้นจากกำหนดจุดศูนย์กลางของกลุ่ม จากนั้นจะกำหนด object แต่ละตัวให้อยู่ในกลุ่มต่างๆซึ่งขั้นตอนนี้แสดงในรูปที่ 2
3. *re-allocation* หลังจากที่มี object ถูกกำหนดเข้าไปในกลุ่มแล้ว จะ update จุดศูนย์กลางของกลุ่มที่ object นี้ถูกกำหนดให้อยู่ก่อนการย้ายกลุ่ม และ update จุดศูนย์กลางของกลุ่มใหม่หลังจาก object นี้ย้ายเข้าไปในกลุ่มแล้ว ขั้นตอนนี้แสดงในรูปที่ 3

ในรูปที่ 2 และ 3 แสดง process ที่ 2 และ 3 ตามลำดับ ซึ่งอัลกอริทึมนี้ประกอบด้วยตัวแปรและฟังก์ชันต่างๆดังนี้

X[i] คือ object i หรือข้อมูล i

X[i,j] คือ ค่าของ attribute j ของ object i

O_prototypes[ ] คือ เก็บ numeric attribute ของจุดศูนย์กลางของกลุ่มข้อมูล

C_prototypes[ ] คือ เก็บ categorical attribute ของจุดศูนย์กลางของกลุ่มข้อมูล

O_prototypes[i,j] คือ ข้อมูลแบบ numeric ที่เป็นสมาชิกของกลุ่มข้อมูล i

C_prototypes[i,j] คือ ข้อมูลแบบ categorical ที่เป็นสมาชิกของกลุ่มข้อมูล i

Distance( ) คือ square Euclidean distance function

Sigma( ) คือ หาค่าได้จาก fn $\delta(\ )$ จากสมการที่ (3)

Clustership[ ] คือ ตัวแปรที่บอกว่า object นี้อยู่ cluster ที่เท่าไร

ClusterCount[ ] คือ จำนวนของ object ใน cluster

SumInCluster[ ] คือ สรุปค่า Numeric ของ object ใน cluster และเคยถูก update numeric attribute ของจุดศูนย์กลางของกลุ่มข้อมูล

FrequencyInCluter[ ] คือ ความถี่ของความแตกต่างของ categorical attribute ใน cluster

HighestFreq( ) คือ หาค่าได้จากบทแทรกที่ 1 เพื่อ update categorical attributed ของจุดศูนย์กลางของกลุ่มข้อมูล

```
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 NumberOf CategoricAttributes
      FrequencyInCluster[cluster,j,X[i,j]] + 1
      C_prototypes[cluster,j] = HighestFreq
      (FrequencyInCluster,cluster,j)
   ENDFOR
ENDFOR
```

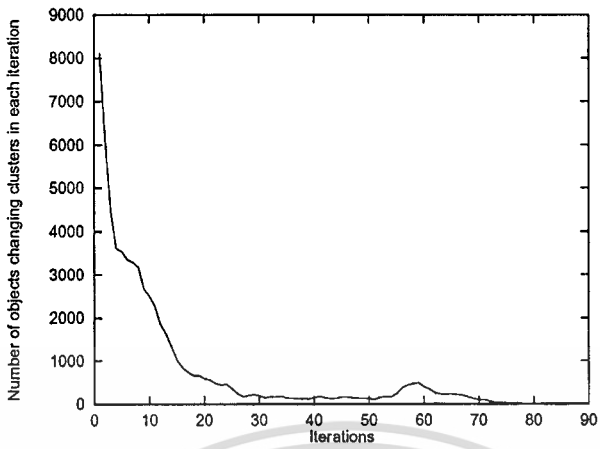
รูปที่ 3.2 แสดงอัลกอริทึมของกระบวนการ *initial allocate*

```
Moves = 0
FOR i = 1 TO NumberOfObjects
   ...
   (To find the cluster whose prototype is the nearest to
   object i. Same as Figure 2)
   ...
   IF (Clustership[i] <> cluster)
      moves + 1
      oldcluster = Clustership[i]
      Clustership[i] = cluster
      ClusterCount[cluster] + 1
      ClusterCount[oldcluster] -1
      FOR j = 1 TO NumberOfNumericAttributes
         SumInCluster[cluster,j] + X[i,j]
         SunInCluster[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 NumberOfCategoricAttributes
         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 3.3 แสดงอัลกอริทึมของกระบวนการ *reallocate*

3.5 ประสิทธิภาพของอัลกอริทึม *k-prototypes*

รูปที่ 4 จะแสดงกราฟที่ได้จากการคำนวณโดยอัลกอริทึม ซึ่งมีข้อมูลจำนวน 75,808 เรคอร์ด และมี 20 attributes ซึ่ง Clusters ได้ 64 กลุ่ม



รูปที่ 3.4 แสดงกราฟของ *k-prototypes* Algorithm

จากกราฟจะเห็นได้ว่าจำนวนของ object ที่เปลี่ยน clusters จะเปลี่ยนแปลงอย่างรวดเร็วในช่วงแรก หลังจากนั้นการเปลี่ยนแปลงจะช้าลง จนเกือบจะไม่เปลี่ยนแปลง ซึ่งแสดงเห็นว่าการทำแบบวนรอบไปเรื่อยๆของขั้นตอน reallocate นั้นสามารถหยุดได้ที่จุดใดจุดหนึ่ง เมื่อการกราฟนั้นเปลี่ยนแปลงเพียงเล็กน้อย ทำให้ช่วยลดเวลาการทำงาน (running time) เมื่อข้อมูลมีจำนวนมาก

ประสิทธิภาพของอัลกอริทึมนี้คือ $O((t+1)kn)$

โดย n คือ จำนวนของ object

k คือ จำนวนของ Cluster

t คือ จำนวนรอบที่วนของขั้นตอน reallocate

โดยทั่วไปแล้ว $k << n$ และ t มักจะไม่เกิน 100 รอบจากการทดลองกับข้อมูลขนาดใหญ่ ดังนั้นอัลกอริทึมนี้จึงมีประสิทธิภาพในการแบ่งกลุ่มข้อมูลที่มีขนาดใหญ่

3.6 ตัวอย่างการทำงานของอัลกอริทึม *k-prototypes* กับข้อมูล

ตัวอย่างการทำงานของอัลกอริทึม *k-protoyypes* กับข้อมูลนี้จะแสดงตัวอย่างขั้นตอนการทำงานและการคำนวณของอัลกอริทึม โดยสมมุติให้ข้อมูลที่จะทำไมนิ่งนั้นได้ผ่านขั้นตอนการเตรียมข้อมูลมาแล้ว ซึ่งข้อมูลที่จะนำมาวิเคราะห์นั้น ประกอบด้วย 3 attributes ได้แก่

- Age
- Gender ประกอบด้วย Male และ Female
- Province ประกอบด้วย Bangkok, Central, North, North East, South

โดยข้อมูลแต่ละ attributes มีชนิดของข้อมูลดังนี้

ตารางที่ 3.1 แสดงชนิดของข้อมูลที่จะนำมาทำไมนิ่ง

Attributes	Type
Age	Numeric
Gender	Categorical
Province	Categorical

ตารางที่ 3.2 แสดงข้อมูลที่จะนำมาทำไมนิ่ง

Age	Gender	Province
16	Female	Central
28	Female	Bangkok
22	Male	North
13	Male	Central
32	Female	Bangkok
21	Female	South
24	Male	North
40	Male	South
18	Female	Central
30	Female	Bangkok

จากอัลกอริทึม จะสามารถแสดงขั้นตอนการทำงานของ *k-prototypes* Algorithm ได้ดังต่อไปนี้

ขั้นตอนที่ 1 ทำการสุ่มเลือกจุดศูนย์กลางของกลุ่มข้อมูล และสมมติว่าต้องการให้มีกลุ่มข้อมูล 3 กลุ่ม และจัดข้อมูลในแต่ละกลุ่มดังต่อไปนี้

ตารางที่ 3.3 แสดงข้อมูลกลุ่มที่ 1

Age	Gender	Province
16	Female	Central
28	Female	Bangkok
22	Male	North

ตารางที่ 3.4 แสดงข้อมูลกลุ่มที่ 2

Age	Gender	Province
13	Male	Central
32	Female	Bangkok
21	Female	South

ตารางที่ 3.5 แสดงข้อมูลกลุ่มที่ 3

Age	Gender	Province
24	Male	North
40	Male	South
18	Female	Central
30	Female	Bangkok

จากนั้นกำหนดจุดศูนย์กลางของแต่ละกลุ่มดังนี้

ตารางที่ 3.6 แสดงจุดศูนย์กลางของข้อมูล 3 กลุ่ม

จุดศูนย์กลางของกลุ่มที่ 1	Age = 16 / Gender = Female / Province = Central
จุดศูนย์กลางของกลุ่มที่ 2	Age = 32 / Gender = Female / Province = Bangkok
จุดศูนย์กลางของกลุ่มที่ 3	Age = 24 / Gender = Male / Province = North

ขั้นตอนที่ 2 นำ object แต่ละตัวมาหาค่า cost function ในกลุ่มต่างๆ โดยสมมติค่า weight ของข้อมูล Categorical ดังนี้ ค่า weight ของ Gender มีค่าเท่ากับ 0.4 และค่า weight ของ Province มีค่าเท่ากับ 0.2 จากสมการที่ (3) จะต้องคำนวณค่า Cost function ของข้อมูลทุกตัวเทียบกับจุดศูนย์กลาง แต่ในตัวอย่างนี้จะขอยกตัวอย่างการจัดกลุ่มข้อมูลกลุ่มที่ 1 โดยเลือก object Age = 28 / Gender = Female / Province = Bangkok เพียงข้อมูลเดียวมาพิจารณา ดังนี้

- เปรียบเทียบกับกลุ่มที่ 1

ตารางที่ 3.7 แสดงการเปรียบเทียบ object กับจุดศูนย์กลางของกลุ่มที่ 1

Object	จุดศูนย์กลางของกลุ่มที่ 1
28	16
Female	Female
Bangkok	Central

Cost function $= (28-16)^2 + 0.4(0) + 0.2(1)$

$= 144 + 0 + 0.2$

$= 144.2$

- เปรียบเทียบกับกลุ่มที่ 2

ตารางที่ 3.8 แสดงการเปรียบเทียบ object กับจุดศูนย์กลางของกลุ่มที่ 2

Object	จุดศูนย์กลางของกลุ่มที่ 2
28	32
Female	Female
Bangkok	Bangkok

Cost function $= (28-32)^2 + 0.4(0) + 0.2(0)$

$= 16$

- เปรียบเทียบกับกลุ่มที่ 3

ตารางที่ 3.9 แสดงการเปรียบเทียบ object กับจุดศูนย์กลางของกลุ่มที่ 2

Object	จุดศูนย์กลางของกลุ่มที่ 3
28	24
Female	Male
Bangkok	North

Cost function $= (28-24)^2 + 0.4(1) + 0.2(1)$

$= 16 + 0.4 + 0.2$

$= 16.6$

เมื่อคำนวณค่า cost function กับจุดศูนย์กลางของทั้ง 3 กลุ่มพบว่า ค่า cost function ของกลุ่มที่ 2 มีค่าต่ำที่สุด ดังนั้นจะกำหนดให้ object Age = 28 / Gender = Female / Province = Bangkok อยู่ในกลุ่มที่ 2

ตารางที่ 3.10 แสดงข้อมูลกลุ่มที่ 1 หลังจากมี object ย้ายกลุ่ม

Age	Gender	Province
16	Female	Central
22	Male	North

ตารางที่ 3.11 แสดงข้อมูลกลุ่มที่ 2 หลังจากมี object ย้ายกลุ่ม

Age	Gender	Province
13	Male	Central
32	Female	Bangkok
21	Female	South
28	Female	Bangkok

ขั้นตอนที่ 3 คือจะทำการ update จุดศูนย์กลางของกลุ่มที่ 1 และ 2 โดยการพิจารณาข้อมูลที่เป็น numeric นั้นจะคำนวณได้โดยการหาค่ามัธยฐาน ส่วนข้อมูลที่เป็น categorical นั้นได้จากการหาค่าฐานนิยม (ถ้ากรณี ไม่สามารถหาค่าฐานนิยมได้ จะสมมติให้ค่าใดค่าหนึ่งเป็นจุดศูนย์กลางแทน) ดังนี้

- Update กลุ่มที่ 1

Age = (16+22)/2 = 19

Gender = (Female, Male)

= Female

Province = (Central, North)

= Central

- Update กลุ่มที่ 2

Age = (13+32+21+28)/4 = 23.5

Gender = (Male, Female, Female, Female)

= Female

Province = (Central, Bangkok, South, Bangkok)

= Bangkok

ดังนั้น จะได้จุดศูนย์กลางของข้อมูลกลุ่มที่ 1 และ 2 ใหม่ดังนี้

ตารางที่ 3.12 สรุปจุดศูนย์กลางของข้อมูล 3 กลุ่มหลังถูก update

จุดศูนย์กลางของกลุ่มที่ 1	Age = 19 / Gender = Female / Province = Central
จุดศูนย์กลางของกลุ่มที่ 2	Age = 23.5 / Gender = Female / Province = Bangkok
จุดศูนย์กลางของกลุ่มที่ 3	Age = 24 / Gender = Male / Province = North

จะต้องทำขั้นตอนที่ 2 และ 3 นี้กับทุก object ดังนั้นจุดศูนย์กลางของแต่ละกลุ่มจะถูก update ไปเรื่อยๆจนกระทั่ง object ทั้งหมดถูกกำหนดให้อยู่ในกลุ่มต่างๆ หลังจากนั้นให้ทำการทดสอบค่า cost function กับจุดศูนย์กลางของกลุ่มอีกครั้ง ถ้าพบว่ามี object ใดใกล้กับจุดศูนย์กลางของกลุ่มอื่นมากกว่า ก็ให้ย้าย object นั้นไปแล้วทำการ update จุดศูนย์กลางใหม่อีกครั้ง ทำจนกระทั่งไม่มีการ object ใดต้องเปลี่ยนกลุ่มอีก

จากการศึกษาหลักการทำงาน และประสิทธิภาพของ *k-prototypes* Algorithm เพื่อใช้เป็นแนวทางในการพัฒนาโครงการนั้น จะเห็นได้ว่า *k-prototypes* Algorithm สามารถที่จะรองรับข้อมูลทั้งประเภท numeric และ categorical ได้ อีกทั้งยังสามารถทำงานกับข้อมูลที่มีจำนวนมาก ได้อย่างมีประสิทธิภาพอีกด้วย ดังนั้นผลลัพธ์ที่จะได้จากการวิเคราะห์ข้อมูลลูกค้านี้ก็คือ จะช่วยให้เจ้าของเว็บไซต์หรือผู้ดูแลเว็บไซต์ เข้าใจความต้องการของลูกค้ากลุ่มต่างๆมากขึ้น ซึ่งจะช่วยให้สามารถกำหนดเป้าหมายและวางแผนงานทางธุรกิจได้ดียิ่งขึ้น

บทที่ 4
การประยุกต์ใช้ดาต้าไมนิ่งกับการจัดกลุ่มลูกค้า

4.1 กำหนดวัตถุประสงค์

เนื่องจากจำนวนลูกค้าที่เข้ามาเยี่ยมชมเว็บไซต์มีจำนวนมากขึ้น แต่ผู้บริหารเว็บไซต์ยังไม่สามารถมองเห็นกลุ่มเป้าหมายหลักที่สามารถทำกำไรให้กับเว็บไซต์ได้อย่างชัดเจน ดังนั้นรายการส่งเสริมการขายที่นำเสนอให้กับลูกค้า บางครั้งยังไม่ได้รับความสนใจเท่าที่ควร

ดังนั้นจึงมีความคิดที่จะนำดาต้าไมนิ่งมาประยุกต์ใช้ โดยมีวัตถุประสงค์เพื่อจัดกลุ่มลูกค้าที่มีลักษณะใกล้เคียงกัน จะได้นำเสนอรายการส่งเสริมการขายหรือสิทธิพิเศษและบริการต่างๆ ให้ตรงตามกลุ่มเป้าหมาย นอกจากจะทำให้สามารถกำหนดเป้าหมายทางการตลาดได้ชัดเจนมากขึ้น

4.2 การเตรียมข้อมูลที่จะนำมาวิเคราะห์

ข้อมูลที่จะนำมาใช้สำหรับการพัฒนาโครงการนั้น ได้นำข้อมูลส่วนหนึ่งมาจากข้อมูลลูกค้าที่ได้สมัครเป็นสมาชิกกับเว็บไซต์มาเป็นข้อมูลในการวิเคราะห์ ซึ่งลักษณะทั่วไปของเว็บไซต์เป็นดังนี้คือ เป็นเว็บไซต์ที่ให้ลูกค้าดาวน์โหลดสิทธิพิเศษผ่านโทรศัพท์มือถือ เพื่อเลือกรับสิทธิพิเศษและบริการต่างๆ จากยี่ห้อสินค้าที่มีอยู่ในเว็บไซต์

ตัวอย่าง วิธีการดาวน์โหลดสิทธิพิเศษและบริการ

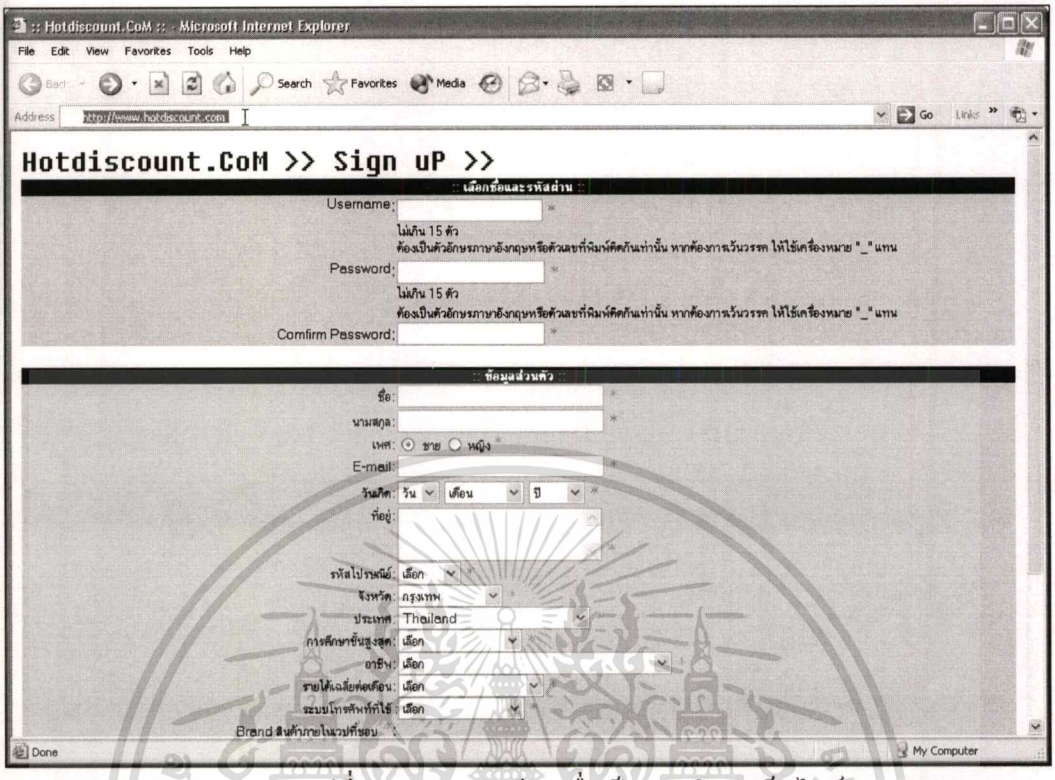
1. เมื่อลูกค้าต้องการใช้สิทธิพิเศษเพื่อซื้อไอศกรีมของ Baskin Robbins 2 ลูกในราคา 55 บาท (จากราคาปกติ 63 บาท)

2. ลูกค้าจะพิมพ์ SMS ในโทรศัพท์มือถือ โดยพิมพ์รหัสพิเศษของ Baskin Robbins คือ BR ส่งไปยังหมายเลขโทรศัพท์ที่กำหนด

3. ระบบจะทำการส่งข้อความตอบกลับมายังโทรศัพท์มือถือ

4. เมื่อได้รับข้อความตอบกลับแล้ว สามารถนำข้อความที่ได้รับไปแสดง ณ ร้านค้า ที่ร่วมรายการ จากตัวอย่างดังกล่าว ลูกค้าสามารถใช้สิทธิพิเศษซื้อไอศกรีม 2 ลูก ในราคาเพียง 55 บาท (จากราคาปกติ 63 บาท) เป็นต้น

สำหรับการสมัครเป็นสมาชิกกับทางเว็บไซต์นั้น ลูกค้าต้องเข้าไปยังเว็บไซต์และลงทะเบียนเพื่อรับสิทธิพิเศษ ซึ่งข้อมูลของลูกค้าที่ได้ลงทะเบียนเพื่อเป็นสมาชิกเหล่านี้สามารถนำมาทำดาต้าไมนิ่งได้



รูปที่ 4.1 การลงทะเบียนเพื่อเป็นสมาชิกกับเว็บไซต์

ข้อมูลทั้งหมดที่ลูกค้าจะต้องกรอก และระบบจะคำนวณและจัดเก็บลงในฐานข้อมูล ประกอบด้วยข้อมูลดังต่อไปนี้

- ชื่อ (Username)
- รหัสผ่าน (Password)
- ยืนยันรหัสผ่าน (Confirm password)
- ชื่อลูกค้า (First name)
- นามสกุล (Last name)
- เพศ (Gender)
- อีเมล์ (Email)
- วันเกิด (Birthday)
- อายุ (Age)
- ที่อยู่ (Address)
- รหัสไปรษณีย์ (Zip code)
- จังหวัด (Province)

- ประเทศ (Country)
- การศึกษาขั้นสูงสุด (Education)
- อาชีพ (Job)
- รายได้เฉลี่ยต่อเดือน (Income)
- ระบบโทรศัพท์ที่ใช้ (Mobile)
- ยี่ห้อสินค้าภายในเว็บที่ชอบ (Brand)

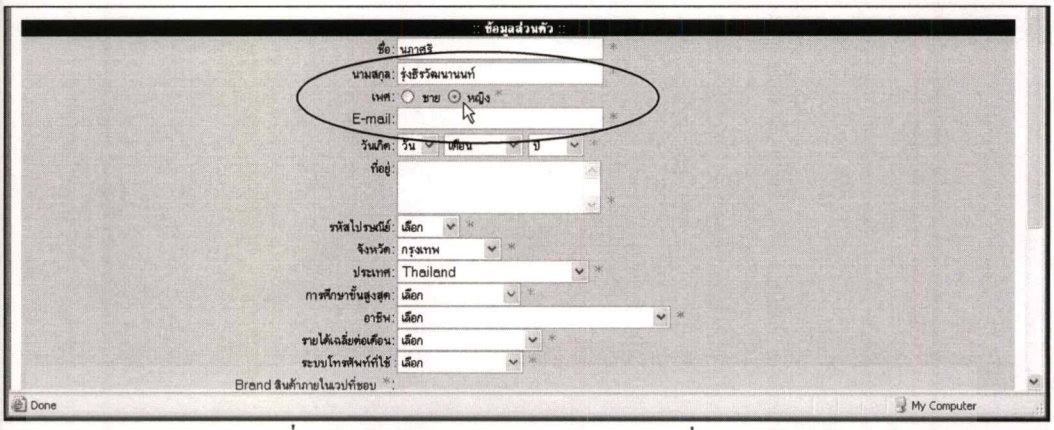
4.2.1 การคัดเลือกข้อมูล

จากวัตถุประสงค์ที่กำหนดไว้คือ เพื่อจัดกลุ่มลูกค้าที่มีลักษณะใกล้เคียงกัน จะได้นำเสนอรายการส่งเสริมการขายหรือสิทธิพิเศษและบริการต่างๆ เมื่อพิจารณาจากวัตถุประสงค์แล้ว การจัดกลุ่มลูกค้าที่มีลักษณะใกล้เคียงกันนั้น สามารถพิจารณาได้จากข้อมูลพื้นฐานของลูกค้า คือ เพศ (Gender), อายุ (Age), รายได้ (Income) และวัตถุประสงค์ได้เน้นถึงการนำเสนอรายการส่งเสริมการขาย ดังนั้นจึงต้องนำข้อมูล ยี่ห้อสินค้า (Brand) มาพิจารณาด้วย ข้อมูลที่ถูกคัดเลือกให้นำมาใช้วิเคราะห์สามารถให้ผลลัพธ์ได้ตรงตามจุดประสงค์ที่กำหนดไว้

4.2.2 การตรวจสอบคุณภาพของข้อมูล

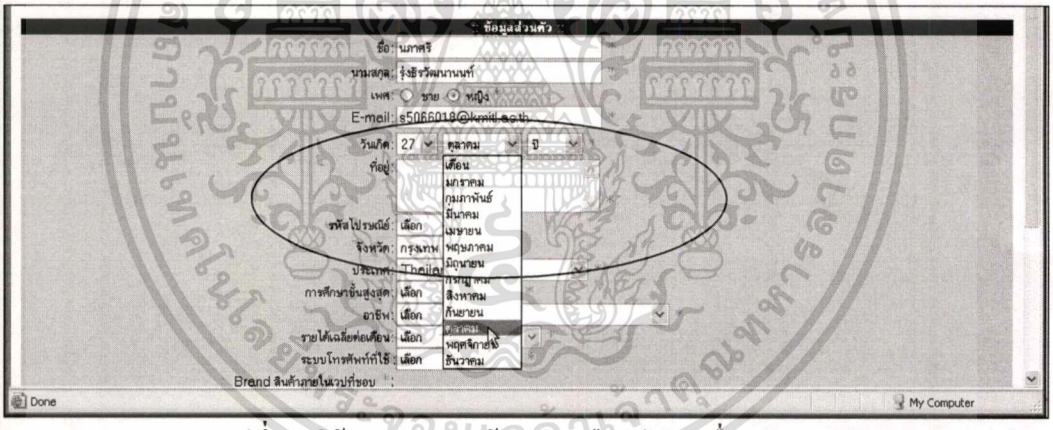
เพื่อให้ได้ผลลัพธ์ที่มีความถูกต้อง แม่นยำ ดังนั้นข้อมูลที่จะนำไปวิเคราะห์จะต้องผ่านกระบวนการเตรียมข้อมูลเพื่อให้เป็นข้อมูลที่มีคุณภาพ จากที่ได้กล่าวข้างต้นแล้วว่า ข้อมูลที่จะนำมาใช้สำหรับการพัฒนาโครงการนั้น ได้นำข้อมูลส่วนหนึ่งมาจากข้อมูลลูกค้าที่ได้สมัครเป็นสมาชิกกับเว็บไซต์มาเป็นข้อมูลในการวิเคราะห์ ซึ่งข้อมูลดังกล่าวเป็นข้อมูลจากฐานข้อมูลของเว็บไซต์เพียงฐานข้อมูลเดียว จึงไม่เกิดปัญหาเรื่องความซ้ำซ้อนของข้อมูล และไม่ต้องทำการลดขนาดของข้อมูล เนื่องจากทางเว็บไซต์เริ่มเก็บข้อมูลของสมาชิกไม่นานนัก ดังนั้นข้อมูลยังมีจำนวนไม่มาก และเพื่อป้องกันการเกิด Missing Value และ Noisy Data ซึ่งอาจส่งผลกระทบต่อการวิเคราะห์ จึงได้ให้ความสำคัญกับการออกแบบหน้าเว็บเพ็จที่ใช้ในการรับข้อมูล โดยจะใช้แบบฟอร์มการรับข้อมูลเช่น Radio Button และ List box เพื่อช่วยลดความผิดพลาดที่เกิดจากการพิมพ์ข้อมูล และยังทำให้สะดวกในการกรอกข้อมูลมากขึ้นด้วย ซึ่งวิธีการเก็บข้อมูลดังกล่าวนี้จะสามารถช่วยลดความผิดพลาดสำหรับการรับข้อมูลที่สำคัญและจำเป็นได้ ตัวอย่างการใช้แบบฟอร์มเพื่อรับข้อมูลต่างๆ เช่น

- ใช้ Radio Button ในการรับข้อมูลเพศ (Gender) ลูกค้าจะต้องเลือกระหว่างเพศชายหรือเพศหญิง เท่านั้น



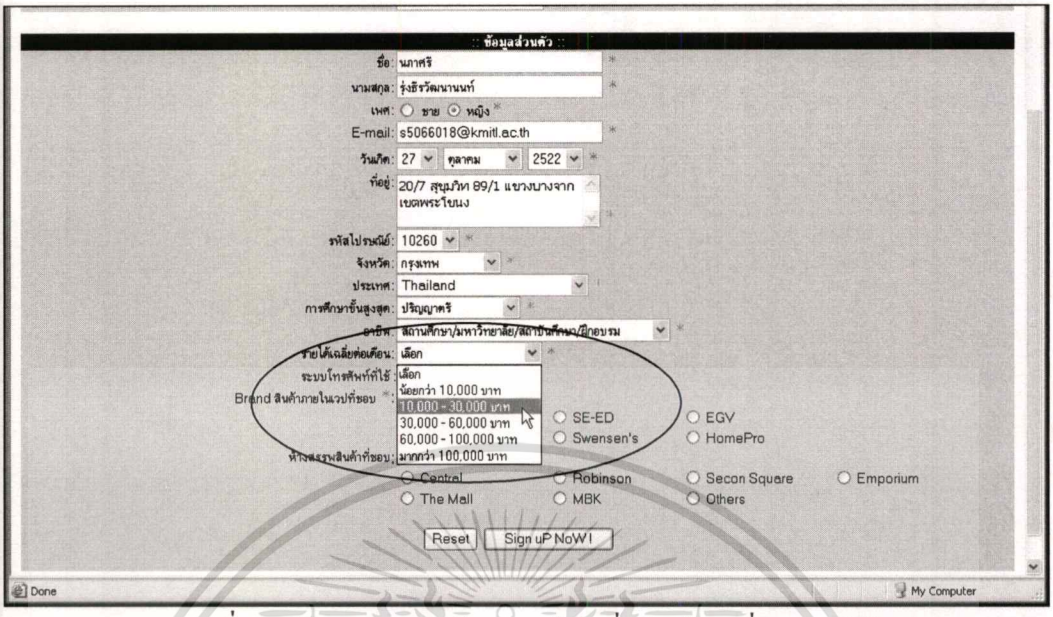
รูปที่ 4.2 ใช้ Radio Button รับข้อมูลเพศ เพื่อลดความผิดพลาด

- ใช้ List box หรือ Combo box ในการรับข้อมูลวันเดือนปีเกิด (Birthday) ดังนั้นลูกค้าไม่ต้องพิมพ์วันเดือนปีเกิดเอง เพียงแต่เลือกข้อมูลที่อยู่ใน List box เท่านั้น จากนั้นเมื่อลูกค้ากดปุ่ม Sign up เพื่อลงทะเบียนสมัครสมาชิก ข้อมูลวันเดือนปีเกิดจะถูกนำมาคำนวณให้เป็นอายุ (Age) และเก็บลงในฐานข้อมูล



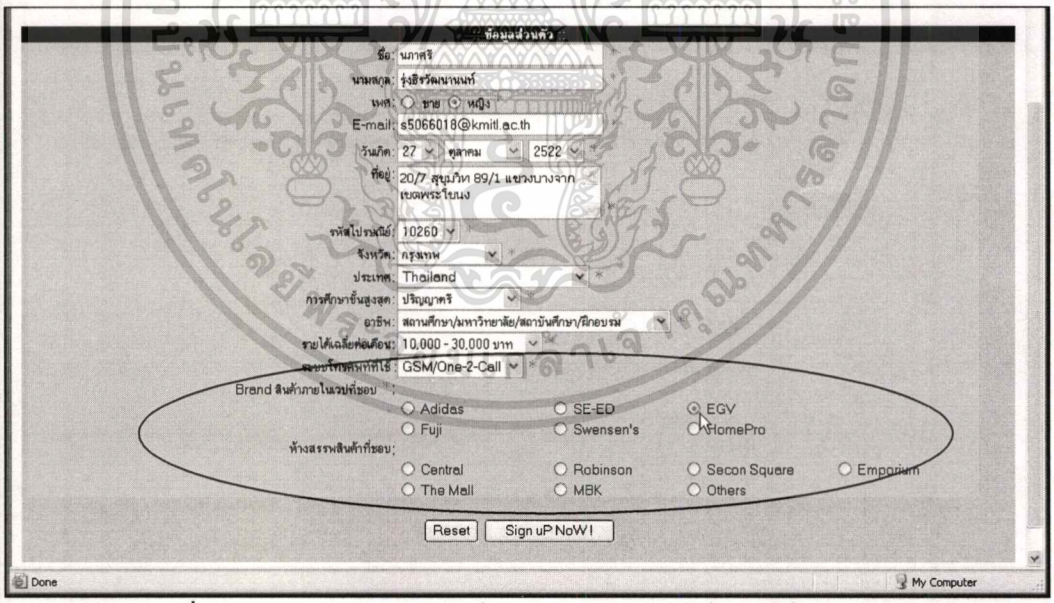
รูปที่ 4.3 ใช้ List Box รับข้อมูลวันเดือนปีเกิด เพื่อลดความผิดพลาด

- ใช้ List box หรือ Combo box ในการรับข้อมูลรายได้ (Income) ข้อมูลรายได้จะแบ่งเป็นรายได้ช่วงต่างๆ ลูกค้าสามารถเลือกได้จาก List box



รูปที่ 4.4 ใช้ List Box รับข้อมูลรายได้เฉลี่ยต่อเดือน เพื่อลดความผิดพลาด

- ใช้ Radio Button ในการรับข้อมูลยี่ห้อสินค้า ลูกค้าสามารถเลือกยี่ห้อสินค้าใดก็ได้เพียงยี่ห้อหนึ่งเท่านั้น



รูปที่ 4.5 ใช้ List Box รับข้อมูลยี่ห้อสินค้าภายในเว็บที่ชอบ เพื่อลดความผิดพลาด

จากการรับข้อมูลข้างต้น จะเห็นว่าข้อมูลที่เราจะนำมาใช้ในการวิเคราะห์สามารถป้องกันการเกิด Missing Value และ Noisy Data ได้กล่าวคือ ข้อมูลที่เราจะนำไปวิเคราะห์ซึ่ง

ประกอบด้วยข้อมูลเพศ (Gender), อายุ (Age), รายได้ (Income) และยี่ห้อสินค้า (Brand) ต่างก็ใช้แบบฟอร์มในการรับข้อมูลเพื่อลดปัญหาการกรอกข้อมูลที่ผิดพลาด ลดการเกิด Noisy Data และข้อมูลข้างต้นเป็นข้อมูลที่เว็บไซต์บังคับให้ลูกค้าต้องกรอก (โดยทำเครื่องหมาย * และแจ้งให้ลูกค้าทราบ) มิฉะนั้นก็จะไม่สามารถลงทะเบียนสมัครสมาชิกได้ ข้อกำหนดดังกล่าวทำให้ข้อมูลที่ได้รับนั้น ลดการเกิดปัญหา Missing Value ได้ ดังนั้นข้อมูลที่เราจะนำมาวิเคราะห์นั้น จะเป็นข้อมูลที่มีคุณภาพ

4.2.3 รูปแบบของข้อมูลที่ถูกจัดเก็บลงในฐานข้อมูล

เมื่อลูกค้ากดปุ่มเพื่อลงทะเบียนสมัครสมาชิกแล้วนั้น ในการเก็บข้อมูลลงฐานข้อมูล ข้อมูลจะถูกปรับเปลี่ยนรูปแบบในการจัดเก็บ ในที่นี้จะขอกล่าวถึงเฉพาะข้อมูลที่จะนำไปพิจารณา มีรายละเอียดดังต่อไปนี้คือ

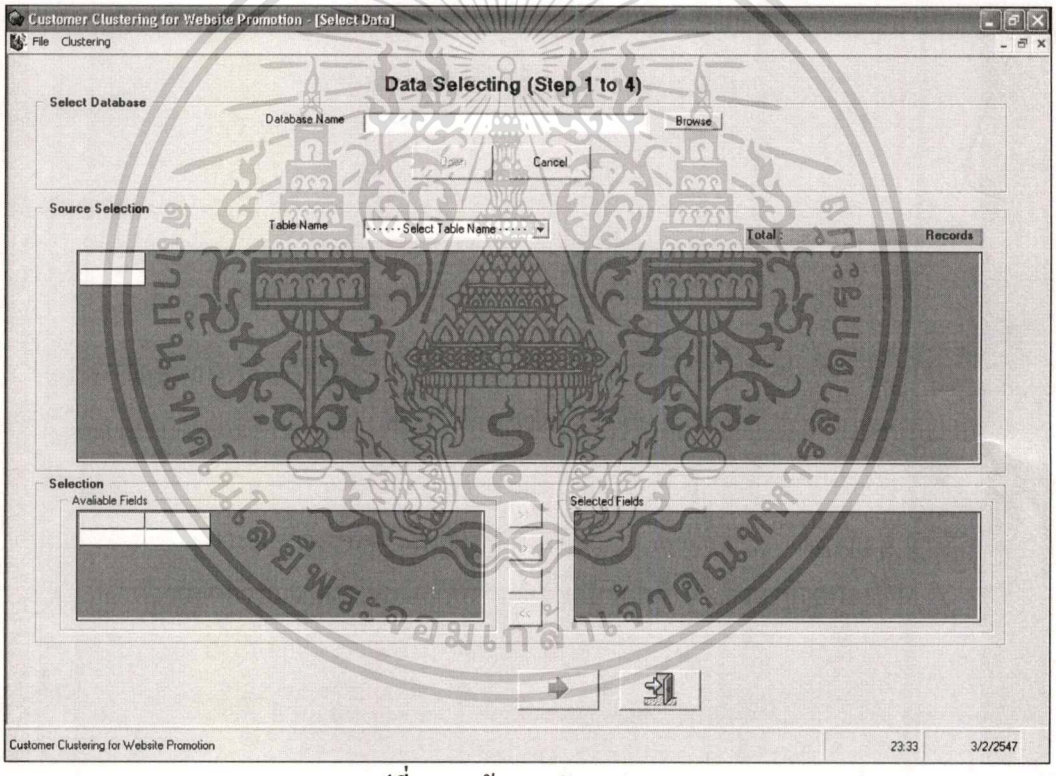
- ข้อมูลเพศ (Gender) เป็นข้อมูลประเภท Categorical จะจัดเก็บในฐานข้อมูลโดยมีชนิดของข้อมูลเป็น Text
  - เพศหญิง (Female) จัดเก็บเป็น 0
  - เพศชาย(Male)จัดเก็บเป็น 1
- ข้อมูลอายุ (Age) คือข้อมูลที่ได้จากการคำนวณจากวันเกิด เป็นข้อมูลประเภท Numeric จะจัดเก็บในฐานข้อมูลโดยมีชนิดของข้อมูลเป็น Number
- ข้อมูลรายได้ (Income) เป็นข้อมูลประเภท Categorical จะจัดเก็บในฐานข้อมูลโดยมีชนิดของข้อมูลเป็น Text
  - รายได้น้อยกว่า 10,000 บาท จัดเก็บเป็น 1
  - รายได้ 10,000 ถึง 30,000 บาท จัดเก็บเป็น 2
  - รายได้ 30,000 ถึง 60,000 บาท จัดเก็บเป็น 3
  - รายได้ 60,000 ถึง 100,000 บาท จัดเก็บเป็น 4
  - รายได้ 100,000 บาทขึ้นไป จัดเก็บเป็น 5
- ข้อมูลยี่ห้อสินค้า (Brand) เป็นข้อมูลประเภท Categorical จะจัดเก็บในฐานข้อมูลโดยมีชนิดของข้อมูลเป็น Text
  - ยี่ห้อ Adidas จัดเก็บเป็น 1
  - ยี่ห้อ Swensen's จัดเก็บเป็น 2
  - ยี่ห้อ EGV จัดเก็บเป็น 3
  - ยี่ห้อ Fuji จัดเก็บเป็น 4
  - ยี่ห้อ SE-ED จัดเก็บเป็น 5

เมื่อเสร็จสิ้นขั้นตอนการเตรียมข้อมูลเรียบร้อยแล้ว จะได้ข้อมูลที่พร้อมจะผ่านกระบวนการทำดาต้าไมนิ่ง

4.3 การนำข้อมูลมาทำดาต้าไมนิ่ง

สำหรับขั้นตอนการทำดาต้าไมนิ่งนั้น จะใช้อัลกอริทึม *k-prototypes* ในการแบ่งกลุ่มข้อมูล เพราะเป็นอัลกอริทึมที่มีความสามารถรองรับข้อมูลทั้งประเภท numeric และ categorical ตรงกับข้อมูลที่ต้องการวิเคราะห์ ดังนั้นจึงได้พัฒนาระบบงาน 'Customer Clustering for Website Promotion' ด้วยอัลกอริทึม *k-prototypes* เพื่อจัดกลุ่มข้อมูลลูกค้าให้ตรงตามจุดประสงค์

เมื่อเข้าสู่โปรแกรมจะปรากฏหน้าจอดังรูปที่ 4.6



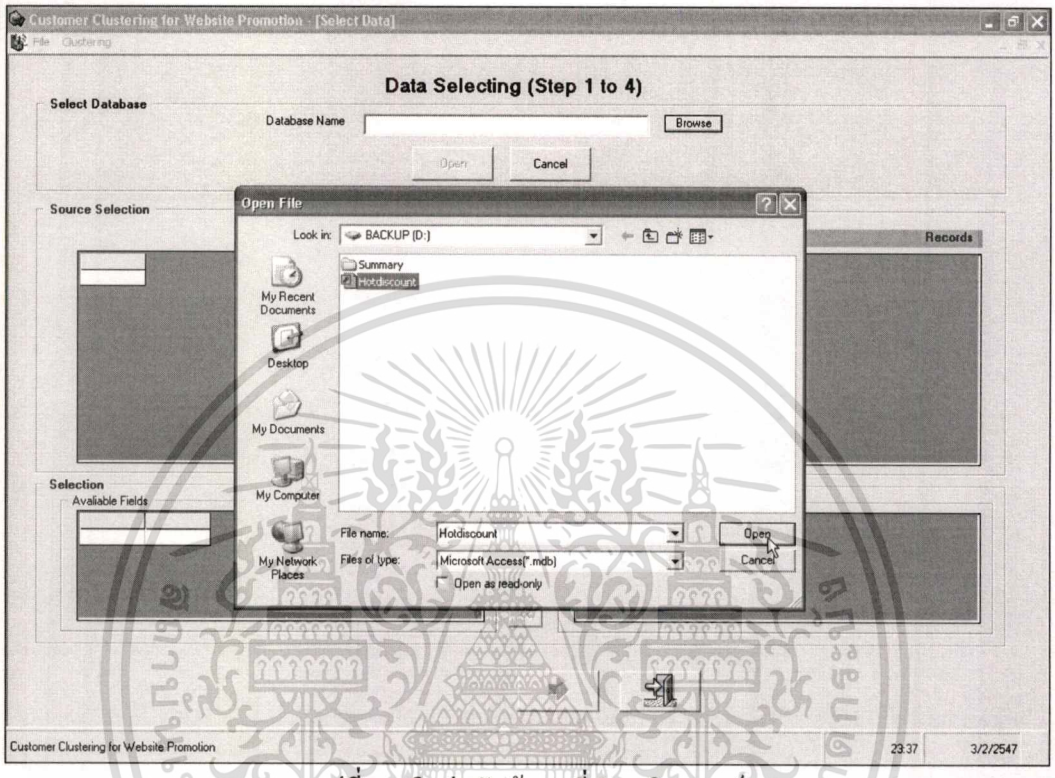
รูปที่ 4.6 หน้าจอหลักของระบบ

หน้าจอหลักคือหน้าจอการเลือกข้อมูล (Data Selecting) จะประกอบด้วย 3 ส่วนคือ

- Select Database เลือกฐานข้อมูล
- Source Selection เลือกตารางที่ต้องการวิเคราะห์
- Selection เลือกฟิลด์ที่ต้องการวิเคราะห์

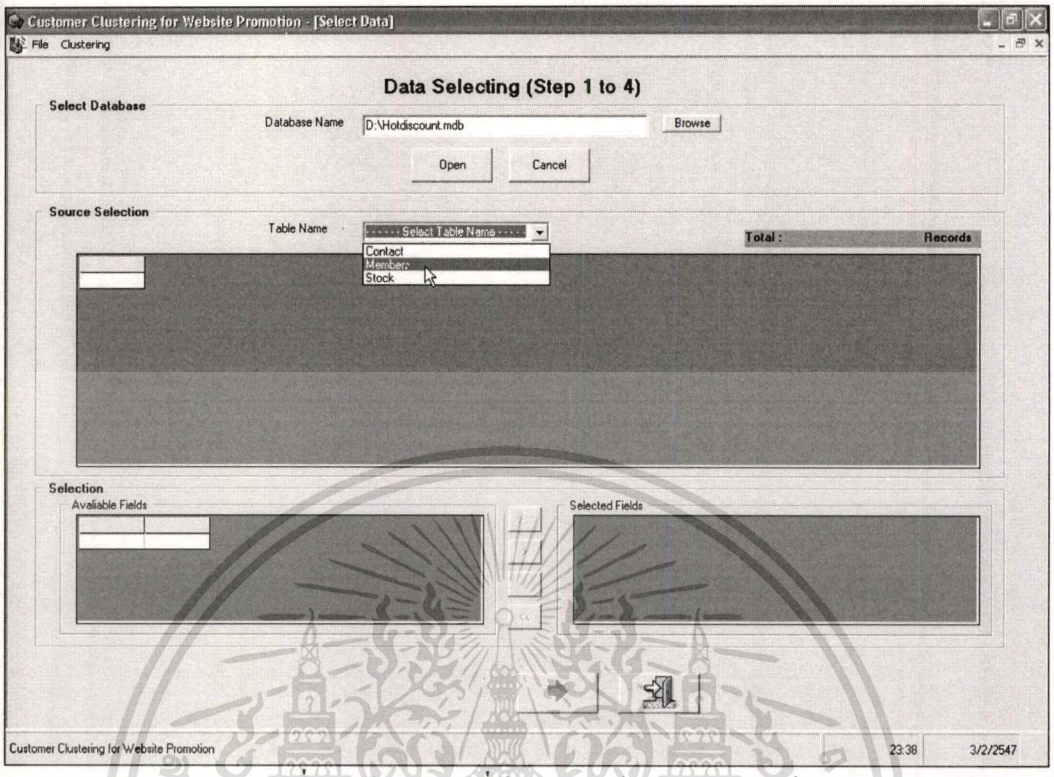
4.3.1 การติดต่อกับข้อมูลที่นำมาวิเคราะห์

กดปุ่ม Browse ในส่วน Select Database เพื่อเลือกไฟล์ฐานข้อมูลนามสกุล *.mdb



รูปที่ 4.7 ติดต่อกับข้อมูลที่นำมาวิเคราะห์

จากนั้นเลือกตารางที่ต้องการนำข้อมูลมาวิเคราะห์

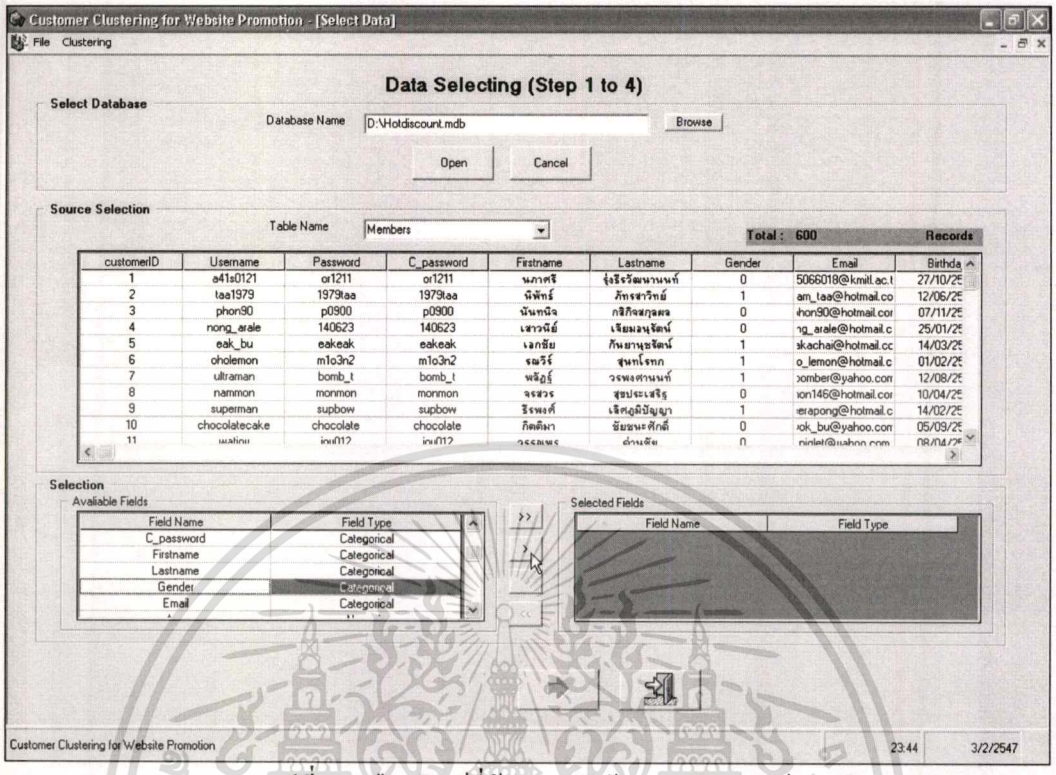


รูปที่ 4.8 เลือกตารางที่ต้องการนำข้อมูลมาวิเคราะห์

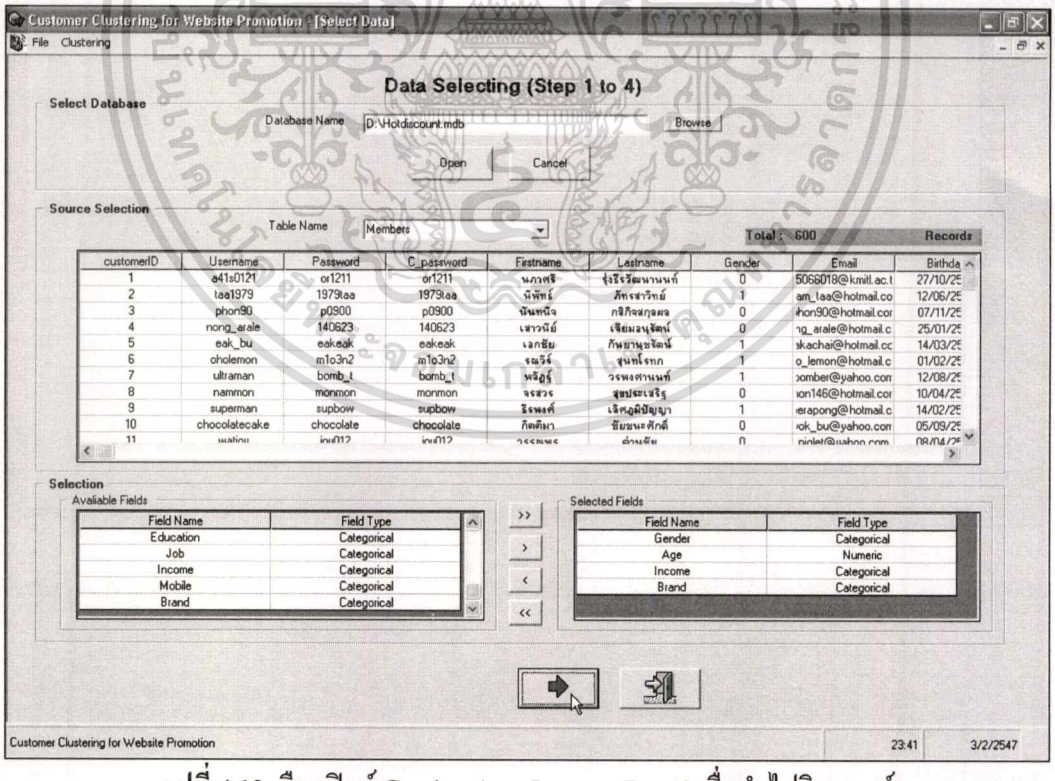
เมื่อเลือกตารางแล้ว ข้อมูลในตารางนั้นจะแสดงใน Source Selection และจะแสดงชื่อฟิลด์และชนิดข้อมูลของฟิลด์นั้นในส่วนของ Selection

4.3.2 การเลือกฟิลด์ที่ต้องการวิเคราะห์

ฟิลด์ในตารางทั้งหมดจะแสดงอยู่ใน Available Fields ฟิลด์ใดที่ต้องการเลือกเพื่อนำไปวิเคราะห์นั้นให้เลือกฟิลด์มาอยู่ใน Selected Fields



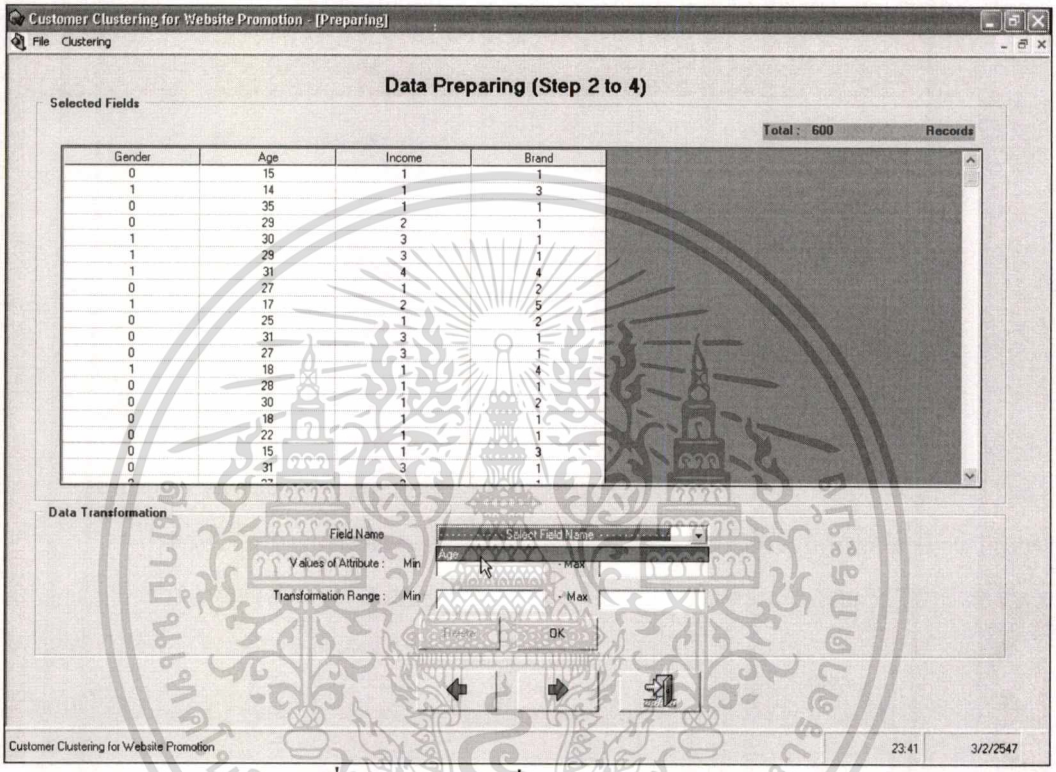
รูปที่ 4.9 เลือกฟิลด์ที่ต้องการนำข้อมูลมาวิเคราะห์



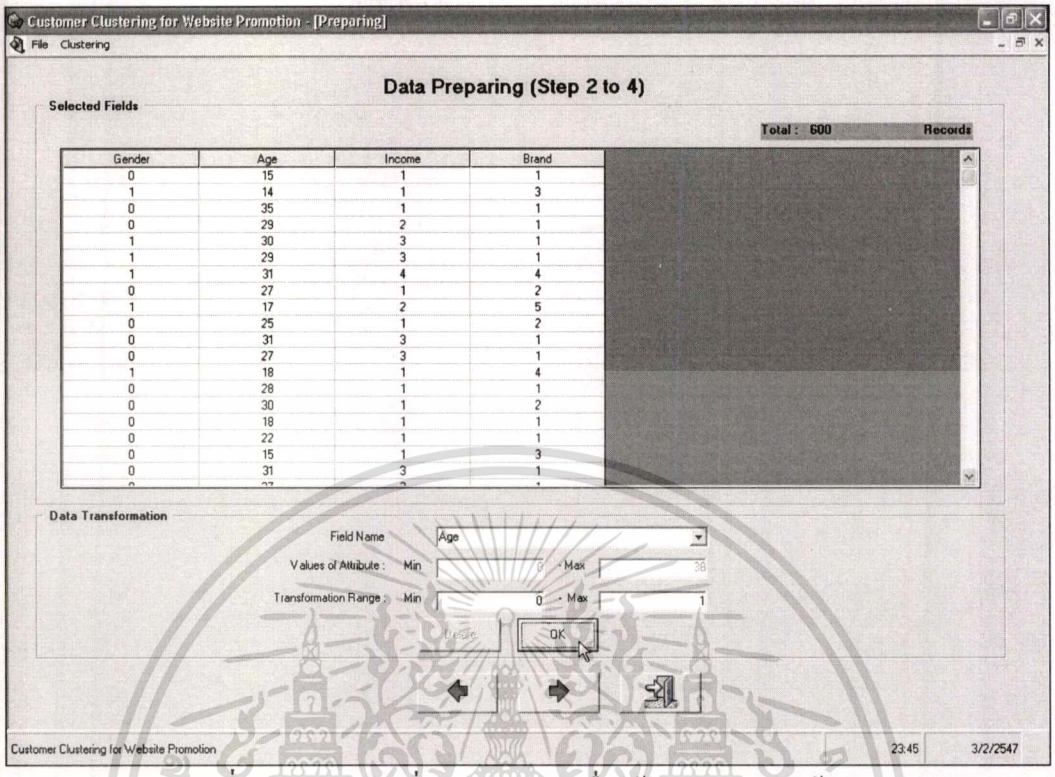
รูปที่ 4.10 เลือกฟิลด์ Gender, Age, Income, Brand เพื่อนำไปวิเคราะห์

4.3.3 การแปลงค่าข้อมูล

หน้าจอที่2 คือการเตรียมข้อมูล (Data Preparing) เป็นการแปลงข้อมูลให้อยู่ในขอบเขตที่ต้องการ ด้วยวิธี Min-Max normalization ข้อมูลที่จะได้แปลงได้นั้นจะต้องเป็นข้อมูลประเภท Numeric เท่านั้น เช่น การแปลงข้อมูลอายุให้อยู่ในขอบเขตที่ต้องการ โดยกำหนดค่าต่ำสุดและค่าสูงสุด



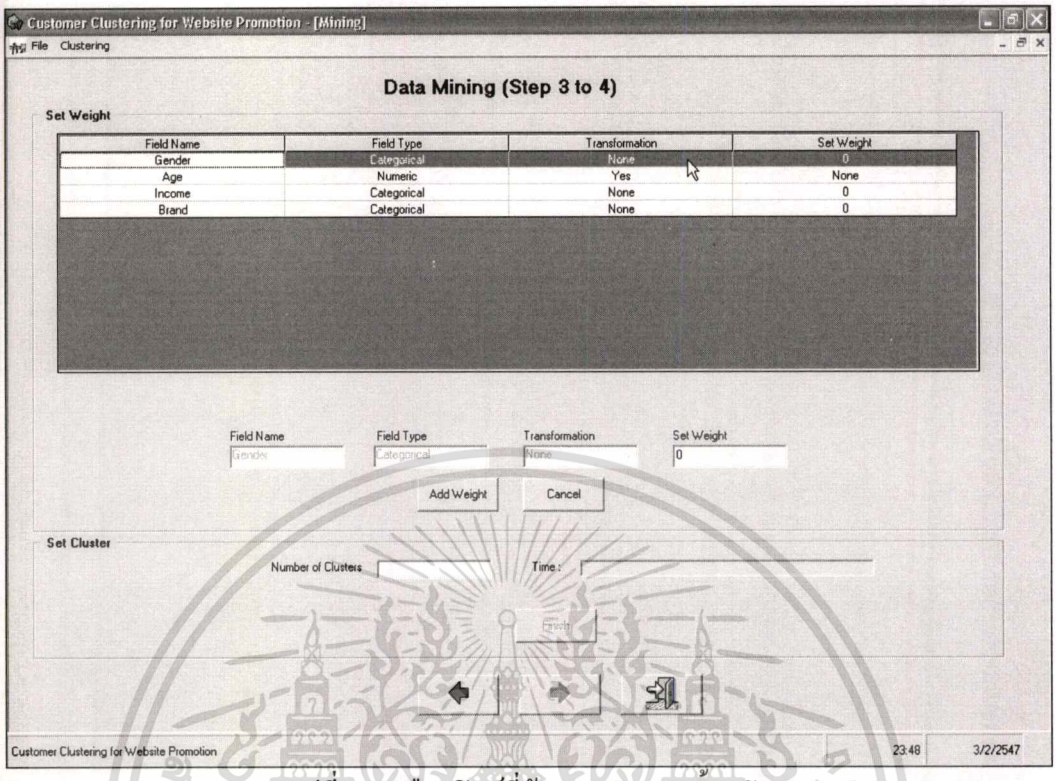
รูปที่ 4.11 เลือกฟิลด์ที่ต้องการแปลงค่าข้อมูล



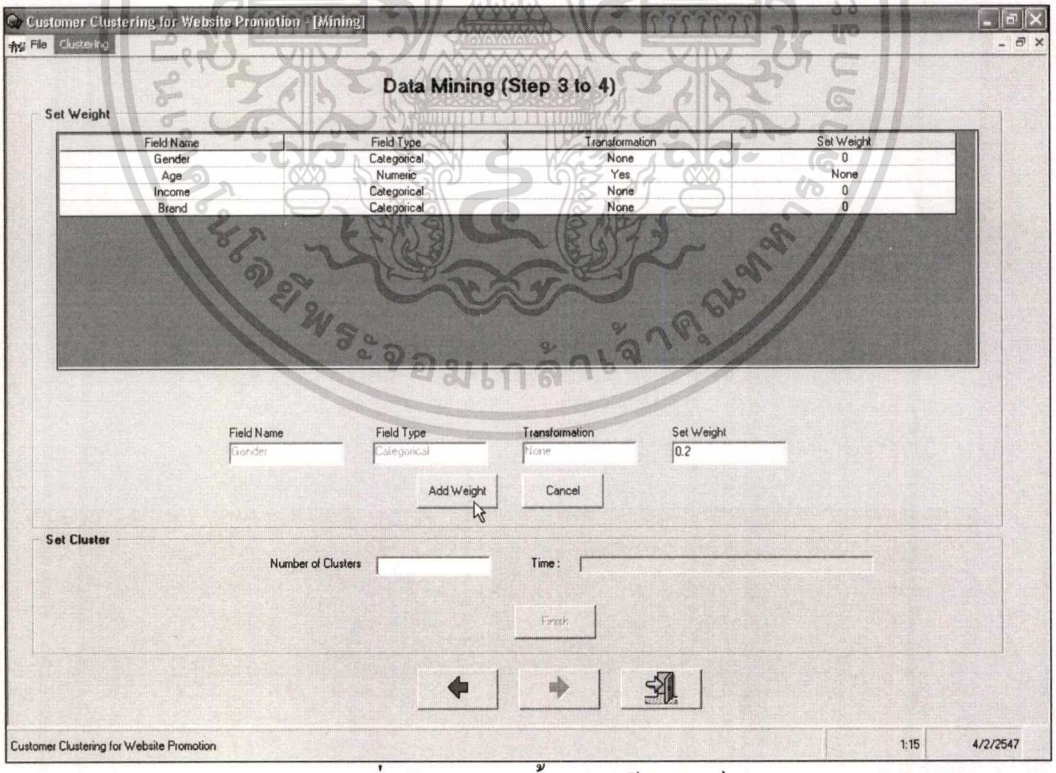
รูปที่ 4.12 กำหนดค่าต่ำสุดและสูงสุดเพื่อใช้ในการแปลงค่าข้อมูล

4.3.4 การกำหนดน้ำหนักให้กับข้อมูลประเภท Categorical

หน้าจอที่3 คือการทำดาต้าไมนิ่ง (Data Mining) ก่อนที่จะกำหนดจำนวนกลุ่มที่ต้องการแบ่งข้อมูลนั้น จากอัลกอริทึมข้อมูลประเภท Categorical นั้นสามารถกำหนดน้ำหนักให้กับข้อมูลได้ เช่น การกำหนดน้ำหนักให้กับฟิลด์เพศ (Gender) ให้เท่ากับ 0.5 เป็นต้น

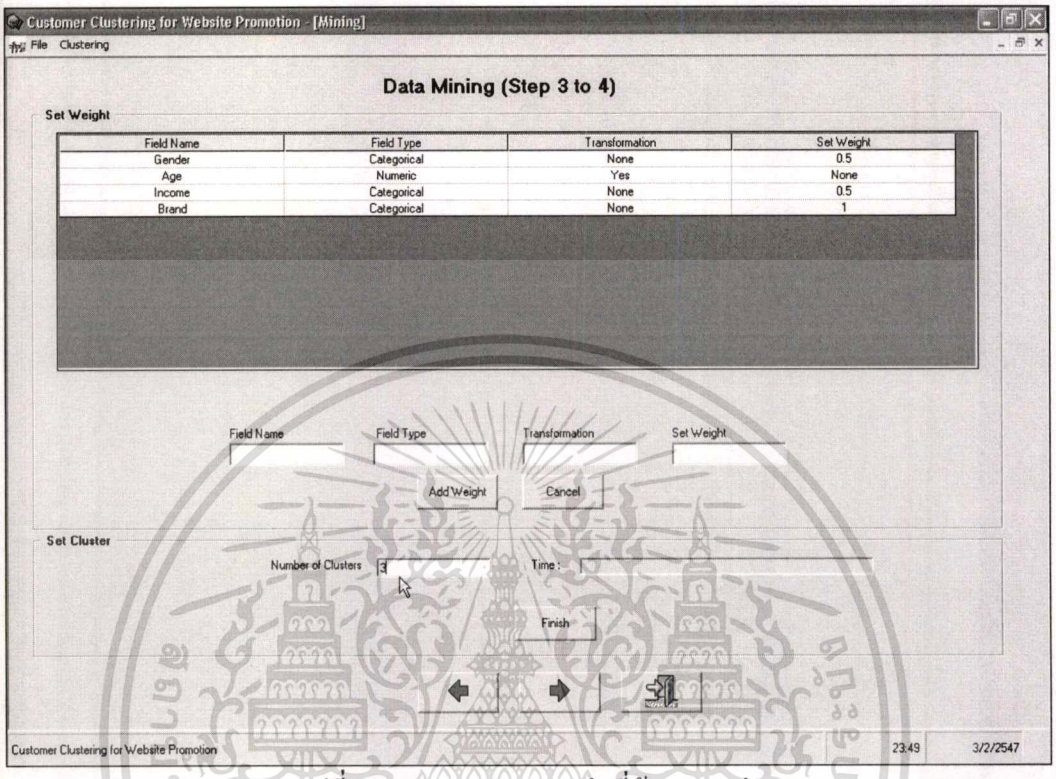


รูปที่ 4.13 เลือกฟิลด์ที่ต้องการกำหนดน้ำหนัก

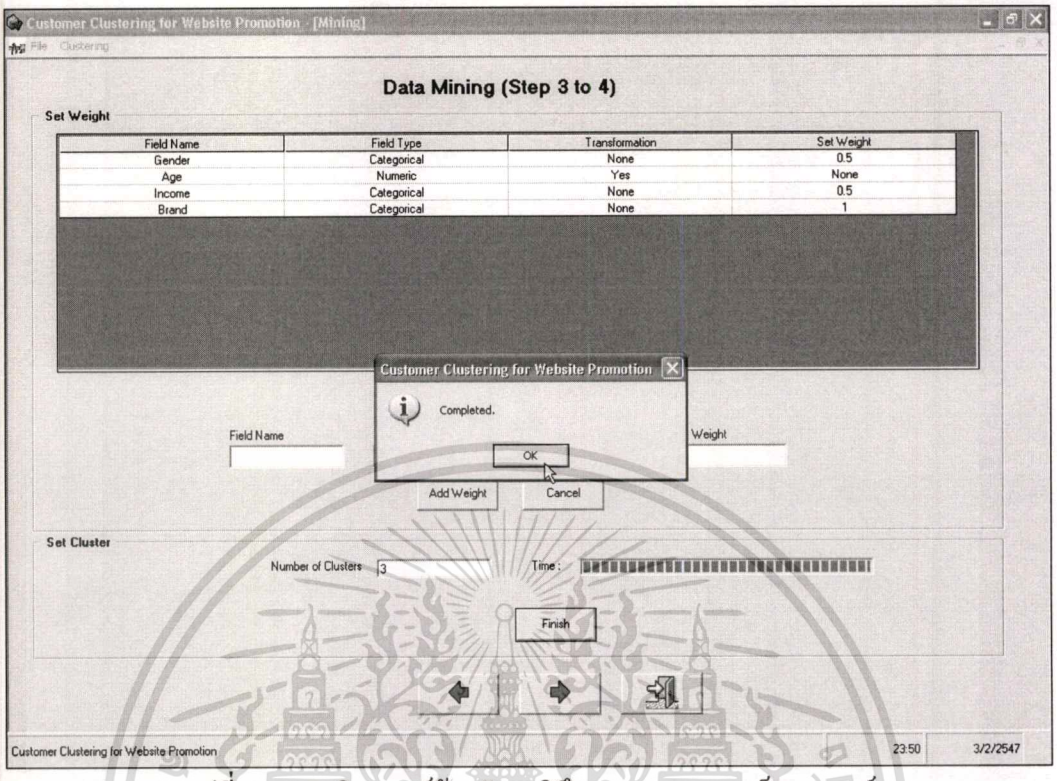


รูปที่ 4.14 กำหนดน้ำหนักให้กับฟิลด์

4.3.5 การกำหนดจำนวนกลุ่มที่ต้องการวิเคราะห์



รูปที่ 4.15 กำหนดจำนวนกลุ่มที่ต้องการแบ่ง

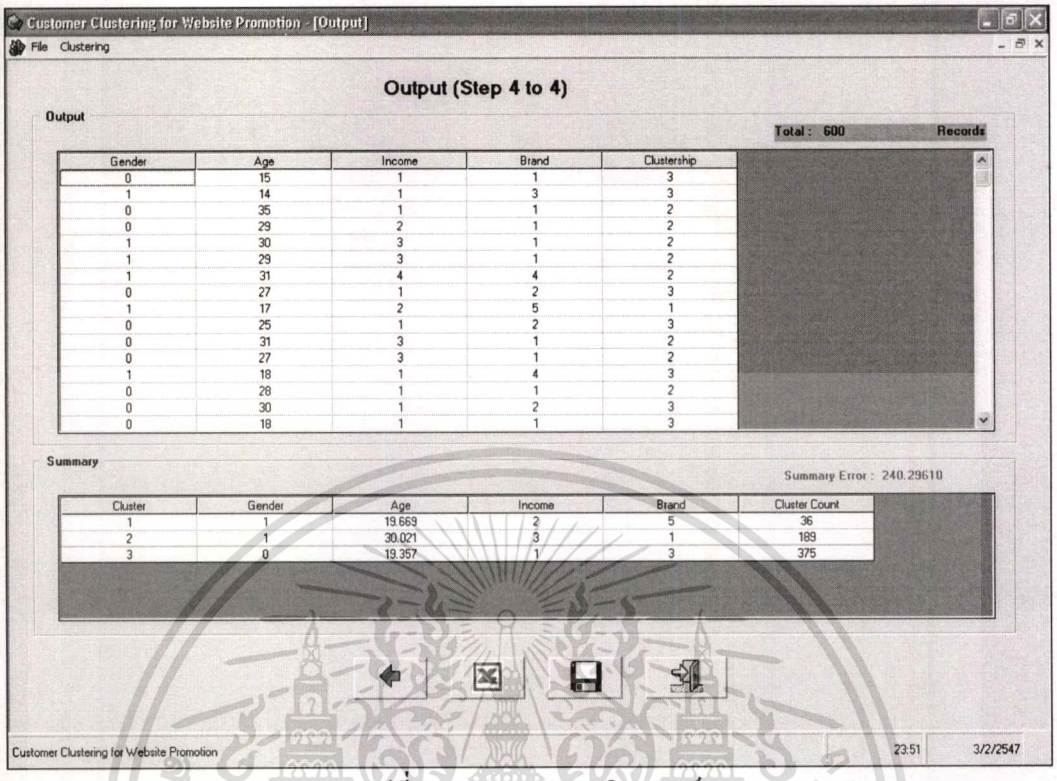


รูปที่ 4.16 การวิเคราะห์ด้วยอัลกอริทึม *k-prototypes* เสร็จสมบูรณ์

4.3.6 การแสดงผล

หลังจากที่ระบบได้ทำการวิเคราะห์ข้อมูลเสร็จเรียบร้อยแล้ว หน้าจอที่ 4 คือการแสดงผล (Output) ซึ่งแบ่งออกเป็น 2 ส่วนคือ

- Output จะแสดงว่าข้อมูลแต่ละเรคอร์ดอยู่ในกลุ่มใด
- Summary สรุปผลการวิเคราะห์ที่ได้ของแต่ละกลุ่มว่าจุดศูนย์กลางมีค่าเท่าใด และจำนวนสมาชิกที่อยู่ในกลุ่มนั้น



รูปที่ 4.17 แสดงผลการวิเคราะห์

4.4 วิเคราะห์ผลการดำเนินงาน

ผลลัพธ์ที่ได้จากการวิเคราะห์ โดยเลือกข้อมูลเพื่อให้ตรงกับวัตถุประสงค์ที่ตั้งไว้คือ ข้อมูลเพศ (Gender) ข้อมูลอายุ(Age) ข้อมูลรายได้(Income) และข้อมูลยี่ห้อสินค้าที่ชอบ(Brand) จากนั้นกำหนดค่าน้ำหนักให้กับข้อมูลเพศ และข้อมูลรายได้เท่ากับ 0.5 ข้อมูลยี่ห้อสินค้าที่ชอบเท่ากับ 1 จากนั้นทำการไมนิ่งด้วยอัลกอริทึม k-prototypes โดยกำหนดให้วิเคราะห์ข้อมูล 600 เรคอร์ด ออกเป็น 3 กลุ่ม ผลลัพธ์ที่ได้พบว่า

- **กลุ่มที่ 1** Gender = 1, Age = 19.669, Income = 2, Brand = 5, Cluster Count = 36
- **กลุ่มที่ 2** Gender = 1, Age = 30.021, Income = 3, Brand = 1, Cluster Count = 189
- **กลุ่มที่ 3** Gender = 0, Age = 19.357, Income = 1, Brand = 3, Cluster Count = 375

เมื่อปรับเปลี่ยนข้อมูลจากรูปแบบที่จัดเก็บ จะอธิบายความหมายได้ดังต่อไปนี้

- **กลุ่มที่ 1** มีลักษณะเป็นเพศชาย
  มีอายุประมาณ 19.669 ปี
  มีรายได้ 10,000 ถึง 30,000 บาท

ยี่ห้อสินค้าที่ชอบคือ SE-ED

มีจำนวนข้อมูลที่อยู่ในกลุ่มที่ 1 เท่ากับ 36

- **กลุ่มที่ 2** มีลักษณะเป็นเพศชาย

  มีอายุประมาณ 30.021 ปี

  มีรายได้ 30,000 ถึง 60,000 บาท

  ยี่ห้อสินค้าที่ชอบคือ Adidas

  มีจำนวนข้อมูลที่อยู่ในกลุ่มที่ 2 เท่ากับ 189

- **กลุ่มที่ 3** มีลักษณะเป็นเพศหญิง

  มีอายุประมาณ 19.357 ปี

  มีรายได้น้อยกว่า 10,000 บาท

  ยี่ห้อสินค้าที่ชอบคือ EGV

  มีจำนวนข้อมูลที่อยู่ในกลุ่มที่ 3 เท่ากับ 375

ผลลัพธ์จากการแบ่งกลุ่มดังกล่าวยังจะต้องอาศัยกระบวนการวิเคราะห์ข้อมูลเชิงธุรกิจอีก เพื่อตัดสินว่าผลลัพธ์ที่ได้นั้น ข้อมูลกลุ่มใดจะสามารถนำไปเป็นข้อมูลเพื่อช่วยตัดสินใจหรือวางแผนการตลาดในทางธุรกิจ

บทที่ 5
สรุปผลการศึกษาและข้อเสนอแนะ

5.1 สรุปผลการศึกษา

จากที่ได้ศึกษาทฤษฎีดาต้าไมนิ่ง ทำให้ได้เรียนรู้ว่าดาต้าไมนิ่งเป็นกระบวนการที่ใช้ค้นหาความรู้จากข้อมูลที่มีอยู่จำนวนมากทำให้ได้สารสนเทศที่มีประโยชน์ และสามารถนำไปช่วยในการบริหารงานและสนับสนุนการตัดสินใจทางด้านธุรกิจ ซึ่งกระบวนการดังกล่าวนี้จะเริ่มตั้งแต่ การกำหนดวัตถุประสงค์ทางธุรกิจ จากนั้นต้องทำการเตรียมข้อมูลที่จะนำมาวิเคราะห์ ซึ่งจะประกอบด้วยขั้นตอนย่อยๆอีก หลังจากได้ข้อมูลเรียบร้อยแล้ว ขั้นตอนต่อไปคือการทำดาต้าไมนิ่ง เมื่อข้อมูลผ่านการทำไมนิ่ง ก็จะได้ผลลัพธ์ที่ต้องนำมาวิเคราะห์ และสุดท้ายคือการนำความรู้หรือสารสนเทศที่ได้ไปใช้ให้เกิดประโยชน์ในทางธุรกิจ

เมื่อได้ศึกษาหลักการดังกล่าวแล้ว จึงได้มีแนวทางในการพัฒนาโครงการโดยนำเสนอถึงการแบ่งกลุ่มข้อมูล หรือคือโอเปอร์เรชั่น Database Segmentation ซึ่งเป็นโอเปอร์เรชั่นหนึ่งในการทำดาต้าไมนิ่ง โดยศึกษาอัลกอริทึม k-prototypes ที่สามารถรองรับข้อมูลทั้งประเภท Numeric และ Categorical ได้และสามารถรองรับข้อมูลที่มีจำนวนมากได้อย่างมีประสิทธิภาพ เมื่อได้ศึกษาอัลกอริทึม k-prototypes แล้วพบว่าสอดคล้องกับข้อมูลที่ต้องการทำดาต้าไมนิ่ง เพราะข้อมูลที่มีอยู่นั้นมีทั้งข้อมูลที่ประเภท Numeric และ Categorical จึงได้วางแนวทางในการพัฒนาระบบด้วยโปรแกรม Microsoft Visual Basic 6.0 และโปรแกรมฐานข้อมูล Microsoft Access โดยกำหนดจุดประสงค์ที่ต้องการศึกษาคือ เพื่อจัดกลุ่มลูกค้าที่มีลักษณะใกล้เคียงกัน จะได้นำเสนอรายการส่งเสริมการขายหรือสิทธิพิเศษและบริการต่างๆ ให้ตรงตามกลุ่มเป้าหมาย เมื่อได้พัฒนาระบบและศึกษาถึงผลลัพธ์ที่ได้นั้น พบว่าอัลกอริทึมนี้สามารถรองรับข้อมูลได้เป็นอย่างดี และผลลัพธ์ที่ได้นั้นสามารถบ่งบอกถึงลักษณะของข้อมูลกลุ่มต่างๆ อันจะเป็นประโยชน์ในการนำเสนอรายการส่งเสริมการขายหรือการวางแผนในเชิงการตลาดได้

5.2 ประโยชน์ที่ได้จากการศึกษาและพัฒนาระบบ

1. ทำให้เข้าใจทฤษฎีของดาต้าไมนิ่งยิ่งขึ้น โดยเฉพาะอย่างยิ่งโอเปอร์เรชั่น Database Segmentation
2. ทำให้ทราบถึงปัญหาและเข้าใจขั้นตอนในการพัฒนาระบบ
3. ทำให้ได้ต้นแบบสำหรับการพัฒนาระบบการจัดกลุ่มข้อมูล ซึ่งสามารถนำไปประยุกต์ใช้ในการพัฒนาระบบจัดกลุ่มข้อมูลในกรณีอื่นๆ

5.3 ข้อเสนอแนะ

1. สำหรับระบบที่พัฒนาขึ้นมานี้ รองรับการนำไปใช้ในการวิเคราะห์ข้อมูลในงานธุรกิจอื่นๆ ได้ เพียงมีไฟล์นามสกุล *.mdb ก็สามารถวิเคราะห์ข้อมูลด้วยระบบนี้ได้
2. เพื่อให้ผลลัพธ์ที่ได้มีความถูกต้องแม่นยำ จึงควรใช้ระบบกับข้อมูลที่อยู่ใน DataWarehouse หรือข้อมูลที่สะอาดแล้ว เนื่องจากกรณีที่ได้ศึกษานี้ ข้อมูลที่นำมาใช้นั้นถูกทำให้สะอาดจากขั้นตอนที่รับข้อมูลจาก user แล้ว จึงไม่มีในส่วนของการทำความสะอาดข้อมูล หากไม่ได้ใช้ข้อมูลที่สะอาด อาจทำให้ผลลัพธ์ที่ได้นั้นคลาดเคลื่อนหรือมีความถูกต้องน้อยลงได้

บรรณานุกรม

ศุภชัย สมพานิช. 2545. **สร้างระบบงานฐานข้อมูลด้วย Visual Basic ฉบับโปรแกรมเมอร์.** พิมพ์ครั้งที่ 1. นนทบุรี: อินโฟเพรส

DATA MINING. [Online]. Available: http://project.cs.kku.ac.th/2544/seminar/day2/413356-4nd413357-5/datamining.doc

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

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

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

ก. การติดตั้งโปรแกรมและการใช้งานเบื้องต้น

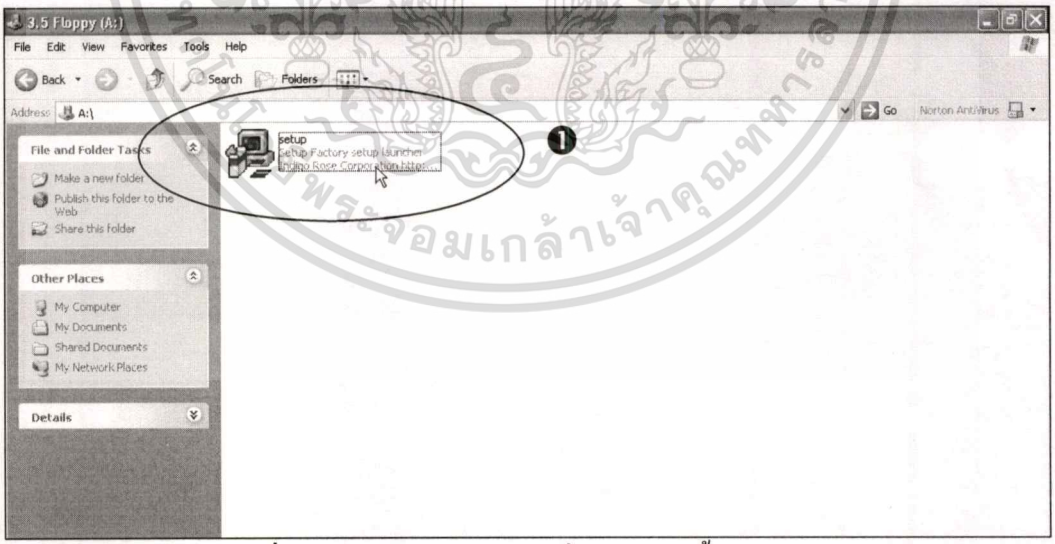
ก.1 การติดตั้งโปรแกรม

สำหรับการติดตั้งโปรแกรม Customer Clustering for Website Promotion สิ่งแรกที่ต้องทำคือการเตรียมเครื่องคอมพิวเตอร์ที่มีคุณสมบัติพร้อมใช้งานโปรแกรม Customer Clustering for Website Promotion ได้ โดยมีคำแนะนำให้ใช้เครื่องคอมพิวเตอร์ที่มีรายละเอียดขั้นต่ำดังนี้คือ

- ระบบปฏิบัติการ Windows 98/98 SE /Me/2000/XP
- เครื่องคอมพิวเตอร์ที่ใช้หน่วยประมวลผลระดับ Pentium 133 MHz เป็นอย่างน้อย แต่ถ้าต้องการรองรับฐานข้อมูลขนาดใหญ่ ขอแนะนำให้ใช้เครื่องที่มีหน่วยประมวลผลกลางมีความเร็วสูงกว่านี้
- หน่วยความจำอย่างน้อย 64 MB หากต้องการให้มีประสิทธิภาพในการทำงานควรมากกว่านี้
- พื้นที่ในฮาร์ดดิสก์ 1 MB หรือมากกว่า สำหรับติดตั้งโปรแกรม ขอแนะนำให้มีพื้นที่ว่างอย่างน้อย50 MB เพื่อเพียงพอในการประมวลผลของโปรแกรม
- จอแสดงผล Super VGA (800 x 600) แสดงผลที่ 256 สี

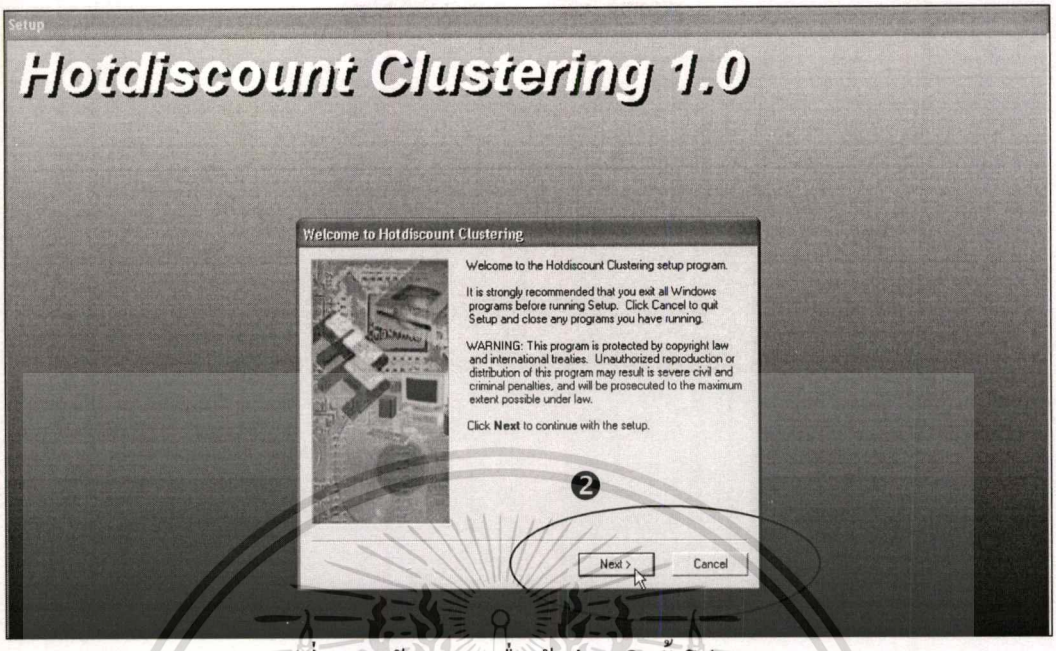
สำหรับการติดตั้งโปรแกรม ก่อนอื่นขอให้ปิดโปรแกรมทั้งหมดที่เปิดไว้ใน Windows หลังจากนั้นให้ใส่แผ่นดิสก์สำหรับติดตั้งโปรแกรมในไดร์ฟ A

1. Double Click ไฟล์ Setup.exe ที่อยู่ในไดร์ฟ A เพื่อทำการติดตั้งโปรแกรม



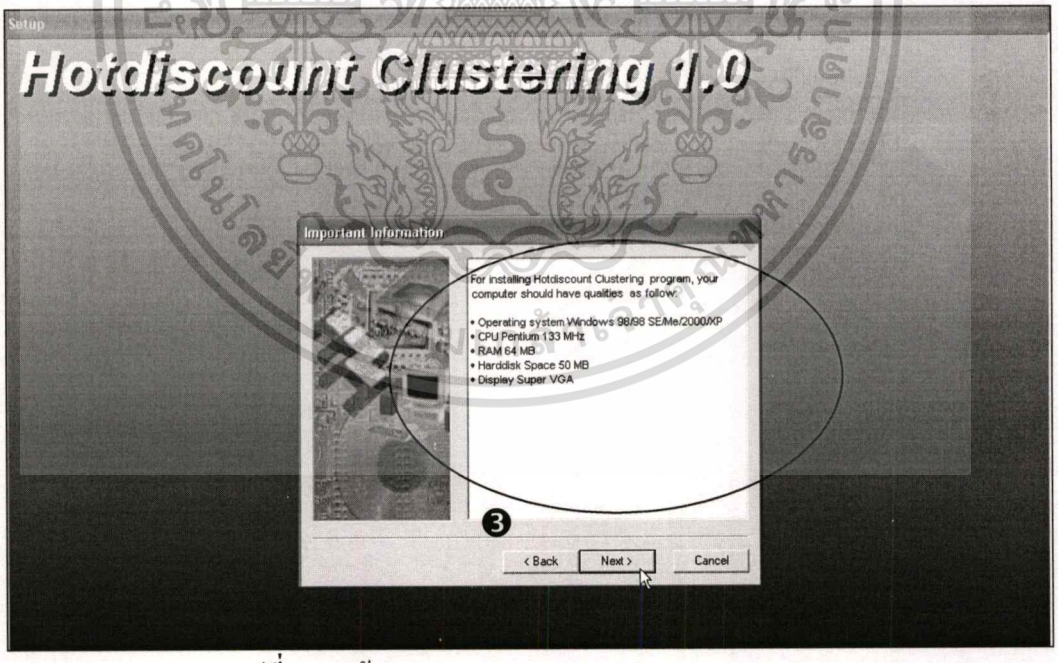
รูปที่ ก.1 เลือกไฟล์ Setup.exe เพื่อทำการติดตั้งโปรแกรม

2. เข้าสู่หน้าจอการติดตั้งโปรแกรม กดปุ่ม Next เพื่อไปสู่ขั้นตอนถัดไป



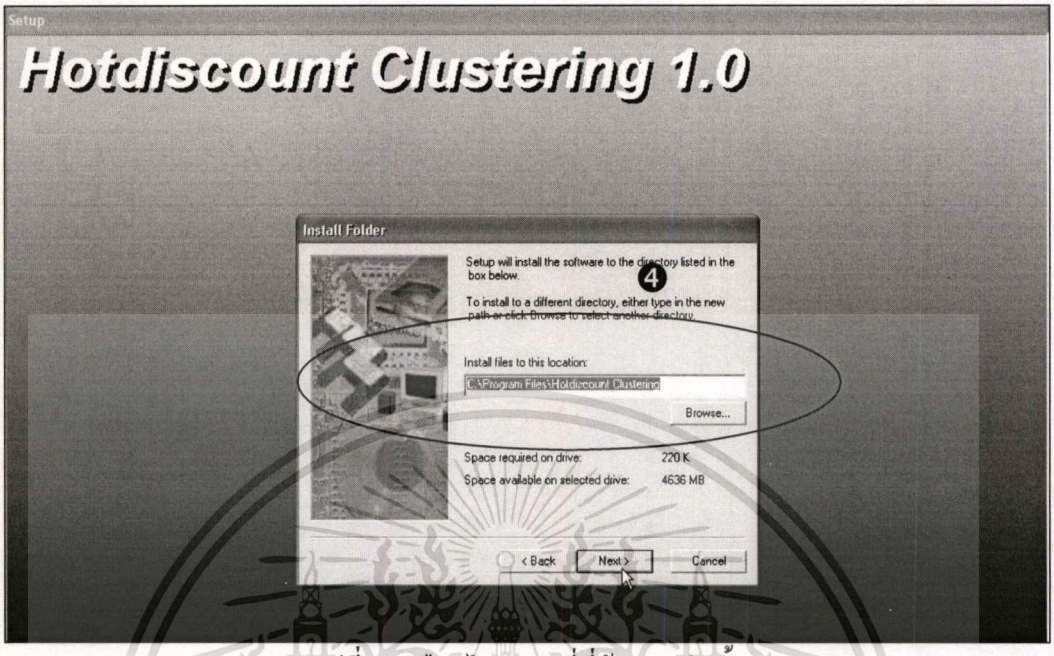
รูปที่ ก.2 หน้าจอแรกเมื่อเข้าสู่การติดตั้งโปรแกรม

3. โปรแกรมจะแสดง System Requirements ของระบบ กดปุ่ม Next เพื่อไปสู่ขั้นตอนถัดไป



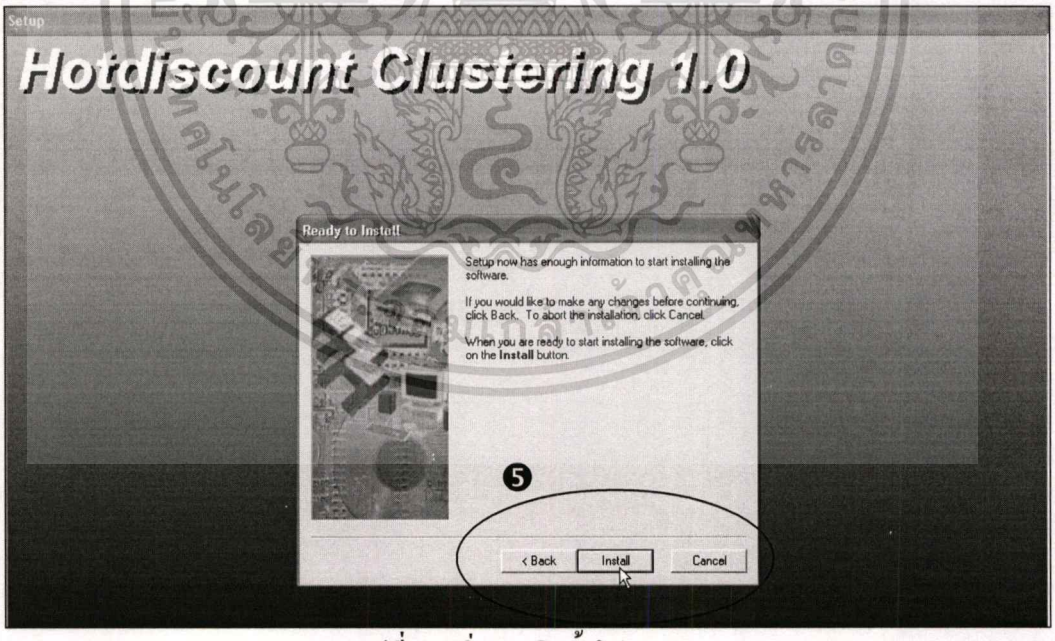
รูปที่ ก.3 หน้าจอแสดง System Requirements ของระบบ

4. เลือกไดเรกทอรี่ที่ต้องการติดตั้งโปรแกรม



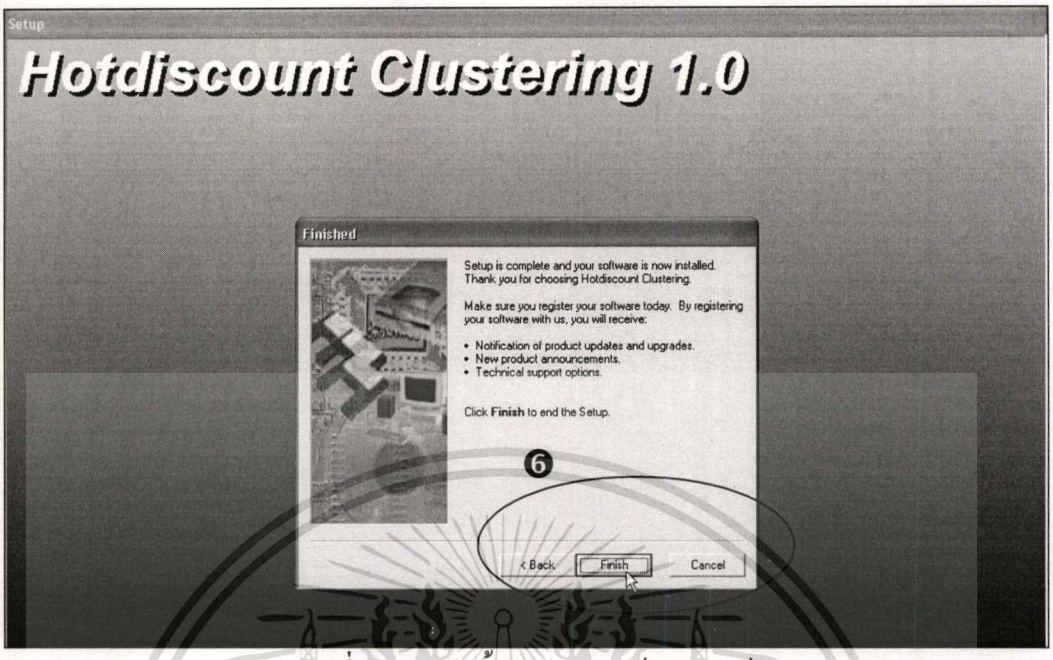
รูปที่ ก.4 เลือกไดเรกทอรี่ที่ต้องการติดตั้ง

5. กดปุ่ม Install เพื่อติดตั้งโปรแกรม



รูปที่ ก.5 เริ่มการติดตั้งโปรแกรม

6. การติดตั้งโปรแกรมเสร็จสมบูรณ์ กดปุ่ม Finish



รูปที่ ก.6 การติดตั้งโปรแกรมเสร็จสมบูรณ์

รูปที่ ก.7 โปรแกรมอยู่ใน Program Files

7. โปรแกรมจะถูกติดตั้งในไดเรกทอรี่ที่กำหนด

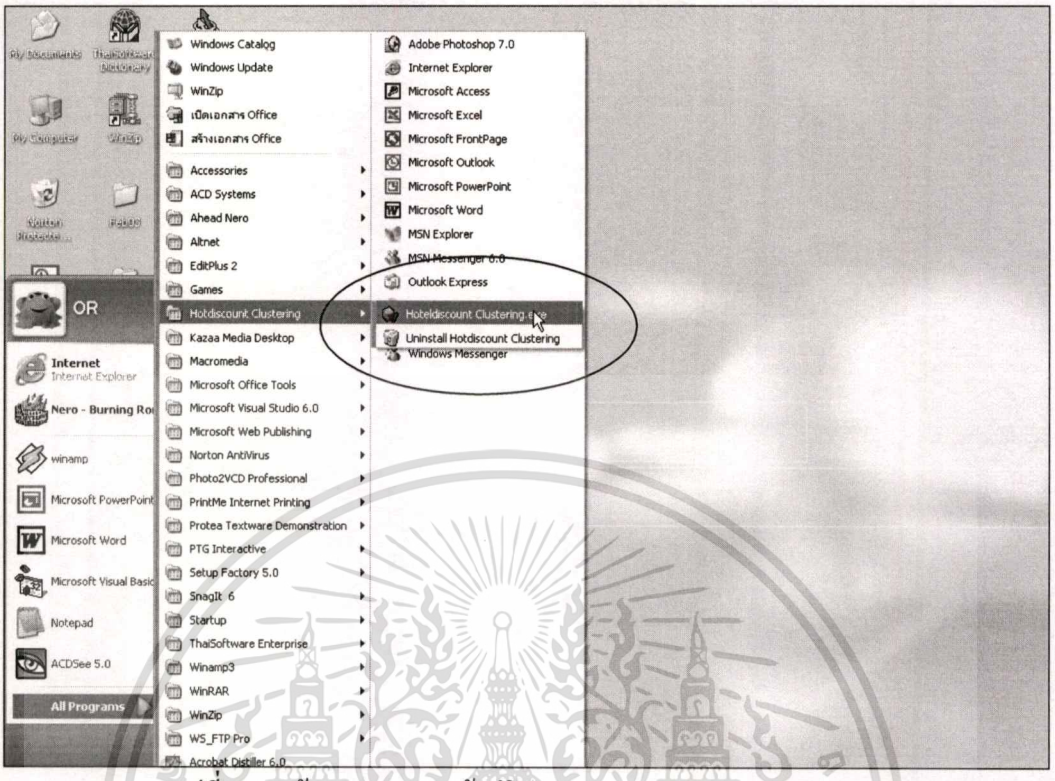
ก.2 การเข้าสู่โปรแกรม

การเข้าสู่โปรแกรม Hotdiscount Clustering สามารถทำได้โดย

1.1.1 Click mouse ที่ Start Menu

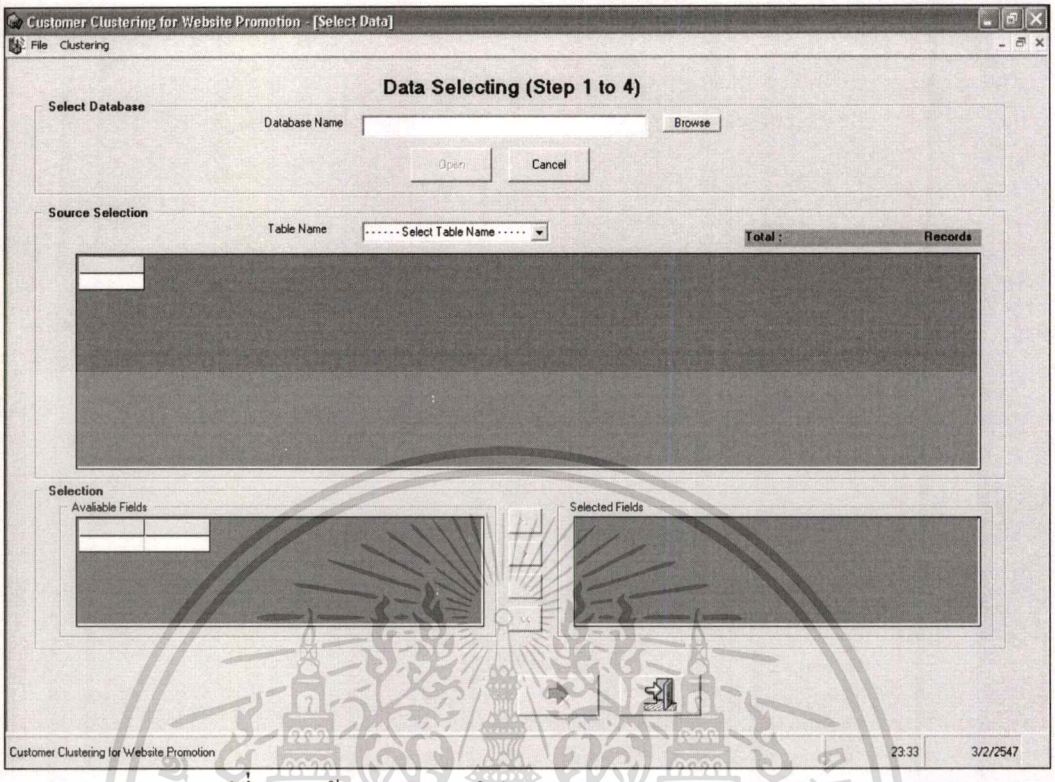
1.1.2 เข้าไปยัง All Programs

1.1.3 จะพบโปรแกรม Hotdiscount Clustering



รูปที่ ก.8 หน้าจอแสดงการเข้าสู่โปรแกรม 'Hotdiscount Clustering'

เมื่อเข้ามาสู่โปรแกรมแล้วจะพบกับหน้าจอหลักของโปรแกรมดังรูป



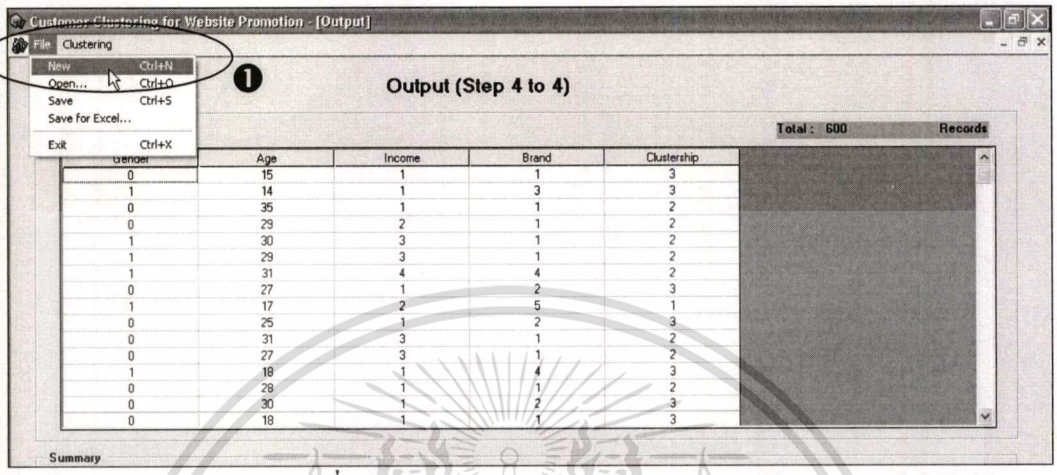
รูปที่ ก.9 หน้าจอหลักของโปรแกรม 'Hotdiscount Clustering'

เมื่อเข้าสู่โปรแกรมจะปรากฏหน้าจอหลักคือหน้าจอการเลือกข้อมูล (Data Selecting) ซึ่งประกอบด้วย 3 ส่วนคือ

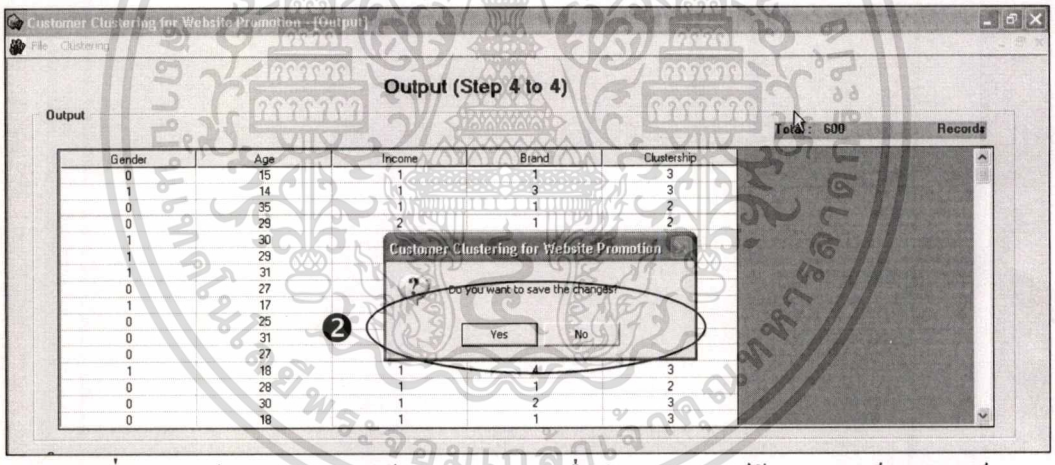
- Select Database การเลือกฐานข้อมูล
- Source Selection การเลือกตารางที่ต้องการวิเคราะห์
- Selection การเลือกฟิลด์ที่ต้องการวิเคราะห์

ก.3 การสร้างโปรเจคใหม่ (New)

เมื่อต้องการสร้างโปรเจคใหม่ให้เลือกคำสั่ง File > New หรือ Ctrl+N ก็ได้



รูปที่ ก.10 หน้าจอแสดงการสร้างโปรเจคใหม่



รูปที่ ก.11 หน้าจอแสดงการสร้างโปรเจคใหม่ เมื่อมีการวิเคราะห์ข้อมูลจนเสร็จสมบูรณ์

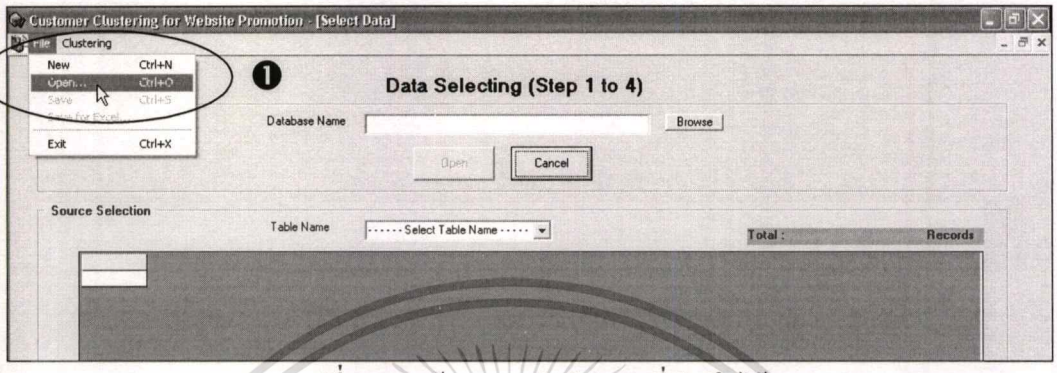
1. เลือกคำสั่ง File > New หรือ Ctrl+N
2. ถ้ากรณีที่ต้องการสร้างเอกสารใหม่ในขณะที่ได้วิเคราะห์ข้อมูลอย่างน้อย 1 ชุดจนเสร็จสมบูรณ์แล้ว หากยังไม่ได้บันทึกข้อมูล โปรแกรมจะถามว่าต้องการ Save ข้อมูลหรือไม่ เลือก

   Yes ถ้าต้องการ Save ข้อมูล

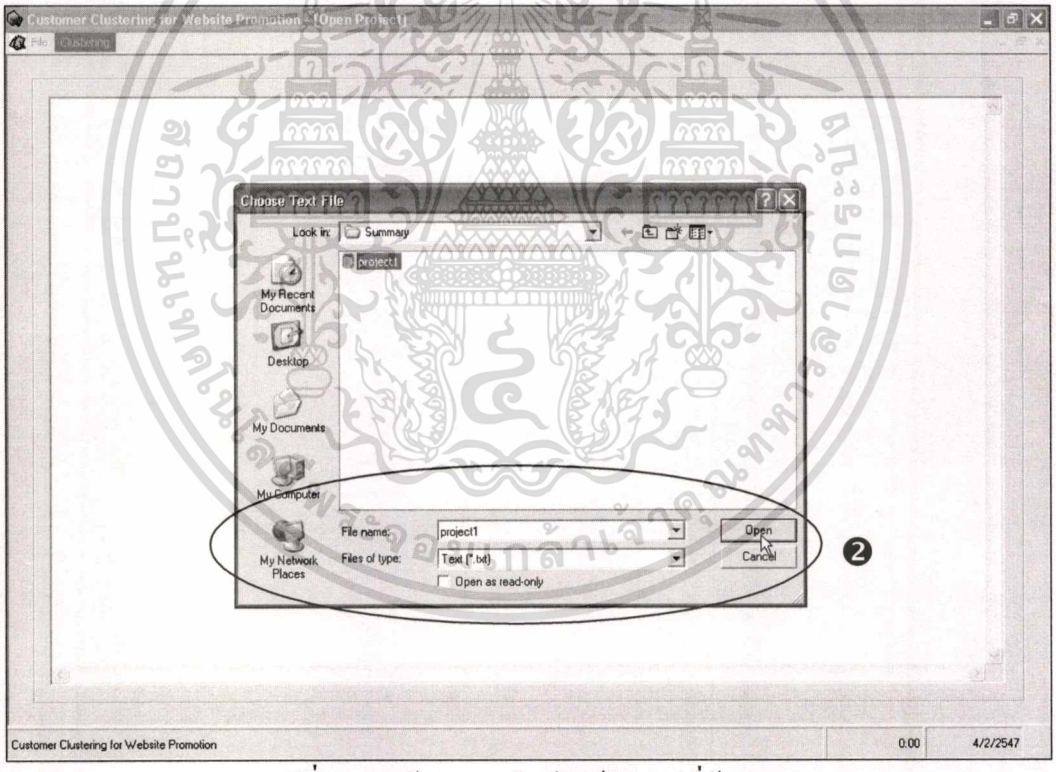
   No ถ้าไม่ต้องการ Save ข้อมูล

ก.4 การเปิดโปรเจคที่บันทึกไว้ (Open)

การเปิดโปรเจคที่ได้บันทึกไว้ สามารถทำได้โดยเลือกคำสั่ง File > Open หรือ Ctrl+O

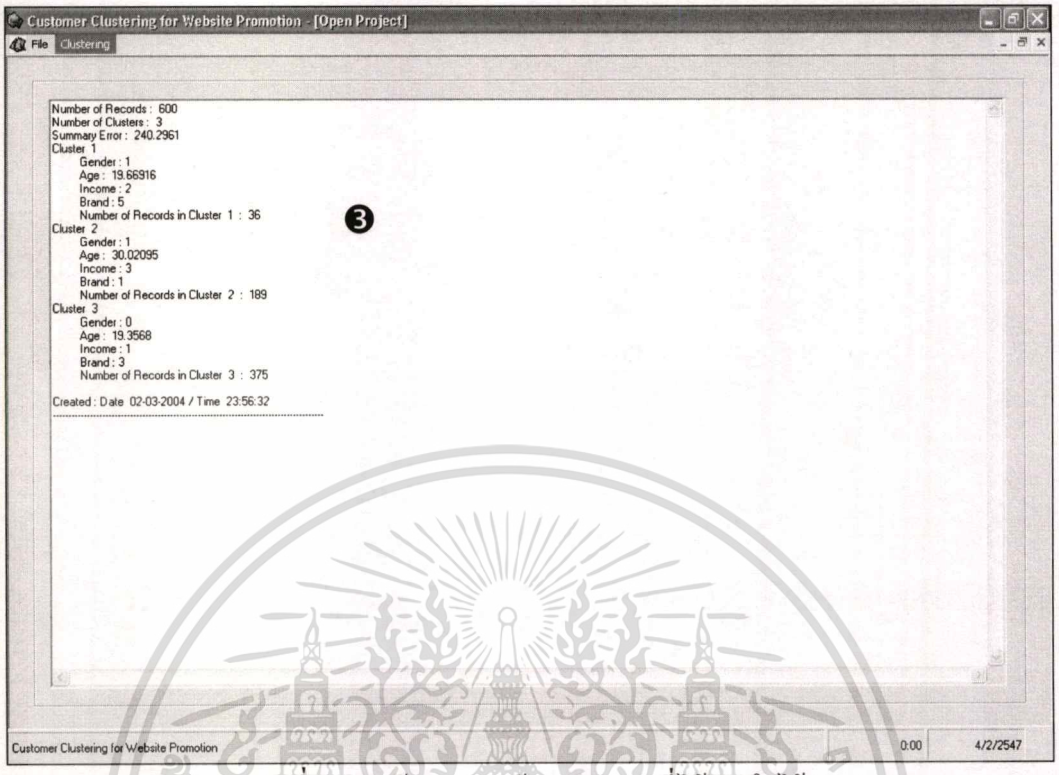


รูปที่ ก.12 หน้าจอการเปิดโปรเจคที่บันทึกไว้



รูปที่ ก.13 หน้าจอการเลือกไฟล์โปรเจคที่ต้องการ

1. เลือกคำสั่ง File > Open หรือ Ctrl+O
2. เลือกไฟล์ฐานข้อมูลนามสกุล *.txt

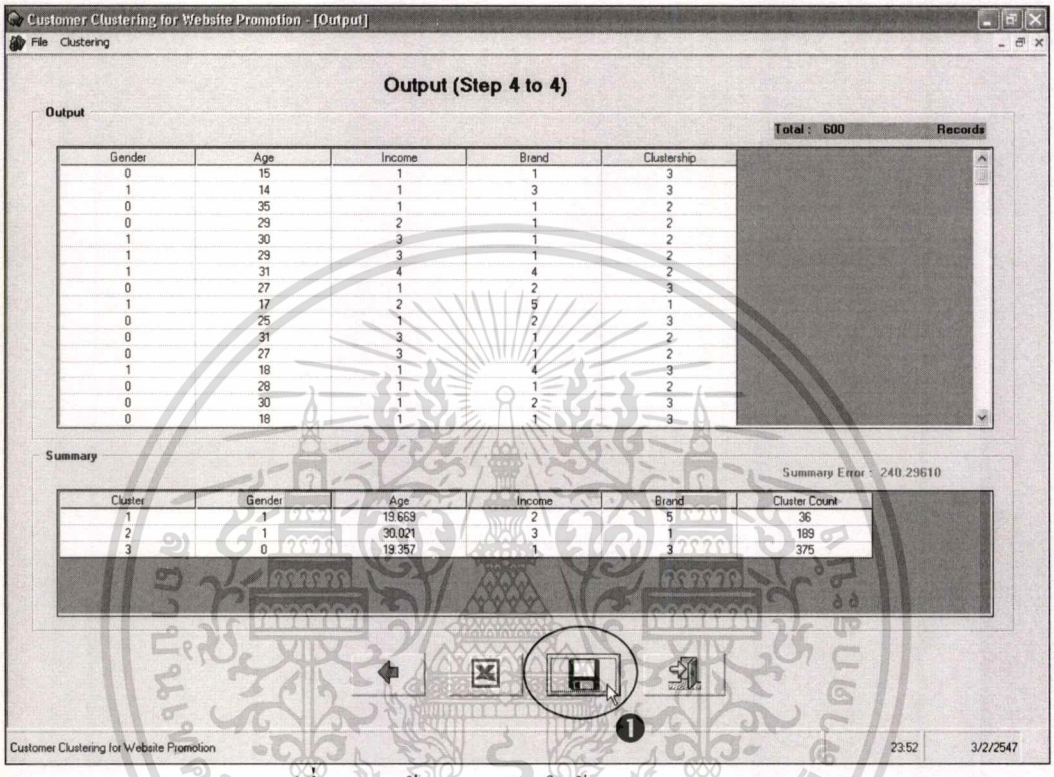


รูปที่ ก.14 หน้าจอแสดงข้อมูลโปรเจคที่ได้บันทึกไว้

3. จะปรากฏข้อมูลสรุปที่ได้บันทึกไว้แสดงที่หน้าจอ

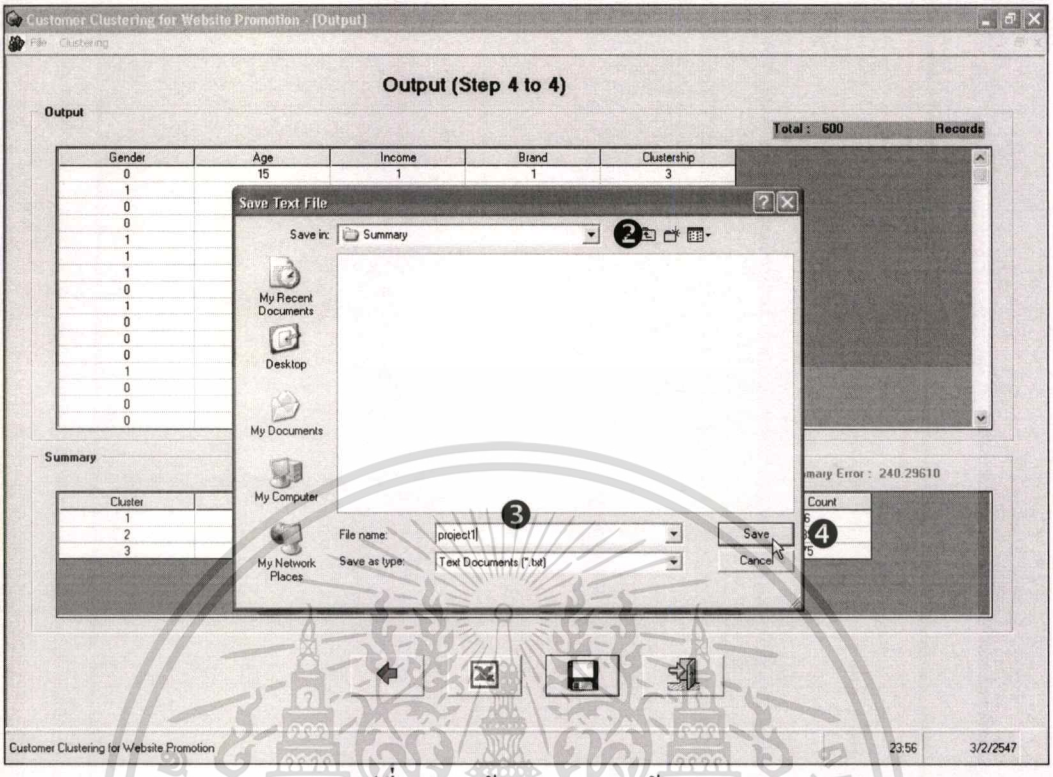
ก.5 การบันทึกโปรเจคแบบ Text File (Save)

เมื่อวิเคราะห์ข้อมูลจนเสร็จสมบูรณ์แล้ว เราสามารถ Save ข้อมูลไว้เพื่อดูผลลัพธ์ที่ได้วิเคราะห์ โดยตั้งชื่อไฟล์ได้ยาวถึง 256 ตัวอักษร และควรเป็นชื่อที่สื่อความหมาย

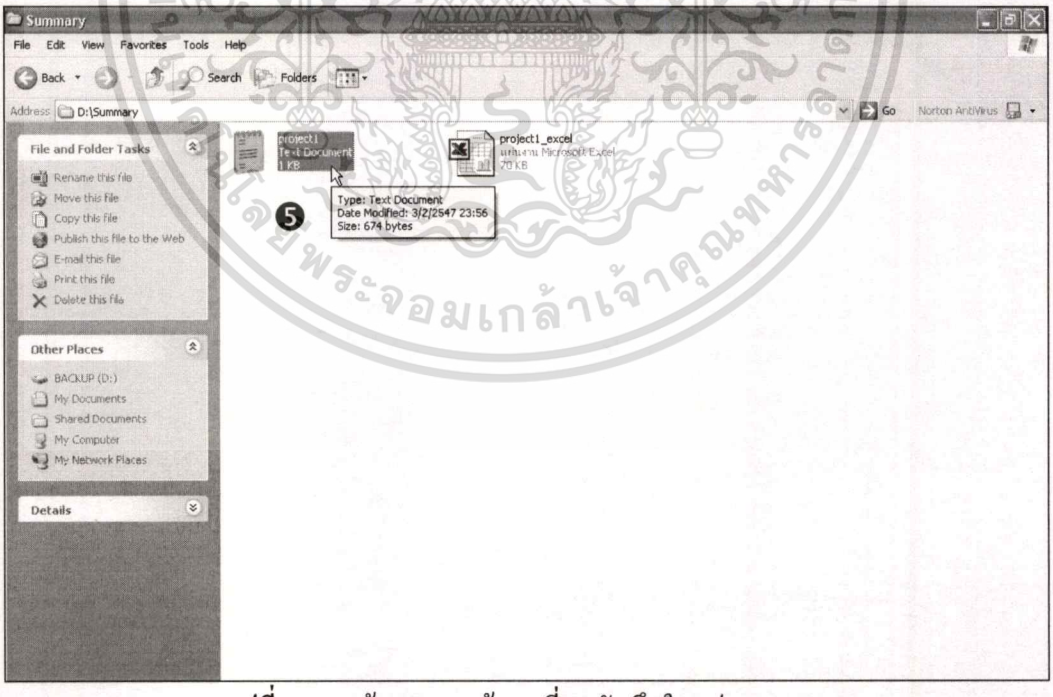


รูปที่ ก.15 หน้าจอการบันทึกข้อมูลแบบ Text File

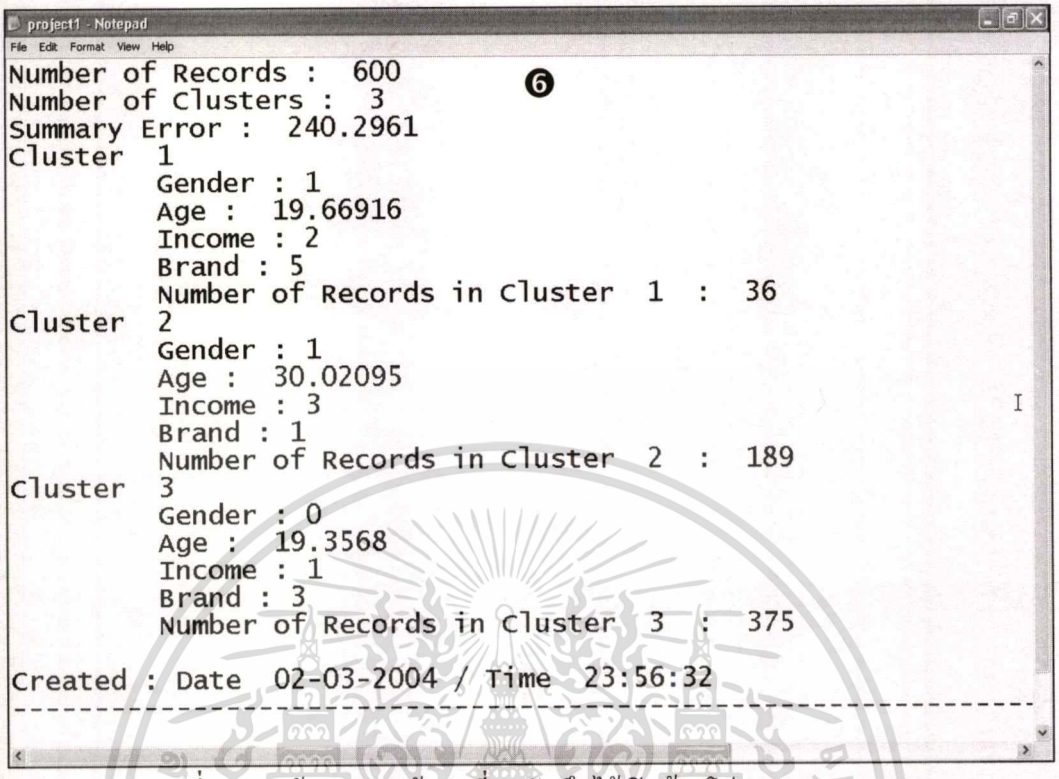
1. เลือกคำสั่ง File > Save หรือ Click mouse ปุ่ม 
2. ระบุตำแหน่งที่ต้องการ Save ข้อมูล
3. ตั้งชื่อไฟล์ที่ต้องการ โดยการ Save นี้จะบันทึกเป็นไฟล์นามสกุล *.txt
4. Click mouse ที่ปุ่ม Save
5. เมื่อ Save เรียบร้อยแล้ว ข้อมูลจะถูกเก็บเป็น Text File ในตำแหน่งที่ระบุไว้



รูปที่ ก.16 หน้าจอการ Save ข้อมูล



รูปที่ ก.17 หน้าจอแสดงข้อมูลที่ถูกบันทึกในรูป Text File

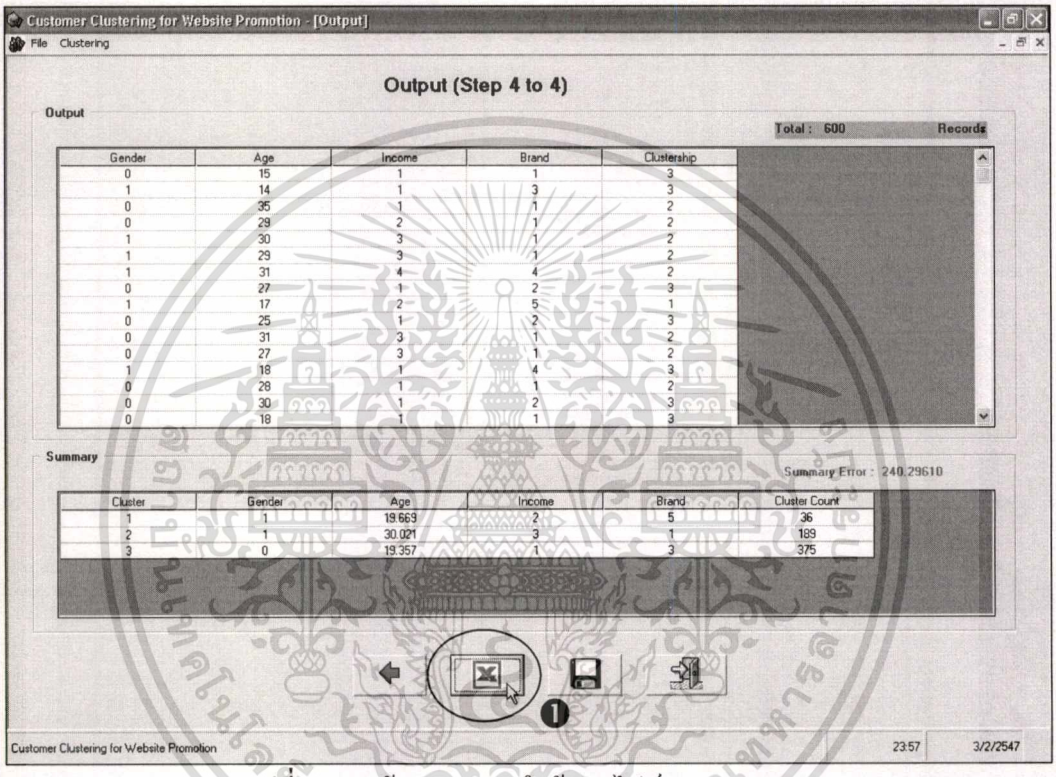


รูปที่ ก.18 หน้าจอแสดงข้อมูลที่ถูกบันทึกไว้เปิดด้วยโปรแกรม Notepad

6. Text File ดังกล่าวสามารถดูข้อมูลที่บันทึกไว้ได้ด้วยโปรแกรม Editor เช่น โปรแกรม Notepad ซึ่งข้อมูลที่บันทึกไว้เป็นข้อมูลสรุป ประกอบด้วยจำนวนเรคอร์ดที่ได้วิเคราะห์ จำนวนกลุ่มที่ได้วิเคราะห์ ค่า Error ที่เกิดขึ้น ข้อมูลจุดศูนย์กลางของแต่ละกลุ่ม และวันที่และเวลาที่วิเคราะห์

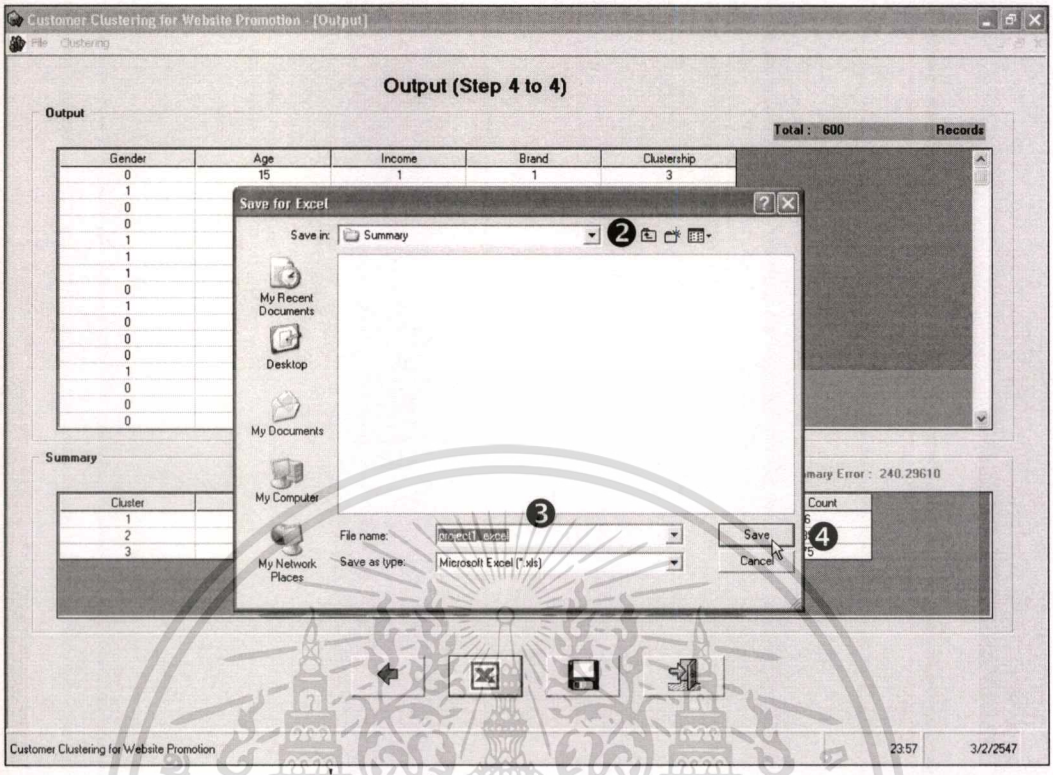
ก.6 การบันทึกโปรเจคแบบ Microsoft Excel (Save for Excel)

เมื่อวิเคราะห์ข้อมูลจนเสร็จสมบูรณ์แล้ว นอกจากจะสามารถ Save ข้อมูลให้เป็น Text File ได้นั้น เรายังสามารถ Save ข้อมูลให้เป็นไฟล์นามสกุล *.xls ไว้เพื่อดูผลลัพธ์ที่ได้วิเคราะห์ และนำข้อมูลดังกล่าวไปใช้กระบวนการอื่นต่อไป โดยตั้งชื่อไฟล์ได้ยาวถึง 256 ตัวอักษร และควรเป็นชื่อที่สื่อความหมาย

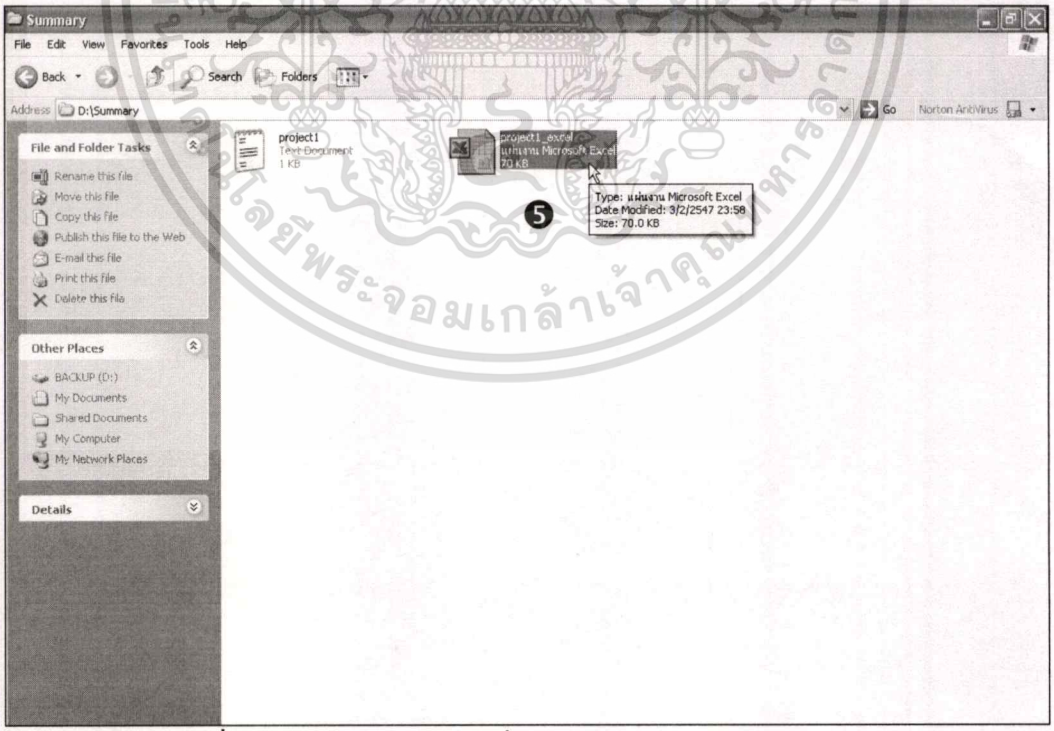


รูปที่ ก.19 หน้าจอการบันทึกข้อมูลไฟล์นามสกุล *.xls

1. เลือกคำสั่ง File > Save for Excel หรือ Click mouse ปุ่ม 
2. ระบุตำแหน่งที่ต้องการ Save ข้อมูล
3. ตั้งชื่อไฟล์ที่ต้องการ โดยการ Save นี้จะบันทึกเป็นไฟล์นามสกุล *.xls
4. Click mouse ที่ปุ่ม Save
5. เมื่อ Save เรียบร้อยแล้ว ข้อมูลจะถูกเก็บเป็นไฟล์นามสกุล *.xls ในตำแหน่งที่ระบุไว้

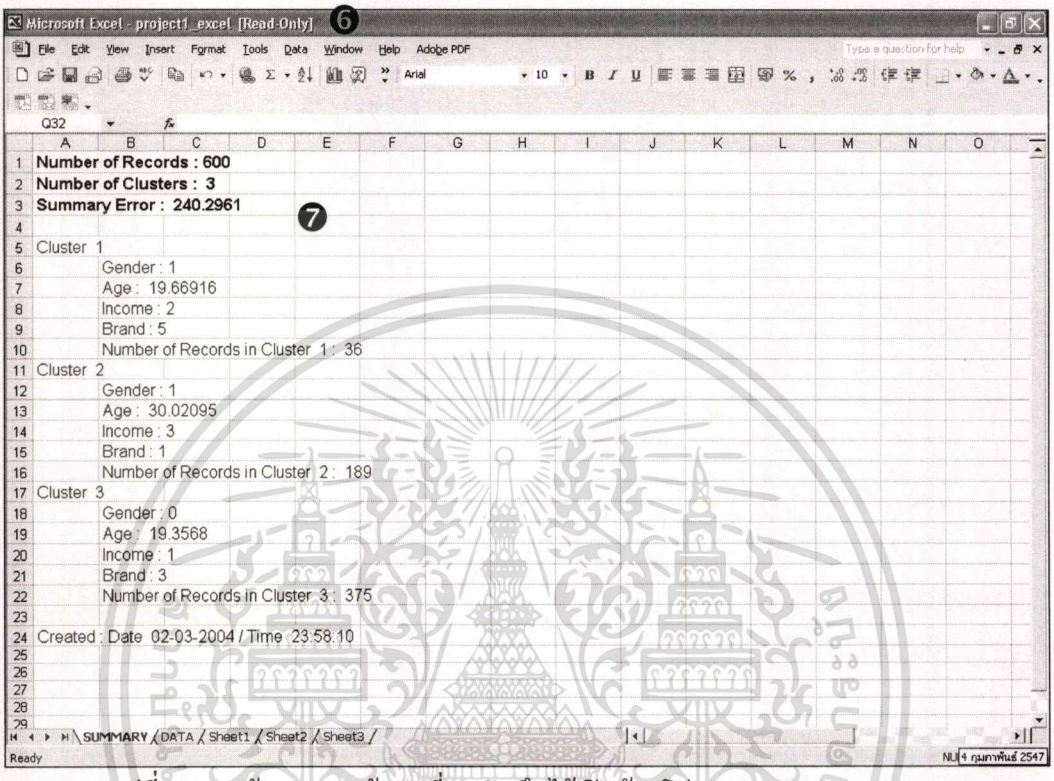


รูปที่ ก.20 หน้าจอการ Save ข้อมูลแบบ Excel



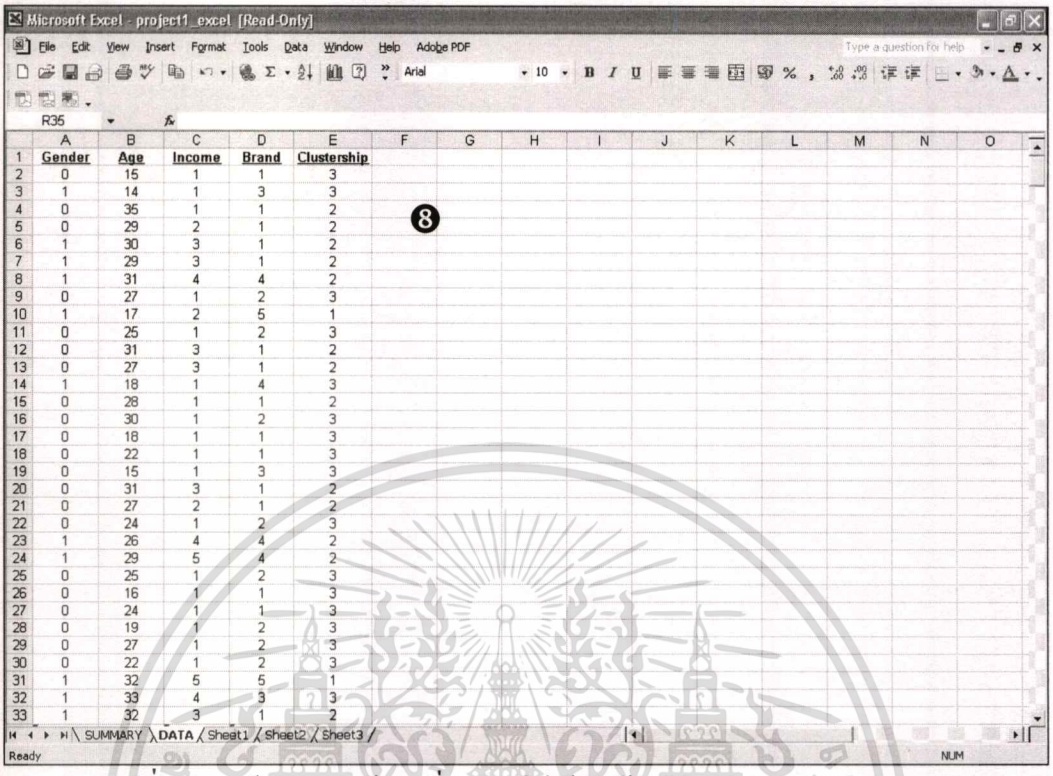
รูปที่ ก.21 หน้าจอแสดงข้อมูลที่ถูกบันทึกในรูปไฟล์นามสกุล *.xls

6. ไฟล์นามสกุล *.xls ดังกล่าวสามารถดูข้อมูลที่บันทึกไว้ได้ด้วยโปรแกรม Microsoft Excel



รูปที่ ก.22 หน้าจอแสดงข้อมูลที่ถูกบันทึกไว้เปิดด้วยโปรแกรม Microsoft Excel

7. เมื่อเปิดไฟล์นามสกุล *.xls ด้วยโปรแกรม Microsoft Excel แล้วข้อมูลที่ถูกบันทึกไว้จะเป็นข้อมูลที่สามารถอ่านได้เพียงอย่างเดียวเท่านั้น ไม่สามารถแก้ไขได้ ข้อมูลจะปรากฏอยู่ใน 2 Sheet คือ Sheet SUMMAREY คือข้อมูลสรุปประกอบด้วย จำนวนเรคอร์ดที่ได้วิเคราะห์ จำนวนกลุ่มที่ได้วิเคราะห์ ค่า Error ที่เกิดขึ้น ข้อมูลจุดศูนย์กลางของแต่ละกลุ่ม และวันที่และเวลาที่วิเคราะห์



Microsoft Excel - project1_excel [Read-Only]

File Edit View Insert Format Tools Data Window Help Adobe PDF

R35

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Gender	Age	Income	Brand	Clustership										
2	0	15	1	1	3										
3	1	14	1	3	3										
4	0	35	1	1	2										
5	0	29	2	1	2										
6	1	30	3	1	2										
7	1	29	3	1	2										
8	1	31	4	4	2										
9	0	27	1	2	3										
10	1	17	2	5	1										
11	0	25	1	2	3										
12	0	31	3	1	2										
13	0	27	3	1	2										
14	1	18	1	4	3										
15	0	28	1	1	2										
16	0	30	1	2	3										
17	0	18	1	1	3										
18	0	22	1	1	3										
19	0	15	1	3	3										
20	0	31	3	1	2										
21	0	27	2	1	2										
22	0	24	1	2	3										
23	1	26	4	4	2										
24	1	29	5	4	2										
25	0	25	1	2	3										
26	0	16	1	1	3										
27	0	24	1	1	3										
28	0	19	1	2	3										
29	0	27	1	2	3										
30	0	22	1	2	3										
31	1	32	5	5	1										
32	1	33	4	3	3										
33	1	32	3	1	2										

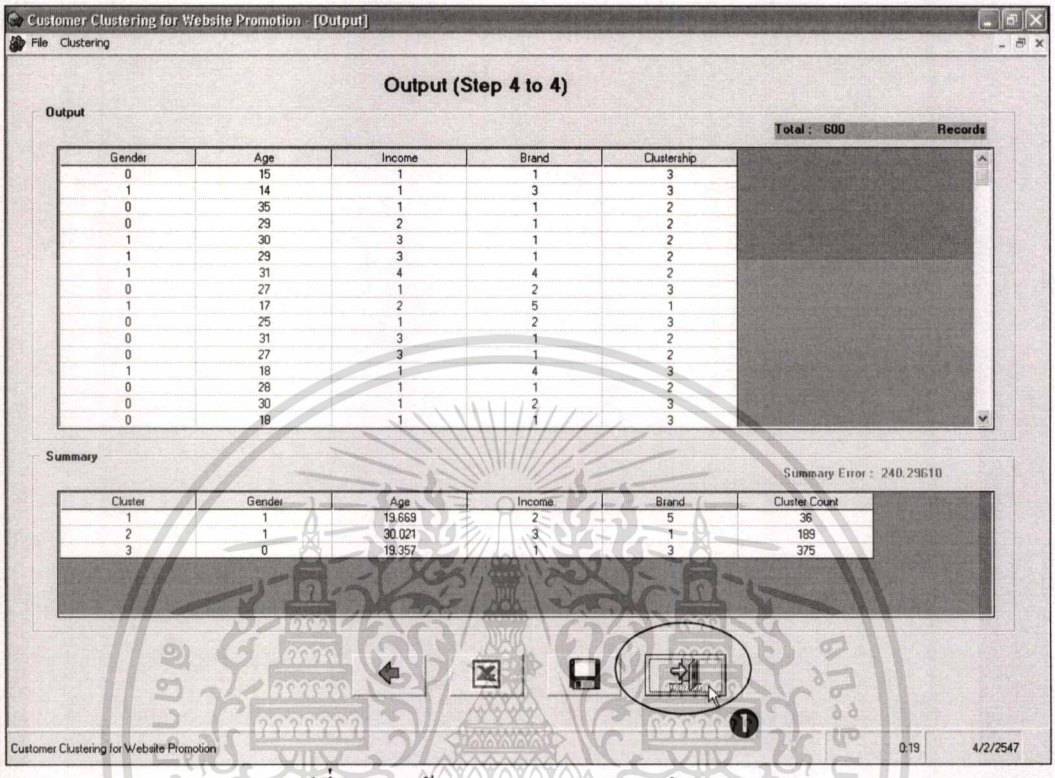
SUMMARY DATA Sheet1 Sheet2 Sheet3

Ready NUM

รูปที่ ก.23 หน้าจอแสดงข้อมูลที่ถูกบันทึกไว้เปิดด้วยโปรแกรม Microsoft Excel

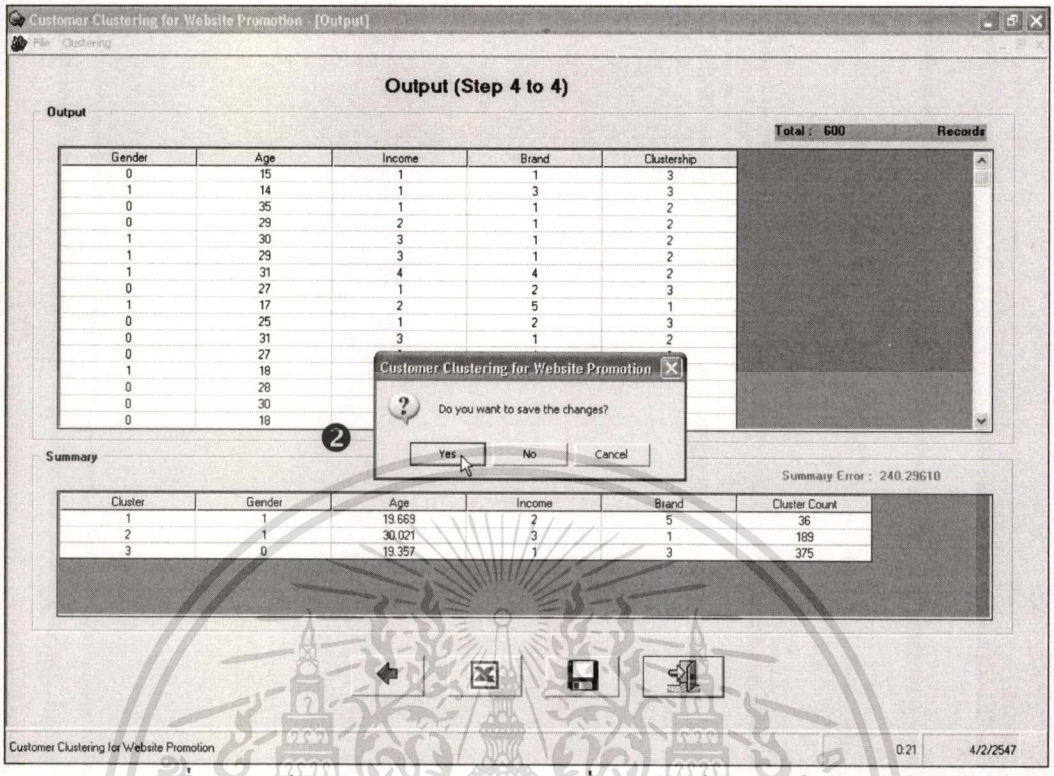
8. Sheet ที่ 2 คือ Sheet DATA ประกอบด้วยฟิลด์ข้อมูลทั้งหมดจากตัวอย่างคือ Gender, Age, Income, Brand และฟิลด์ที่บอกว่าเรคอร์ดนั้นอยู่ที่กลุ่มใด คือ Clustership

ก.7 การออกจากโปรแกรม (Exit)



รูปที่ ก.24 หน้าจอแสดงการออกจากโปรแกรม

1. เลือกคำสั่ง File > Exit หรือ Ctrl+X หรือ Click mouse ที่ปุ่ม 



รูปที่ ก.25 หน้าจอการออกจากโปรแกรมเมื่อวิเคราะห์ข้อมูลเสร็จสมบูรณ์

2. เมื่อต้องการออกจากโปรแกรมในขณะที่ได้วิเคราะห์ข้อมูลอย่างน้อย 1 ชุดจนเสร็จสมบูรณ์แล้ว หากยังไม่ได้บันทึกข้อมูล โปรแกรมจะถามว่าต้องการ Save ข้อมูลหรือไม่ เลือก

Yes ถ้าต้องการ Save ข้อมูล

No ถ้าไม่ต้องการ Save ข้อมูล

Cancel ถ้าต้องการยกเลิกคำสั่ง

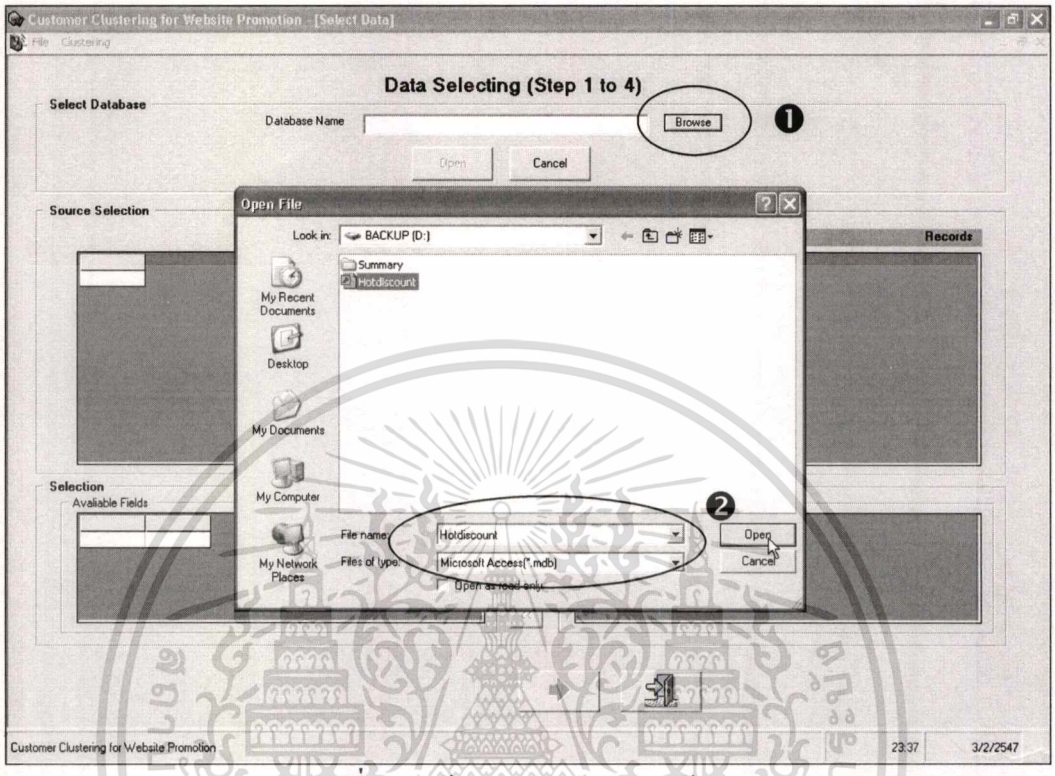
ข. การทำงานของโปรแกรม

ข.1 การทำงานขั้นที่ 1 การเลือกข้อมูล (Data Selecting)

รูปที่ ข.1 หน้าจอแสดงขั้นตอนที่ 1 การเลือกข้อมูล

1. เมื่อเข้าสู่โปรแกรม การวิเคราะห์ข้อมูลนั้นจะต้องผ่านกระบวนการ 4 ขั้นตอนคือ
   - Step 1 - Data Selecting
   - Step 2 - Data Preparing
   - Step 3 – Data Mining
   - Step 4 –Output
2. ใน Step 1 หน้าจอหลักคือ หน้าจอการเลือกข้อมูล (Data Selecting) จะประกอบด้วย 3 ส่วนคือ
   - Select Database เลือกฐานข้อมูล
   - Source Selection เลือกตารางที่ต้องการวิเคราะห์
   - Selection เลือกฟิลด์ที่ต้องการวิเคราะห์

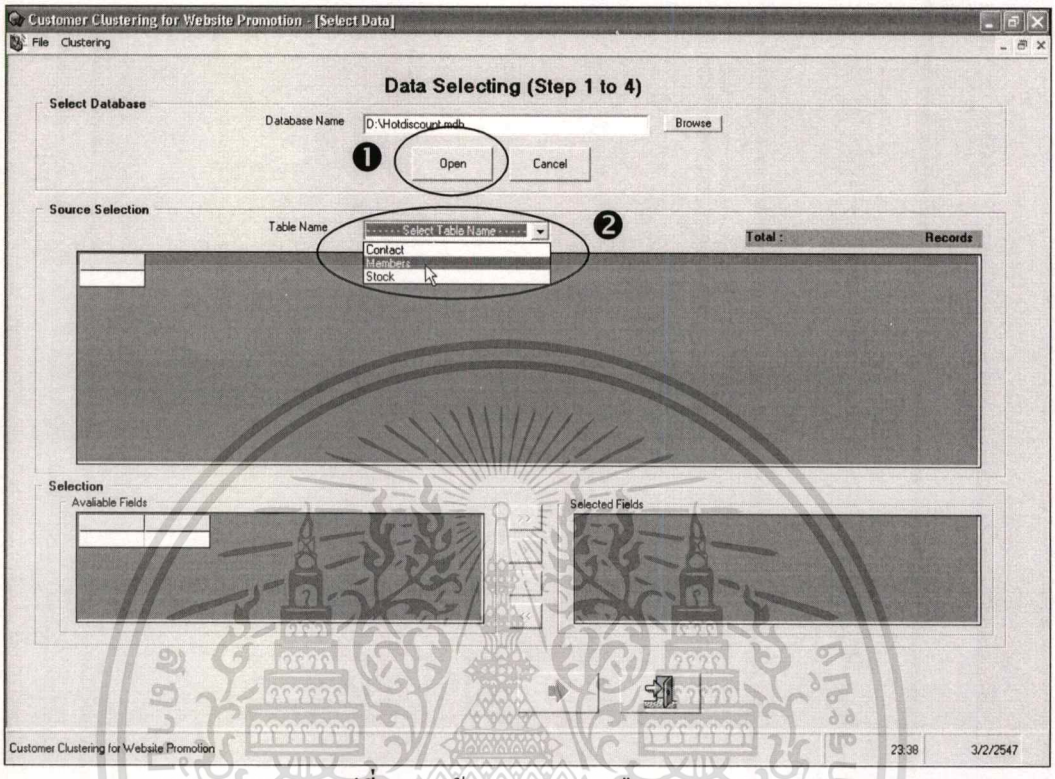
ข.1.1 การติดต่อกับฐานข้อมูล



รูปที่ ข.2 หน้าจอการติดต่อกับฐานข้อมูล

1. Click mouse ที่ปุ่ม Browse ในส่วน Select Database เพื่อเลือกไฟล์ฐานข้อมูลนามสกุล *.mdb
2. Click mouse ที่ปุ่ม Open เพื่อแสดงข้อมูลในฐานข้อมูลที่เลือก

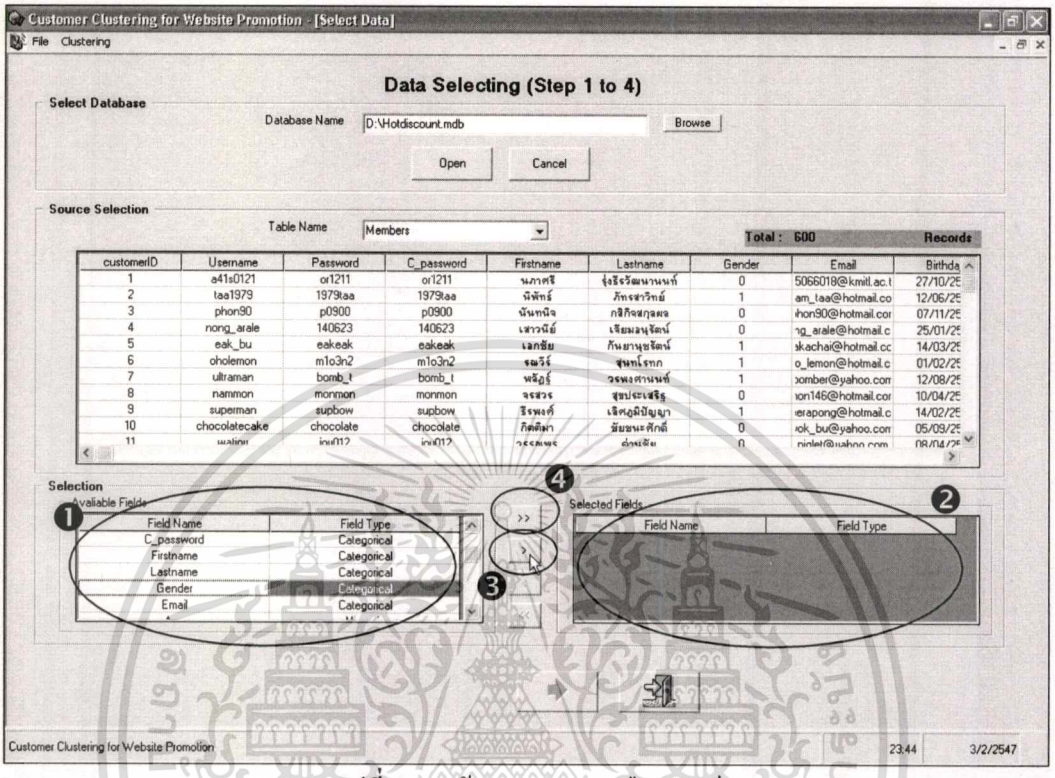
ข.1.2 การเลือกตารางที่ต้องการวิเคราะห์



รูปที่ ข.3 หน้าจอแสดงการเลือกตาราง

1. หลักจาก Click mouse ที่ปุ่ม Open แล้วที่ Table Name จะแสดงชื่อตารางทั้งหมดที่อยู่ในฐานข้อมูลนั้น
2. จากนั้นเลือกชื่อตารางที่ต้องการ ในที่นี้ขอยกตัวอย่างขั้นตอนการทำงาน เพื่อให้ตรงตามวัตถุประสงค์ที่ตั้งไว้ ในกรณีนี้จึงเลือกตารางชื่อ 'Members' เพื่อมาทำดาต้าไมนิ่ง แต่ถ้าวัตถุประสงค์มีการเปลี่ยนแปลง ก็สามารถเลือกตารางอื่นเพื่อมาวิเคราะห์ได้ตามความเหมาะสม

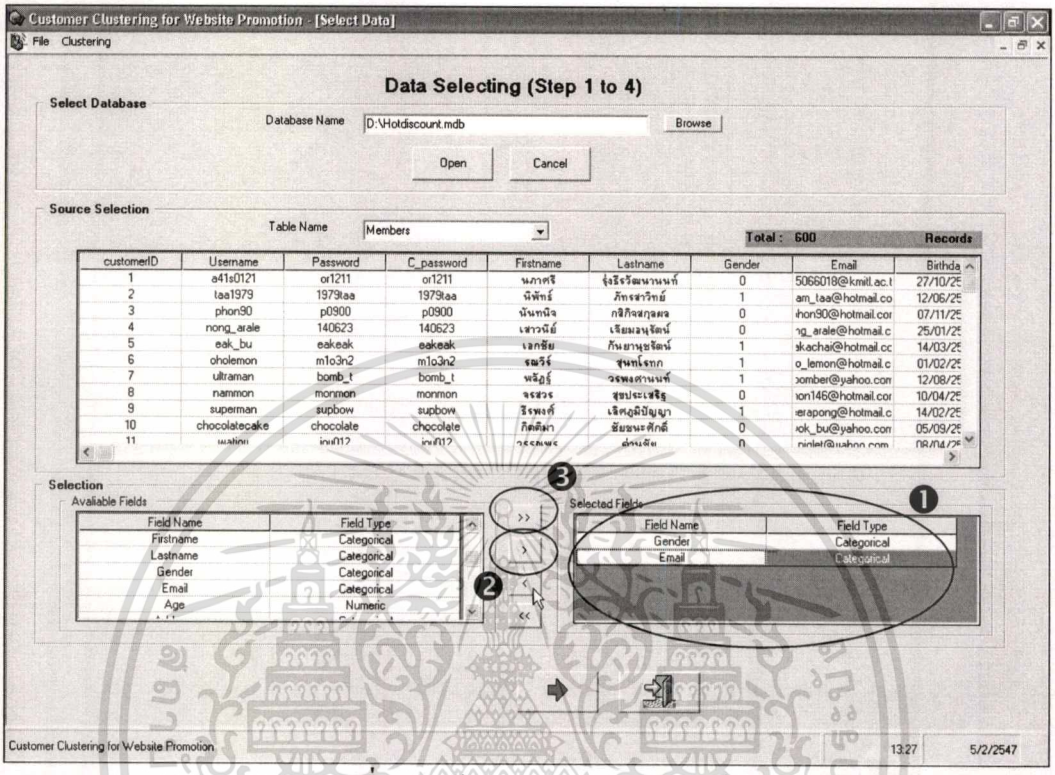
ข.1.3 การเลือกฟิลด์ที่ต้องการวิเคราะห์



รูปที่ ข.4 หน้าจอแสดงการเลือกฟิลด์

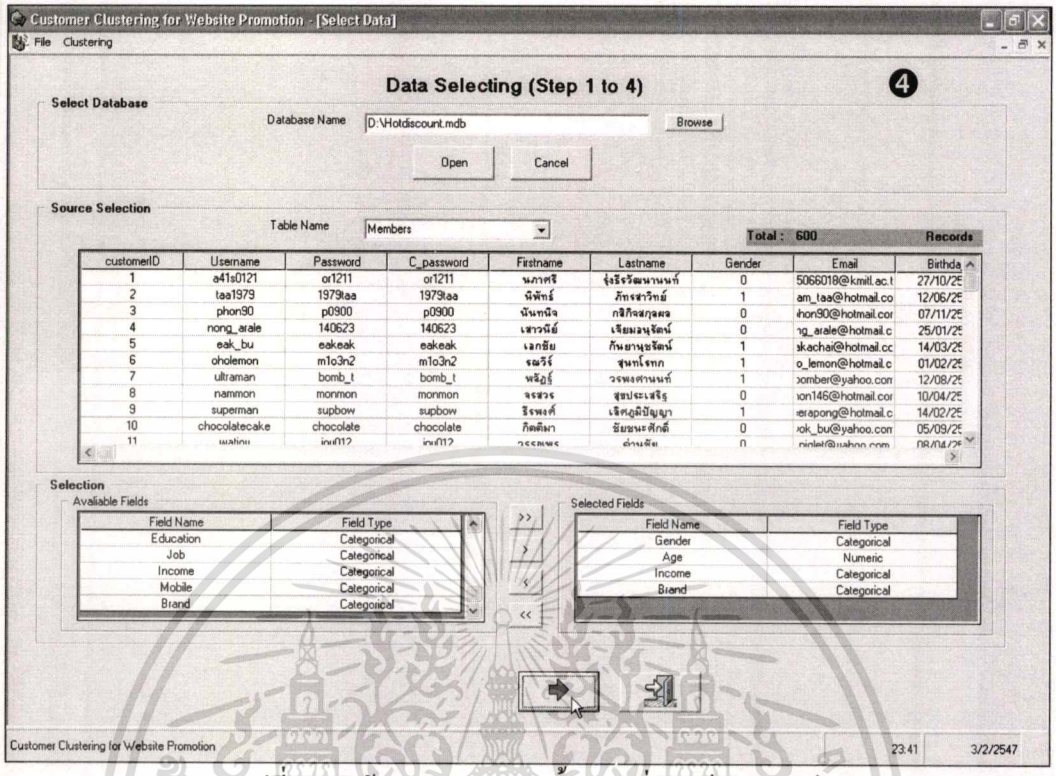
1. เมื่อเลือกตารางเรียบร้อยแล้ว ในส่วน Source Selection จะแสดงข้อมูลทุกเรคอร์ดที่อยู่ในตารางนั้น และจำนวนเรคอร์ด
2. และในส่วน Selection นั้นแบ่งออกเป็น 2 ส่วนย่อยคือ
   - Available Fields แสดงฟิลด์และประเภทของฟิลด์นั้นทั้งหมด
   - Selected Fields เป็นส่วนที่จะแสดงฟิลด์ที่ต้องการนำไปวิเคราะห์
3. วิธีการเลือกฟิลด์ที่ต้องการนำไปวิเคราะห์สามารถทำได้โดย Click mouse ที่ฟิลด์ที่ต้องการ แล้ว Click mouse ที่ปุ่ม > ฟิลด์ที่เลือกจะไปปรากฏที่ Selected Field
4. ถ้าต้องการเลือกฟิลด์ทั้งหมดสามารถทำได้โดย Click mouse ที่ปุ่ม >>

ข.1.4 การยกเลิกฟิลด์



รูปที่ ข.5 หน้าจอแสดงการยกเลิกฟิลด์

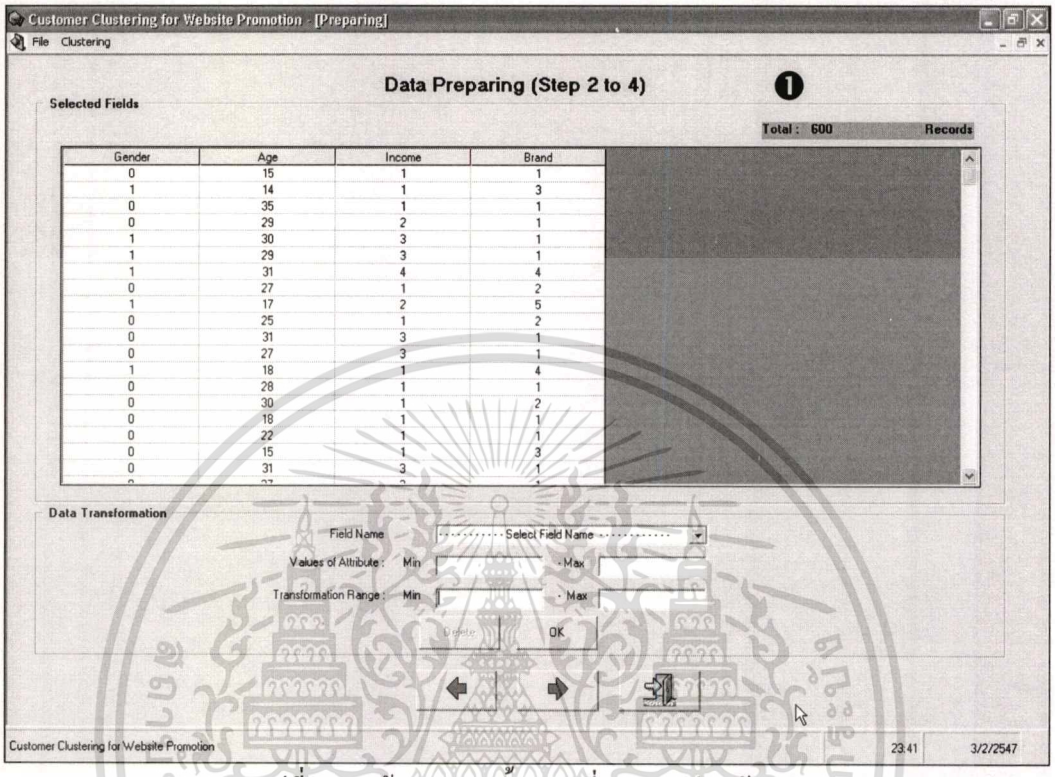
1. วิธีการยกเลิกเลือกฟิลด์เลือกไปแล้วสามารถทำได้โดย Click mouse ที่ฟิลด์ที่ต้องการใน Selected Fields
2. Click mouse ที่ปุ่ม < ฟิลด์ที่เลือกจะถูกยกเลิก
3. ถ้าต้องการกเลิกฟิลด์ทั้งหมดสามารถทำได้โดย Click mouse ที่ปุ่ม <<



รูปที่ ข.6 หน้าจอแสดงการทำขั้นตอนที่ 1 เสร็จสมบูรณ์

4. หลังจากที่ทำ Step 1 เสร็จเรียบร้อยแล้ว Click mouse ที่ปุ่ม  เพื่อไปทำงานที่ Step 2

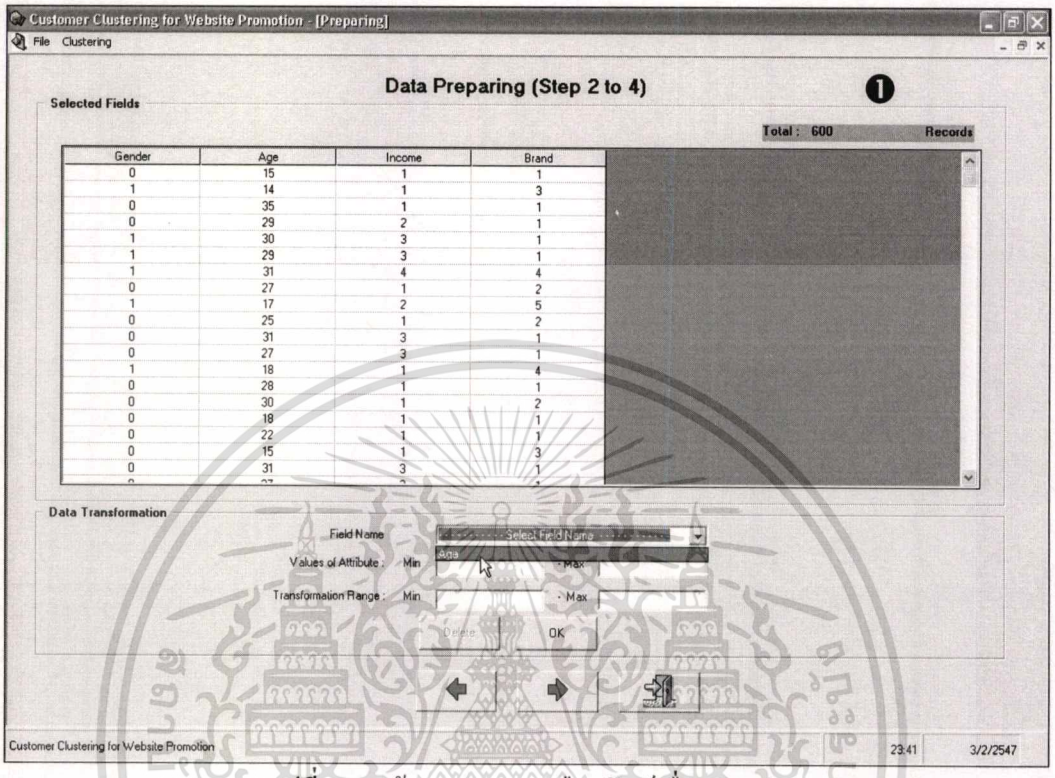
ข.2 การทำงานขั้นที่ 2 การเตรียมข้อมูล (Data Preparing)



รูปที่ ข.7 หน้าจอแสดงขั้นตอนที่ 2 การเตรียมข้อมูล

1. Step 2 คือการเตรียมข้อมูล (Data Preparing) เป็นการแปลงข้อมูลให้อยู่ในขอบเขตที่ต้องการ ด้วยวิธี Min-Max normalization ข้อมูลที่จะได้แปลงได้นั้นจะต้องเป็นข้อมูลประเภท Numeric เท่านั้น เช่น การแปลงข้อมูลอายุให้อยู่ในขอบเขตที่ต้องการ โดยกำหนดค่าต่ำสุดและค่าสูงสุด ในขั้นตอนนี้แบ่งออกเป็น 2 ส่วนย่อยคือ
    - Selected Fields จะแสดงฟิลด์ที่เลือกจาก Step 1 และข้อมูลทุกเรคอร์ดของฟิลด์นั้นๆ
    - Data Transformation เป็นส่วนที่เลือกฟิลด์ที่ต้องการแปลงค่า และกำหนดค่าต่ำสุดและสูงสุด

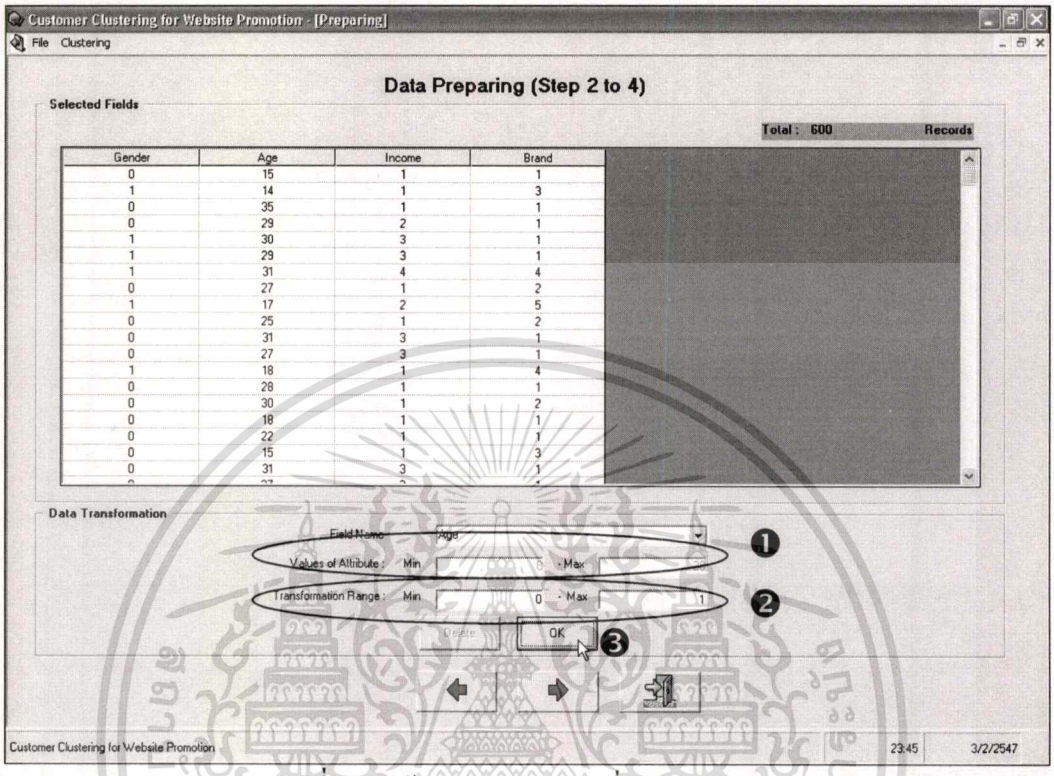
ข.2.1 การเลือกฟิลด์ที่ต้องการแปลงค่า



รูปที่ ข.8 หน้าจอแสดงการเลือกฟิลด์เพื่อแปลงค่า

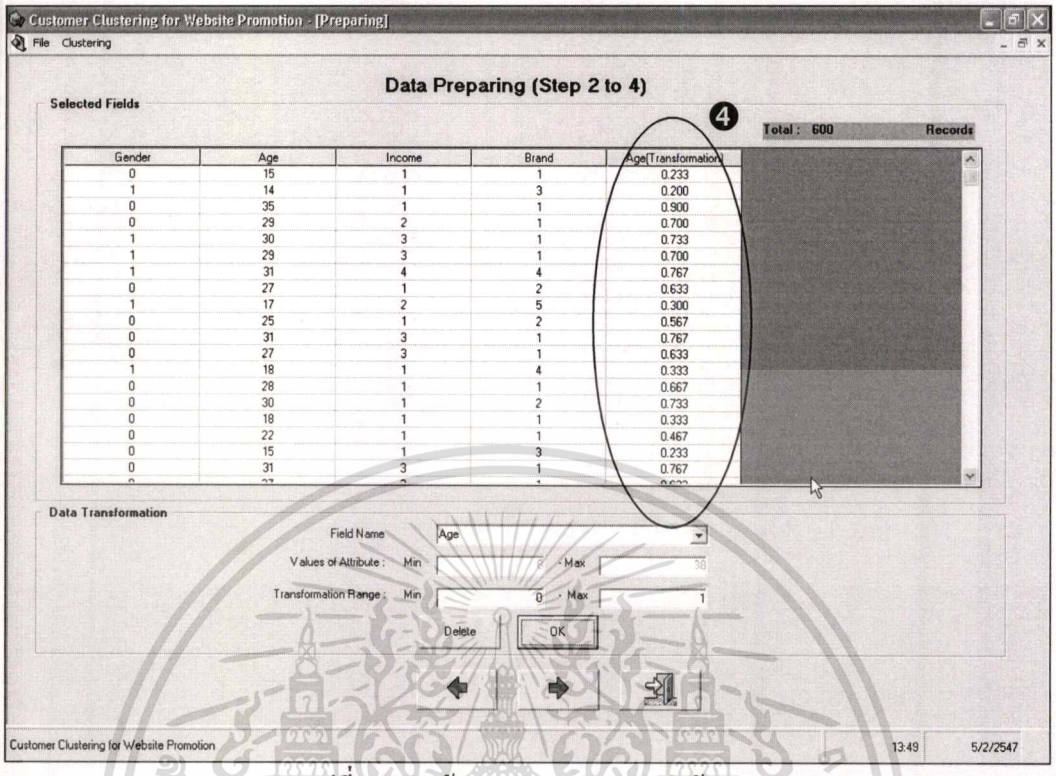
1. เลือกฟิลด์ใน Fields Name ซึ่งจะแสดงเฉพาะฟิลด์ที่เป็นข้อมูลประเภท Numeric เท่านั้น

ข.2.2 การกำหนดค่าต่ำสุดและสูงสุดในการแปลงค่า



รูปที่ ข.9 หน้าจอการกำหนดค่าต่ำสุดและสูงสุด

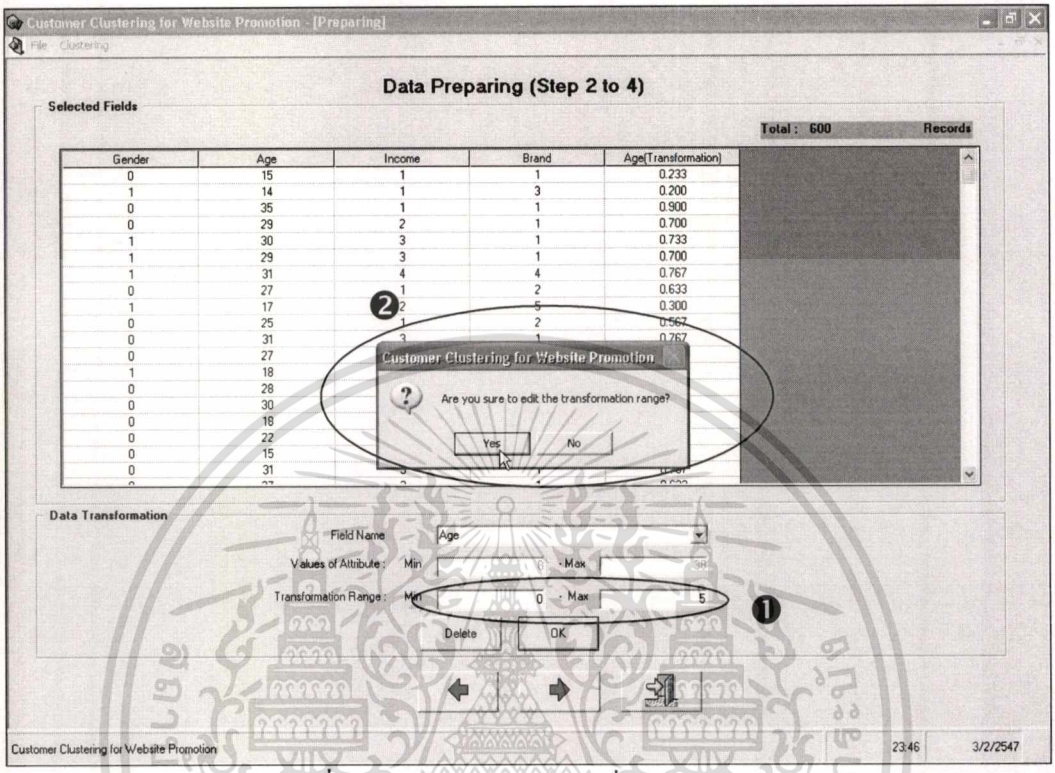
1. หลังจากเลือกฟิลด์ที่ต้องการแปลงค่าแล้ว ใน Values of Attribute จะแสดงค่าต่ำสุดและค่าสูงสุดของข้อมูลฟิลด์นั้นเช่น เลือก Field 'Age' มีข้อมูลต่ำสุดมีค่า 8 และข้อมูลสูงสุดมีค่า 38 เป็นต้น
2. จากนั้นระบุค่าต่ำสุดและสูงสุดที่ต้องการให้ข้อมูลถูกแปลงค่าในช่อง Transformation Range
3. Click mouse ที่ปุ่ม OK เพื่อทำการแปลงค่าข้อมูล



รูปที่ ข.10 หน้าจอแสดงการแปลงค่าข้อมูล

4. หลังจากแปลงค่าข้อมูลเรียบร้อยแล้ว จะแสดงข้อมูลที่ถูกแปลงค่าใน Selected Fields

ข.2.3 การแก้ไขค่าต่ำสุดและสูงสุด



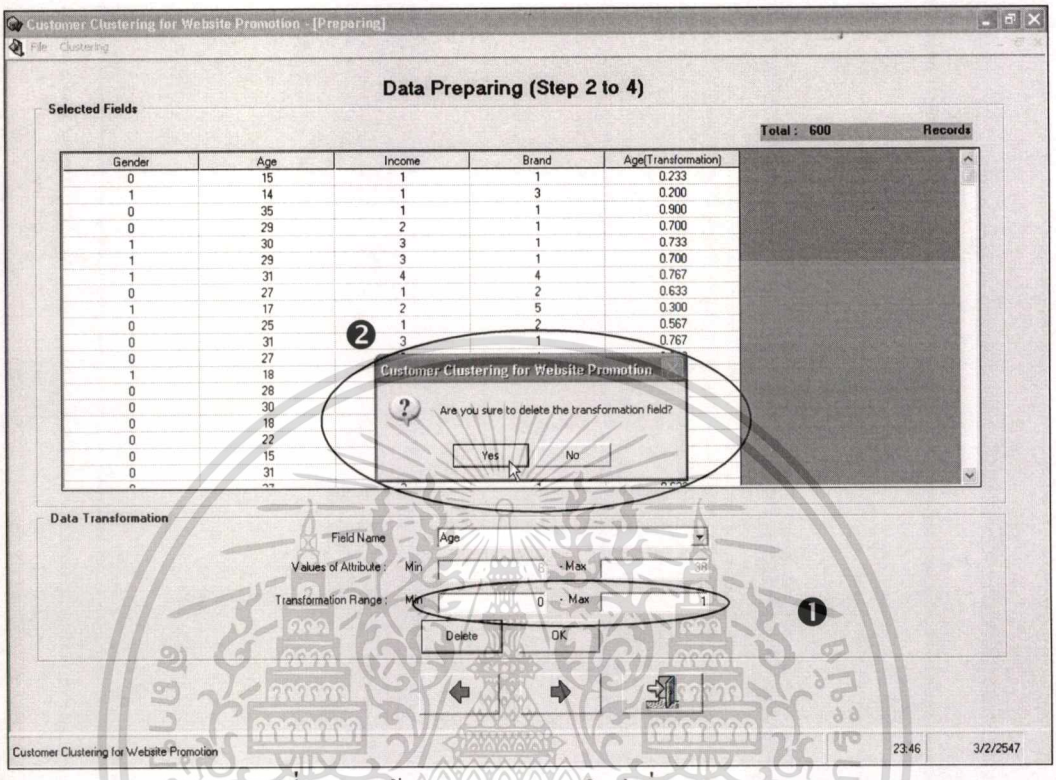
รูปที่ ข.11 หน้าจอการแก้ไขค่าต่ำสุดและสูงสุด

1. การแก้ไขค่าต่ำสุดสูงหลังจากที่ได้แปลงค่าข้อมูลไปแล้ว สามารถทำได้โดย เลือกฟิลด์ที่ต้องการแก้ไขใน Filed Name จะทำให้ค่า Values of Attribute ของข้อมูลและ Transformation Range จะแสดงค่าต่ำสุดสูงสุดที่ถูกกำหนดไว้แสดงออกมา
2. สามารถแก้ไขค่าต่ำสุดและค่าสูงสุดได้ โดยจะถามว่าต้องการแก้ไขหรือไม่ เลือก

   Yes ถ้าต้องการแก้ไขค่าต่ำสุดและค่าสูงสุด

   No ถ้าไม่ต้องการแก้ไขค่าต่ำสุดและค่าสูงสุด

ข.2.4 การยกเลิกค่าต่ำสุดและสูงสุด

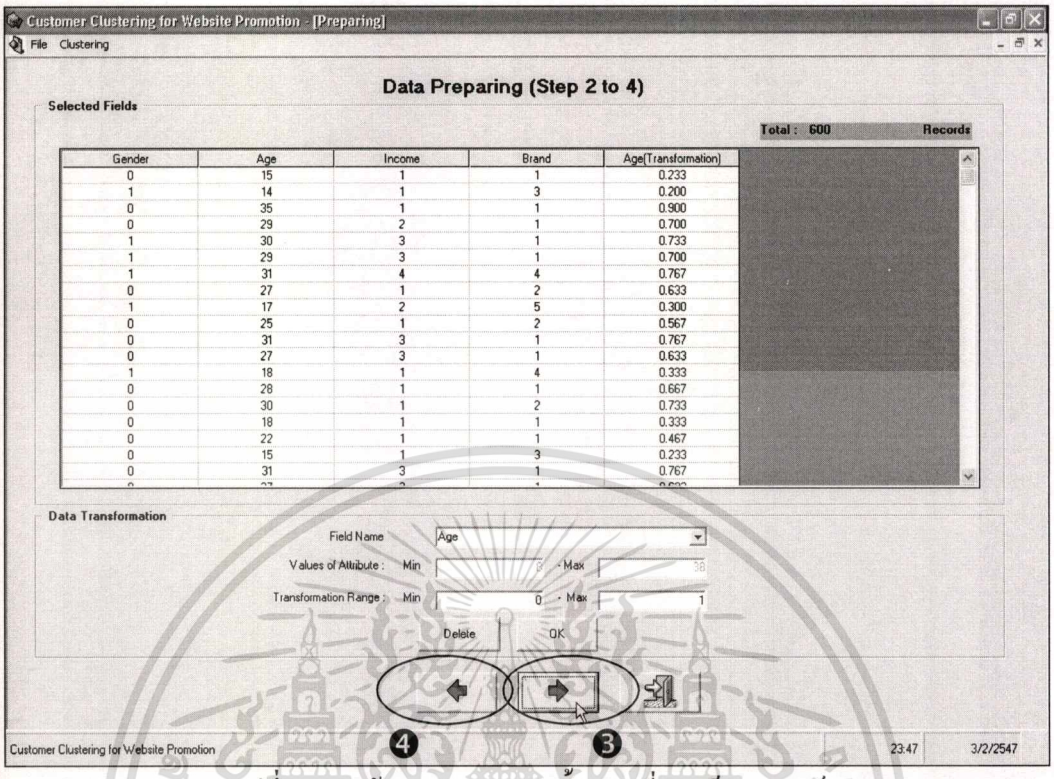


รูปที่ ข.12 หน้าจอแสดงการยกเลิกค่าต่ำสุดและสูงสุด

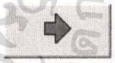
1. การยกเลิกค่าต่ำสุดสูงหลังจากที่ได้แปลงค่าข้อมูลไปแล้ว สามารถทำได้โดย เลือกฟิลด์ที่ต้องการแก้ไขใน Filed Name จะทำให้ค่า Values of Attribute ของข้อมูลและ Transformation Range จะแสดงค่าต่ำสุดสูงสุดที่ถูกกำหนดไว้แสดงออกมา
2. สามารถยกเลิกค่าต่ำสุดและค่าสูงสุดได้ โดยจะถามว่าต้องการยกเลิกหรือไม่ เลือก

   Yes ถ้าต้องการยกเลิกค่าต่ำสุดและค่าสูงสุด

   No ถ้าไม่ต้องการยกเลิกค่าต่ำสุดและค่าสูงสุด

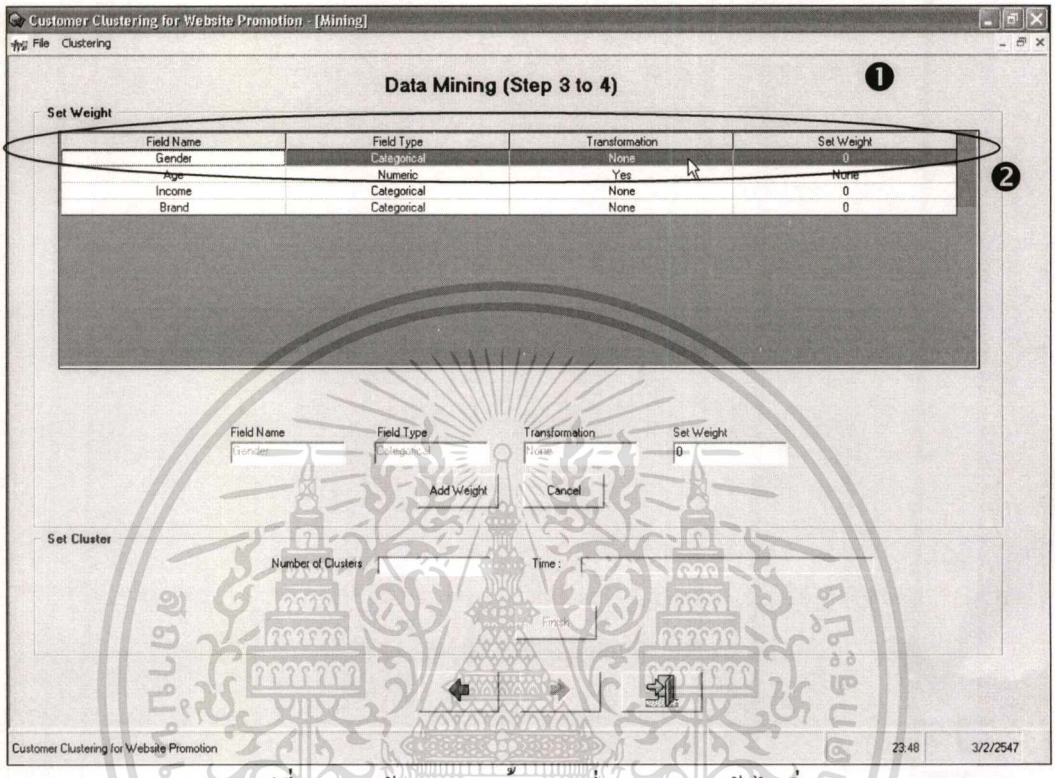


รูปที่ ข.13 หน้าจอแสดงการทำขั้นตอนที่ 2 เสร็จสมบูรณ์

3. หลังจากที่ทำ Step 2 เสร็จเรียบร้อยแล้ว Click mouse ที่ปุ่ม  เพื่อไปทำงานที่ Step 3
4. หากต้องการย้อนกลับไปแก้ไขข้อมูลใน Step 1 สามารถทำได้โดย Click mouse ที่ปุ่ม 

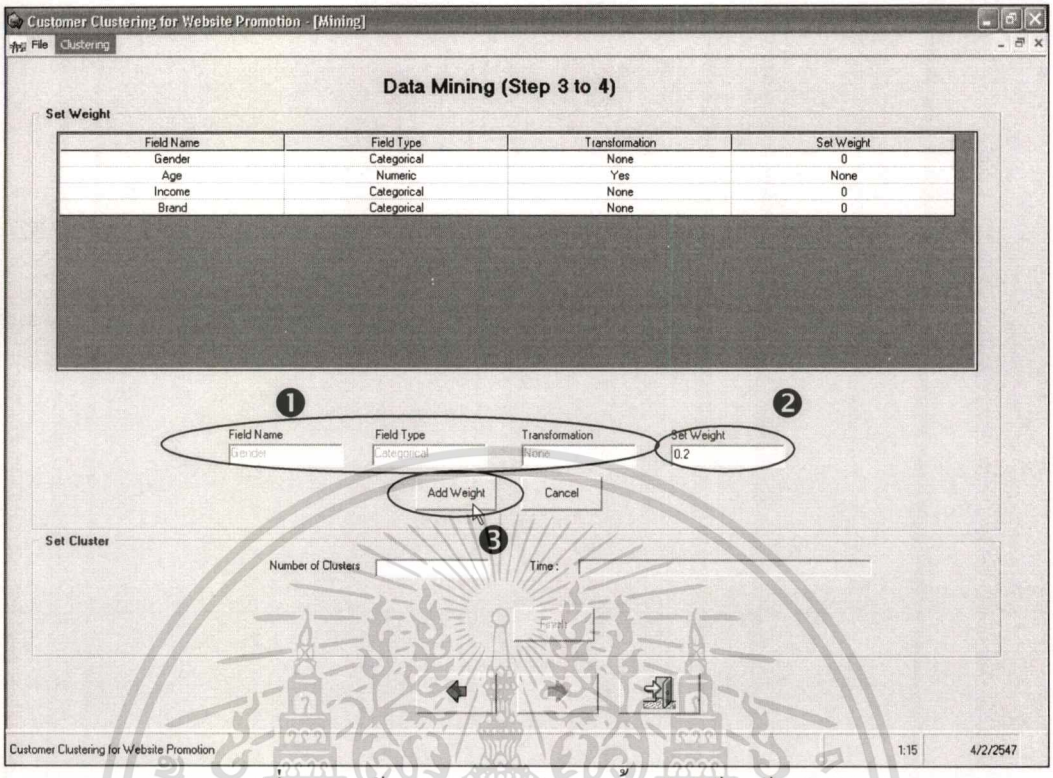
ข.3 การทำงานขั้นที่ 3 การทำไมนิ่งด้วยอัลกอริทึม k-prototypes (Data Mining)

ข.3.1 การกำหนดน้ำหนักให้กับข้อมูลประเภท Categorical

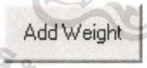


รูปที่ ข.14 หน้าจอแสดงขั้นตอนที่ 3 การทำดาต้าไมนิ่ง

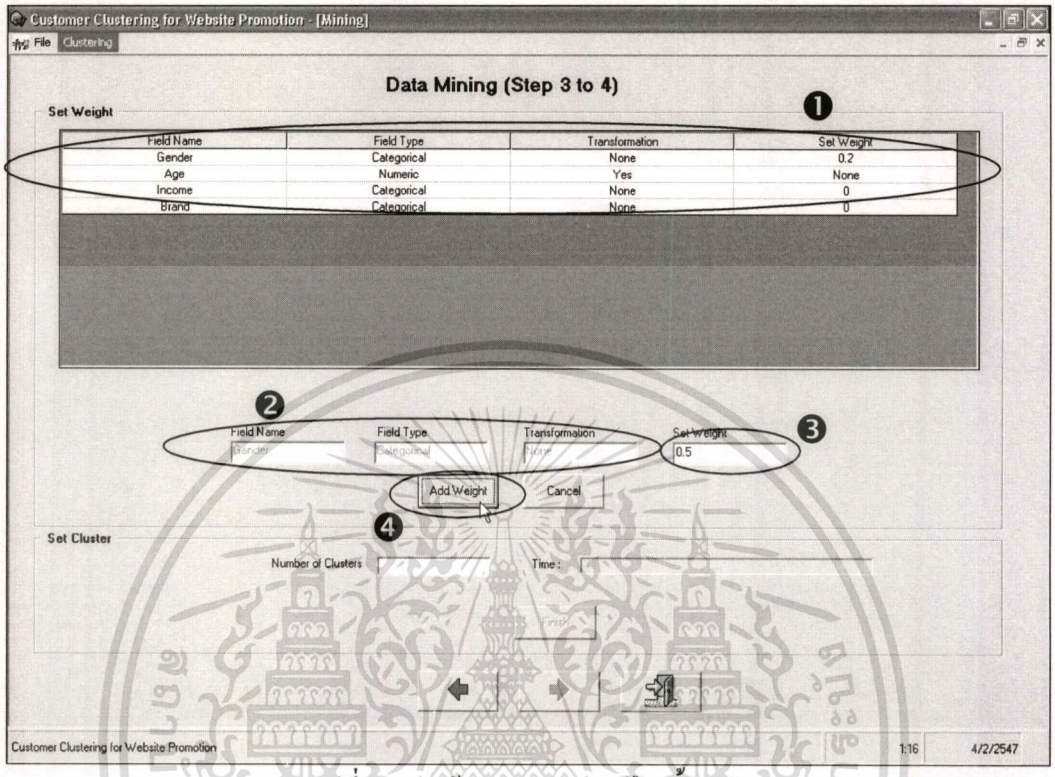
1. Step 3 คือการทำดาต้าไมนิ่ง (Data Mining) ก่อนที่จะกำหนดจำนวนกลุ่มที่ต้องการแบ่งข้อมูลนั้น จากอัลกอริทึมข้อมูลประเภท Categorical นั้นสามารถกำหนดน้ำหนักให้กับข้อมูลได้ เช่น การกำหนดน้ำหนักให้กับฟิลด์เพศ (Gender) ให้เท่ากับ 0.5 เป็นต้น ในขั้นตอนที่ 3 นี้จะแบ่งออกเป็น 2 ส่วนคือ
   - Set Weight จะกำหนดน้ำหนักให้กับข้อมูลประเภท Categorical
   - Set Clusters กำหนดจำนวนกลุ่มที่ต้องการวิเคราะห์
2. วิธีการกำหนดน้ำหนักสามารถทำได้โดย Click mouse ที่ข้อมูลที่ต้องการใน Set Weight



รูปที่ ข.15 หน้าจอแสดงการกำหนดน้ำหนักให้กับข้อมูล

3. ข้อมูลที่ถูก Click mouse จะปรากฏอยู่ใน Field Name, Field Type และ Transformation สำหรับในช่อง Set Weight ค่าน้ำหนักจะถูกกำหนดค่า Default ให้เท่ากับ 0
4. ระบุน้ำหนักที่ต้องการให้กับฟิลด์ที่เลือก
5. Click mouse ที่ปุ่ม  เพื่อบันทึกค่าน้ำหนัก

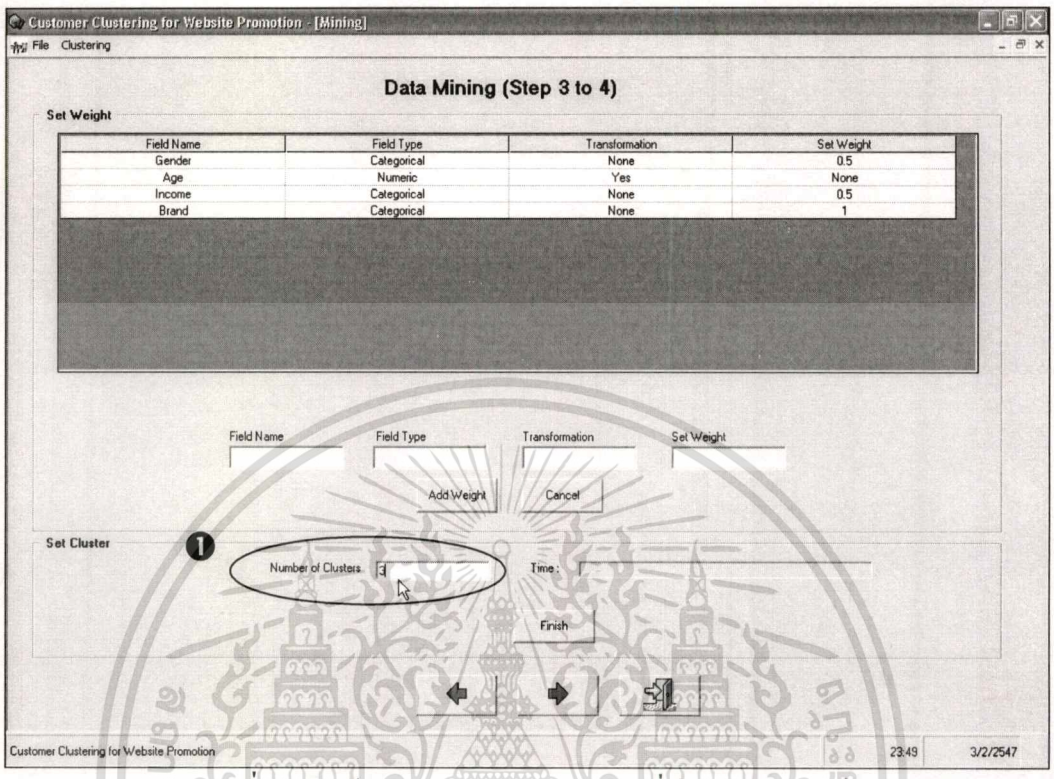
ข.3.2 การแก้ไขน้ำหนัก



รูปที่ ข.16 หน้าจอแสดงการแก้ไขน้ำหนัก

1. น้ำหนักที่กำหนดให้กับฟิลด์ไปแล้วนั้นสามารถแก้ไขได้โดย Click mouse ที่ข้อมูลที่ต้องการแก้ไขใน Set Weight
2. ข้อมูลที่ถูก Click mouse จะปรากฏอยู่ใน Field Name, Field Type และ Transformation สำหรับในช่อง Set Weight จะแสดงค่าน้ำหนักที่ถูกกำหนดไว้
3. ระบุน้ำหนักใหม่ที่ต้องการ
4. Click mouse ที่ปุ่ม Add Weight เพื่อบันทึกค่าน้ำหนักใหม่

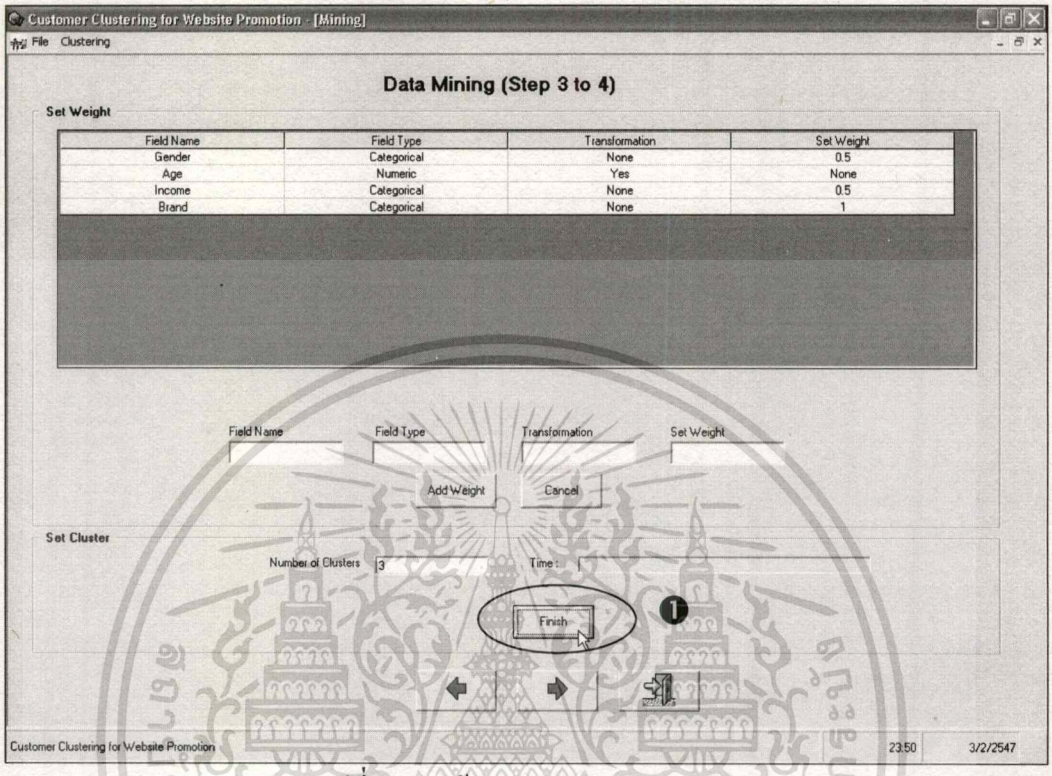
ข.3.3 การกำหนดจำนวนกลุ่มที่ต้องการวิเคราะห์



รูปที่ ข.17 หน้าจอการกำหนดจำนวนกลุ่มที่ต้องการวิเคราะห์

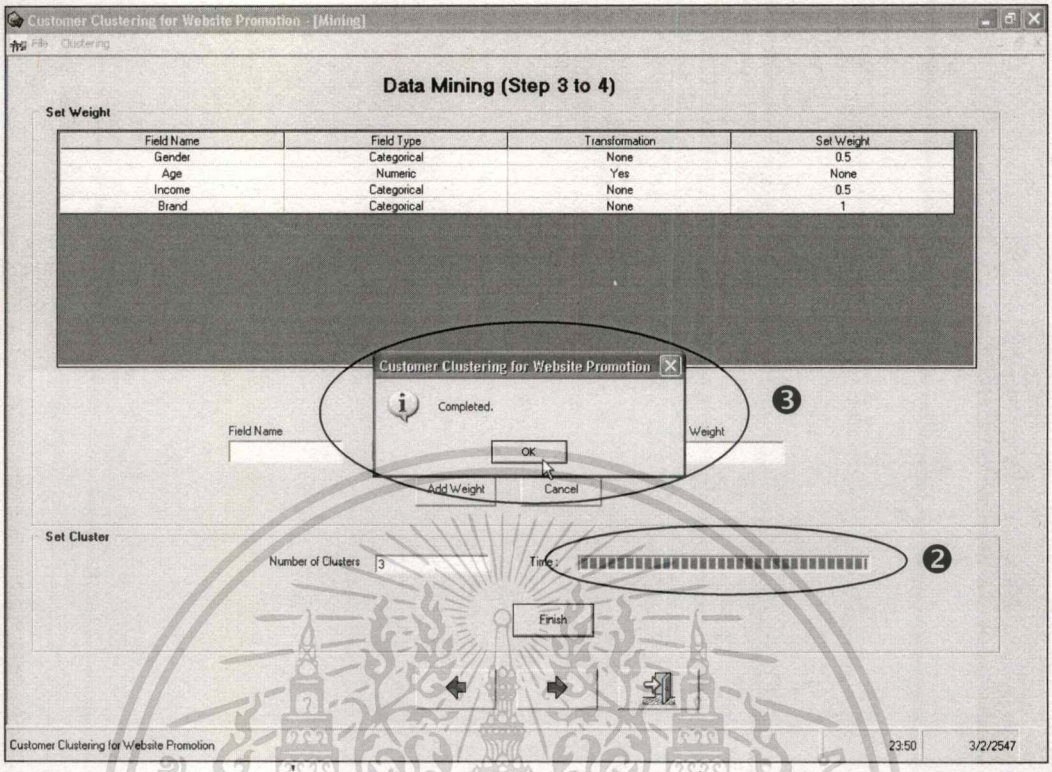
1. หลังจากที่กำหนดค่าน้ำหนักเรียบร้อยแล้ว การกำหนดจำนวนกลุ่มที่ต้องการวิเคราะห์สามารถทำได้โดย ระบุจำนวนกลุ่มที่ต้องการใน Number of Clusters
2. จำนวนกลุ่มที่ระบุ จะต้องเป็นเลขจำนวนเต็มบวกที่มากกว่า 2 เพื่อให้การวิเคราะห์มีประสิทธิภาพ

ข.3.4 การประมวลผล



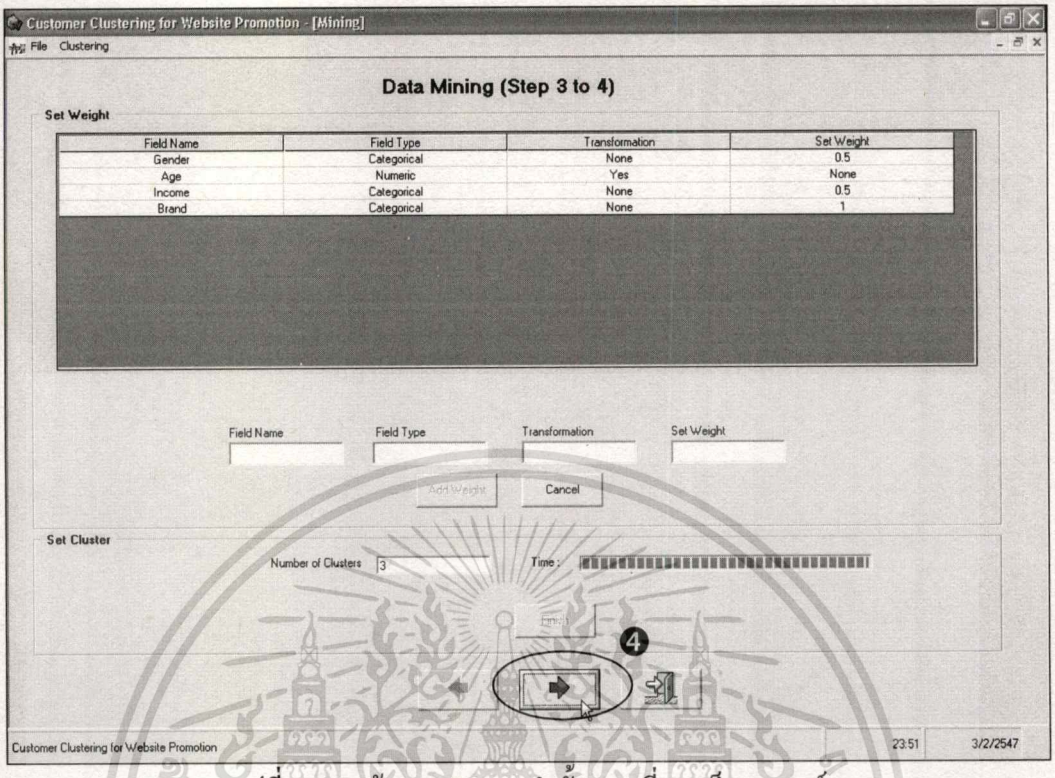
รูปที่ ข.18 หน้าจอแสดงการประมวลผล

1. เมื่อระบุข้อมูลทุกอย่างเรียบร้อยแล้ว การประมวลผลโปรแกรม สามารถทำได้โดย Click mouse ที่ปุ่ม Finish
2. เมื่อกดปุ่ม Finish แล้วระยะเวลาการประมวลผลสามารถสังเกตได้จากแถบ Time ที่ปรากฏอยู่



รูปที่ ข.19 หน้าจอแสดงการประมวลผลเสร็จสมบูรณ์

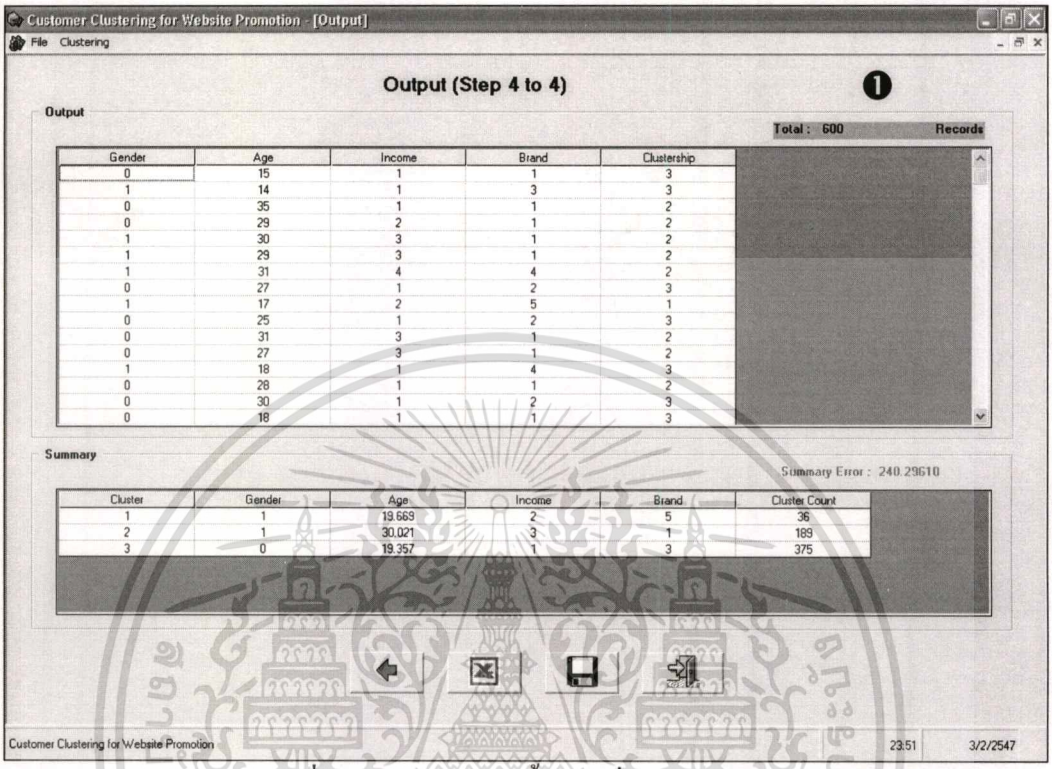
3. เมื่อโปรแกรมทำการประมวลผลเสร็จแล้ว จะปรากฏข้อความ Complete เพื่อบอกให้ทราบว่าการประมวลผลได้เสร็จสิ้นสมบูรณ์แล้ว



รูปที่ ข.20 หน้าจอแสดงการทำขั้นตอนที่ 3 เสร็จสมบูรณ์

4. หลังจากที่ทำ Step 3 เสร็จเรียบร้อยแล้ว Click mouse ที่ปุ่ม _____ เพื่อไปดูผลการวิเคราะห์ที่ Step 4

ข.4 การทำงานขั้นที่ 4 การแสดงผล (Output)



รูปที่ ข.21 หน้าจอแสดงขั้นตอนที่ 4 การแสดงผล

1. Step 4 คือการแสดงผล (Output) ซึ่งแบ่งออกเป็น 2 ส่วนคือ
   - Output จะแสดงว่าฟิลด์ข้อมูลทั้งหมดจากตัวอย่างคือ Gender, Age, Income, Brand และฟิลด์ที่บอกว่าเรคอร์ดนั้นอยู่ที่กลุ่มใด คือ Clustership
   - Summary สรุปผลการวิเคราะห์ที่ได้ของแต่ละกลุ่มว่าจุดศูนย์กลางมีค่าเท่าใด และจำนวนสมาชิกที่อยู่ในกลุ่มนั้น และสามารถดูค่า Error ที่เกิดขึ้นได้ที่ Summary Error
2. หลังจากได้ผลลัพธ์แล้ว สามารถบันทึกข้อมูลทั้งหมดได้ 2 วิธีคือ
   - วิธีที่ 1 บันทึกข้อมูลเป็น Text File ซึ่งจะบันทึกข้อมูลในส่วน Summary เท่านั้น หลังจาก Save แล้ว สามารถเปิดดูข้อมูลดังกล่าวได้ในโปรแกรม โดยเลือกคำสั่ง File > Open หรือเปิดดูข้อมูลในโปรแกรม Editor เช่น Notepad เป็นต้น วิธีการ Save ข้อมูลเป็น Text File สามารถทำได้โดยเลือกคำสั่ง File > Save หรือ Click mouse ที่ปุ่ม 

- วิธีที่ 2 บันทึกข้อมูลเป็นไฟล์นามสกุล *.xls ซึ่งจะบันทึกข้อมูลทั้งในส่วน Output และ Summary หลังจาก Save แล้ว สามารถเปิดดูข้อมูลดังกล่าวได้ในโปรแกรม Microsoft Excel วิธีการ Save ข้อมูลเป็นไฟล์นามสกุล *.xls สามารถทำได้โดย เลือกคำสั่ง File > Save for Excel หรือ Click mouse ที่ปุ่ม 

3. ถ้าต้องการย้อนกลับไปแก้ไขข้อมูลใน Step 1 ถึง 3 สามารถทำได้โดย Click mouse ที่ปุ่ม 
4. ถ้าต้องการเริ่มการทำงานใหม่โดยยกเลิกข้อมูลที่กำหนดไว้ทั้งหมด สามารถทำได้โดย เลือกคำสั่ง File > New หรือ Ctrl+N
5. ถ้าต้องการปิดโปรแกรม สามารถทำได้โดย เลือกคำสั่ง File > Exit หรือ Ctrl+X หรือ Click mouse ที่ปุ่ม 

ค. ความหมายของ Warning Massage และวิธีการแก้ไข

การใช้โปรแกรมในขั้นตอนต่างๆนั้นบางครั้งอาจเกิดข้อผิดพลาดได้ ซึ่งในขั้นตอนต่างๆ จะปรากฏข้อความเตือนขึ้นเมื่อเกิดข้อผิดพลาดที่อาจส่งผลต่อการวิเคราะห์ข้อมูล ดังนั้นเพื่อให้ทราบถึงสาเหตุ และคำแนะนำถึงแนวทางการแก้ปัญหาเพื่อผู้ใช้ให้สามารถวิเคราะห์ข้อมูลในขั้นตอนต่อไปได้

Warning Massage ทั้งหมดนี้สามารถค้นหาคำอธิบายได้ง่าย เพราะถูกแสดงโดยแบ่งตามขั้นตอนต่างๆ ผู้ใช้สามารถดูได้จากขั้นตอนที่เกิดและหมายเลขที่ปรากฏ

ตารางที่ ค.1 แสดง Warning Message ที่เกิดขึ้นในขั้นตอนที่ 1

Step 1 (Data Selecting)	สาเหตุ	การแก้ไข
Warning 11: Please select the database name.	ยังไม่ได้ระบุฐานข้อมูลที่ต้องการวิเคราะห์	ให้กดปุ่ม Browse เพื่อเลือกฐานข้อมูลที่ต้องการวิเคราะห์
Warning 12: This field was selected.	เลือกฟิลด์ซ้ำกับฟิลด์ที่มีอยู่แล้วใน Selected fields	เลือกฟิลด์ใน Available fields ตัวอื่น

ตารางที่ ค.2 แสดง Warning Message ที่เกิดขึ้นในขั้นตอนที่ 2

Step 2 (Data Preparing)	สาเหตุ	การแก้ไข
Warning 21: Please select field name.	ยังไม่ได้ระบุชื่อฟิลด์ที่ต้องการแปลงข้อมูล (Transformation)	ให้เลือกชื่อฟิลด์ที่ต้องการแปลงข้อมูล (Transformation)
Warning 22: Please set the transformation range.	ยังไม่ได้ระบุขอบเขตต่ำสุดหรือสูงสุดของข้อมูลที่ต้องการแปลง	ให้ระบุขอบเขตต่ำสุดหรือสูงสุดของข้อมูลที่ต้องการแปลงใน Transformation Range

ตารางที่ ค.3 แสดง Warning Message ที่เกิดขึ้นในขั้นตอนที่ 3

Step 3 (Data Mining)	สาเหตุ	การแก้ไข
Warning 31: Please number the weight.	ยังไม่ได้ระบุค่า weight ให้กับฟิลด์ประเภท Categorical ที่เลือก	ให้ระบุค่า weight กับฟิลด์ที่เลือก
Warning 32: The numeric cannot set the weight.	ข้อมูลประเภท Numeric ไม่สามารถระบุค่า weight ได้	สามารถระบุค่า weight เฉพาะฟิลด์ประเภท Categorical เท่านั้น
Warning 33: Please number the clusters.	ยังไม่ได้ระบุจำนวนกลุ่มที่ต้องการ	ระบุจำนวนกลุ่มที่ต้องการ
Warning 34: The number must be a integer at least two.	จำนวนกลุ่มที่ระบุไม่ใช่เลขจำนวนเต็มหรือมีค่าน้อยกว่า 2	จำนวนกลุ่มจะต้องระบุเป็นเลขจำนวนเต็มตั้งแต่ 2 ขึ้นไปเท่านั้น

ประวัติผู้เขียน

ชื่อผู้เขียน	นางสาวนภาศรี รุ่งธีรวัฒนานนท์
วันเดือนปีเกิด	27 ตุลาคม 2522
สถานที่เกิด	กรุงเทพมหานคร
ประวัติการศึกษา	สำเร็จการศึกษาระดับมัธยมศึกษาจากโรงเรียนเตรียมอุดมศึกษาพัฒนาการ ปี 2540 สำเร็จการศึกษาระดับปริญญาตรีจากคณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยกรุงเทพ ปี 2544
ประวัติการทำงาน	อาจารย์พิเศษ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยกรุงเทพ (ปี 2545-ปัจจุบัน)