เหมืองข้อมูล

DATA MINING

นายเกริกชัย วัฒนาประสาทกุล
นายทศพล กาญจะโนสถ
นายธันยวัฒน์ จิตติพลังศรี

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตร์บัณฑิต
ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ปีการศึกษา 2547

เหมืองข้อมูล

DATA MINING

โดย

นายเกริกชัย วัฒนาประสาทกุล

นายทศพล กาญจะโนสถ

นายธันยวัฒน์ จิตติพลังศรี

อาจารย์ที่ปรึกษา

ดร.วรวัฒน์ ลิ้มโภคา

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตร์บัณฑิต

ภาควิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ปีการศึกษา 2547

ปริญญานิพนธ์ปีการศึกษา 2547

ภาควิชา วิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เรื่อง เหมืองข้อมูล

DATA MINING

ผู้จัดทำ

1. นายเกริกชัย	วัฒนาประสาทกุล	รหัสประจำตัว 44010034
2. นายทศพล	กาญจะโนสถ	รหัสประจำตัว 44010179
3. นายธันยวัฒน์	จิตติพลังศรี	รหัสประจำตัว 44010212

อาจารย์ที่ปรึกษา

(ดร.วรวัฒน์ ลิ้มโภคา)

เหมืองข้อมูล

นายเกริกชัย วัฒนาประสาทกุล 44010034
นายทศพล กาญจะโนสถ 44010179
นายธันยวัฒน์ จิตติพลังศรี 44010212
ดร.วรวัฒน์ ลิ้มโภคา อาจารย์ที่ปรึกษา
ปีการศึกษา 2547

บทคัดย่อ

เป็นที่ยอมรับกันว่า การทำธุรกิจในปัจจุบันนั้น นอกจาก 5M ซึ่งคือ คน วัตถุดิบ เงิน เครื่องจักร และวิธีการทำงานแล้ว ยังต้องมีข้อมูลเป็นปัจจัยหลักอีก แต่ปัญหาในการทำธุรกิจในปัจจุบันนั้นไม่ได้อยู่ที่ว่า มีข้อมูลหรือไม่ แต่ปัญหากลับกลายเป็นว่า จะสามารถทำอย่างไรให้ข้อมูลที่มีนั้นเป็นข้อมูลที่มีประโยชน์ต่อการทำงาน และการตัดสินใจทางธุรกิจ จึงเป็นที่มาของการทำ Data Mining ซึ่งเป็นการค้นหาความรู้ (Knowledge) ใหม่ๆ จากข้อมูลที่มีอยู่ โดยทางทีมงานได้นำ Data Mining มาประยุกต์ใช้กับ 3 ปัญหาทางธุรกิจเครื่องปรับอากาศ ซึ่งคือ การค้นหาสาเหตุความไม่พอใจในการบริการลูกค้าโดยการทำ learning decision tree ,สาเหตุการซื้อสินค้าเครื่องปรับอากาศของลูกค้า โดยการทำ Clustering และการพยากรณ์จำนวนการซื้อสินค้าเครื่องปรับอากาศโดยใช้ Neural Network โดยปัญหาแรก และปัญหาที่สองจะจำลองข้อมูลขึ้นมาโดยการ Simulate แต่ปัญหาสุดท้ายใช้ข้อมูลจริงซึ่งได้จากบริษัท Saijo Denki (International) จำกัด, กรมอุตุนิยมวิทยา และสำนักงานคณะกรรมการพัฒนาการเศรษฐกิจ และสังคมแห่งชาติ ซึ่งทางทีมงานได้คาดหวังว่าผลการทำงานจะได้ สาเหตุความไม่พอใจของลูกค้าต่อการบริการของบริษัท, สาเหตุของการซื้อสินค้า และ Model ในการพยากรณ์จำนวนการซื้อสินค้าเครื่องปรับอากาศ โดยได้ทำการทดลอง Clustering ด้วยวิธี Simulate 5 การทดลองได้ผลดังนี้ การทดลองที่ 1 ได้ ค่าความเบี่ยงเบน 4.79% การทดลองที่ 2 ได้ ค่าความเบี่ยงเบน 2.9% การทดลองที่ 3 ได้ ค่าความเบี่ยงเบน 23.53% การทดลองที่ 4 ได้ ค่าความเบี่ยงเบน 4.46% การทดลองที่ 5 ได้ ค่าความเบี่ยงเบน 16.73% และใช้ข้อมูลจริงได้ 2 Segment นอกจากนี้ได้ทำการทดลอง Decision Tree 5 การทดลองด้วยการ Simulate เช่นกัน โดยได้ผลดังนี้ การทดลองที่ 1 ได้ตรงกับที่คาดหวัง 100% การทดลองที่ 2 ได้ตรงกับที่คาดหวัง 100% การทดลองที่ 3 ได้ตรงกับที่คาดหวัง 100% การทดลองที่ 4 ได้ตรงกับที่คาดหวัง 92.5% การทดลองที่ 1 ได้ตรงกับที่คาดหวัง 90.3%

DATA MINING

Mr. Kerkchai Watanaprasartkul
Mr. Thosapon Kanchanosoth
Mr. Thunyawat Chittiphalungsri
Dr. Vorawat Limpoka Advisor
Academic Year 2004

ABSTRACT

It is widely accepted that only 5M resources, Man, Material, Money, Machine and Method, for business nowadays is not enough. Running business needs information and knowledge. Moreover, business problem is not whether the company has the information or not. The key problem is how to make the information the company has useful and help decision making. Thus, it comes to "Data Mining" which able to extracts the knowledge from the database. It helps the company creates more competitive advantages by creates more knowledge. In this project, we implements Data Mining to three Air-Conditioning business problems. First, we try to extract the scenarios that the customers do not appreciate the customer service department by using learning decision tree. Second, we try to discover the air-conditioner buying factors by using clustering. Third, we try to predict numbers of air-conditioner by using the information from Metrology Department, Saijo Denki (International) and Sapapat. The last one is done by Neural Network. We have an experiment on Clustering based on simulating data and real data collected by questionnaire. We simulate 5 cases on Clustering. The first experiment shows the variance of 4.79%. The second one shows the variance of 2.9%. The third one shows the variance of 23.53%. The forth one shows the variance of 4.46%. The last one shows the variance of 16.73%. Moreover, after clustering the real data, it reveals the 2 main customer segments. Furthermore, we simulate 5 cases on Decision tree. The correction of the first decision tree is 100% also the second and the third one. The forth one is 92.5% and the last one is 90.3%.

กิตติกรรมประกาศ

ปริญญานิพนธ์ฉบับนี้คงไม่อาจเสร็จได้ด้วยดี หากไม่ได้รับความช่วยเหลือ และร่วมมือจากหลายๆ ฝ่ายด้วยกัน บุคคลแรกที่ต้องกล่าวถึงเพราะเป็นส่วนสำคัญที่ทำให้ปริญญานิพนธ์นี้เสร็จลงได้ก็คือ อาจารย์ วรวัฒน์ ลิ้มโภคา อาจารย์ที่ปรึกษาปริญญานิพนธ์ ที่ให้ความเอาใจใส่ แนะนำ และช่วยเหลือเสมอมา ซึ่งต้องขอขอบพระคุณเป็นอย่างมาก

และต้องขอขอบพระคุณบุคคลสำคัญที่สุดที่ทำให้ข้าพเจ้ามีวันนี้ ก็คือ บิดา มารดา อันเป็นที่เคารพรักยิ่ง ซึ่งได้เลี้ยงดูข้าพเจ้ามาเป็นอย่างดี พร้อมทั้งให้โอกาสในการศึกษาอย่างเต็มที่ และยังให้กำลังใจ เอาใจใส่เสมอมา ในทุก ๆ ด้านอันหาที่เปรียบมิได้ ข้าพเจ้าขอระลึกในพระคุณอันสุดประมาณ และขอกราบขอบพระคุณมา ณ ที่นี้

เกริกชัย วัฒนาประสาทกุล
ทศพล กาญจะโนสถ
ธันยวัฒน์ จิตติพลังศรี

สารบัญ

สารบัญตาราง

สารบัญรูปภาพ

บทที่ 1
บทนำ

1.1 ความสำคัญและที่มา

ในปัจจุบัน บริษัทที่ประกอบธุรกิจส่วนใหญ่ในหลายๆประเทศจะมีการเก็บรวบรวมข้อมูลต่างๆ เอาไว้จำนวนมากในรูปแบบของไฟล์ในคอมพิวเตอร์ แต่ในประเทศไทยนั้น มีบริษัททางธุรกิจขนาดใหญ่เพียงไม่กี่บริษัทที่มีการเก็บรวบรวมข้อมูลในคอมพิวเตอร์ เนื่องด้วยบริษัทส่วนใหญ่ในไทยนั้นมักจะเก็บข้อมูลในรูปแบบของกระดาษและไม่เล็งเห็นถึงประโยชน์ของการเก็บรวบรวมข้อมูล จึงทำให้ในบางครั้งเมื่อต้องการหาข้อมูล อาจหาข้อมูลไม่พบหรือใช้เวลานานในการหาข้อมูลแต่ละครั้งซึ่งยังรวมไปถึงการจัดการกับข้อมูลที่มีอยู่ด้วย ทางผู้จัดทำจึงมีความคิดที่ว่าจะทำอย่างไรให้ข้อมูลที่เก็บรวบรวมไว้นั้นจะมีประโยชน์มากกว่าการเก็บและค้นหาข้อมูล จึงได้ทำการศึกษาดาต้าไมนิ่ง (Data Mining) และนำมาวิเคราะห์หาความรู้ (Knowledge) ที่เรายังไม่ทราบจากฐานข้อมูลขนาดใหญ่ เพื่อเป็นประโยชน์ต่อบริษัทในทางธุรกิจอีกทางหนึ่ง ซึ่งอาจทำให้บริษัทหลายๆบริษัทในไทยได้เล็งเห็นถึงความสำคัญในการเก็บข้อมูลไว้ในคอมพิวเตอร์และสามารถใช้ข้อมูลที่มีอยู่ให้เกิดประสิทธิภาพสูงสุด

1.2 วัตถุประสงค์ของงานวิจัย

1.2.1 เพื่อศึกษาถึงกระบวนการดาต้าไมนิ่งว่าเป็นอย่างไรและสามารถนำเทคนิคนี้ไปใช้กับข้อมูลได้อย่างไร

1.2.2 เพื่อเป็นการใช้ฐานข้อมูลที่มีอยู่เป็นจำนวนมากให้เกิดประโยชน์สูงสุด อีกทั้งยังเป็นการเพิ่มคุณค่าให้กับข้อมูลที่มีอยู่นอกเหนือจากการเก็บและค้นหา

1.2.3 เพื่อให้มีความรู้ความสามารถในการวิเคราะห์ปัญหาต่างๆได้อย่างมีประสิทธิภาพมากขึ้น

1.2.4 เพื่อสามารถนำความรู้ที่ได้ศึกษามาไปประยุกต์ใช้งานกับปัญหาได้จริง และสามารถสร้างโปรแกรมขึ้นมาเพื่อวิเคราะห์ผลลัพธ์ได้อย่างมีประสิทธิภาพ

1.3 ขอบเขตของงานวิจัย

งานวิจัยนี้ได้นำดาต้าไมนิ่งมาประยุกต์ใช้กับปัญหาทางธุรกิจ โดยยกธุรกิจที่เกี่ยวกับเครื่องปรับอากาศมาเป็นปัญหา ซึ่งเป็นปัญหาพื้นฐานที่อาจนำไปประยุกต์ใช้กับธุรกิจประเภทอื่นได้เช่นกัน โดยจะทำการศึกษาและ นำอัลกอริธึม (Algorithm) ที่เกี่ยวข้องกับเหมืองข้อมูล (Data Mining) มาทดสอบกับข้อมูลที่เตรียมไว้ เพื่อหาความแม่นยำ ซึ่งไม่ได้คาดหวังว่าจะต้องได้ความแม่นยำ 100% ในทุกๆกรณี ทำการศึกษาและออกแบบงาน หรือปัญหาที่จะทำการทดสอบว่าต้องใช้ข้อมูลอะไรบ้างและต้องการอะไรจากการใช้ ดาต้าไมนิ่ง กับข้อมูลที่มีอยู่

โดยงานวิจัยนี้จะสร้างฐานข้อมูลตัวอย่าง เพื่อใช้ทำการทดสอบ และสร้างโปรแกรมในการ เจนเนอเรต (Generate) ข้อมูลตามเงื่อนไขทดสอบต่างๆ สร้างโปรแกรมที่ใช้ทำดาต้าไมนิ่งสำหรับปัญหาทางธุรกิจที่ได้เลือกมา 3 ปัญหา ได้แก่ 1. หาว่าลูกค้าไม่ชอบสินค้า จากการบริการ มี รูปแบบอะไรบ้าง

2. หาว่าลูกค้าซื้อสินค้าเพราะอะไร และ 3. หาความสัมพันธ์ของตัวแปรต่างๆ เพื่อคาดการณ์ยอดขายสินค้าในอนาคต

นอกจากนั้นในโครงงานนี้ยังถือว่าเป็นโครงงานที่ทดลองสร้าง เพื่อศึกษาความเป็นไปได้ในการใช้งาน ดังนั้นจึงมีข้อจำกัดของข้อมูลบางอย่าง เช่น อาจจะมีการจำกัดชนิดของข้อมูล การจำกัดจำนวนข้อมูลที่ไม่สามารถรองรับข้อมูลปริมาณมาก ๆ ได้ แต่ถึงอย่างไรก็ยังเพียงพอต่อการทดสอบอย่างแน่นอน

1.4 ผลที่คาดว่าจะได้รับ

สามารถสร้างโปรแกรมคาต้าไมนิ่งโดยนำอัลกอริธึมที่ได้ศึกษามาไปประยุกต์ใช้แก้ปัญหาทางธุรกิจเครื่องปรับอากาศทั้ง 3 ปัญหาได้อย่างมีประสิทธิภาพ และแสดงผลลัพธ์ที่ได้จากการใช้โปรแกรมเพื่อนำไปวิเคราะห์ในการปรับปรุงธุรกิจต่อไป

1.5 วิธีการดำเนินงาน

1. ทำการศึกษาคาต้าไมนิ่งว่าคืออะไรและมีกระบวนการทำงานอย่างไรบ้าง รวมไปถึงอัลกอริธึมต่างๆที่ใช้ในคาต้าไมนิ่ง
2. นำเอาความรู้ที่ได้ไปทำการศึกษาและออกแบบปัญหาที่จะทำว่าต้องใช้ข้อมูลอะไรบ้างและต้องการอะไรจากการใช้คาต้าไมนิ่งกับข้อมูลที่มีอยู่ หลังจากนั้นจึงพิจารณาเลือกอัลกอริธึมที่เหมาะสมกับข้อมูลที่มีเพื่อนำไปประยุกต์ใช้ในการวิเคราะห์หาความรู้จากข้อมูลที่มีอยู่โดยได้แบ่งเป็น 2 ส่วนได้แก่ วิเคราะห์ข้อมูลที่ได้จากการจำลองขึ้นเอง และวิเคราะห์ข้อมูลจริงที่ได้จากการทำแบบสำรวจของประชาชนตามสถานที่ต่างๆ
3. พัฒนาโปรแกรมคาต้าไมนิ่งและทำการทดลองกับข้อมูลที่มีอยู่เพื่อหาความรู้ที่เรายังไม่ทราบซึ่งมีความสัมพันธ์กับข้อมูลที่มีอยู่จากฐานข้อมูล

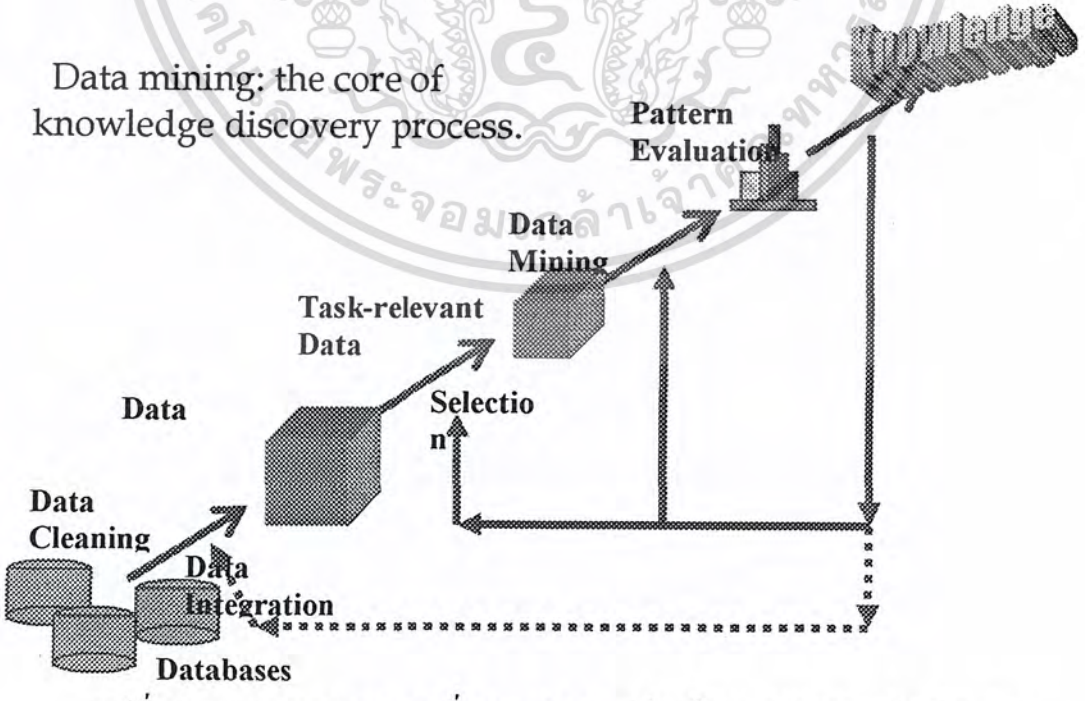
บทที่ 2

Overview Data Mining

คือขบวนการทำงานที่เรียกว่า process ที่สกัดข้อมูล (Extract data) จากฐานข้อมูลขนาดใหญ่ (Large Information) เพื่อให้ได้สารสนเทศ (Useful Information) ที่เรายังไม่รู้ (Unknown data) โดยเป็นสารสนเทศที่มีเหตุผล (Valid) และสามารถนำไปใช้ได้ (Actionable) ซึ่งเป็นสิ่งสำคัญในการที่จะช่วยการตัดสินใจในการทำธุรกิจ

รูปที่ 2-1 แสดงเปรียบเทียบการทำ Data Mining

โดยที่คาต้าไมนิ่งนั้นเป็น 1 ในกระบวนการของ KDD (Knowledge Discovery in Databases) ซึ่ง KDD คือ การค้นหาข้อมูลที่มีประโยชน์ ที่อาจจะซ่อนอยู่ในระบบฐานข้อมูลที่มีขนาดใหญ่ๆ เพื่อนำผลลัพธ์ที่ได้ออกมาไปใช้ประโยชน์ เช่น นำไปใช้ในการประกอบการตัดสินใจทางด้านธุรกิจ (decision support system) หรือทำให้องค์กรธุรกิจเข้าใจพฤติกรรมของลูกค้า หรือผู้บริโภคได้ดีขึ้น ทำให้สามารถรักษาลูกค้าเก่าไว้ได้ และ อาจจะหาลูกค้าใหม่ได้มากขึ้น เมื่อระบบธุรกิจนั้นๆ มีการแข่งขันสูง เป็นต้น



รูปที่ 2-2 กระบวนการของ KDD ซึ่ง Data Mining เป็น 1 ในกระบวนการหลักของ KDD

โดยทั่วไป กระบวนการ Data Mining จะประกอบด้วย 7 ขั้นตอนหลักๆ คือ

ขั้นตอนที่ 1 **กำหนดผลลัพธ์ทางธุรกิจ** : ระบุ ผลลัพธ์ทางธุรกิจที่ต้องเตรียมและจากนั้นก็หาวิธีที่จะแปลง ผลลัพธ์ทางธุรกิจ ให้กลายเป็นคำถามหรือชุดของคำถามซึ่งคาด้าไมนิ่งสามารถเตรียมได้

ผลลัพธ์ทางธุรกิจ ควรจะตรงกับความต้องการโดยประกอบด้วย

- คำอธิบายที่ชัดเจนของปัญหาที่ได้เตรียมไว้
- ความเข้าใจในข้อมูลซึ่งอาจจะตรงประเด็น
- มุมมองสำหรับการนำผลลัพธ์ที่ได้ไปใช้ในธุรกิจ

ขั้นตอนที่ 2 **กำหนด Data Model ที่จะใช้** : กำหนดข้อมูลที่จะใช้โดยจะใช้บางส่วนของคลังข้อมูล

โดย Data Model พื้นฐานจะประกอบด้วย

- แหล่งข้อมูลที่จะใช้

ระบุที่เก็บข้อมูลที่แท้จริงว่าอยู่ที่ใด

- ชนิดของข้อมูล

กำหนดว่าข้อมูลมีโครงสร้างอย่างไร

- ข้อมูลที่มีอยู่

แสดงตารางหรือไฟล์ข้อมูลและ field ที่ข้อมูลเก็บอยู่

- คำอธิบายข้อมูล

ประกอบด้วยชื่อและคำอธิบายของ field ของชื่อเหล่านี้

โดย data model ที่ต้องการจะอยู่ในรูปแบบของไฟล์ๆเดียวหรือตารางในฐานข้อมูลซึ่ง 1 record ต่อ 1 ลูกค้าหรือแผนกหรืออะไรก็ตามที่เป็นเป้าหมายในการสืบค้น ซึ่งแต่ละ record จะประกอบด้วยตัวแปร 1 ตัวหรือมากกว่านั้นซึ่งแต่ละตัวแปรอาจได้มาจากแหล่งข้อมูลที่แตกต่างกัน ซึ่งใน application ทางธุรกิจข้อมูลส่วนใหญ่จะเป็น

- Transaction data คือข้อมูลที่ใช้งานโดยจะถูกสร้างขึ้นในแต่ละครั้งที่มีผลกระทบเกิดขึ้นกับเป้าหมาย
- Relationship data คือข้อมูลที่ประกอบด้วยข้อมูลที่มีความสัมพันธ์ที่ไม่เปลี่ยนแปลงซึ่งเกี่ยวกับลูกค้า ผลิตภัณฑ์ อุปกรณ์ รายการในบัญชี และกระบวนการทำงาน
- Demographic data ประกอบด้วย ข้อมูลเฉพาะของแต่ละบุคคลโดยมาจากแหล่งข้อมูลภายนอก เช่น อายุ เพศ

ข้อดีหลักๆของการใช้ data model คือจะเตรียมวิธีในการมองว่าคาด้าไมนิ่งจะสามารถถูกนำมาใช้กับธุรกิจได้อย่างไร

ขั้นตอนที่ 3 **หาแหล่งข้อมูลและกระทำการกับข้อมูลก่อนการทำในกระบวนการ :** หา แหล่งข้อมูลและกระทำการกับข้อมูลก่อนการทำในกระบวนการที่มีอยู่ใน data model โดยจะประกอบด้วยขั้นตอนการระบุ การรวบรวม การกรอง และการรวมข้อมูลดิบเข้าไปในรูปแบบที่ต้องการใน data model และ mining function ที่ได้ถูกเลือกไว้

Data preprocessing

- ถ้าข้อมูลไม่ได้มาจากคลังข้อมูล ข้อมูลนั้นต้องผ่านกระบวนการ cleansing ,aggregated , transforming และ filtering ก่อนเสมอ

ขั้นตอนที่ 4 **ประเมินค่าของ data model :** ประกอบด้วย 3 ขั้นตอนดังนี้

- ขั้นตอนแรก

Visual inspection ประกอบด้วยการเลือกข้อมูลที่จะเป็น input ด้วย visualizing tool โดยอาจจะนำไปสู่การตรวจพบการกระจายของข้อมูลที่ไม่น่าเป็นไปได้ ตัวอย่าง เช่น การ join ตารางผิดในระหว่างขั้นตอนการเตรียมข้อมูลที่สามารถเป็นผลลัพธ์ในตัวแปรที่เก็บค่าที่แท้จริงซึ่งเป็นของ field ที่ต่างกันได้

- ขั้นตอนที่ 2

จะเกี่ยวข้องกับการความขัดแย้งกันของข้อมูลและการแก้ปัญหาค่า error เนื่องจากการกระจายตัวของข้อมูลที่ผิดปกติที่ได้พบภายในขั้นตอนแรกสามารถทำให้เกิดการเก็บรวบรวมข้อมูลที่ผิดพลาด โดยค่าที่ไม่เกี่ยวข้องหรือค่าที่ผิดนั้นจะทำให้เกิดผลลัพธ์ที่ไม่ตรงกับความต้องการ

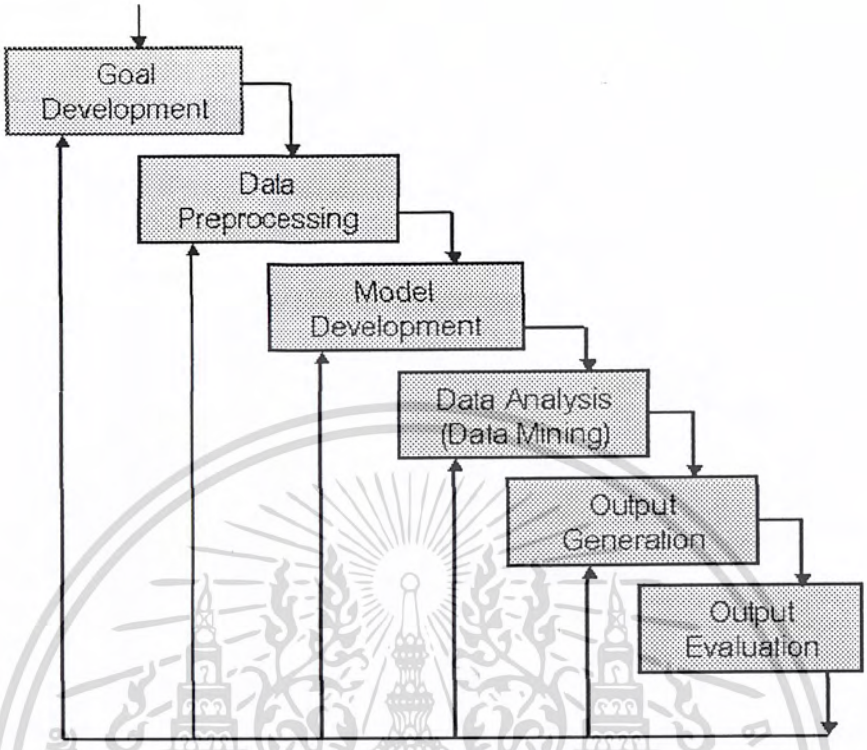
- ขั้นตอนสุดท้าย

เป็นการเลือกลักษณะหรือตัวแปรเพื่อที่จะให้คาด้าไมนิ่งทำงาน

ขั้นตอนที่ 5 **เลือกเทคนิคหรือ Algorithm ในการทำคาด้าไมนิ่ง :** ในขั้นตอนนี้จะเป็นการกำหนดเทคนิคที่เหมาะสมหรือนำเทคนิคต่างๆมารวมกันเพื่อนำไปใช้และ วิธีที่ในการนำเทคนิคดังกล่าวไปประยุกต์ใช้กับข้อมูล

ขั้นตอนที่ 6 **ทำการแปลผลลัพธ์ที่ได้ :** ในขั้นตอนนี้จะถูกดำเนินการโดยผู้เชี่ยวชาญ เนื่องจากผลลัพธ์ที่ได้นั้นสามารถแปลไปเป็นข้อมูลได้ในหลายแนวทางซึ่ง บางครั้งก็เป็นสิ่งที่ยากในการแปลผลลัพธ์ที่ได้

ขั้นตอนที่ 7 นำผลลัพธ์ที่ได้ไปประยุกต์ใช้ในทางธุรกิจ



รูปที่ 2-3 กระบวนการของ Data Mining

ความแตกต่างระหว่าง Data Mining และ Data warehouse

Data Mining เป็นกระบวนการซึ่งข้อมูลจะถูกวิเคราะห์โดยอัตโนมัติเพื่อค้นหารูปแบบ ที่สำคัญที่ทำนายค่าเชิงสถิติ ตัวอย่าง เช่น ระบบ Data Mining อาจจะทำการ process กลุ่มข้อมูลที่บันทึกบัตร credit และระบุรูปแบบการซื้อที่มีลักษณะเข้าข่ายเป็นธุรกิจที่ฉ้อโกง

ส่วน Data warehouse เป็นกระบวนการซึ่งองค์กรทางธุรกิจได้เก็บรวบรวมและจัดการกับข้อมูลที่มีปริมาณมากๆ ซึ่งข้อมูลใน Data warehouse จะถูกใช้เพื่อเป็น input เข้าสู่ระบบโปรแกรม Data Mining อีกที

บทที่ 3

Data Mining Algorithms

Classical Technique

Statistics – Linear Regression; Statistics for Prediction

เป็นการคาดการณ์ทางสถิติโดยสมมติสมการเส้นตรง

Prediction = a + b x Predictor

ตัวอย่างในการนำไปใช้ เช่น การคาดการณ์ปริมาณเงินฝากของลูกค้า เช่น

Y = $1,000 + 0.01 x customer annual's income

Nearest Neighborhood

เป็นการคาดการณ์โดยใช้สิ่งรอบข้างเป็นตัวกำหนด โดยยึดสมมติฐานที่ว่า "สิ่งที่อยู่ใกล้กัน จะมีค่าที่ถูกคาดการณ์เหมือนกัน"

ตัวอย่างการนำไปใช้ คือ ตลาดหุ้น เช่น เมื่อข้อมูล 9 ช่วงเวลาแรก เหมือนกัน เวลาที่ 10 ก็น่าจะเหมือนกันเช่นกัน เพราะถือว่า อยู่ในกลุ่มข้อมูลเดียวกัน หรือ ใช้ใน text retrieval

Classification

เป็นการแบ่งหมวดหมู่โดยทำการกำหนดสิ่งที่เป็นลักษณะเด่นในแต่ละหมวดหมู่ ซึ่งจะแบ่งข้อมูลตามความคล้ายคลึงกันจากตัวอย่างข้อมูลที่มีอยู่ (supervised learning)

Association Rules

เป็นการหาความสัมพันธ์ระหว่าง data items ด้วยกันเองซึ่งมีพื้นฐานมาจากการเกิดขึ้นร่วมกันหรือพร้อมกันในฐานข้อมูล

Rule Induction

เป็นการค้นหา รูปแบบ ที่น่าสนใจทุกๆ รูปแบบ ใน database โดยอาศัย 2 ตัวบ่งชี้ ได้แก่ Accuracy ซึ่งบ่งบอกว่า rule นั้นถูกต้อง หรือไม่ และ Coverage ซึ่งเป็นตัวบ่งบอกว่า กฎนั้นได้ถูกใช้ไปบ่อยแค่ไหน

Neural Networks

Neural Network สามารถนำมาใช้ใน Data Mining โดยใช้ในการทำนายค่า การแยกประเภท และการแบ่งกลุ่มได้ และถูกนำไปใช้อย่างกว้างขวาง เช่น การนำไปทำนายค่าทางด้านการเงิน การวินิจฉัยโรค การ Segmentation กลุ่มลูกค้า เป็นต้น หลักการของ Neural Network คือการประสานช่องว่างระหว่างมนุษย์ และคอมพิวเตอร์ โดยจำลองการเชื่อมต่อระหว่างเซลล์ประสาทของมนุษย์มาไว้บนคอมพิวเตอร์ ซึ่งมีหลักการ คือการเรียนรู้ จากตัวอย่างที่มีอยู่ หรือกล่าวได้ว่าเรียนรู้จากประสบการณ์ ซึ่งจะเห็นได้ว่ากระบวนการดังกล่าวต่างจาก von Neumann machines ซึ่งมีพื้นฐานการประมวลผลพื้นฐานทางด้านข้อมูลของมนุษย์ ซึ่งก็มีข้อเสียคือ ผลลัพธ์ที่ได้จาก Neural Network คือ ค่าถ่วงน้ำหนัก (Weight) ภายในเครือข่าย ซึ่งค่าดังกล่าว จะไม่บอกเหตุผลว่า ทำไมถึงได้คำตอบเช่นนี้

โดย neural network แก้ปัญหาได้หลักๆ 3 ประเภท คือ

- ปัญหาที่ไม่สามารถสร้าง Algorithmic Solution
- ปัญหาที่เรามีตัวอย่างของ behavior เยอะมากๆ
- ปัญหาที่เราต้องการดึงโครงสร้าง ออกจากข้อมูลที่มีอยู่

ปัญหาที่เหมาะกับการใช้ Neural Network จะมี 3 ประการ คือ

- สามารถกำหนด input ที่ชัดเจนเท่านั้น กล่าวคือ ต้องทราบว่าคุณลักษณะของข้อมูลอันไหนเป็นคุณลักษณะที่สำคัญ
- สามารถกำหนด output ที่ชัดเจน ซึ่งจะต้องทราบว่าทำนายค่าอะไร
- ประสบการณ์ต้องมีอย่างเพียงพอ กล่าวคือต้องมีตัวอย่างมากพอในการเรียนรู้

อีกหนึ่งคุณลักษณะของ Neural Network คือจะทำงานได้ดีที่สุดเมื่อมีการกำหนดช่วง input และ output ระหว่าง 0 และ 1 ด้วยเหตุนี้จึงต้องมีการนวด (massaging) ค่าของข้อมูล

โดย Neural Network จะต้องเรียนรู้ผ่านทางการ Training โดยการ Train นั้นจะสิ้นสุดเมื่อค่าถ่วงน้ำหนักนั้นไม่เปลี่ยนแปลงมาก หรือจนกระทั่งเทรนครบตามจำนวนรอบที่กำหนดไว้ (หนึ่งรอบเท่ากับการเทรนจนหมดชุดเทรนหนึ่งครั้ง) ซึ่งในทางปฏิบัติจะให้ Neural Network ทำงานบนชุดทดสอบ (test set) ซึ่งมันไม่เคยพบมาก่อนเพื่อให้แน่ใจว่า Neural Network ได้เรียนรู้รูปแบบที่ดีที่สุดในชุดเทรนแล้ว เมื่อ Neural Network ถูกเทรนจนได้ผลเป็นที่น่าพอใจแล้วก็จะได้ Model Neural Network ที่ใช้การได้ โดยค่า output ที่ได้จาก Model นี้จะเป็นตัวเลขระหว่าง 0 ถึง 1 ดังนั้นเราจะต้องแปลงค่านี้กลับ (unmassage)

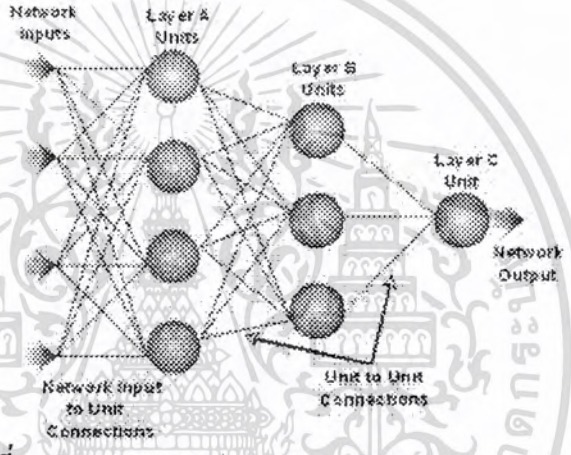
การใช้ Neural Network กับ Data Mining

การสร้าง Model สำหรับการแยกประเภท และการทำนายค่า ดังนี้

1. กำหนด input และ output
2. นวด input และ output ให้มีค่าอยู่ระหว่าง 0 และ 1
3. สร้าง Neural Network ที่มี topology ที่เหมาะสมกับงานนั้น

4. Train Neural Network โดยใช้ตัวอย่างจากชุดเทรน
5. ทดสอบ Neural Network โดยใช้ชุดทดสอบซึ่งเป็นอิสระจากชุดเทรน (ข้อมูลในชุดทดสอบจะต้องไม่ซ้ำกับข้อมูลที่อยู่ในชุดเทรน) ถ้าได้ผลไม่เป็นที่น่าพอใจอาจจะต้องเทรนใหม่ หรือเปลี่ยนชุดเทรน Topology และค่า Parameter ต่างๆ)
6. นำ Model ที่ได้ไปใช้ในการทำนาย หรือแยกประเภท

Neural Network ประกอบด้วยหลายๆ Unit แบ่งเป็นหลายๆ Layer โดย Layer ที่อยู่ด้านซ้ายสุดจะถูกเรียกว่า Input Layer และด้านขวาสุดจะถูกเรียกว่า Output Layer และ Layer ตรงกลางถูกเรียกว่า Hidden Layer โดยจากรูปจะเห็นได้ว่า Neural Network มีการไหลของข้อมูลจาก Input เพียงทิศทางเดียว และไม่มี Cycle อยู่ใน Network เลย ซึ่งจะเป็นแบบที่ถูกเรียกว่า Feed Forward Neural Network



รูปที่ 3-1 Feed Forward Neural Network

โดย Hidden Layer ทำให้ Neural Network สามารถจดจำ รูปแบบ ต่างๆ ได้มากขึ้น แต่การเพิ่มขนาดของ Hidden Layer ก็สามารถทำให้ Neural Network ทำงานได้ดีขึ้น แต่อาจทำให้เกิด Overfitting ได้

โดยการสร้าง Feed Forward Neural Network จะต้องมีการกำหนดสิ่งต่างๆ ดังต่อไปนี้

- Unit แต่ละ Unit มีพฤติกรรมอย่างไร อะไรคือ Activation Function
- Topology ของ Neural Network เป็นอย่างไร
- Neural Network ใช้ Algorithm อะไรในการเรียนรู้

Unit ของ Neural Network

แบ่งเป็น 2 ส่วน คือ Combination Function ซึ่งจะรวม Input ต่างๆ เข้ามาเป็นค่าเดียว และ Input แต่ละตัวจะมีค่าถ่วงน้ำหนักของมันเอง ซึ่ง Combination Function ที่นิยมคือ Weighted Sum ซึ่งคือการที่ Input แต่ละตัวถูกคูณด้วยค่าถ่วงน้ำหนักของมัน จากนั้นจึงนำผลคูณแต่ละตัวที่ได้นั้นมาบวกกัน

ส่วนที่สองคือ Transfer Function ซึ่งจะทำหน้าที่แปลงค่าที่ได้จาก Combination Function ไปเป็น Output ของ Unit ตัวอย่างของ Transfer Function ได้แก่ Sigmoid Function, Hyperbolic Tangent Function เป็นต้น

Feed Forward Neural Network

เป็น Topology หนึ่งของ Neural Network โดยแบ่งเป็น 3 Layer ดังที่ได้กล่าวไปแล้ว ซึ่งคือ Input Layer, Hidden Layer, Output Layer ตามลำดับ ซึ่งโดยทั่วไปแล้ว Hidden Layer เพียง 1 Layer ก็เพียงพอแล้ว ซึ่งถ้า Hidden Layer มีจำนวนมากเกินไป ก็จะทำให้เกิดการจดจำ (memorize) มากกว่า การเรียนรู้ (generalize) ดังนั้นจึงไม่ควรตั้งขนาดของ Hidden Layer ที่เยอะเกินไป

ซึ่ง Hidden Layer และ Output Layer นั้นจะมีทางเข้าด้านบนอีกหนึ่งทางเข้า ซึ่งเป็น Input คงที่เท่ากับ 1 เสมอ และมีค่าถ่วงน้ำหนักด้วย จุดประสงค์เพื่อทำการ Bias ซึ่งจะช่วยให้ Neural Network เข้าใจรูปแบบ ต่างๆ ได้ดีขึ้น

Neural Network เรียนรู้โดยการใช้ Back Propagation

การ Train Neural Network คือกระบวนการในการหาค่าถ่วงน้ำหนักที่ดีที่สุดสำหรับ Input แต่ละตัว โดยมีเป้าหมายอยู่ที่การใช้ชุด Train เพื่อสร้างค่าถ่วงน้ำหนักที่ทำให้ Output ของ Neural Network ที่ค่าใกล้เคียงกับ Output ที่ต้องการมากที่สุด และเป็นจำนวนหลายตัวอย่างที่สุดเท่าที่จะทำได้ Algorithm ที่นิยมใช้ในการ Train คือ Back Propagation ซึ่งมี 3 ส่วนหลักๆ คือ

- ขั้นตอนการทำ Feed Forward Pattern ของ Input ที่ใช้ในการ train
- ขั้นตอนการส่งค่าความผิดพลาดย้อนกลับ (Backpropagation of error)
- ขั้นตอนการปรับค่าถ่วงน้ำหนัก

Activation Function

เป็น function ที่ควรจะเป็น Function ต่อเนื่อง สามารถทำการ Diff ได้ และควรเป็นฟังก์ชั่นเพิ่ม นอกจากนั้นเพื่อให้การคำนวณเป็นไปอย่างมีประสิทธิภาพ ก็ควรมี ค่าอนุพันธ์ (Derivative) ที่ง่ายต่อการคำนวณ ซึ่งตัวอย่างของ Activation Function คือ Binary Sigmoid Function ซึ่งมี Range อยู่ช่วง (0,1)ซึ่งนิยามได้ดังนี้

$$f(x) = \frac{1}{1-e^{-x}}$$

$$f'(x) = f(x)[1-f(x)]$$

และ Bipolar Sigmoid Function ซึ่งมี Range อยู่ในช่วง (-1, 1) ซึ่งนิยามได้ดังนี้

$$f(x) = \left(\frac{2}{1+e^{-x}}\right) - 1$$

$$f'(x) = \frac{1}{2}[1+f(x)][1-f(x)]$$

Algorithm การ Train

เราสามารถนำเอา Activation Function ที่ได้กล่าวไปนั้น มาใช้ใน Algorithm Back Propagation มาตรฐานที่ให้ไว้ในหัวข้อนี้ได้ ซึ่งมีขั้นตอนการทำงานดังนี้

ขั้นตอน 0 กำหนดค่าถ่วงน้ำหนักโดยการสุ่มค่า

ขั้นตอน 1 ในขณะที่เงื่อนไขการ Train เป็นเท็จ ทำขั้นตอนที่ 2 – 9

ขั้นตอน 2 สำหรับแต่ละคู่การ Train ของค่า Input และค่าเป้าหมาย (Training Pair) ทำขั้นตอนที่ 3 ถึง 8

Feed Forward:

ขั้นตอน 3 แต่ละ Input Unit รับสัญญาณ Input และกระจายสัญญาณนี้ไปยังทุก Unit ใน Layer ถัดไป ทุก Unit

ขั้นตอน 4 แต่ละ Hidden Unit หาค่า Input ตามสูตร

$$Z\_in_j = v_{0j} + sum(x_i v_{ij})$$

ใช้ Activation Function คำนวณหาค่า Output

$$Z_j = f(z\_in_j)$$

และส่งสัญญาณนี้ไปยังทุก unit ใน Layer ถัดไป

ขั้นตอน 5 แต่ละ Output Unit หาค่า Input สุทธิตามสูตร

$$Y\_in_k = w_{0k} + sum(z_j w_{jk})$$

และใช้ Activation Function หาค่าสัญญาณ Output

$$Y_k = f(y\_in_k)$$

Back Propagation

ขั้นตอน 6 แต่ละ Output Unit รับค่าเป้าหมายที่สอดคล้องกับ Input Patterm นั้น แล้วคำนวณหาเทอมข้อมูลของค่าความผิดพลาด δ ตามสูตร

$$\delta_k = (t_k - y_k)f'(y\_in_k)$$

คำนวณค่าถ่วงน้ำหนักที่จะต้องปรับเปลี่ยน (Weight Correction Term) เพื่อใช้ในการ update w_{jk} ตามสูตร

$$\Delta w_{jk} = \alpha \delta_k z_j$$

คำนวณค่า Bias ที่ต้องปรับเปลี่ยน (Bias Correction Term) เพื่อใช้ในการ update w_{0k} ตามสูตร

$$\Delta w_{0k} = \alpha \delta_k$$

และส่ง δ_k ไปยังทุก Unit ใน Layer ที่ต่ำกว่าติดกัน

ขั้นตอน 7 แต่ละ Hidden Unit ทำการรวมค่าเดลต้าที่ได้รับจาก Unit ใน Output Layer

$$\delta\_in_j = sum(\delta_k w_{jk})$$

แล่วคูณด้วยอนุพันธ์ของ Activation Function เพื่อคำนวณหาค่าเทอมข้อมูลของค่าความผิดพลาด

$$\delta_j = \delta\_in_j f'(z\_in_j)$$

คำนวณค่าถ่วงน้ำหนักที่จะต้องปรับเปลี่ยน เพื่อใช้ในการ update v_{ij} ตามสูตร

$$\Delta v_{ij} = \alpha \delta_j x_i$$

คำนวณค่า Bias ที่ต้องปรับเปลี่ยน เพื่อใช้ในการ update v_{0j} ตามสูตร

$$\Delta v_{0j} = \alpha \delta_j$$

ปรับเปลี่ยนค่าถ่วงน้ำหนัก และ bias

ขั้นตอน 8 แต่ละ Output Unit Update Bias และค่าถ่วงน้ำหนัก ตามสูตรดังนี้

$$w_{jk}(new) = w_{jk}(old) + \Delta w_{jk}$$

ขั้นตอน 9 ตรวจสอบเงื่อนไขการหยุดเทรน

การปรับเปลี่ยนค่าถ่วงน้ำหนักโดยใช้โมเมนตัม

ในวิธีการ Back Propagation ที่มีการใช้โมเมนตัม การเปลี่ยนแปลงค่าถ่วงน้ำหนักจะอยู่ในทิศทางซึ่งเป็นผลรวมค่า gradient ในปัจจุบัน กับค่า Gradient ก่อนหน้านั้น วิธีการปรับเปลี่ยนค่าถ่วงน้ำหนักแบบนี้จะให้ผลดีเมื่อมีข้อมูลสำหรับการเทรนที่แตกต่างจากข้อมูลสำหรับการเทรนที่แตกต่างจากข้อมูลส่วนใหญ่มากๆ (และอาจเป็นข้อมูลที่ไม่ถูกต้อง) นอกจากนี้การลู่เข้าสู่ผลลัพธ์ที่ต้องการก็อาจจะเร็วขึ้นได้เมื่อมีการใช้เทอมโมเมนตัม

สูตรที่ใช้ในการปรับเปลี่ยน โมเมนตัม คือ

$$w_{jk}(t+1) = w_{jk}(t) + \alpha\delta_k z_i + \mu[\, w_{jk}(t) - w_{jk}(t-1)] \,,$$

หรือ $$\Delta w_{jk}(t+1) = \alpha\delta_k z_i + \mu\Delta w_{jk}(t)$$

และ

$$v_{jk}(t+1) = v_{ij}(t) + \alpha\delta_j z_i + \mu[\, w_{ij}(t) - w_{ij}(t-1)] \,,$$

หรือ $$\Delta w_{ij}(t+1) = \alpha\delta_j z_i + \mu\Delta w_{ij}(t)$$

เมื่อพารามิเตอร์โมเมนตัม μ มีค่าอยู่ในช่วง (0,1) การใช้โมเมนตัมจะทำให้ Neural Network ปรับเปลี่ยนค่าถ่วงน้ำหนักได้น้อยลงเมื่อพบ รูปแบบ ที่มีรูปแบบไม่สอดคล้องกับ รูปแบบ ส่วนใหญ่

สัญญาณ output สุทธิ (net input) ที่ Y_k รับเข้ามาจะแทนด้วย y_in_k :

$$y\_in_j = w_{0k} + \text{sum}(z_j w_{jk})$$

สัญญาณ output (activation) ของ Y_k จะแทนด้วย y_k:

$$y_k = f(y\_in_k)$$

สัญลักษณ์ที่ใช้

x เวกเตอร์ของ Input : $X = (x_1, \dots, x_i, \dots, x_n)$

t เวกเตอร์ของค่าเป้าหมาย : $T = (t_1, \dots, t_i, \dots, t_n)$

δ_k ข้อมูลเกี่ยวกับค่าความผิดพลาดของ Unit Y_k ที่ถูกส่งย้อนกลับไปยัง Hidden Unit ที่เชื่อมอยู่กับ Unit Y_k ใช้ในการปรับค่าถ่วงน้ำหนัก w_{jk}

δ_j ข้อมูลเกี่ยวกับค่าความผิดพลาดของ Unit Z_j ที่ถูกส่งย้อนกลับไปยัง Input Unit ที่เชื่อมอยู่กับ Unit Z_j ใช้ในการปรับค่าถ่วงน้ำหนัก v_{ij}

α อัตราการเรียนรู้ (Learning Rate)

X_i Input unit i

V_{0j} ค่า Bias ของ Hidden Unit l

Z_j Hidden unit j:

สัญญาณ input สุทธิ (net input) ที่ Z_j รับเข้ามาจะแทนด้วย z_in_j :

$$z\_in_j = v_{0j} + sum(x_i v_{ij})$$

สัญญาณ output (activation) ของ Z_j จะแทนด้วย Z_j:

$$z_j = f(z\_in_j)$$

w_{0k} ค่า Bias ของ output unit k

Y_k output unit k:

Genetic Algorithm

Genetic Algorithm เป็น Algorithm ที่อิงกับธรรมชาติของการพัฒนาโครโมโซม ซึ่งแบ่งออกเป็น 2 ลักษณะการพัฒนาคือ Cross-Over และ Mutate ซึ่ง solution นี้มักใช้กับงานที่มี Domain ของคำตอบใหญ่มากๆ และต้องการคำตอบที่ near-Optimum โดยตัวอย่างของปัญหาดังกล่าว คือ การหาจุดสูงสุดของ Graph ด้านล่าง

รูปที่ 3-2 แสดง Graph ที่นำ Genetic Algorithm ไปใช้หาจุดสูงสุด

โดย Algorithm มีขั้นตอนการทำงานดังนี้

1. **[Start]** สร้าง population ซึ่งมี n chromosomes โดยวิธี
2. **[Fitness]** ประเมินความมีค่าของแต่ละ Chromosome ด้วย fitness *f(x*
3. **[New population]** สร้าง population ใหม่โดยทำ step ด้านล่างซ้ำไปเรื่อยๆ จนกว่าจะได้ population ที่สมบูรณ์
    1. **[Selection]** เลือก 2 parent chromosomes จาก population ตามค่า Fitness ที่แต่ละ Chromosome มีโดยอาศัยพื้นฐานที่ว่า Fitness ยิ่งสูง ยิ่งมีโอกาสถูกเลือกมากขึ้น
    2. **[Crossover]** ทำการ crossover จาก 2 parent จนเกิดลูกซึ่งเป็นผลจากการ crossover เมื่อ probability ในการ crossover สูงกว่าที่กำหนด ถ้าไม่แล้ว ให้ลูกมีคุณสมบัติเดียวกับ parent ทุกประการ
    3. **[Mutation]** ทำการ mutate ให้เกิด offspring เมื่อ probability ในการ mutation สูงกว่าที่กำหนด
    4. **[Accepting]** ทำการวาง offspring ที่เกิดมาใหม่ใน population
4. **[Replace]** ใช้ population ใหม่นี้ในการทำงานต่อไป
5. **[Test]** ถ้าเงื่อนไขในการจบการทำงานนั้นถูกต้องให้จบการทำงาน ถ้าไม่แล้วให้ทำงานจนกว่าจะเกิด population ที่ดีที่สุด
6. **[Loop]** ไป step 2

การทำ Cross-Over นั้นมีวิธีการทำที่เรียกว่า Roulette-Wheel Method ซึ่งจะนำ fitness ของข้อมูลทั้งหมด มาเฉลี่ย เพื่อสร้างเป็นวงล้อ แล้ว Random วิ่งไปตามล้อเรื่อยๆ แล้วค่อยๆ ดึงข้อมูลออกมา ซึ่งการทำงานดังกล่าว จะทำให้ ข้อมูลที่ดี จะอยู่รอด ซึ่งเป็นไปตามกฎทางชีววิทยาของชาลส์ ดาร์วิน ซึ่งจะเป็นการพัฒนาคุณภาพของโครโมโซม

ส่วนลักษณะการ Cross-Over มีลักษณะต่างๆ ดังนี้

Single Point Cross-Over คือการที่ส่วนต้นของลูกนั้นเอามาจาก parent 1 และส่วนที่เหลือนั้นเอามาจาก parent อื่นๆ เช่น

11001001 + 11011**111** = **11001111**

Two Point Cross-Over คือการที่ cross-over point สองจุดนั้นถูกเลือกขึ้นมา ดังนั้น ลูกจะมีลักาณะเหมือนกับ parent 1 จำนวน สอง ส่วน และที่เหลือ จะเป็นไปตาม parent ที่เหลือ เช่น

11001**011** + **11**011**111** = **11011111**

Uniform Cross Over คือลักษณะของ parent สืบทอดไปยัง ลูก อย่างสุ่ม (Randomly) เช่น

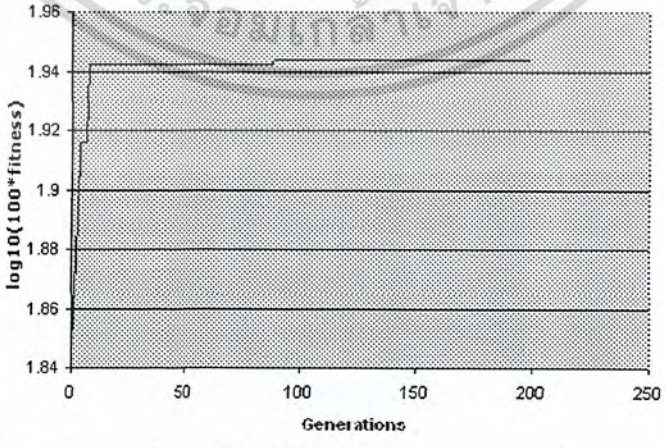
11001**011** + **11**011**101** = 11011111

Arithmetic Cross Over คือการสร้าง ลูกจาก Arithmetic Operation เช่น

11001011 + 11011111 = 11001001 (AND)

ส่วนการทำ mutation นั้น มักจะทำขึ้นโดยการ Random ซึ่งเป็นการเปลี่ยนลักษณะทางพันธุกรรมของคำตอบไปอย่างชัดเจน เพื่อหลีกเลี่ยง Local Maxima เช่น

11001001 => **1**0001001



รูปที่ 3-3 แสดงกราฟความก้าวหน้าของ Fitness รวมของแต่ละ Generation

ค่าที่ดีที่สุดของ Population จะเป็นไปตามลักษณะดังกล่าว โดยจะสังเกตได้ว่า จำนวน Generation นั้นบ่งบอกถึง ความดีขึ้นเรื่อยๆ ของคำตอบ

Parameter Of Genetic Algorithm

Cross Over Rate เป็น ค่าความน่าจะเป็นในการเกิด Cross Over เช่น 80% – 90%

Mutation Rate เป็น ค่าความน่าจะเป็นในการเกิด Mutation เช่น 0.05% - 1%

Population Size เป็นขนาดของ set of solution เช่น 20 - 30

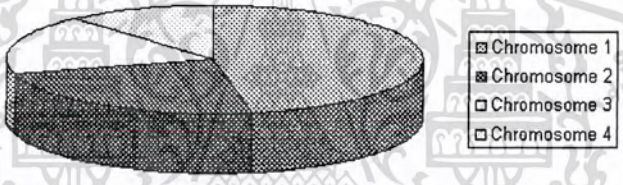
Selection เป็น method ในการเลือกคู่เพื่อทำการ cross-over เช่น Roulette Wheel Selection

Encoding

Cross Over and Mutation Type

Selection Method

Roulette Wheel Method



รูปที่ 3-4 Roulette Wheel Method

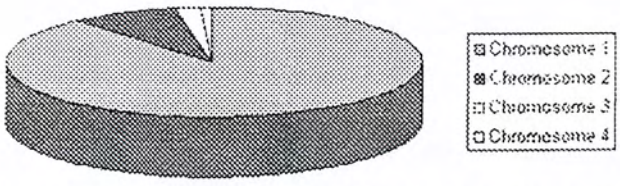
เป็นวิธีในการเลือกคู่ในการทำ Cross-Over โดยมี Algorithm ดังนี้

1. **[Sum]** คำนวณผลรวมของ Fitness ทั้งหมดใน population
2. **[Select]** สร้าง random number ในช่วง ***(0,S) - r.***
3. **[Loop]** วิ่งผ่านผลรวมของ Fitness ใน population ทั้งหมดหมุนไปเรื่อยๆ จนกว่าจะไปตกในตัวเลขเดียวกับที่ random ขึ้นมา

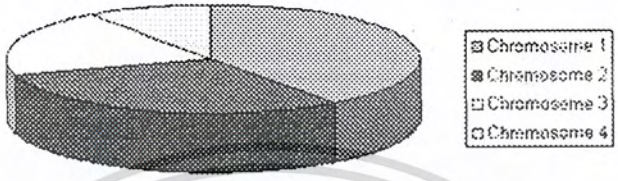
ซึ่งการทำงานดังกล่าว Chromosome ที่มี Fitness สูงกว่า ก็จะมีโอกาสในการถูกเลือกที่สูงกว่า

Rank Selection

เป็นวิธีที่นำมาแก้ปัญหาในบางจุดของ Roulette Wheel Selection เช่น เมื่อ Chromosome หนึ่งๆ มีค่า Fitness สูงกว่า Chromosome ที่สูงกว่า Chromosome อื่นมากๆ ซึ่งเมื่อทำมาเป็น Roulette Wheel แล้วเป็น 80 – 90% เป็น Wheel ทำให้ Chromosome อื่น ไม่มีโอกาสในการถูกเลือก Rank Selection จึงได้เกิดขึ้นมาเพื่อแก้ปัญหาดังกล่าว โดย นำ Fitness มาเรียงตามลำดับน้อยไปหามากแล้ว ทำ Fitness แต่ละตัวใหม่ โดยตัวที่น้อยที่สุด มีค่า 1, 2 ไล่ไปเรื่อยๆ จนถึงตัวที่มีค่าสูงสุด และกระทำการเลือก ตามลักษณะเหมือนกับ Roulette Wheel



รูปที่ 3-5 สถานการณ์ของ Fitness ต่างๆ ก่อนการทำ Ranking Selection



รูปที่ 3-6 สถานการณ์ของ Fitness ต่างๆ หลังการทำ Ranking Selection

ตัวอย่าง Application ของ Genetic Algorithm

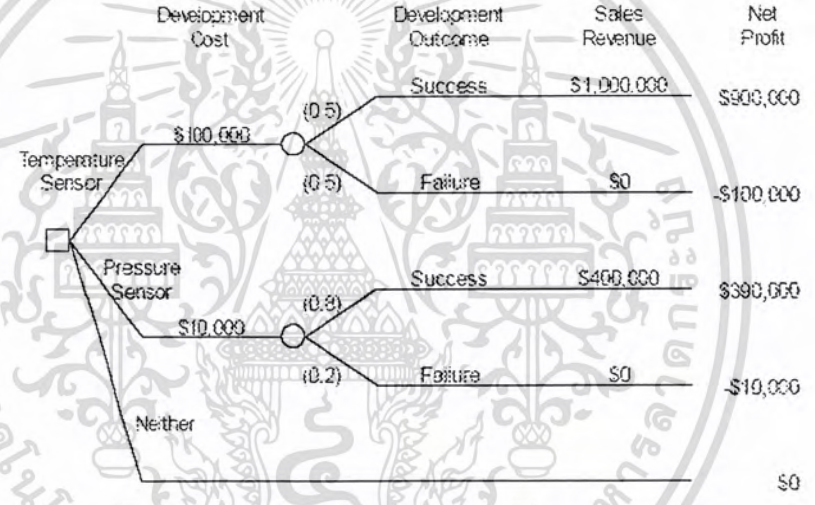
- Nonlinear dynamical systems – การคาดการณ์ และการวิเคราะห์ข้อมูล
- ออกแบบ neural networks, ทั้ง architecture และ weights
- เส้นทางในการเดินทางของหุ่นยนต์
- Evolving LISP programs (genetic programming)
- Strategy planning
- หารูปร่างของ protein molecules
- TSP and sequence scheduling
- Functions ในการสร้างรูปภาพ

Decision tree

เป็นการแบ่งกลุ่มของข้อมูลออกมาเป็น Data Structure รูปต้นไม้ ซึ่งถูกมองว่า เป็นการแบ่ง segmentation ของ original dataset ซึ่งการแบ่ง segment นั้นต้องแบ่งโดยมีจุดประสงค์เฉพาะเจาะจง

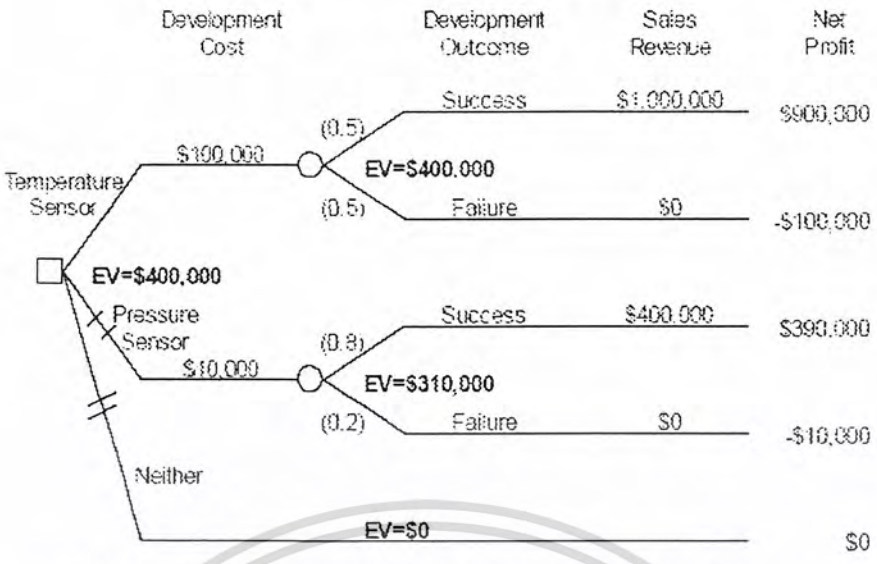
การใช้ Decision Tree สามารถใช้ได้ใน Exploration และ Data Preprocessing ซึ่งสามารถเป็นการ parse ขั้นต้นให้กับ neural networks, nearest neighbor and normal statistics routine และสามารถใช้ Decision Trees สำหรับการ Prediction เช่น การสร้าง predictive model สำหรับแต่ละสินค้า

ตัวอย่างการใช้งาน เช่น เมื่อบริษัทขายสัญญาณมือถือต้องการทำ promotion ส่งเสริมการขาย โดยการดูว่า ลูกค้ากลุ่มบน นั้นมีรายได้มากขึ้น ซึ่งสมควรส่งเสริมการขาย หรือดูว่าลูกค้ากลุ่มล่างนั้นมีโอกาสในการขยายจำนวนมากขึ้น จึงสมควรส่งเสริมการขายให้มากขึ้นหรือไม่



รูปที่ 3-7 ตัวอย่าง Decision Tree

ดังรูปที่ 3-7 เป็นตัวอย่างหนึ่งของ Decision Tree ซึ่งอ่านจากซ้ายไปขวา Node ที่อยู่ซ้ายสุด ถูกเรียกว่า root node ซึ่งกิ่งก้านสาขาต่างๆ นั้นถูกกระจายจากบนสุดของต้นไม้ ซึ่งคือ root node ลงไปเรื่อยๆ ซึ่งกิ่งก้านดังกล่าว ถูกกระจายลงมาตามลักษณะของการตัดสินใจ ซึ่งวงกลมใน tree เรียกว่า chance node ตัวเลขที่อยู่ในวงเล็บข้างๆ นั้นคือ โอกาสในการเกิดขึ้นของกิ่งนั้นๆ และส่วนขวาสุดของรูป ถูกเรียกว่า endpoint ซึ่งแต่ละ endpoint แสดงถึง แต่ละผลลัพธ์ของ ลำดับการตัดสินใจที่เกิดขึ้น



รูปที่ 3-8 ตัวอย่าง Decision Tree สำหรับเหตุการณ์ที่ตัดสินใจได้ไม่แน่นอน

Expected Value สำหรับเหตุการณ์ที่ตัดสินใจได้ไม่แน่นอน ถูกคำนวณโดย คูณ แต่ละ ผลลัพธ์ที่น่าจะเป็นไปได้ ด้วยค่าความน่าจะเป็นในการเกิดขึ้น แล้วพวกกันทั้งหมด เช่น Pressure Sensor Expected Value = (0.8 x 390,000) + (0.2 x -10,000) = 310,000 เป็นต้น และในกรณีที่มีหลายๆ ทางเลือก ก็จะเลือกอันที่มี Expected Value สูงที่สุด เช่น Expected Value ของ root node คือ 400,000 ซึ่งเกิดจากการหาค่าที่มากที่สุดของ 400,000$, 310,000$, 0$ ซึ่งในการเลือกทั่วไปแล้ว ให้เลือกค่าที่มากที่สุด แต่ในกรณีที่เลือกต้นทุนให้เลือกค่าที่น้อยที่สุด

การคำนวณหาค่า Expected Value จาก node ล่างสุดขึ้นมา เป็น Bottom-Up Methodology นั้นจะถูกเรียกว่า Decision Tree Rollback และลำดับของการตัดสินใจเป็นขั้นๆ จนกว่าจะถึงคำตอบจะถูกเรียกว่า Decision Strategy

Certainty Equivalent คือการที่ 2 สิ่งมีค่า Expected Value เท่ากัน

Risk Attitude: ถ้า Certainty Equivalent ในเทอมของกำไรนั้นมีค่าน้อยกว่า กำไรที่ตั้งเป้าไว้ นั้นจะถูกเรียกว่า risk averse และถ้ามีค่าเท่ากันจะเรียกว่า risk neutral และถ้ามีค่ามากกว่าจะเรียกว่า risk seeking

Utilities Function นั้นจะประมวลผลออกมาเป็นตัวเลขที่คำนวณจาก Certainty Equivalent ใช้สำหรับการเลือกว่า สมควรจะเสี่ยงหรือไม่

CART – Growing a forest and picking the best tree

CART ย่อมาจาก Classification and Regression Trees เป็น algorithm ในการทำ Data Exploration และ Data Prediction

วิธีการคือ การทำให้ tree มีความซับซ้อนมากๆ แล้ว prune ต้นไม้กลับให้เป็น optimally general tree

CHAID เป็น Algorithm ที่คล้ายๆ กับ CART แต่ต่างกันที่วิธีในการ split ข้อมูล

Clustering

Clustering of Data เป็นวิธีในการจัดกลุ่มข้อมูลที่เหมือนๆ กันให้อยู่ในกลุ่มที่เล็กลง เช่น การรวมกลุ่มลูกบอลสีเดียวกัน เป็นต้น

Clustering Algorithm นั้นเป็นวิธีการในการหากลุ่มที่เหมือนกันของข้อมูล หนึ่งในวิธีนั้นคือการหา centroid ของข้อมูล ซึ่งหมายถึง จุดที่มีค่ากลางของกลุ่มหนึ่งๆ แล้วจึงทำการหาสมาชิกของกลุ่มนั้นๆ โดยใช้ centroid เป็นตัวกลางในการหาแล้วใช้วิธีการทางสถิติยศาสตร์เข้าช่วย

Clustering Algorithm นั้นถูกใช้อย่างแพร่หลายตัวอย่างเช่น Pattern Recognition, Artificial Intelligence, survey of markets, survey of products, survey of sales program และ R&D เป็นต้น

Clustering คือกระบวนการในการแบ่งกลุ่มของข้อมูล หรือวัตถุให้เป็นกลุ่มย่อยที่มีความหมาย หรือที่เรียกว่า Cluster โดย Clustering เป็น unsupervised classification เนื่องจากไม่มีการ รู้ว่าแต่ละกลุ่มย่อยสื่อถึงอะไร ก่อนที่จะทำการ clustering ซึ่งจะตรงข้ามกับ Classification ซึ่งเป็น Supervised Classification

Clustering ที่ดีจะต้องได้ Cluster ที่มีคุณภาพสูง ซึ่งหมายถึง
Intra-class similarity ซึ่งคือความเหมือนกันภายใน Cluster สูง และ
Inter-class similarity ซึ่งคือความเหมือนกันภายนอก Cluster ต่ำ

คุณภาพของ Clustering นั้นถูกวัดโดยความ Function ที่ Represent ความเหมือนใน cluster และความแตกต่างภายนอก cluster

Clustering Techniques สามารถแบ่งได้ออกเป็น 4 กลุ่มย่อย ซึ่งคือ

Partitioning เป็นการแบ่ง cluster โดย partition criterion ซึ่งเมื่อได้กลุ่มของ cluster ก็จะแบ่งตามความเหมือนต่อ partition criterion ดังกล่าวออกเป็น cluster ย่อย

Hierarchy algorithm เป็นลำดับของ partition ซึ่งคือการที่แต่ละวัตถุ merge กันไปเรื่อยๆ จนกระทั่งสุดท้าย ได้ 1 cluster ใหญ่ ซึ่งคือ set ทั้งหมดของวัตถุ

Density-based เป็นการหาพื้นที่ของข้อมูลที่มีความหนาแน่นมากกว่า threshold ที่กำหนดจากแต่ละ cluster กล่าวคือเป็นการแบ่งข้อมูลตามความหนาแน่นข้อมูล

Grid-based เป็นการแบ่งพื้นที่ของข้อมูลออกเป็น cell ย่อยแล้ว แบ่งข้อมูลตาม cell data space นั้นๆ กล่าวคือเป็นการแบ่งเชิงพื้นที่

K-means Algorithm

เป็น unsupervised learning algorithm หนึ่งที่มีความง่าย และสามารถแก้ปัญหา clustering ได้ กระบวนการแก้ปัญหาคือการที่แบ่งกลุ่มของข้อมูลให้อยู่ใน set ที่ได้กำหนดขึ้น สมมติว่าเป็น k กลุ่ม โดยหลักการคือการกำหนด centroid k จุด แต่ละ centroid สำหรับแต่ละ cluster ซึ่งในทางปฏิบัติแล้วสมควร

แบ่งให้แต่ละ centroid อยู่ห่างกันมากที่สุด หลังจากนั้นให้จัดแต่ละจุดที่เหลือ เข้าแต่ละ set ที่ได้แบ่งขึ้นตาม centroid โดยอาศัยหลักการที่ว่ามีระยะทางต่ำสุดไปที่ centroid นั้นๆ หลังจากนั้นให้คำนวณ centroid ของแต่ละ set ขึ้นมาใหม่ และกระทำการเหมือนเดิมไปเรื่อยๆ ซึ่ง centroid ก็จะย้ายจุดไปเรื่อยๆ จนกระทั่ง centroid จะไม่ย้ายตำแหน่ง

โดยวิธีการดังกล่าว คือ

1. กำหนด k จุด โดยแต่ละจุดเป็น centroid แต่ละ cluster
2. กำหนดให้แต่ละจุดไปอยู่ในแต่ละกลุ่มโดยอาศัยหลักการที่ว่าอยู่ใกล้ centroid นั้นๆ ที่สุด
3. เมื่อทุกๆ จุดมีกลุ่มอยู่แล้ว ก็ให้คำนวณหา centroid อันใหม่
4. ทำขั้นตอนที่ 2 และ 3 ไปเรื่อยๆ จนกระทั่ง centroid นั้นไม่ขยับแล้ว

Hierarchical Clustering

Example Minimal Spanning Tree Method

จัดกลุ่มของอาหารตามตารางด้านล่าง

Food item #	**Protein content, P**	**Fat content, F**
Food item #1	1.1	60
Food item #2	8.2	20
Food item #3	4.2	35
Food item #4	1.5	21
Food item #5	7.6	15
Food item #6	2.0	55
Food item #7	3.9	39

ตารางที่ 3-1 ตัวอย่างข้อมูลที่จะทำการจัดกลุ่ม

ซึ่งขั้นแรกจะเป็นการหา centroid ของข้อมูลโดยขั้นแรก สมมติฐานว่า ทุกๆ ข้อมูลยังเป็น centroid อยู่

ขั้นที่สองทำการหา ระยะทาง ระหว่างข้อมูลทุกๆ ความเป็นไปได้ ตาม Euclidean metric ซึ่งมีสูตรการคำนวณดังนี้

$$d = \sqrt{\sum_{i=1}^{k} (p_i - q_i)^2}$$

โดย d คือระยะทาง

P และ q คือข้อมูลของทั้งสองจุด ในทุกๆ แง่มุม จำนวนแง่มุม = k

ซึ่งในขั้นตอนที่สองจะนำสมการดังกล่าว มาใช้ในการหาระยะทางของทุกๆ คู่ เช่น ระหว่าง item#1,#2

$$d_{c1,c2} = \sqrt{(1.1-8.2)^2+(60-20)^2}$$

$$= \sqrt{(-7.1)^2+(40)^2}$$

$$= \sqrt{50.41+1600}$$

$$= \sqrt{1650.41}$$

$$= 40.62$$

ก็จะได้ตาราง ดังนี้

Cluster number	C1	C2	C3	C4	C5	C6	C7
C1	0	40.62	25.19	39.00	45.46	5.08	21.18
C2	known	0	15.52	6.77	5.03	35.54	19.48
C3	known	known	0	14.25	20.28	20.12	4.01
C4	known	known	known	0	8.55	34.00	18.19
C5	known	known	known	known	0	40.39	24.28
C6	known	known	known	known	known	0	16.11
C7	known	known	known	known	known	known	0

ตารางที่ 3-2 ผลลัพธ์จากการหาระยะทางของทุกๆ คู่

ซึ่งจะเห็นว่า จุดที่มีระยะทางน้อยที่สุด คือ C3 และ C7 ซึ่งมีค่า 4.01 ก็ทำการรวม 2 จุดดังกล่าว เข้าด้วยกัน โดย centroid คือ P = (4.2 + 3.9) / 2 และ F = (35 + 39) / 2

Cluster number	Protein content, P	Fat content, F
C1	1.1	60
C2	8.2	20
C37	4.05	37
C4	1.5	21

C5	7.6	15
C6	2.0	55

ตารางที่ 3-3 ผลลัพธ์ที่ได้จากการรวม Clusterที่ 3 และ 7

Step 3 ทำ Step 2 ไปเรื่อยๆ จนกว่าจะมีจำนวน cluster = 4 ซึ่งก็จะคำนวณหา Euclidean Table ใหม่ คือ

Cluster number	C1	C2	C37	C4	C5	C6
C1	0	40.62	23.18	39.00	45.46	5.08
C2	known	0	17.49	6.77	5.03	35.54
C37	known	known	0	16.20	22.28	18.11
C4	known	known	known	0	8.55	34.00
C5	known	known	known	known	0	40.26
C6	known	known	known	known	known	0

ตารางที่ 3-4 เป็นการหาผลต่างที่น้อยที่สุดระหว่าง แต่ละ Cluster

และทำการรวมจุด ใกล้สุด ซึ่งคือ C2 และ C5 ก็ทำการรวมสองจุดดังกล่าว ก็จะเป็นตารางดังรูปด้านล่าง

Cluster number	Protein content, P	Fat content, F
C1	1.1	60
C25	7.9	17.5
C37	4.05	37
C4	1.5	21
C6	2.0	55

ตารางที่ 3-5 ผลลัพธ์ที่ได้จากการรวม Clusterที่ 2 และ 5

ซึ่งเมื่อทำไปเรื่อยๆ จะได้คำตอบเป็น

Cluster number	Protein content, P	Fat content, F
C16	1.55	57.50
C25	7.9	17.5

C37	4.05	37
C4	1.5	21

ตารางที่ 3-6 ผลลัพธ์สุดท้ายจากการทำ Clustering

ซึ่งเป็นกลุ่มของ Cluster ที่ถูก group แล้วเหลือ 4 กลุ่มด้วยกัน

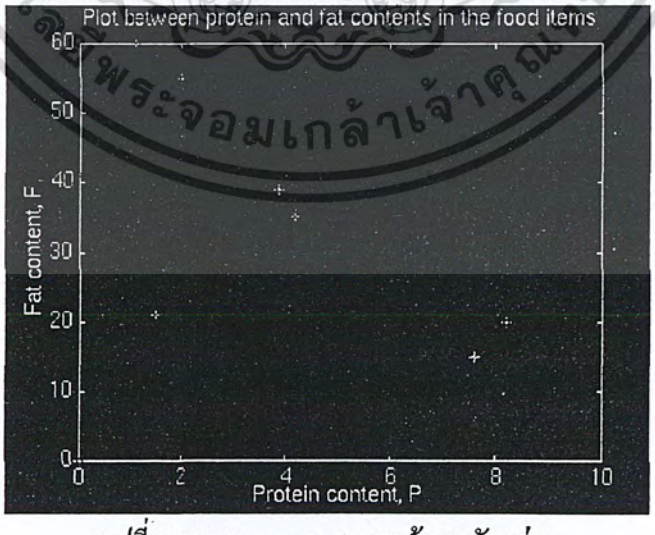
Ex K-Mean Clustering Problem

ตัวอย่างของข้อมูลเริ่มต้น คือ

Food item #	Protein content, P	Fat content, F
Food item #1	1.1	60
Food item #2	8.2	20
Food item #3	4.2	35
Food item #4	1.5	21
Food item #5	7.6	15
Food item #6	2.0	55
Food item #7	3.9	39

ตารางที่ 3-7 ตัวอย่างข้อมูลเริ่มต้นในการทำ K-Mean Clustering

เพื่อความง่ายในการเลือกจุดที่ห่างที่สุดระหว่างกันก็ plot graph ออกมาเป็นดังรูปที่ 3-9



รูปที่ 3-9 การ plot graph จากข้อมูลตัวอย่าง

ซึ่งจะเห็นได้ว่า จุด 1 กับ 2, 1 กับ 3, 1 กับ 4, 1 กับ 5, 2 กับ 3, 2 กับ 4, 3 กับ 4 นั้นมีค่ามากที่สุด ซึ่งก็จะเลือก 4 centroid ออกมาคือ

Cluster number	Protein content, P	Fat content, F
C1	1.1	60
C2	8.2	20
C3	4.2	35
C4	1.5	21

ตารางที่ 3-8 แสดงผลจากการเลือก 4 centroid

ซึ่งเราก็มองว่า จุดที่ 1 ใกล้กับจุดที่ 6 มากที่สุด จึงรวมสองจุดนี้เข้าด้วยกัน แล้วหา centroid ใหม่ รวมทั้งจุด 2 กับ 5 และ จุด 3 กับ 7 ซึ่งใกล้กันมากที่สุดด้วย จึงรวมเข้าด้วยกันเป็นตารางด้านล่าง

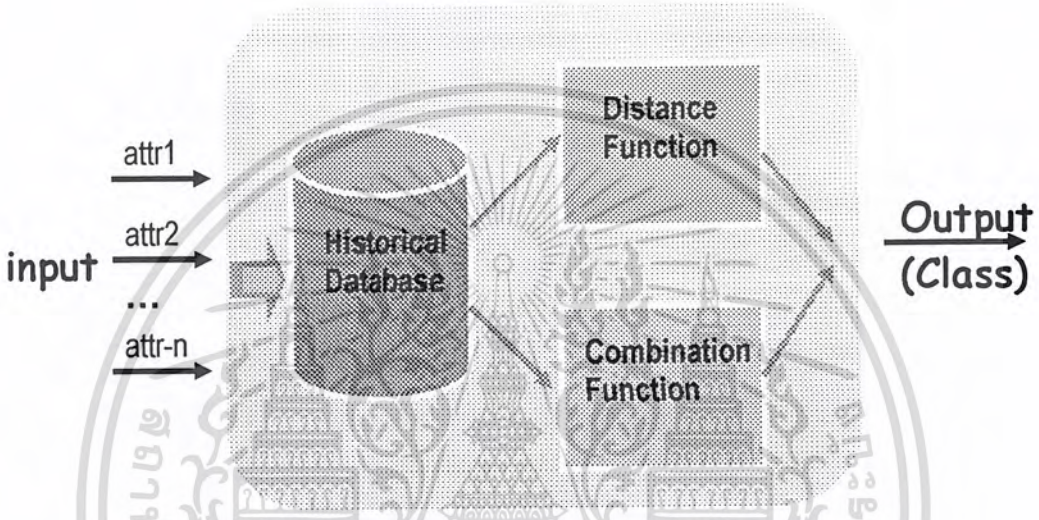
Cluster number	Protein content, P	Fat content, F
C16	1.55	57.50
C25	7.9	17.5
C37	4.05	37
C4	1.5	21

ตารางที่ 3-9 แสดงผลจากการรวมจุดที่ 1,6 จุดที่ 2,5 จุดที่ 3,7

Memory - Based Reasoning (MBR)

โดยอัลกอริธึมนี้ จะเรียนรู้เกี่ยวกับเทคนิคการจัดเป็นประเภทๆและคาดเดาค่าทางตัวเลขซึ่งมีพื้นฐานจาก ประสบการณ์ตรงหรือฐานข้อมูลเก่า โดย

- จะถือว่ากรณีที่ใกล้เคียงกันหลายๆกรณีจาก record ที่ผ่านการทำแล้ว(neighbor)เป็นกรณีเดียวกัน
- หลังจากนั้น ก็จะใช้ประโยชน์ในการจัดเป็นประเภทๆจากกรณีของ neighbor นี้กับการจัดเป็นประเภทๆของ record อื่นๆที่ยังไม่รู้



รูปที่ 3-10 แสดงถึงโครงสร้างของ Memory Based Reasoning

ขั้นตอนการทำงานของ MBR

1 เลือก training set

2 หา distance function เป็นการวัดค่าความใกล้เคียงกันระหว่าง object 2 object ที่ใกล้เคียงกันที่สุด ซึ่งจะมีค่าที่ห่างกันน้อยที่สุด

โดยมี 4 คุณสมบัติ คือ

- $d(A,B) \geq 0$
- $d(A,A) = 0$
- $d(A,B) = d(B,A)$
- $d(A,B) \leq d(A,C) + d(C,B)$

3 เลือกจำนวนของ neighbors ที่ใกล้ที่สุด

4 หา combination function ซึ่งถูกนำมาใช้ในกรณีที่ใช้ค่าที่ใกล้เคียงกันหลายๆค่า โดยมี 2 วิธี คือ

- Majority voting (democracy voting)
- Weighted voting (shareholder voting)

ข้อดี

- ได้ผลลัพธ์ที่เข้าใจได้โดยง่าย
- สามารถใช้ได้กับข้อมูลที่ไม่มีเกณฑ์ในการแบ่ง รวมไปถึงข้อมูลที่ไม่มีความสัมพันธ์กันเลย
- ง่ายในการนำไปใช้
- เหมาะสมในการนำไปใช้ประมาณค่าที่หายไป

ข้อเสีย

- ต้องใช้พื้นที่ในการเก็บ training set จำนวนมาก
- ผลลัพธ์จะขึ้นอยู่กับ distance function , combination function และ จำนวนของ neighbors

Market Basket Analysis

เป็นอัลกอริธึมที่ตรวจสอบรายการซื้อขายที่ยาวๆเพื่อที่จะกำหนดสินค้าที่ถูกซื้อพร้อมกันบ่อยๆ โดยจะใช้ชื่อของมัน(สินค้า)จากความคิดของคนใน supermarket ที่วางสินค้าทั้งหมดลงใน shopping cart(a market basket) ผลที่ได้จะเป็นประโยชน์กับบริษัทที่ขายสินค้า ซึ่งขายในร้าน เมนู catalog หรือขายตรงไปยังลูกค้า

Market Basket Analysis ถูกใช้กำหนดสินค้าที่ขายไปพร้อมกัน input ที่เข้ามาโดยปกติจะเป็นรายการซื้อขายซึ่งแต่ละคอลัมน์จะแสดงสินค้าและแต่ละแถว แสดงการขาย(จำนวนที่ขายได้) หรือลูกค้า ขึ้นอยู่กับเป้าหมายของการวิเคราะห์เพื่อที่จะหาสินค้าที่ขายด้วยกัน ณ เวลาเดียวกัน หรือลูกค้าคนเดียวกัน

ตัวอย่างเช่น การวิเคราะห์ market basket ของบันทึกการขายของ supermarket อาจแสดงให้เห็นว่าตะกร้ารถเข็นที่บรรจุชีสอยู่ ก็น่าจะมีหัวหอมคองบรรจุอยู่ด้วย พนักงานขายก็สามารถใช้ข้อมูลนี้ในการจัดชั้นวางของใหม่ หรือใช้เป็นเป้าหมายในการโฆษณาได้

Statistical Data Mining

การสรุปสถิติไม่ได้เป็นอัลกอริธึมของ machine learning แต่มันก็เป็นส่วนสำคัญในกระบวนการวิเคราะห์ กลไกที่ใช้สำรวจสถิติจะเก็บสถิติพื้นฐานเกี่ยวกับข้อมูล รวมไปถึงค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐานและความถี่ นอกจากนี้รายงานก็จะแสดงโดยแผนภูมิความถี่สำหรับแต่ละประเภท ข้อความ และ yes/no

Discriminate Analysis

การวิเคราะห์แบบแยกความแตกต่างเป็นกลไกการสำรวจอย่างหนึ่งซึ่งใช้เปรียบเทียบชุดข้อมูลกับข้อมูลที่มีอยู่ มันไม่ได้เป็นการแบ่งประเภทโดยตรง เพราะไม่ได้มี attribute ที่เป็นเป้าหมาย การแยกความแตกต่างจะตอบคำถามที่ว่าอะไรแบ่งแยกชุดข้อมูลที่ถูกเลือกจากชุดข้อมูลที่เหลือ ตัวอย่างเช่น การแยกความแตกต่างสามารถใช้ให้เป็นประโยชน์ขึ้นอยู่กับการเลือกมุมมองของจุดข้อมูลจากกราฟ การแยกความแตกต่างจะหากฎที่สามารถใช้คาดเดาได้ถ้าจุดข้อมูลที่ให้มาอยู่ในชุดข้อมูลที่ถูกเลือก หรืออยู่ที่อื่นในชุดข้อมูล

Find Dependencies

อัลกอริธึมในการหาเมืองขึ้น จะค้นพบความสัมพันธ์ในการตัดสินใจ , ความสัมพันธ์กันน้อย , ความสัมพันธ์ในหลายมิติ , ความสัมพันธ์ที่ไม่เป็นเชิงเส้นในข้อมูล ซึ่งมันยังรวบรวมคุณสมบัติในทางสถิติของความสัมพันธ์ที่พบด้วย

โดยจะช่วยแก้ปัญหาหลักๆดังนี้

- ค้นพบชุดของตัวแปรอิสระที่มีอิทธิพลหลักๆต่อตัวแปรที่เป็นเป้าหมายโดย
- จะกรองจุดที่ห่างไกลมากๆ ซึ่งไม่ match กับข้อมูลที่เหลือออก

Link Analysis

กลไกการสำรวจการวิเคราะห์การเชื่อมโยงจะแสดงและเสนอรูปแบบที่ซับซ้อนของความเกี่ยวพันกันระหว่างแต่ละค่าของ attribute ที่แบ่งประเภทและที่เป็น Boolean ทั้งหมด ผลของการวิเคราะห์จะถูกแสดงในรูปกราฟของการเชื่อมโยง object หลายๆตัว ผลลัพธ์ของวิธีนี้จะช่วยให้เข้าใจได้ดีขึ้นในโครงสร้างของข้อมูลที่ค้นหาซึ่งถูกซ่อนไว้และช่วยให้แยกรูปแบบที่น่าสนใจสำหรับการสืบหาต่อไปได้รวดเร็ว

ตัวอย่างของการใช้อัลกอริธึมนี้

ฐานข้อมูลการตลาด : แสดงให้เห็นลักษณะนิสัยพื้นฐานของลูกค้าประจำ ซึ่งจะแสดงถึงพฤติกรรมการซื้อ

การวิเคราะห์ในการสื่อสาร : แสดงให้เห็นรูปแบบการสื่อสารหลักและปัญหาbottleneckในเครือข่าย

บทที่ 4
การนำ Data Mining มาประยุกต์ใช้ในปัญหาทางธุรกิจ

Problem 1 หาว่าลูกค้าไม่ชอบสินค้า จากการ บริการ มี รูปแบบ อะไรบ้าง

Business Goal:

จัดลำดับชั้นความพึงพอใจต่อตัวสินค้าของลูกค้าในมุมมองทางด้านการบริการหลังการขาย โดยแบ่งเป็น level ได้ว่า ดีมาก, ดี, พอใจ, แย่, แย่มาก

Solution:

ใช้แบบสอบถามประจำปี

ใช้ Decision Tree ในการเรียนรู้ รูปแบบ ของการ บริการ ที่ลูกค้าไม่ชอบ โดย training set จะประกอบด้วย

Input: อัตราการ service,
อัตราการ service หลังติดตั้ง,
อัตราการ service ที่ไม่หายขาด,
จำนวนปีหลังติดตั้ง

Output: ตัวเลขประเมินความพึงพอใจของลูกค้า ตัวแปรคุณภาพ

สิ่งที่ได้จาก Decision Tree คือ criteria ว่า อัตราการ service, อัตราการ service หลังติดตั้ง เท่าไหร่ และ อัตราการ service ที่ไม่หายขาด เท่าไหร่ หรือ รูปแบบ อย่างไรที่ทำให้ ตัวแปรคุณภาพต่ำกว่าที่ต้องการ

Problem 2 หาว่า ลูกค้าซื้อสินค้าเพราะอะไร

Business Goal:

จัดกลุ่มของลูกค้าได้ว่า ซื้อสินค้าจากอะไร และขนาดของกลุ่มเป็นเท่าไหร่

Solution:

ใช้แบบสอบหลังติดตั้ง,แบบสอบถามประจำปี

พยายามสร้าง Segment

Input: ตราสินค้า, คุณภาพ, ประหยัดไฟ,
การฟอกอากาศ, เงียบ,
ความสวยงาม, ราคา, การบรกการหลังการขาย,
อื่นๆ

สิ่งที่ได้จาก Clustering แบบ Hierarchical Method โดยใช้ Minimal Spanning Tree คือ set ของตัวเลขประเมินราคา, คุณภาพ, ระบบฟอกอากาศ, ประหยัดไฟ, เงียบ, อื่นๆ โดยแต่ละ set ก็จะมีตัวเลขบอกว่า set กลุ่มนี้มีขนาดเท่าไหร่ อย่างไร

หลังจากนั้นจะทำ Algorithm ในการประเมินศักยภาพของแต่ละกลุ่มลูกค้า ว่าควรจะกระตุ้นการขายในกลุ่มลูกค้ากลุ่มใด โดยพิจารณาจาก กำไรของกลุ่มสินค้านั้นๆ เป็นต้น

วิธีการทำงาน

ใช้ Clustering Techniques แบบ hierarchical Clustering แบบ Single-Link Stage โดยมี function ในการหา distance คือ ระยะทาง ระหว่างข้อมูลทุกๆ ความเป็นไปได้ ตาม Euclidean metric ซึ่งมีสูตรการคำนวณดังนี้

$$d = \sqrt{\sum_{i=1}^{k} (p_i - q_i)^2}$$

โดย d คือระยะทาง

P และ q คือข้อมูลของทั้งสองจุด ในทุกๆ แง่มุม จำนวนแง่มุม = k

ซึ่งแง่มุมที่พิจารณาได้แก่ ราคา, คุณภาพ, การฟอกอากาศ, ประหยัดไฟ, เงียบ, อื่นๆ และได้ผลลัพธ์เป็น cluster และจำนวนใน cluster และรายละเอียดตัวแปรต่างๆ ของ cluster นั้นๆ

ซึ่งสมควรมี function ในการประเมินความน่าลงทุน หรือในด้านอื่นๆ ด้วย เช่น กลุ่มไหนมีกำไรสูงสุด กลุ่มไหนมีคู่แข่งน้อยราย หรือมากราย เป็นต้น

Problem 3 หาความสัมพันธ์ของตัวแปรต่างๆ เพื่อคาดการณ์ยอดขายสินค้าในอนาคต

Business Goal:

สามารถพยากรณ์ยอดขายเครื่องปรับอากาศในแต่ละรุ่น ตามตัวแปรที่กำหนดได้

Solution:

ใช้การเก็บข้อมูลจากบริษัท Saijo Denki ประมาณ 10 ปี โดยแบ่งตามรุ่นการขายต่างๆ

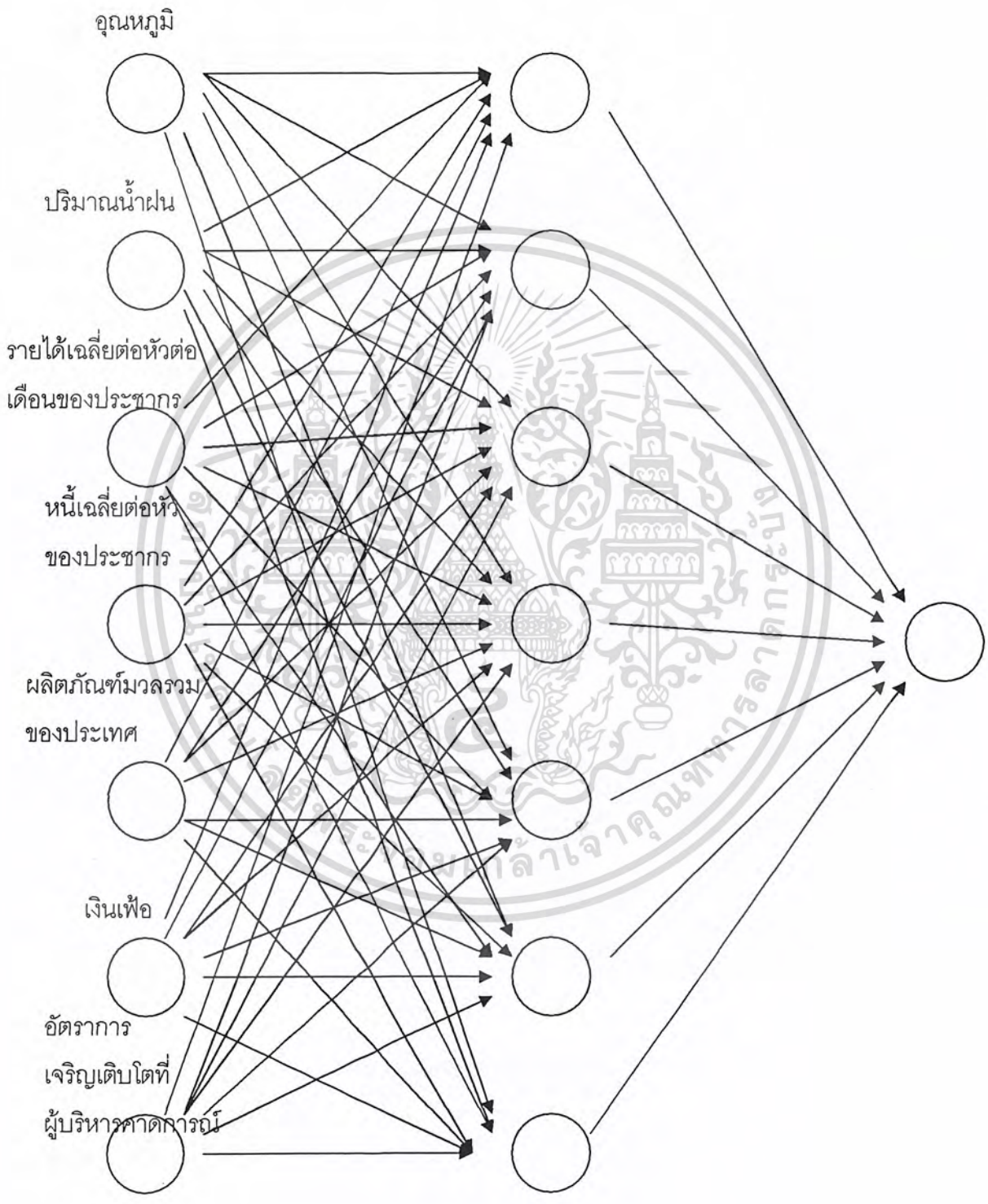
ใช้ Neural Network ในการหาความสัมพันธ์ โดย training set จะประกอบด้วย

Input: อุณหภูมิ
ปริมาณน้ำฝน
รายได้เฉลี่ยต่อหัวต่อเดือนของประชากร
หนี้เฉลี่ยต่อหัวของประชากร
ผลิตภัณฑ์มวลรวมของประเทศ
เงินเฟ้อ
อัตราการเจริญเติบโตที่ผู้บริหารคาดการณ์

Output: ยอดขายเครื่องปรับอากาศรวมในโซนนั้นๆ

สิ่งที่ได้จาก Neural Network คือ ความสัมพันธ์ของตัวแปร สภาพอากาศ, สถานะทางการเงินของประชาชนในพื้นที่จังหวัดนั้น และตัวเลขทางเศรษฐกิจที่สำคัญของประเทศ เพื่อคาดการณ์ยอดขายสินค้าในอนาคตในการหายอดขาย เครื่องปรับอากาศ ในแต่ละรุ่น ในโซนต่างๆ

วิธีการทำงาน



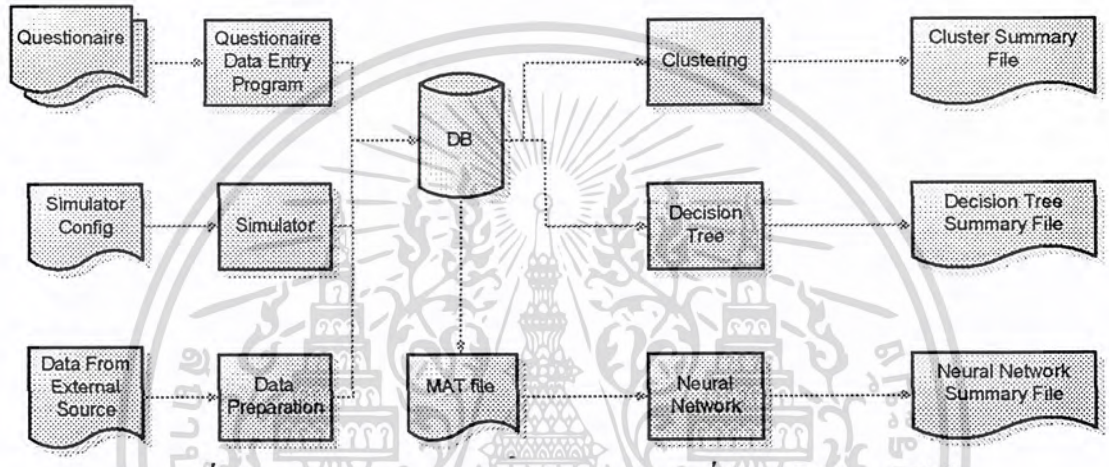
รูปที่ 4-1 ตัวอย่าง วิธีการทำงานของ Neural Network

การแก้ปัญหาในส่วนนี้ใช้ Neural Network ซึ่งมี Topology ดังรูป และมีการเรียนรู้แบบ Back Propagation

สรุปกระบวนการนำ Data Mining แก้ปัญหาทั้ง 3 ข้อ

ปัญหา Data Mining ที่ทีมงานได้นำมาลองใช้ มี 3 ปัญหาดังนี้

1. หาว่าลูกค้าไม่ชอบสินค้า จากการ บริการ มี รูปแบบ อะไรบ้าง
2. หาว่าลูกค้าซื้อสินค้าเพราะอะไร
3. หาความสัมพันธ์ของตัวแปรต่างๆ เพื่อคาดการณ์ยอดขายสินค้าในอนาคต



รูปที่ 4-2 แสดงการทำงานของการทำ Data Mining ที่ทีมงานได้นำมาลองใช้

วิธีการแก้ปัญหาที่ 1 (หาว่าลูกค้าไม่ชอบสินค้า จากการ บริการ มี รูปแบบ อะไรบ้าง)

Input ของปัญหาที่ 1 ได้มาจากการจำลองข้อมูลขึ้นมาโดย Simulator โดยมีขั้นตอนการนำข้อมูลมาใช้ดังต่อไปนี้

1. จัดทำรายละเอียดของการ Simulate ข้อมูล เป็น File Simulate.xls
2. หลังจากนั้นทำการ Simulate โดย Simulator โดย Simulator จะทำการนำข้อมูลที่ได้จากการ Simulate เข้าสู่ Database เอง

จากนั้นทำการสร้าง Decision Tree ซึ่งเป็น Data Mining Algorithm ที่ได้ถูกเลือกขึ้นมาใช้ในการหา Service Pattern ที่ทำให้ลูกค้ามีความรู้สึกไม่พอใจได้ โดยจะเก็บผลลัพธ์ไว้ใน File DecisionTree_ผลลัพธ์.txt

วิธีการแก้ปัญหาที่ 2 (หาว่าลูกค้าซื้อสินค้าเพราะอะไร)

Input ของปัญหาที่ 2 ได้นำมาจาก 2 แหล่ง คือ แบบสอบถามซึ่งเป็นการเก็บข้อมูลจริง และ Simulator ซึ่งเป็นการจำลองหาข้อมูลที่ได้กำหนดไว้ล่วงหน้า เพื่อทดสอบ Algorithm ว่าถูกต้อง หรือไม่

Input Method 1

1. จัดทำแบบสอบถาม (Questionaire)
2. เก็บข้อมูลเข้า Database โดยการทำ Data Entry ที่ส่วนของ Questionaire Data Entry Program ซึ่งเขียนด้วย Microsoft Access 2003
3. แปลงเป็น File Questionaire.csv
4. ทำการ Import เข้าสู่ Database

Input Method 2

1. จัดทำรายละเอียดของการ Simulate ข้อมูล เป็น File Simulate.xls
2. หลังจากนั้นทำการ Simulate โดย Simulator โดย Simulator จะทำการนำข้อมูลที่ได้จากการ Simulate เข้าสู่ Database เอง

ในส่วนของ Database ใช้ mySql version 4.0.17 ในการเก็บข้อมูล และใช้โปรแกรม SQLyog ในการดูแล และจัดการข้อมูล โดยเป็น GUI Program (Graphic User Interface Program) สำหรับ mySql version ดังกล่าว

จากนั้นทำการ Clustering ซึ่งเป็น Data Mining Algorithm ที่ได้ถูกเลือกขึ้นมาใช้ในการแบ่ง Segment ของกลุ่มลูกค้า และเก็บผลลัพธ์ที่ได้ใน ตาราง Cluster ในฐานข้อมูล datamining

วิธีการแก้ปัญหาที่ 3 (หาความสัมพันธ์ของตัวแปรต่างๆ เพื่อคาดการณ์ยอดขายสินค้าในอนาคต)

Input ของปัญหาที่ 3 ได้มาจากการนำข้อมูลจริงมาจากกรมอุตุนิยมวิทยา สำนักงานคณะกรรมการพัฒนาการเศรษฐกิจและสังคมแห่งชาติ สำนักนายกรัฐมนตรี และบริษัทซัยโจ เดนกิ อินเตอร์เนชั่นแนล จำกัด

โดยข้อมูลที่ได้จากกรมอุตุนิยมวิทยา คือ ข้อมูลปริมาณน้ำฝน และอุณหภูมิเฉลี่ยของแต่ละจังหวัด ในปี 2001, 2002, 2003 มีข้อมูลอยู่ 48 จังหวัด

ข้อมูลที่ได้จากสำนักงานคณะกรรมการพัฒนาการเศรษฐกิจและสังคมแห่งชาติ สำนักนายกรัฐมนตรี คือ ข้อมูลรายได้เฉลี่ยต่อหัวประชากร และข้อมูลรายได้ ค่าใช้จ่าย ภาวะหนี้สินของประชากร ซึ่งมีข้อมูลในปี 1996, 1998, 2000, 2002 มีข้อมูลอยู่ 76 จังหวัด

ข้อมูลที่ได้จากบริษัทซัยโจ เดนกิ อินเตอร์เนชั่นแนล จำกัด คือจำนวนยอดขายต่อปี ต่อพื้นที่ของบริษัท ซึ่งมีข้อมูลในปี 2000, 2001, 2002 มีอยู่ทั้งหมด 45 จังหวัด

โดยข้อมูลดังกล่าวมีขั้นตอนในการนำไปใช้งานดังนี้

1. เก็บข้อมูลจากกรมอุตุนิยมวิทยา สำนักงานคณะกรรมการพัฒนาการเศรษฐกิจและสังคมแห่งชาติ สำนักนายกรัฐมนตรี และบริษัทซัยโจ เดนกิ อินเตอร์เนชั่นแนล จำกัด
2. นำข้อมูลดังกล่าวมาแปลงให้เป็นข้อมูลที่สามารถนำไปใช้ต่อได้โดย ทำให้เป็น File Excel เนื่องจากข้อมูลเดิมเป็น File Text (*.txt) แล้วจากนั้น ใช้ VB Application ในการดึงข้อมูลเข้าสู่ Database (mySql)

3. จากนั้น Export ข้อมูลจาก Database ออกมาเป็น *.csv เพื่อให้สื่อสารกับ Application อื่นได้

4. จากนั้น Import *.csv ดังกล่าว เข้าสู่ Matlab จึงได้ *.mat File ที่สามารถนำไปคำนวณโดย Matlab ได้

จากนั้นทำการ Train Neural Network ที่สร้างขึ้น โดยใช้ Matlab 7.0 ในการ Train Neural Network ดังกล่าว

หลังจากได้ Neural Network ที่สามารถทำงานดังกล่าวได้แล้ว ก็ Export Weight ต่างๆ และค่า Bias ของ Neural Network ออกมาเพื่อเก็บข้อมูลดังกล่าวไว้เพื่อใช้ประโยชน์ต่อไป

บทที่ 5
Simulator : โปรแกรมดาต้าไมนิ่งในส่วนของการจำลองข้อมูล

เนื้อหาในบทนี้เกี่ยวกับการทำงานของโปรแกรมดาต้าไมนิ่งในส่วนของการจำลองข้อมูล ซึ่งโปรแกรมนี้ จะทำการ generate data ให้กับตาราง database 2 ตารางได้แก่ ตาราง Product และService โดยรายละเอียดจะอธิบายในส่วนถัดๆไป ส่วน Data ที่ได้ออกมาจะนำไปใช้ประโยชน์ในการทำการทดลอง Datamining ว่าสามารถได้ผลลัพธ์ตามที่คาดไว้หรือไม่ อย่างไร

5.1 การออกแบบ และสร้าง Database ให้กับโปรแกรมดาต้าไมนิ่งในส่วนของการจำลองข้อมูล

ใช้ tool ช่วยในการสร้าง Database และแปลงข้อมูลใน Database ไปเป็นไฟล์ .csv tool ที่ใช้มีชื่อว่า SQLyog Job Agent (SJA) ซึ่งใช้คู่กันกับ MySql

Database Table ที่จำเป็นต้องใช้งานมีดังนี้

1) เก็บข้อมูลคะแนนปัจจัยในการเลือกซื้อเครื่องปรับอากาศ

Database Table ชื่อ PRODUCT มี Field ดังนี้

Field	Type	ความหมาย
BUY_BRAND	int(11)	เก็บคะแนนปัจจัยชื่อเสียงของตราสินค้า
BUY_QUALITY	int(11)	เก็บคะแนนปัจจัยคุณภาพ - อายุการใช้งาน
BUY_SAVE_ENERGY	int(11)	เก็บคะแนนปัจจัยประหยัดไฟ
BUY_AIR_PURIFY	int(11)	เก็บคะแนนปัจจัยระบบฟอกอากาศ
BUY_QUIET	int(11)	เก็บคะแนนปัจจัยความเงียบ
BUY_BEAUTY	int(11)	เก็บคะแนนปัจจัยความสวยงาม
BUY_PRICE	int(11)	เก็บคะแนนปัจจัยราคา
BUY_SERVICE	int(11)	เก็บคะแนนปัจจัยการบริการหลังการขาย
BUY_OTHER	int(11)	เก็บคะแนนปัจจัยอื่นๆ

ตารางที่ 5.1-1 ตาราง PRODUCT

2) เก็บข้อมูลความพอใจของลูกค้าต่อการบริการ

Database Table ชื่อ SERVICE มี Field ดังนี้

Field	Type	ความหมาย
SERV_RATE	int(11)	เก็บคะแนนความพอใจต่อความถี่ของการบริการ
SERV_AFTER_SETUP_RATE	int(11)	เก็บคะแนนความพอใจต่อความถี่ของการบริการหลังติดตั้งไม่นาน
SERV_NOT_FINISH_RATE	int(11)	เก็บคะแนนความพอใจต่อความถี่ของการบริการที่ไม่เสร็จในครั้งเดียว
NO_YEAR	int(11)	เก็บอายุการใช้งานของเครื่องปรับอากาศ
SERV_POINT	int(11)	เก็บคะแนนความพอใจของลูกค้าการบริการในภาพรวม

ตารางที่ 5.1-2 ตาราง SERVICE

3) เก็บข้อมูล หลังจากทำ Clustering แล้ว Database Table ชื่อ CLUSTER มี Field ดังนี้

Field	Type	ความหมาย
AMOUNT_CLUSTER	int(11)	เก็บจำนวนกลุ่มที่ต้องการ Cluster
BUY_BRAND	text	เก็บผลลัพธ์ปัจจัยชื่อเสียงของตราสินค้า
BUY_QUALITY	text	เก็บผลลัพธ์ปัจจัยคุณภาพ - อายุการใช้งาน
BUY_SAVE_ENERGY	text	เก็บผลลัพธ์ปัจจัยประหยัดไฟ
BUY_AIR_PURIFY	text	เก็บผลลัพธ์ปัจจัยระบบฟอกอากาศ
BUY_QUIET	text	เก็บผลลัพธ์ปัจจัยความเงียบ
BUY_BEAUTY	text	เก็บผลลัพธ์ปัจจัยความสวยงาม
BUY_PRICE	text	เก็บผลลัพธ์ปัจจัยราคา
BUY_SERVICE	text	เก็บผลลัพธ์ปัจจัยการบริการหลังการขาย
BUY_OTHER	text	เก็บผลลัพธ์ปัจจัยอื่นๆ

ตารางที่ 5.1-3 ตาราง CLUSTER

4) เก็บข้อมูลจากแบบสอบถาม Database Table ชื่อ QUESTIONNAIRE มี Field ดังนี้

Field	Type	ความหมาย
ID	int(11)	เป็น index ของตาราง
Occupation	text	อาชีพ
Sex	text	เพศ
Age	text	อายุ
Income	text	รายได้ต่อเดือน
HomeType	text	การพักอาศัย
BrandName	int(11)	คะแนนปัจจัยชื่อเสียงของตราสินค้า
Quality	int(11)	คะแนนปัจจัยคุณภาพ - อายุการใช้งาน
EnergySaving	int(11)	คะแนนปัจจัยประหยัดไฟ
AP	int(11)	คะแนนปัจจัยระบบฟอกอากาศ
Quiet	int(11)	คะแนนปัจจัยความเงียบ
Beauty	int(11)	คะแนนปัจจัยความสวยงาม
Price	int(11)	คะแนนปัจจัยราคา
AfterSale	int(11)	คะแนนปัจจัยการบริการหลังการขาย
Other	int(11)	คะแนนปัจจัยอื่นๆ

ตารางที่ 5.1-4 ตาราง QUESTIONNAIRE

5.2 การทำงานของโปรแกรมดาต้าไมนิ่งในส่วนของการจำลองข้อมูล

โปรแกรมในส่วนนี้ จะทำการ generate data ให้กับตาราง database 2 ตารางได้แก่ ตาราง Product และServiceในตาราง Product จะกำหนดจำนวนที่จะ generate และเปอร์เซนต์ของข้อมูลในแต่ละ field ของตารางได้ และสามารถกำหนดเงื่อนไขความสัมพันธ์ของข้อมูลระหว่าง field ได้ เช่น กำหนดให้ generate ทั้งหมด 100 records คะแนน Quality เท่ากับ 1 มี 50 records(50%) โดยใน 50 records นี้ จะเลือกคะแนน Price เท่ากับ 2 ทั้งหมด 25 records (50%ของคะแนน Quality เท่ากับ1) เป็นต้น โดยเงื่อนไขสามารถมีได้มากกว่า 1 เงื่อนไขในตาราง Service จะกำหนดจำนวนที่จะ generate เงื่อนไขได้เช่นเดียวกับตาราง Product

5.3 Pseudo Code การทำงานของโปรแกรมดาต้าไมนิ่งในส่วนของการจำลองข้อมูล

เลือก ตาราง database ระหว่าง Product กับ Service

ถ้า เลือก Product {

เปิดหน้าต่าง ที่ใช้ Generate ตาราง database Product

เริ่มต้น มีเพียงส่วนกำหนดข้อมูลเบื้องต้นเท่านั้นที่ Enable

ถ้า กดปุ่ม OK {

ตรวจสอบ input ทั้งหมดในส่วนข้อมูลเบื้องต้น

ถ้า ถูกต้อง {

รับจำนวนที่จะทำการGenerate จากช่อง Number of products เก็บไว้ในตัวแปร pd_no

รับจำนวนเปอร์เซ็นต์ของคะแนนที่จะ generate จากส่วนกำหนดข้อมูลเบื้องต้น ตรง % of Points และ เก็บไว้ใน Array

Disable ส่วนกำหนดข้อมูลเบื้องต้นยกเว้น ปุ่ม Cancel

Enable ส่วนกำหนด เงื่อนไข

} Else แจ้ง error

}

ถ้ากดปุ่ม Cancel {

Clear ข้อมูลใน pd_no และ Array ที่เก็บ % of Points ไว้

Clear List เงื่อนไขทั้งหมด

Clear ช่องรับ input ในส่วนกำหนดข้อมูลเบื้องต้น ตรง % of Points

Disable ส่วนกำหนดเงื่อนไข

Enable ส่วนกำหนดข้อมูลเบื้องต้น

}

ถ้ากดปุ่ม Insert {

ตรวจสอบ เงื่อนไขที่ต้องการ Insert

ถ้า ถูกต้อง {

เก็บเงื่อนไข ไว้ใน Condition List และแสดงตรงส่วนแสดงเงื่อนไขทั้งหมด

} Else แจ้ง error

}

ถ้ากดปุ่ม Delete {

ลบเงื่อนไขที่เลือกไว้ออกจาก Condition List และแสดง เงื่อนไขที่เหลือตรงส่วนแสดงเงื่อนไขทั้งหมด

}

ถ้ากดปุ่ม Save conditions {

เปิดหน้าต่างเพื่อใช้ในการกำหนดชื่อ, ที่อยู่ของไฟล์ที่จะ save

ถ้า เลือกไฟล์ได้ถูกต้อง {

ให้นำค่า Number of products ที่เก็บไว้ในตัวแปร pd_no และเงื่อนไขต่างๆ ที่เก็บอยู่ใน Condition List เขียนลงในไฟล์เป้าหมาย

} Else แจ้ง error

}

ถ้ากดปุ่ม Generate Data {

ถ้า มีเงื่อนไข {

Loop

หาจำนวนที่ต้อง Generate ของเงื่อนไข

Loop

กำหนด query ตามเงื่อนไข

ทำการ execute query ลงตาราง Database Product

Until ครบจำนวนที่ต้อง Generate ของเงื่อนไข

Until ครบจำนวนเงื่อนไข

}

ถ้า จำนวนที่ต้อง Generate ยังไม่ครบ {

Loop

กำหนด query โดย random ค่าใน query

ทำการ execute query ลงตาราง Database Product

Until ครบจำนวนที่เหลือที่ต้อง Generate

}

}

ถ้ากดปุ่ม Delete Data {

กำหนด query ="DELETE * FROM PRODUCT"

ทำการ execute query

}

ถ้ากดปุ่ม Cluster {

Run Program Cluster

}

}

ถ้าเลือก Service {

เปิดหน้าต่าง ที่ใช้ Generate ตาราง database Service

ถ้า กดปุ่ม Insert {

ตรวจสอบ เงื่อนไขที่ต้องการ Insert

ถ้า ถูกต้อง {

เก็บเงื่อนไข ไว้ใน Condition List และแสดงตรงส่วนแสดงเงื่อนไขทั้งหมด

} Else แจ้ง error

}

ถ้ากดปุ่ม Delete {

ลบเงื่อนไขที่เลือกไว้ออกจาก Condition List และแสดง เงื่อนไขที่เหลือตรงส่วนแสดง เงื่อนไขทั้งหมด

}

ถ้ากดปุ่ม Save conditions {

เปิดหน้าต่างเพื่อใช้ในการกำหนดชื่อ, ที่อยู่ของไฟล์ที่จะ save

ถ้า เลือกไฟล์ได้ถูกต้อง {

นำ Number of Customer และเงื่อนไขต่างๆ ที่เก็บอยู่ใน Condition List เขียนลงใน ไฟล์เป้าหมาย

} Else แจ้ง error

}

ถ้ากดปุ่ม Generate Data {

ตรวจสอบ Number of customers มา ถูกต้องหรือไม่

ถ้าถูกต้อง {

เก็บ Number of customers ลงตัวแปร cus_no

ถ้า มีเงื่อนไข {

Loop

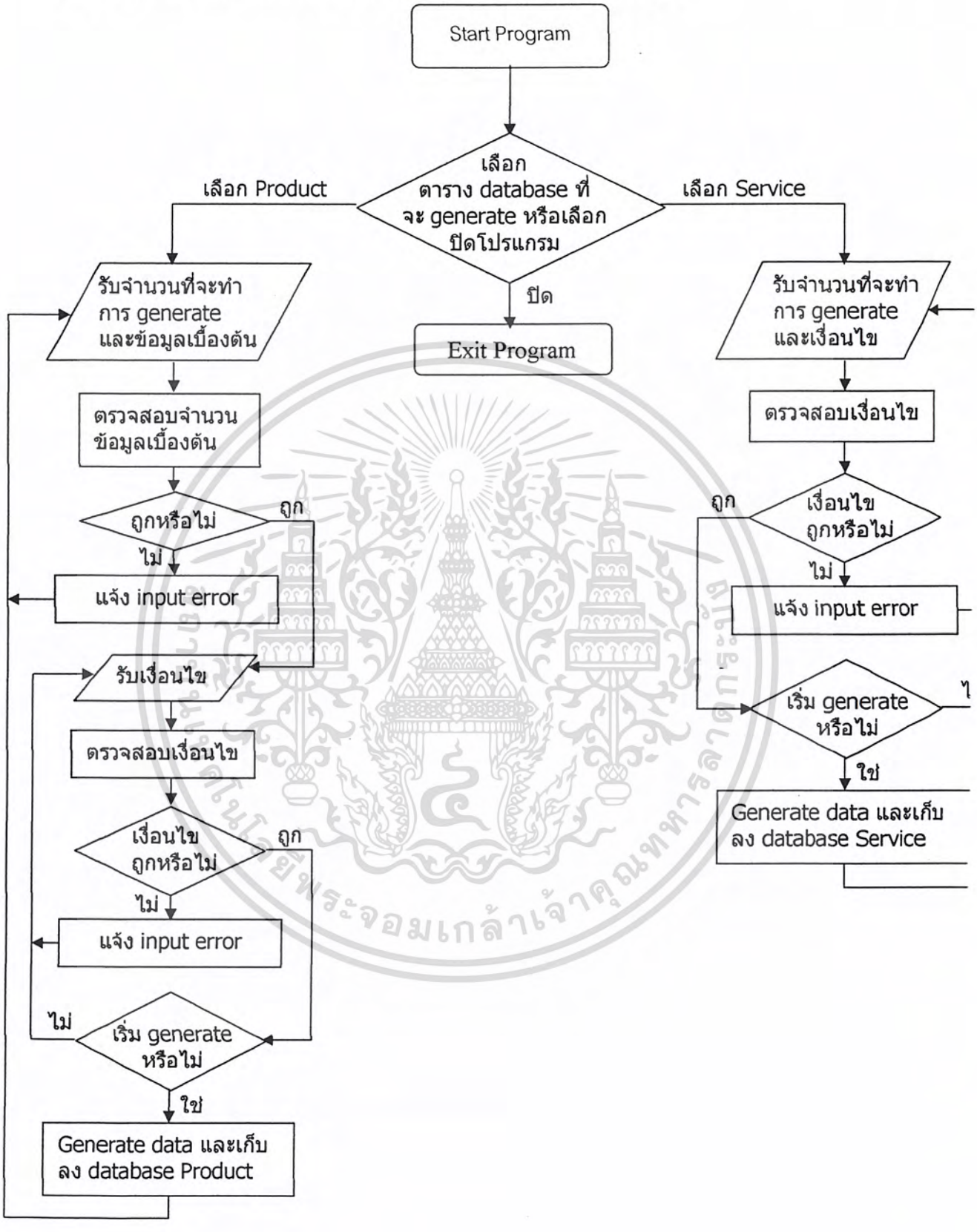
หาจำนวนที่ต้อง Generate ของเงื่อนไข

Loop

กำหนด query ตามเงื่อนไข

```
                    ทำการ execute query ลงตาราง Database Product
                    Until ครบจำนวนที่ต้อง Generate ของเงื่อนไข
                Until ครบจำนวนเงื่อนไข
            }
            ถ้า จำนวนที่ต้อง Generate ยังไม่ครบ {
                Loop
                    กำหนด query โดย random ค่าใน query
                    ทำการ execute query ลงตาราง Database Product
                Until ครบจำนวนที่เหลือที่ต้อง Generate
            }
        }
    }
    ถ้ากดปุ่ม Delete Data {
        กำหนด query ="DELETE * FROM SERVICE"
        ทำการ execute query
    }
}
```

5.4 Flowchart การทำงานของโปรแกรมดาต้าไมนิ่งในส่วนของการจำลองข้อมูล



รูปที่ 5.4-1 Flowchart การทำงานของโปรแกรมดาต้าไมนิ่งในส่วนของการจำลองข้อมูล

บทที่ 6

Clustering

Clustering of Data เป็นวิธีในการจัดกลุ่มข้อมูลที่เหมือนๆ กันให้อยู่ในกลุ่มที่เล็กลง เช่น การรวมกลุ่มลูกบอลสีเดียวกัน เป็นต้น

Clustering Algorithm นั้นเป็นวิธีการในการหากลุ่มที่เหมือนกันของข้อมูล หนึ่งในวิธีนั้นคือการหา centroid ของข้อมูล ซึ่งหมายถึง จุดที่มีค่ากลางของกลุ่มหนึ่งๆ แล้วจึงทำการหาสมาชิกของกลุ่มนั้นๆ โดยใช้ centroid เป็นตัวกลางในการหาแล้วใช้วิธีการทางสถิติยศาสตร์เข้าช่วย

Clustering Algorithm นั้นถูกใช้อย่างแพร่หลายตัวอย่างเช่น Pattern Recognition, Artificial Intelligence, survey of markets, survey of products, survey of sales program และ R&D เป็นต้น

Clustering คือกระบวนการในการแบ่งกลุ่มของข้อมูล หรือวัตถุให้เป็นกลุ่มย่อยที่มีความหมาย หรือที่เรียกว่า Cluster โดย Clustering เป็น unsupervised classification เนื่องจากไม่มีการ รู้ว่าแต่ละกลุ่มย่อยสื่อถึงอะไร ก่อนที่จะทำการ clustering ซึ่งจะตรงข้ามกับ Classification ซึ่งเป็น Supervised Classification

Clustering ที่ดีจะต้องได้ Cluster ที่มีคุณภาพสูง ซึ่งหมายถึง
Intra-class similarity ซึ่งคือความเหมือนกันภายใน Cluster สูง และ
Inter-class similarity ซึ่งคือความเหมือนกันภายนอก Cluster ต่ำ

คุณภาพของ Clustering นั้นถูกวัดโดยความ Function ที่ Represent ความเหมือนใน cluster และความแตกต่างภายนอก cluster

Clustering Techniques สามารถแบ่งได้ออกเป็น 4 กลุ่มย่อย ซึ่งคือ

Partitioning เป็นการแบ่ง cluster โดย partition criterion ซึ่งเมื่อได้กลุ่มของ cluster ก็จะแบ่งตามความเหมือนต่อ partition criterion ดังกล่าวออกเป็น cluster ย่อย

Hierarchy algorithm เป็นลำดับของ partition ซึ่งคือการที่แต่ละวัตถุ merge กันไปเรื่อยๆ จนกระทั่งสุดท้าย ได้ 1 cluster ใหญ่ ซึ่งคือ set ทั้งหมดของวัตถุ

Density-based เป็นการหาพื้นที่ของข้อมูลที่มีความหนาแน่นมากกว่า threshold ที่กำหนดจากแต่ละ cluster กล่าวคือเป็นการแบ่งข้อมูลตามความหนาแน่นข้อมูล

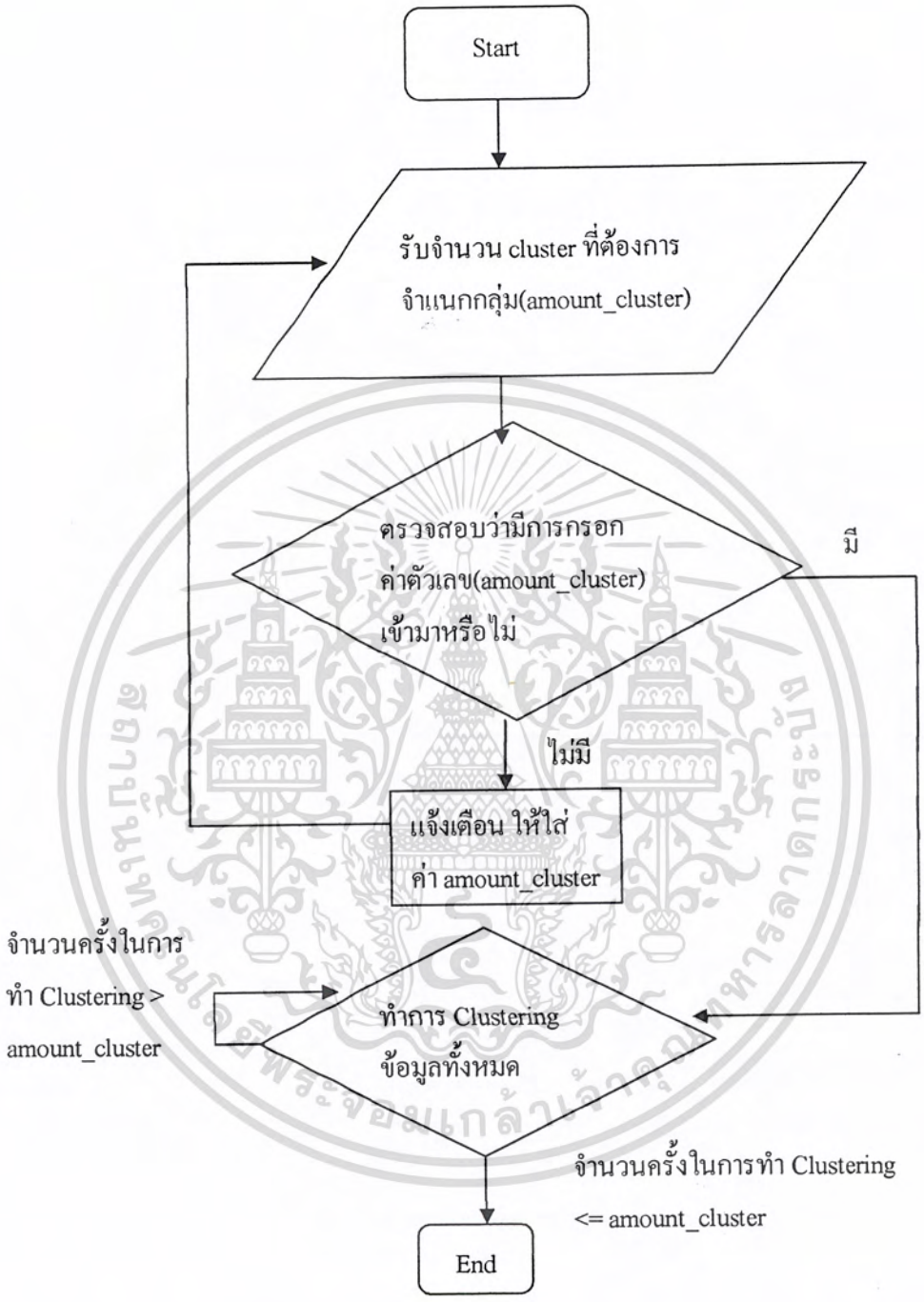
Grid-based เป็นการแบ่งพื้นที่ของข้อมูลออกเป็น cell ย่อยแล้ว แบ่งข้อมูลตาม cell data space นั้นๆ กล่าวคือเป็นการแบ่งเชิงพื้นที่

โดยในโครงงานนี้ได้ประยุกต์ใช้ Clustering แบบ Hierachy algorithm ในการแก้ปัญหา ซึ่งรายละเอียด และตัวอย่างของการ Clustering ได้กล่าวไปแล้วในส่วนของ Data Mining Algorithms

6.1 Pseudo code การทำงานของโปรแกรมดาต้าไมนิ่งในส่วนของการคลัสเตอร์ข้อมูล

```
        Get amount_cluster
Loop
        Get Data From Table Product to LinkList DataElement
        Find 2 Elements that contains Minimum Euclidean Distance calculated every pair of Elements
        Calculate Mean of 2 elements
        Count (frequency of elements in every cluster)
        Remove 2 Elements from LinkList DataElement
        AddLast New Mean Element
Until   count( LinkList DataElement ) <= amount_cluster
```

6.2 Flowchart การทำงานของโปรแกรมดาต้าไมนิ่งในส่วนของการคลัสเตอร์ข้อมูล



รูปที่ 6.2-1 Flowchart การทำงานของโปรแกรมดาต้าไมนิ่งในส่วนของการคลัสเตอร์ข้อมูล

6.3 การประเมินผลการทำ Clustering

การประเมินผลการทำงานของ Cluster จะเป็นการเปรียบเทียบระหว่างผลลัพธ์ที่ได้มาจากโปรแกรม และผลที่คาดว่าจะได้ ซึ่งได้จากการคำนวณ

การหา Variance ของ Segment หนึ่งๆ

$$V_{Segment} = \frac{\sum_{i=1}^{n} |t_i - a_i|}{n}$$

$V_{Segment}$ คือ ค่าเบี่ยงเบนที่เกิดจากการทำงานของโปรแกรม

n คือ จำนวนตัวแปร (parameter) ที่เกี่ยวข้องกับ Segment นั้นๆ

t_i คือ ข้อมูลที่เกิดจากการคำนวณ

a_i คือ ข้อมูลที่เกิดการจากทำงานของโปรแกรม

โดยผลที่ได้จะมีค่าอยู่ระหว่าง 0 ถึง 4 ซึ่ง จะสามารถแปลงข้อมูลเป็น % ได้โดย สมการต่อไปนี้

$$VP_{Segment}(\%) = V_{Segment} \times 20$$

$VP_{Segment}$ คือ ค่าเบี่ยงเบนที่เกิดจากการทำงานของโปรแกรม (%)

$$V_{Clustering} = \frac{\sum_{i=1}^{n} w_i v_i}{\sum_{i=1}^{n} w_i}$$

$V_{Clustering}$ คือ ค่าเบี่ยงเบนที่เกิดจากการทำงานของโปรแกรมจากทั้งการ Clustering

n คือ จำนวน Segment ที่เกี่ยวข้องกับการ Clustering นั้นๆ

w_i คือ จำนวนข้อมูลใน Segment นั้นๆ

v_i คือ ค่าเบี่ยงเบนที่เกิดจากการทำงานของโปรแกรมของ Segment นั้นๆ (%)

โดยผลลัพธ์ที่ได้จะมีค่าอยู่ระหว่าง 0 ถึง 100 เปอร์เซนต์

6.4 การทดลอง

การทดลองประกอบด้วย 6 การทดลอง

โดยการทดลองที่ 1 จะเป็นการทำ Clustering จากข้อมูลจริงที่ได้จากการสำรวจความคิดเห็นของประชาชนตามสถานที่ต่างๆ จำนวน 198 คน และการทดลองที่ 2 ถึงการทดลองที่ 6 นั้นจะเป็นการทำ Clustering จากข้อมูลที่ได้จากการ Simulation ในเงื่อนไขต่างๆ จำนวน 1000 ข้อมูล

การทดลองที่ 1

ฐานข้อมูลจริง ที่ได้จากการสำรวจความคิดเห็นของประชาชนตามสถานที่ต่างๆ โดยในการทดลองนี้ได้เลือกจำนวน Cluster ที่ใช้ในการจำแนกกลุ่ม เป็น 5 Cluster โดยผลลัพธ์ (Output) ที่ได้เป็นดังนี้

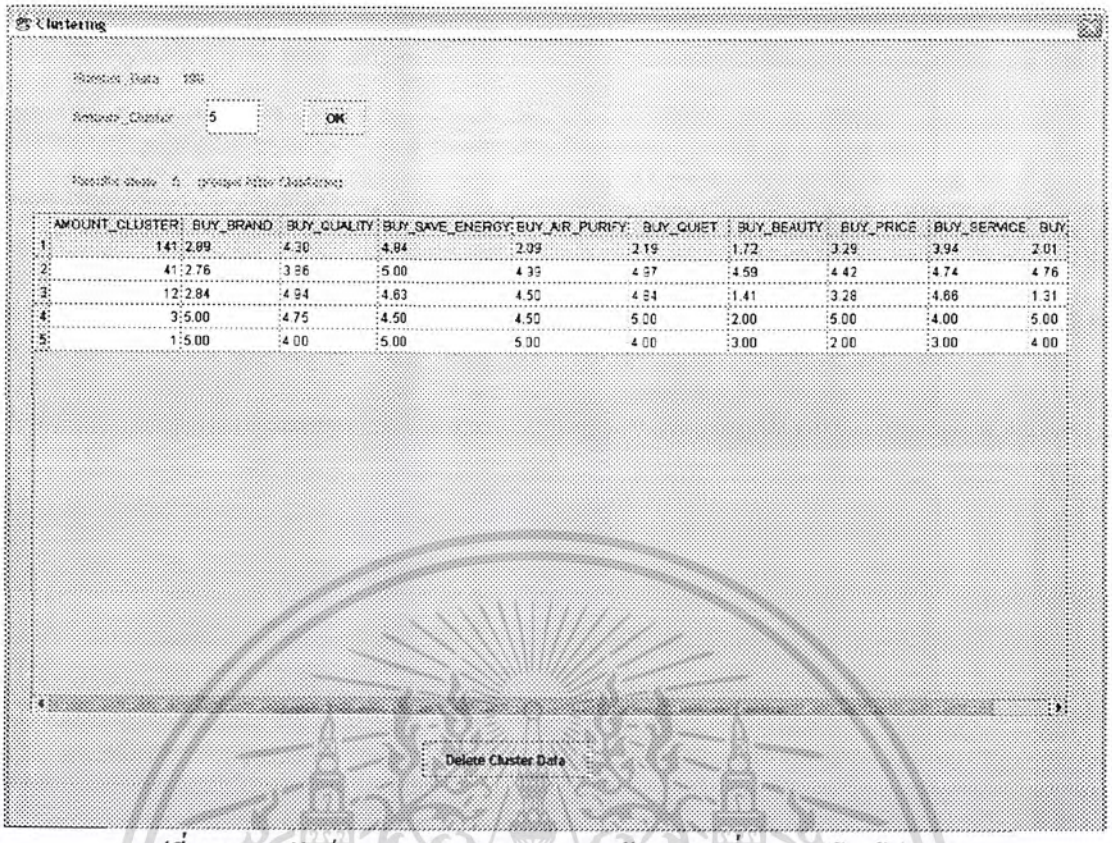
ผลลัพธ์ที่ได้จากการ Clustering ข้อมูลจริง

จำนวนของคลัสเตอร์ (Number of cluster): 5

จำนวนของข้อมูลจริง : 198

กลุ่ม	Size	Brand	Quality	Energy	Air Purify	Quiet	Beauty	Price	Service	Others
1	141	2.89	4.3	4.84	2.09	2.19	1.72	3.29	3.94	2.01
2	41	2.76	3.86	5	4.99	4.97	4.59	4.42	4.74	4.76
3	12	2.84	4.94	4.63	4.5	4.84	1.41	3.28	4.66	1.31
4	3	5	4.75	4.5	4.5	5	2	5	4	5
5	1	5	4	5	5	4	3	2	3	4

ตารางที่ 6.4-1 ผลลัพธ์ของการทำ Clustering จากข้อมูลจริง



รูปที่ 6.4-1 ผลลัพธ์ของการทำ Clustering จากข้อมูลจริงที่แสดงผลโดยโปรแกรม

การทดลองที่ 2

เป้าหมายในการจำลองข้อมูล (Simulation Goal)

Size แต่ละ Segment มีขนาดไม่เท่ากัน	No
ภายใน Segment มีคุณสมบัติที่ Random อยู่	Yes
ทุก Segment มีคุณสมบัติที่ Random อยู่	No
มีส่วนของการตลาดที่ Random อยู่	No
มี Segment ที่มีขนาดเล็กมาก (ต่ำกว่า 10%)	No

การแบ่งส่วนการตลาด (Market Segmentation)

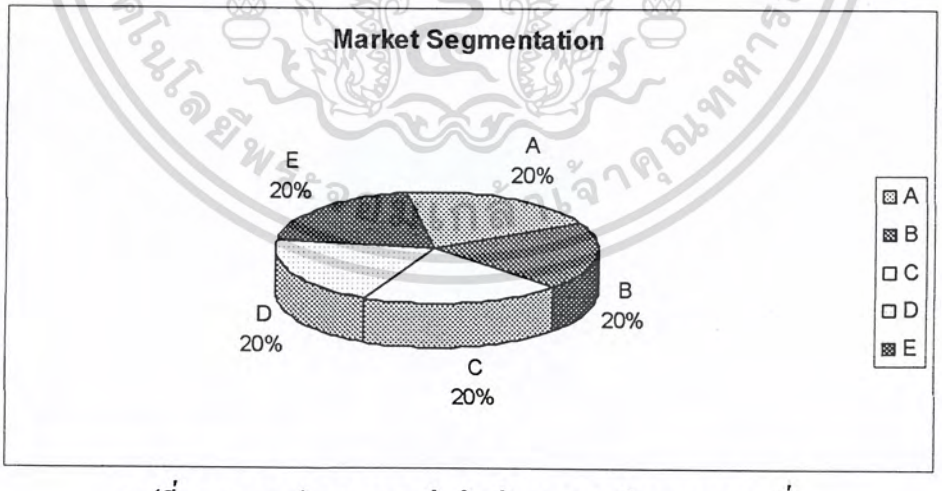
กลุ่มที่ 1 20 % ดูคุณภาพ ราคา ประหยัดไฟ ระบบฟอกอากาศและความเงียบเป็นหลัก สนใจการบริการหลังการขายปานกลาง ส่วนปัจจัยที่เหลือไม่สนใจ

กลุ่มที่ 2 20 % ดูคุณภาพและยี่ห้อเป็นหลัก ไม่สนใจความสวยงามและอื่นๆ ปัจจัยที่เหลือสนใจบ้าง

กลุ่มที่ 3 20 % ดูยี่ห้อ การบริการหลังการขาย ความเงียบ คุณภาพและการประหยัดพลังงานเป็นหลัก โดยสนใจสองปัจจัยแรกก่อน ส่วนปัจจัยที่เหลือสนใจบ้าง

กลุ่มที่ 4 20 % ดูปัจจัยโดยรวมเป็นหลัก แต่จะเน้นที่คุณภาพและประหยัดไฟก่อน

กลุ่มที่ 5 20 % ดูราคา คุณภาพ อื่นๆและยี่ห้อเป็นหลัก ปัจจัยที่เหลือสนใจน้อยถึงปานกลาง



รูปที่ 6.4-2 กราฟแสดงการแบ่งส่วนในการตลาดในการทดลองที่ 2

การแมพอินพุท (Input Mapping)

กลุ่ม	Size	Brand	Quality	Energy	Air Purify	Quiet	Beauty	Price	Service	Others
1	200	ต่ำมาก	สูงมาก	สูงมาก	สูงมาก	สูงมาก	ต่ำมาก	สูงมาก	ปานกลาง	ต่ำมาก
2	200	สูงมาก	สูงมาก	เฉลี่ย	เฉลี่ย	เฉลี่ย	ต่ำมาก	เฉลี่ย	เฉลี่ย	ต่ำมาก
3	200	สูงมาก	สูง	สูง	เฉลี่ย	สูง	เฉลี่ย	เฉลี่ย	สูงมาก	เฉลี่ย
4	200	สูง	สูงมาก	สูงมาก	สูง	สูง	สูง	สูง	สูง	สูง
5	200	สูงมาก	สูงมาก	ปานกลาง	ปานกลาง	ปานกลาง	ต่ำ	สูงมาก	ปานกลาง	สูงมาก

ตารางที่ 6.4-2 การแมพอินพุทจากการแบ่งส่วนการตลาดในการทดลองที่ 2

การจำลองอินพุท (Input Simulation)

จำนวนของข้อมูลที่จำลอง (Number of sampling): 1000

กลุ่ม	Size	Brand	Quality	Energy	Air Purify	Quiet	Beauty	Price	Service	Others
1	200	1	5	5	5	5	1	5	3	1
2	200	5	5	Random	Random	Random	1	Random	Random	1
3	200	5	4	4	Random	4	Random	Random	5	Random
4	200	4	5	5	4	4	4	4	4	4
5	200	5	5	3	3	3	2	5	3	5

ตารางที่ 6.4-3 การจำลองอินพุทที่ใช้ในการทดลองที่ 2

ผลลัพธ์ที่คาดหวังไว้ (Output Expected)

จำนวนของคลัสเตอร์ (Number of cluster): 5

Maximum Variance (%): 10

กลุ่ม	Size	Brand	Quality	Energy	Air Purify	Quiet	Beauty	Price	Service	Others
1	200	1	5	5	5	5	1	5	3	1
2	200	5	5	3	3	3	1	3	3	1
3	200	5	4	4	3	4	3	3	5	3
4	200	4	5	5	4	4	4	4	4	4
5	200	5	5	3	3	3	2	5	3	5

ตารางที่ 6.4-4 แสดงผลลัพธ์ที่คาดหวังไว้ในการทดลองที่ 2

ผลลัพธ์ที่แท้จริง (Output Actual)

จำนวนของคลัสเตอร์ (Number of cluster): 5

กลุ่ม	Size	Brand	Quality	Energy	Air Purify	Quiet	Beauty	Price	Service	Others
1	200	1	5	5	5	5	1	5	3	1
2	200	5	5	2.32	1.31	1.49	1	1.52	1.65	1
3	200	5	4	4	1.47	4	4	2	5	2.47
4	200	4	5	5	4	4	4	4	4	4
5	200	5	5	3	3	3	2	5	3	5

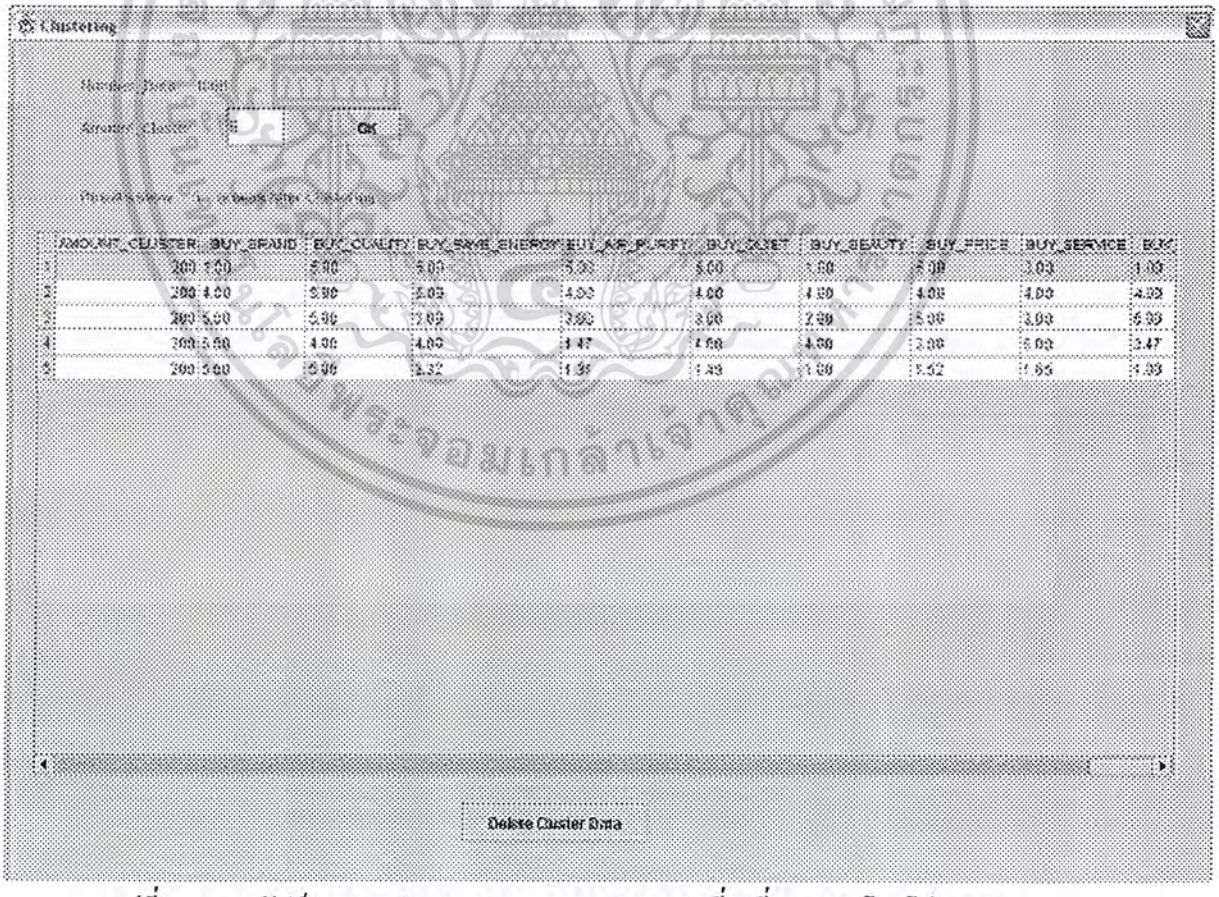
ตารางที่ 6.4-5 แสดงผลลัพธ์ที่ได้ที่แท้จริงจากการทดลองที่ 2

ผลวิเคราะห์การทดลอง

กลุ่ม	Size	**Variance (%)**	**Achieve [Y/N]**
1	200	0	Y
2	200	14.95	N
3	200	9.02	Y
4	200	0	Y
5	200	0	Y

ตารางที่ 6.4-6 ผลการวิเคราะห์ค่าเบี่ยงเบน (Variance) ในการทดลองที่ 2

Actual Variance (%): 4.79%



รูปที่ 6.4-3 ผลลัพธ์ของการทำ Clustering จากการทดลองที่ 2 ที่แสดงผลโดยโปรแกรม

การทดลองที่ 3

เป้าหมายในการจำลองข้อมูล (Simulation Goal)

Size แต่ละ Segment มีขนาดไม่เท่ากัน	Yes
ภายใน Segment มีคุณสมบัติที่ Random อยู่	Yes
ทุก Segment มีคุณสมบัติที่ Random อยู่	No
มีส่วนของการตลาดที่ Random อยู่	No
มี Segment ที่มีขนาดเล็กมาก (ต่ำกว่า 10%)	No

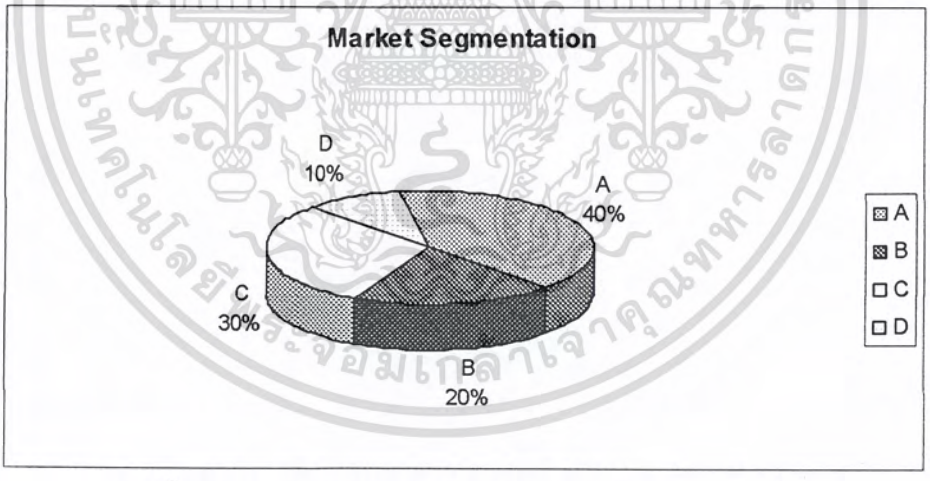
การแบ่งส่วนการตลาด (Market Segmentation)

กลุ่มที่ 1 40 % ดูคุณภาพ ระบบฟอกอากาศ ประหยัดไฟและความเงียบเป็นหลักโดยไม่ค่อยสนใจ factor อื่นๆ

กลุ่มที่ 2 30 % ดูราคา ยี่ห้อ คุณภาพและการประหยัดพลังงานเป็นหลัก โดยการบริการหลังการขายเป็นปัจจัยรองลงมา ส่วนปัจจัยอื่นๆไม่ค่อยมีผล

กลุ่มที่ 3 20 % ดูคุณภาพ และการประหยัดพลังงานเป็นหลัก ส่วนปัจจัยอื่นมีผลในการตัดสินใจบ้าง

กลุ่มที่ 4 10 % ดูระบบฟอกอากาศ และความเงียบเป็นหลัก โดยไม่สนใจราคา แต่ปัจจัยอื่นมีผลบ้าง



รูปที่ 6.4-4 กราฟแสดงการแบ่งส่วนในการตลาดในการทดลองที่ 3

การแมพอินพุท (Input Mapping)

กลุ่ม	Size	Brand	Quality	Energy	Air Purify	Quiet	Beauty	Price	Service	Others
1	400	ต่ำมาก	สูงมาก	สูงมาก	สูงมาก	สูงมาก	ต่ำมาก	ต่ำมาก	ต่ำมาก	ต่ำมาก
2	300	สูงมาก	สูงมาก	สูงมาก	ต่ำมาก	ต่ำมาก	ต่ำมาก	สูงมาก	สูง	ต่ำมาก
3	200	เฉลี่ย	สูงมาก	สูงมาก	เฉลี่ย	เฉลี่ย	เฉลี่ย	เฉลี่ย	เฉลี่ย	เฉลี่ย
4	100	เฉลี่ย	เฉลี่ย	เฉลี่ย	สูงมาก	สูงมาก	เฉลี่ย	ต่ำมาก	เฉลี่ย	เฉลี่ย

ตารางที่ 6.4-7 การแมพอินพุทจากการแบ่งส่วนการตลาดในการทดลองที่ 3

การจำลองอินพุท (Input Simulation)

จำนวนของข้อมูลที่จำลอง (Number of sampling): 1000

กลุ่ม	Size	Brand	Quality	Energy	Air Purify	Quiet	Beauty	Price	Service	Others
1	400	1	5	5	5	5	1	1	1	1
2	300	5	5	5	1	1	1	5	4	1
3	200	Random	5	5	Random	Random	Random	Random	Random	Random
4	100	Random	Random	Random	5	5	Random	1	Random	Random

ตารางที่ 6.4-8 การจำลองอินพุทที่ใช้ในการทดลองที่ 3

ผลลัพธ์ที่คาดหวังไว้ (Output Expected)

จำนวนของคลัสเตอร์ (Number of cluster) : 4

Maximum Variance (%) : 10

กลุ่ม	Size	Brand	Quality	Energy	Air Purify	Quiet	Beauty	Price	Service	Others
1	400	1	5	5	5	5	1	1	1	1
2	300	5	5	5	1	1	1	5	4	1
3	200	3	5	5	3	3	3	3	3	3
4	100	3	3	3	5	5	3	1	3	3

ตารางที่ 6.4-9 แสดงผลลัพธ์ที่คาดหวังไว้ในการทดลองที่ 3

ผลลัพธ์ที่แท้จริง (Output Actual)

จำนวนของคลัสเตอร์ (Number of cluster): 4

กลุ่ม	Size	Brand	Quality	Energy	Air Purify	Quiet	Beauty	Price	Service	Others
1	400	1	5	5	5	5	1	1	1	1
2	300	5	5	5	1	1	1	5	4	1
3	198	3	5	5	3.78	2.33	3.07	3.62	2.54	4.5
4	102	2.6	2.03	3.30	4.25	4.5	3.54	1.5	3.77	3.47

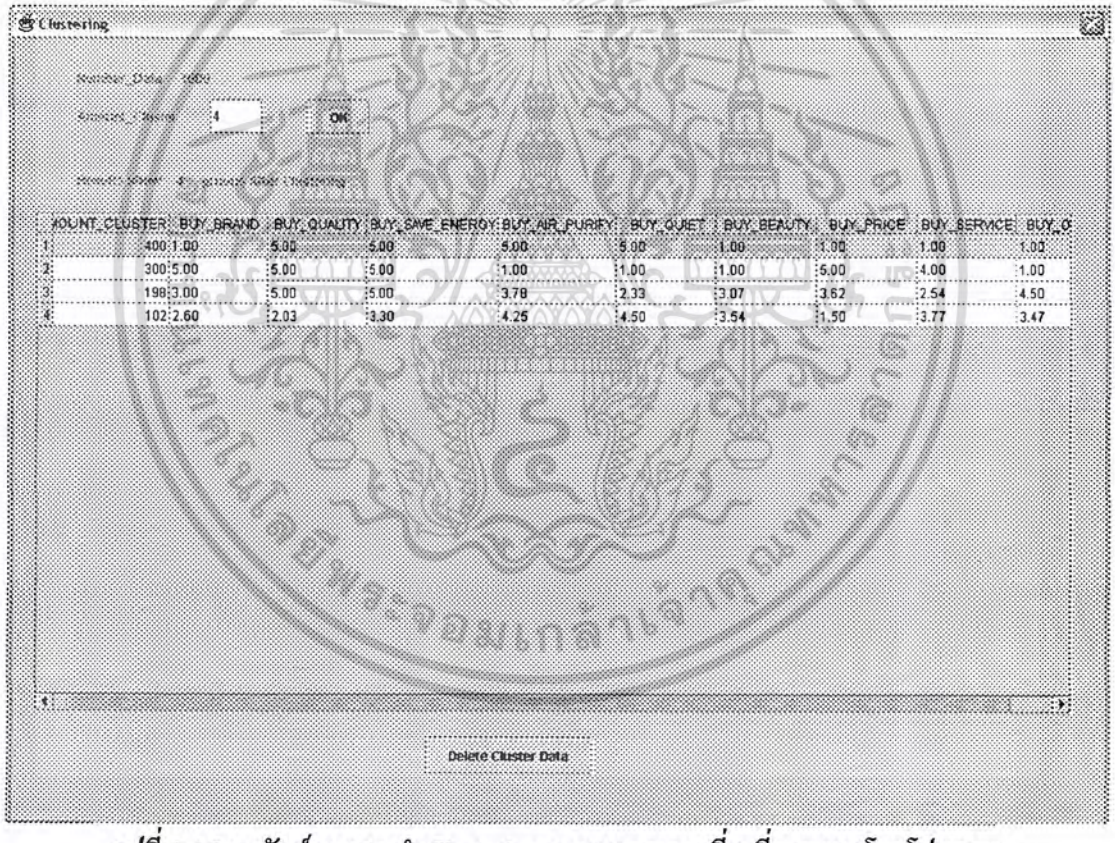
ตารางที่ 6.4-10 แสดงผลลัพธ์ที่ได้ที่แท้จริงจากการทดลองที่ 3

ผลวิเคราะห์การทดลอง

กลุ่ม	Size	**Variance (%)**	**Achieve [Y/N]**
1	400	0	Y
2	300	0	Y
3	198	8.67	Y
4	102	11.56	N

ตารางที่ 6.4-11 ผลการวิเคราะห์ค่าเบี่ยงเบน (Variance) ในการทดลองที่ 3

Actual Variance (%): 2.90%



รูปที่ 6.4-5 ผลลัพธ์ของการทำ Clustering จากการทดลองที่ 3 ที่แสดงผลโดยโปรแกรม

การทดลองที่ 4

เป้าหมายในการจำลองข้อมูล (Simulation Goal)

Size แต่ละ Segment มีขนาดไม่เท่ากัน	Yes
ภายใน Segment มีคุณสมบัติที่ Random อยู่	Yes
ทุก Segment มีคุณสมบัติที่ Random อยู่	Yes
มีส่วนของการตลาดที่ Random อยู่	No
มี Segment ที่มีขนาดเล็กมาก (ต่ำกว่า 10%)	No

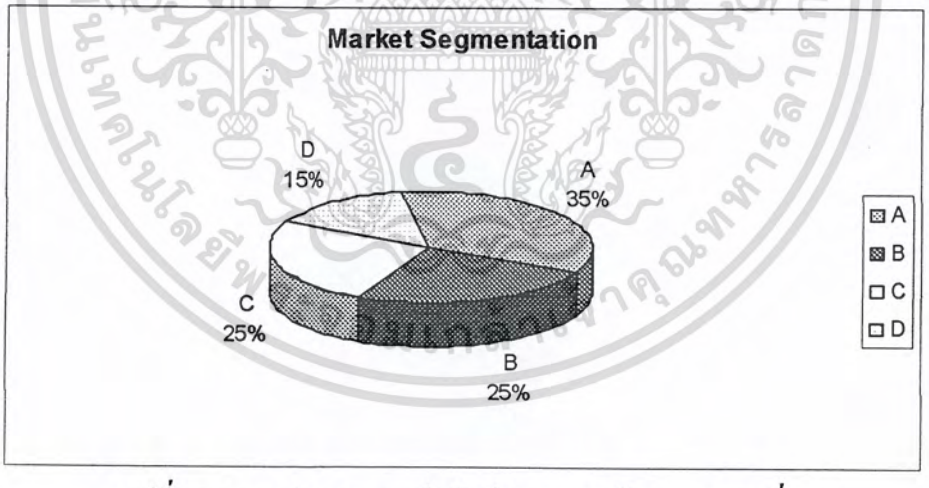
การแบ่งส่วนการตลาด (Market Segmentation)

กลุ่มที่ 1 35 % ดูคุณภาพ ประหยัดไฟและความเงียบเป็นหลัก โดยดูการฟอกอากาศและยี่ห้อรองลงมา และปัจจัยอื่นๆ บ้าง

กลุ่มที่ 2 25 % ดูคุณภาพ และการประหยัดพลังงานเป็นหลัก โดยไม่สนใจเรื่องความสวยงาม ส่วนปัจจัยอื่นมีผลในการตัดสินใจบ้าง

กลุ่มที่ 3 25 % ดูราคา และการประหยัดพลังงานเป็นหลัก ดูเรื่องคุณภาพเป็นรอง และปัจจัยอื่นๆ บ้าง

กลุ่มที่ 4 15 % ดูระบบฟอกอากาศ ยี่ห้อและความเงียบเป็นหลัก โดยสนใจคุณภาพและการบริการหลังการขายรองลงมา ไม่สนใจราคาและความสวยงาม แต่ปัจจัยอื่นมีผลบ้าง



รูปที่ 6.4-6 กราฟแสดงการแบ่งส่วนในการตลาดในการทดลองที่ 4

การแมพอินพุท (Input Mapping)

กลุ่ม	Size	Brand	Quality	Energy	Air Purify	Quiet	Beauty	Price	Service	Others
1	350	เฉลี่ย	สูงมาก	เฉลี่ย	สูง	สูงมาก	เฉลี่ย	เฉลี่ย	เฉลี่ย	เฉลี่ย
2	250	เฉลี่ย	สูงมาก	สูงมาก	เฉลี่ย	เฉลี่ย	ต่ำมาก	เฉลี่ย	เฉลี่ย	เฉลี่ย
3	250	เฉลี่ย	สูง	สูงมาก	เฉลี่ย	เฉลี่ย	เฉลี่ย	สูงมาก	เฉลี่ย	เฉลี่ย
4	150	สูงมาก	สูง	เฉลี่ย	สูงมาก	สูงมาก	ต่ำ	ต่ำ	สูง	เฉลี่ย

ตารางที่ 6.4-12 การแมพอินพุทจากการแบ่งส่วนการตลาดในการทดลองที่ 4

การจำลองอินพุท (Input Simulation)

จำนวนของข้อมูลที่จำลอง (Number of sampling): 1000

กลุ่ม	Size	Brand	Quality	Energy	Air Purify	Quiet	Beauty	Price	Service	Others
1	350	Random	5	Random	4	5	Random	Random	Random	Random
2	250	Random	5	5	Random	Random	1	Random	Random	Random
3	250	Random	4	5	Random	Random	Random	5	Random	Random
4	150	5	4	Random	5	5	2	2	4	Random

ตารางที่ 6.4-13 การจำลองอินพุทที่ใช้ในการทดลองที่ 4

ผลลัพธ์ที่คาดหวังไว้ (Output Expected)

จำนวนของคลัสเตอร์ (Number of cluster): 4

Maximum Variance (%): 15

กลุ่ม	Size	Brand	Quality	Energy	Air Purify	Quiet	Beauty	Price	Service	Others
1	350	3	5	3	4	5	3	3	3	3
2	250	3	5	5	3	3	1	3	3	3
3	250	3	4	5	3	3	3	5	3	3
4	150	5	4	3	5	5	2	2	4	3

ตารางที่ 6.4-14 แสดงผลลัพธ์ที่คาดหวังไว้ในการทดลองที่ 4

ผลลัพธ์ที่แท้จริง (Output Actual)

จำนวนของคลัสเตอร์ (Number of cluster): 4

กลุ่ม	Size	Brand	Quality	Energy	Air Purify	Quiet	Beauty	Price	Service	Others
1	498	3.25	4.63	5	1.81	2.35	2.01	3.68	2.93	2.41
2	471	3.86	4.75	2.07	4.25	5	1.88	1.45	4.29	2.64
3	29	3.14	5	1.41	4	5	3.56	4.51	2.43	4.31
4	2	1	4.5	5	2	1	2	5	1	1

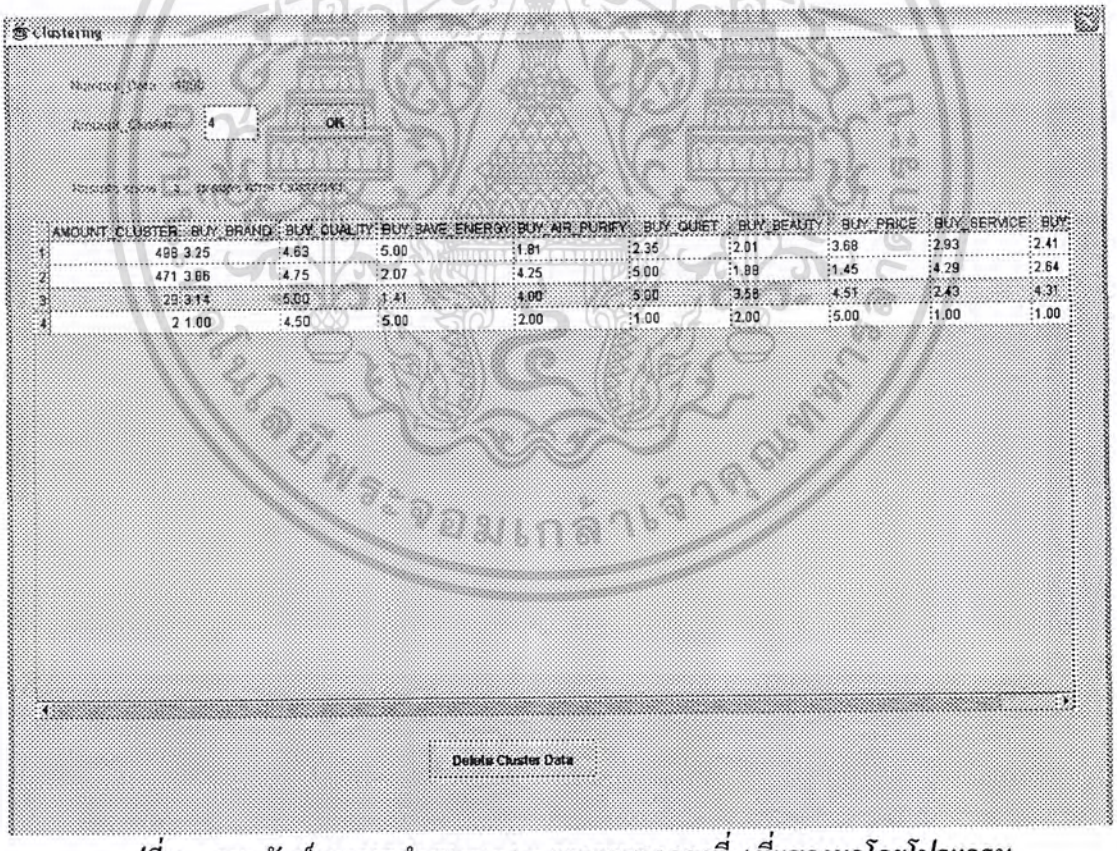
ตารางที่ 6.4-15 แสดงผลลัพธ์ที่ได้ที่แท้จริงจากการทดลองที่ 4

ผลวิเคราะห์การทดลอง

กลุ่ม	Size	Variance (%)	Achieve [Y/N]
1	498	21.76	N
2	471	25.27	N
3	29	23.69	N
4	2	47.78	N

ตารางที่ 6.4-16 ผลการวิเคราะห์ค่าเบี่ยงเบน (Variance) ในการทดลองที่ 4

Actual Variance (%): 23.53%



	AMOUNT	CLUSTER	BUY_BRAND	BUY_QUALITY	BUY_SAVE_ENERGY	BUY_AIR_PURIFY	BUY_QUIET	BUY_BEAUTY	BUY_PRICE	BUY_SERVICE	BUY
1	498	3.25	4.63	5.00	1.81	2.35	2.01	3.68	2.93	2.41	
2	471	3.66	4.75	2.07	4.25	5.00	1.89	1.45	4.29	2.64	
3	29	3.14	5.00	1.41	4.00	5.00	3.58	4.51	2.43	4.31	
4	2	1.00	4.50	5.00	2.00	1.00	2.00	5.00	1.00	1.00	

รูปที่ 6.4-7 ผลลัพธ์ของการทำ Clustering จากการทดลองที่ 4 ที่แสดงผลโดยโปรแกรม

การทดลองที่ 5

เป้าหมายในการจำลองข้อมูล (Simulation Goal)

Size แต่ละ Segment มีขนาดไม่เท่ากัน	Yes
ภายใน Segment มีคุณสมบัติที่ Random อยู่	Yes
ทุก Segment มีคุณสมบัติที่ Random อยู่	No
มีส่วนของการตลาดที่ Random อยู่	Yes
มี Segment ที่มีขนาดเล็กมาก (ต่ำกว่า 10%)	No

การแบ่งส่วนการตลาด (Market Segmentation)

กลุ่มที่ 1 25 % ลูกค้าเล็งเห็นถึง คุณภาพ ประหยัดพลังงานและระบบฟอกอากาศ เป็นสำคัญ ส่วนปัจจัยด้านอื่นๆที่เหลือสนใจปานกลาง

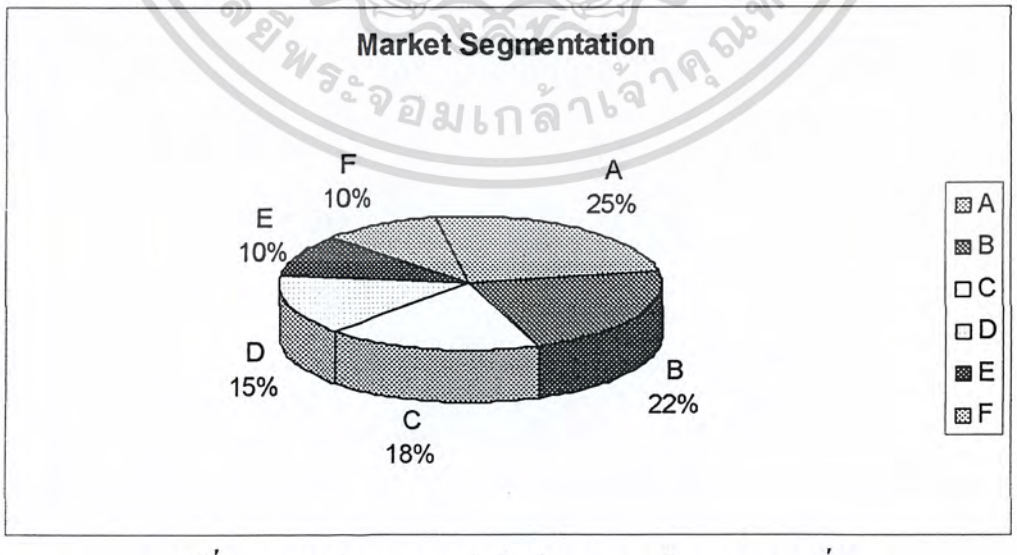
กลุ่มที่ 2 22 % ลูกค้าเล็งเห็นถึงราคา การบริการหลังการขายและคุณภาพเป็นสำคัญ สนใจเรื่องความสวยงามค่อนข้างน้อย ส่วนปัจจัยที่เหลือ สนใจบ้าง

กลุ่มที่ 3 18 % ลูกค้าเล็งเห็นถึงความสำคัญของปัจจัยโดยรวมทั้งหมด ซึ่งจะเน้นที่ยี่ห้อ ราคาและคุณภาพก่อนปัจจัยอื่นๆ

กลุ่มที่ 4 15 % ลูกค้าเล็งเห็นถึงราคา ยี่ห้อ คุณภาพ และประหยัดพลังงานเป็นสำคัญ ซึ่งจะเน้นที่ราคาก่อน ส่วนปัจจัยอื่นๆที่เหลือ สนใจค่อนข้างน้อย

กลุ่มที่ 5 10 % ลูกค้าเล็งเห็นถึง คุณภาพ ประหยัดพลังงานและความเงียบเป็นสำคัญ ไม่สนใจเรื่องความสวยงามและอื่นๆ ส่วนปัจจัยที่เหลือ สนใจปานกลาง

กลุ่มที่ 6 10 % Random



รูปที่ 6.4-8 กราฟแสดงการแบ่งส่วนในการตลาดในการทดลองที่ 5

การแมพอินพุท (Input Mapping)

กลุ่ม	Size	Brand	Quality	Energy	Air Purify	Quiet	Beauty	Price	Service	Others
1	250	ปานกลาง	สูงมาก	สูงมาก	สูงมาก	ปานกลาง	ปานกลาง	ปานกลาง	ปานกลาง	ปานกลาง
2	220	เฉลี่ย	สูงมาก	เฉลี่ย	เฉลี่ย	เฉลี่ย	ต่ำ	สูงมาก	สูงมาก	เฉลี่ย
3	180	สูงมาก	สูงมาก	สูงมาก	สูง	สูง	สูง	สูงมาก	สูง	สูง
4	150	สูง	สูง	สูง	ต่ำ	ต่ำ	ต่ำ	สูงมาก	ต่ำ	ต่ำ
5	100	ปานกลาง	สูงมาก	สูงมาก	ปานกลาง	สูงมาก	ต่ำมาก	ปานกลาง	ปานกลาง	ต่ำมาก
6	100	เฉลี่ย	เฉลี่ย	เฉลี่ย	เฉลี่ย	เฉลี่ย	เฉลี่ย	เฉลี่ย	เฉลี่ย	เฉลี่ย

ตารางที่ 6.4-17 การแมพอินพุทจากการแบ่งส่วนการตลาดในการทดลองที่ 5

การจำลองอินพุท (Input Simulation)

จำนวนของข้อมูลที่จำลอง (Number of sampling): 1000

กลุ่ม	Size	Brand	Quality	Energy	Air Purify	Quiet	Beauty	Price	Service	Others
1	250	3	5	5	5	3	3	3	3	3
2	220	Random	5	Random	Random	Random	2	5	5	Random
3	180	5	5	5	4	4	4	5	4	4
4	150	4	4	4	2	2	2	5	2	2
5	100	3	5	5	3	5	1	3	3	1
6	100	Random	Random	Random	Random	Random	Random	Random	Random	Random

ตารางที่ 6.4-18 การจำลองอินพุทที่ใช้ในการทดลองที่ 5

ผลลัพธ์ที่คาดหวังไว้ (Output Expected)

จำนวนของคลัสเตอร์ (Number of cluster): 6

Maximum Variance (%): 15

กลุ่ม	Size	Brand	Quality	Energy	Air Purify	Quiet	Beauty	Price	Service	Others
1	250	3	5	5	5	3	3	3	3	3
2	220	3	5	3	3	3	2	5	5	3
3	180	5	5	5	4	4	4	5	4	4
4	150	4	4	4	2	2	2	5	2	2
5	100	3	5	5	3	5	1	3	3	1
6	100	3	3	3	3	3	3	3	3	3

ตารางที่ 6.4-19 แสดงผลลัพธ์ที่คาดหวังไว้ในการทดลองที่ 5

ผลลัพธ์ที่แท้จริง (Output Actual)

จำนวนของคลัสเตอร์ (Number of cluster): 6

กลุ่ม	Size	Brand	Quality	Energy	Air Purify	Quiet	Beauty	Price	Service	Others
1	250	3	5	5	5	3	3	3	3	3
2	220	3.09	5	1.88	2	1	2	5	5	5
3	180	5	5	5	4	4	4	5	4	4
4	150	4	4	4	2	2	2	5	2	2
5	100	3	5	5	3	5	1	3	3	1
6	100	1.34	1.65	2.02	1	1	5	2.79	1	5

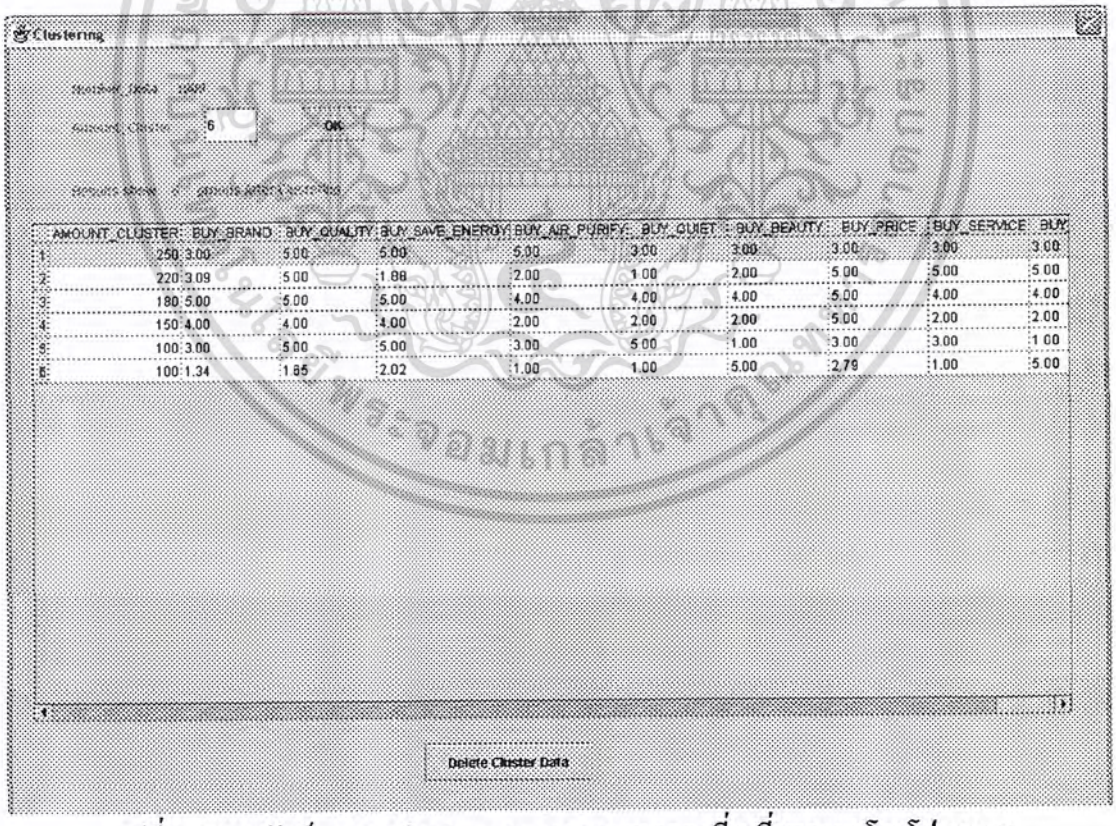
ตารางที่ 6.4-20 แสดงผลลัพธ์ที่ได้ที่แท้จริงจากการทดลองที่ 5

ผลวิเคราะห์การทดลอง

กลุ่ม	Size	Variance (%)	Achieve [Y/N]
1	250	0	N
2	220	13.8	Y
3	180	0	N
4	150	0	N
5	100	0	N
6	100	14.2	Y

ตารางที่ 6.4-21 ผลการวิเคราะห์ค่าเบี่ยงเบน (Variance) ในการทดลองที่ 5

Actual Variance (%): 4.46%



	AMOUNT_CLUSTER	BUY_BRAND	BUY_QUALITY	BUY_SAVE_ENERGY	BUY_AIR_PURIFY	BUY_QUIET	BUY_BEAUTY	BUY_PRICE	BUY_SERVICE	BUY
1	250	3.00	5.00	5.00	5.00	3.00	3.00	3.00	3.00	3.00
2	220	3.09	5.00	1.86	2.00	1.00	2.00	5.00	5.00	5.00
3	180	5.00	5.00	5.00	4.00	4.00	4.00	5.00	4.00	4.00
4	150	4.00	4.00	4.00	2.00	2.00	2.00	5.00	2.00	2.00
5	100	3.00	5.00	5.00	3.00	5.00	1.00	3.00	3.00	1.00
6	100	1.34	1.85	2.02	1.00	1.00	5.00	2.79	1.00	5.00

รูปที่ 6.4-9 ผลลัพธ์ของการทำ Clustering จากการทดลองที่ 5 ที่แสดงผลโดยโปรแกรม

การทดลองที่ 6

เป้าหมายในการจำลองข้อมูล (Simulation Goal)

Size แต่ละ Segment มีขนาดไม่เท่ากัน	Yes
ภายใน Segment มีคุณสมบัติที่ Random อยู่	Yes
ทุก Segment มีคุณสมบัติที่ Random อยู่	Yes
มีส่วนของการตลาดที่ Random อยู่	Yes
มี Segment ที่มีขนาดเล็กมาก (ต่ำกว่า 10%)	Yes

การแบ่งส่วนการตลาด การแบ่งส่วนการตลาด (Market Segmentation)

กลุ่มที่ 1 30 % ลูกค้าเล็งเห็นถึง คุณภาพ ประหยัดพลังงาน ความเงียบและระบบฟอกอากาศเป็นสำคัญ โดยที่มองในเรื่องของราคาและความสวยงามค่อนข้างน้อย ส่วนปัจจัยที่เหลือสนใจในระดับเฉลี่ย

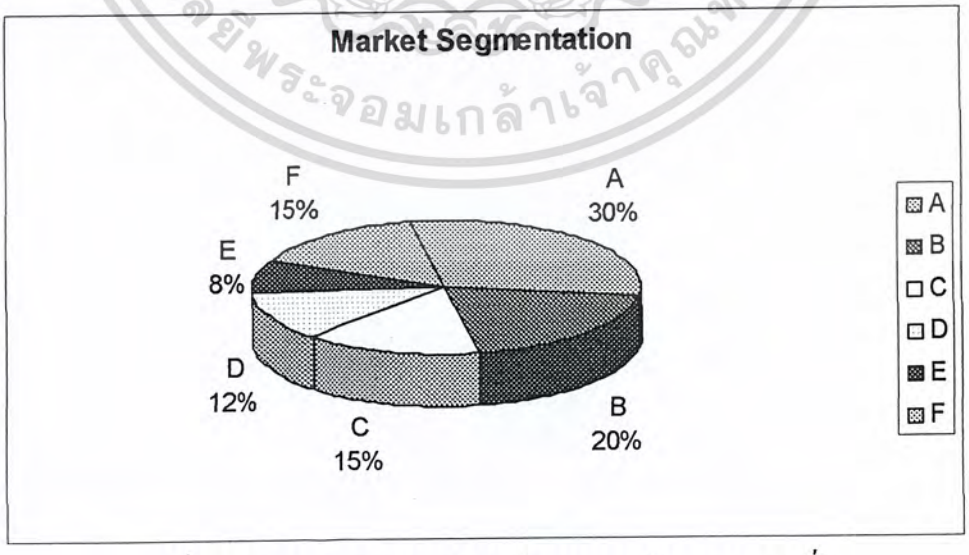
กลุ่มที่ 2 20 % ลูกค้าเล็งเห็นถึง คุณภาพ ประหยัดพลังงาน ความเงียบ ยี่ห้อและระบบฟอกอากาศเป็นสำคัญ แต่จะมองที่ 3 ปัจจัยแรกก่อน ส่วนปัจจัยที่เหลือ สนใจในระดับเฉลี่ย

กลุ่มที่ 3 15 % ลูกค้าเล็งเห็นถึงความสำคัญของปัจจัยโดยรวมทั้งหมด ซึ่งราคา ความสวยงามและอื่นๆนั้น สนใจในระดับเฉลี่ย

กลุ่มที่ 4 15 % Random

กลุ่มที่ 5 12 % ลูกค้าเล็งเห็นถึงปัจจัยโดยรวมทั้งหมดอยู่ในระดับเฉลี่ย ซึ่งจะเน้นที่ราคาและคุณภาพก่อนปัจจัยอื่นๆ

กลุ่มที่ 6 8 % ลูกค้าเล็งเห็นถึงราคา เป็นสำคัญ โดยที่คุณภาพและประหยัดไฟเป็นปัจจัยที่รองลง ส่วนปัจจัยที่เหลือ สนใจในระดับเฉลี่ย



รูปที่ 6.4-10 กราฟแสดงการแบ่งส่วนในการตลาดในการทดลองที่ 6

การแมพอินพุท (Input Mapping)

กลุ่ม	Size	Brand	Quality	Energy	Air Purify	Quiet	Beauty	Price	Service	Others
1	300	เฉลี่ย	สูงมาก	สูงมาก	สูงมาก	สูงมาก	ต่ำ	ต่ำ	เฉลี่ย	เฉลี่ย
2	200	สูง	สูงมาก	สูงมาก	สูง	สูงมาก	เฉลี่ย	เฉลี่ย	เฉลี่ย	เฉลี่ย
3	150	สูงมาก	สูงมาก	สูงมาก	สูง	สูง	เฉลี่ย	เฉลี่ย	สูงมาก	เฉลี่ย
4	150	เฉลี่ย	เฉลี่ย	เฉลี่ย	เฉลี่ย	เฉลี่ย	เฉลี่ย	เฉลี่ย	เฉลี่ย	เฉลี่ย
5	120	เฉลี่ย	สูงมาก	เฉลี่ย	เฉลี่ย	เฉลี่ย	เฉลี่ย	สูงมาก	เฉลี่ย	เฉลี่ย
6	80	เฉลี่ย	สูง	สูง	เฉลี่ย	เฉลี่ย	เฉลี่ย	สูงมาก	เฉลี่ย	เฉลี่ย

ตารางที่ 6.4-22 การแมพอินพุทจากการแบ่งส่วนการตลาดในการทดลองที่ 6

การจำลองอินพุท (Input Simulation)

จำนวนของข้อมูลที่จำลอง (Number of sampling): 1000

กลุ่ม	Size	Brand	Quality	Energy	Air Purify	Quiet	Beauty	Price	Service	Others
1	300	Random	5	5	5	5	2	2	Random	Random
2	200	4	5	5	4	5	Random	Random	Random	Random
3	150	5	5	5	4	4	Random	Random	5	Random
4	150	Random	Random	Random	Random	Random	Random	Random	Random	Random
5	120	Random	5	Random	Random	Random	Random	5	Random	Random
6	80	Random	4	4	Random	Random	Random	5	Random	Random

ตารางที่ 6.4-23 การจำลองอินพุทที่ใช้ในการทดลองที่ 6

ผลลัพธ์ที่คาดหวังไว้ (Output Expected)

จำนวนของคลัสเตอร์ (Number of cluster) : 6

Maximum Variance (%) : 15

กลุ่ม	Size	Brand	Quality	Energy	Air Purify	Quiet	Beauty	Price	Service	Others
1	300	3	5	5	5	5	2	2	3	3
2	200	4	5	5	4	5	3	3	3	3
3	150	5	5	5	4	4	3	3	5	3
4	150	3	3	3	3	3	3	3	3	3
5	120	3	5	3	3	3	3	5	3	3
6	80	3	4	4	3	3	3	5	3	3

ตารางที่ 6.4-24 แสดงผลลัพธ์ที่คาดหวังไว้ในการทดลองที่ 6

ผลลัพธ์ที่แท้จริง (Output Actual)

จำนวนของคลัสเตอร์ (Number of cluster): 6

กลุ่ม	Size	Brand	Quality	Energy	Air Purify	Quiet	Beauty	Price	Service	Others
1	500	3.59	4.75	4.75	4.5	3.75	2	4.25	1.95	3.21
2	350	4.5	5	5	4	4.5	4.5	3.5	4.5	2.75
3	143	2.5	1.43	1.69	1.86	1.4	2.3	2.83	2.35	3.62
4	3	2	1.25	3	2.5	2	3	1.75	1	1.5
5	2	1	3	1	1	1	1	4	2	1.5
6	2	3	3	1.5	3	1	1	1	3	4.5

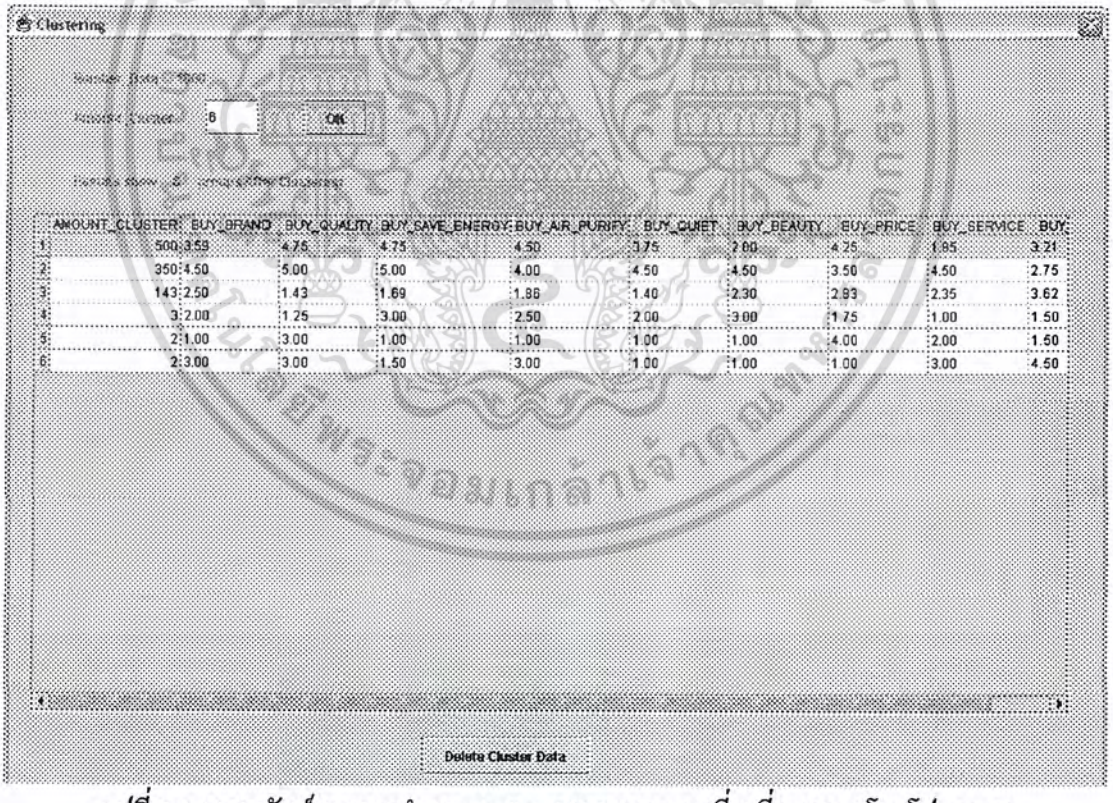
ตารางที่ 6.4-25 แสดงผลลัพธ์ที่ได้ที่แท้จริงจากการทดลองที่ 6

ผลวิเคราะห์การทดลอง

กลุ่ม	Size	Variance (%)	Achieve [Y/N]
1	500	14.1	Y
2	350	10.56	Y
3	143	40.58	N
4	3	20	N
5	2	31.11	N
6	2	31.11	N

ตารางที่ 6.4-26 ผลการวิเคราะห์ค่าเบี่ยงเบน (Variance) ในการทดลองที่ 6

Actual Variance (%): 16.73%



	AMOUNT_CLUSTER	BUY_BRAND	BUY_QUALITY	BUY_SAVE_ENERGY	BUY_AIR_PURIFY	BUY_QUIET	BUY_BEAUTY	BUY_PRICE	BUY_SERVICE	BUY
1	500	3.59	4.75	4.75	4.50	3.75	2.00	4.25	1.95	3.21
2	350	4.50	5.00	5.00	4.00	4.50	4.50	3.50	4.50	2.75
3	143	2.50	1.43	1.69	1.86	1.40	2.30	2.93	2.35	3.62
4	3	2.00	1.25	3.00	2.50	2.00	3.00	1.75	1.00	1.50
5	2	1.00	3.00	1.00	1.00	1.00	1.00	4.00	2.00	1.50
6	2	3.00	3.00	1.50	3.00	1.00	1.00	1.00	3.00	4.50

รูปที่ 6.4-11 ผลลัพธ์ของการทำ Clustering จากการทดลองที่ 6 ที่แสดงผลโดยโปรแกรม

บทที่ 7

Decision Tree

Decision Tree เป็นกระบวนการเรียนรู้ประเภท Inductive Learning ที่ง่ายต่อการนำไปใช้ โดย Decision Tree จะรับ input จากสถานการณ์ต่างๆ ซึ่งเป็นกลุ่มของ Attribute และให้ผลลัพธ์เป็นผลลัพธ์ของการตัดสินใจ (Decision) ซึ่งคือ ค่าที่เป็นการทำนาย จาก input ดังกล่าวโดย Input สามารถเป็นได้ทั้ง Discrete และ Continuous ตัวอย่างของ Discrete-valued function คือ Classification Learning และตัวอย่างของ Continuous Function คือ Regression

Decision Tree จะเข้าถึงคำตอบโดยการทำการตัดสินใจไปเรื่อยๆ ตาม Tree โดยแต่ละ Node ภายในจะเทียบได้กับการทดสอบของค่าหนึ่งๆ และกิ่งของ Node นั้นๆ เปรียบได้เสมือนกับ ผลลัพธ์ที่เป็นไปได้ของการทดสอบนั้นๆ โดยแต่ละ Leaf Node จะเป็นผลลัพธ์ของ Decision Tree ดังกล่าว

โดยรายละเอียด และตัวอย่างของ Decision Tree ได้กล่าวไปแล้วในส่วนของ Data Mining Algorithms

7.1 การทำงานของคำสั่งต่างๆ

Treefit

เป็นการสร้าง Tree-Model โดย Regression หรือ Classification

คำสั่ง T = treefit(X,y)

T = treefit(X,y,'param1',val1,'param2',val2,...)

รายละเอียด

T = treefit(X,y) เป็นคำสั่งที่ใช้ในการสร้าง Decision Tree T สำหรับการคาดการณ์ค่า y โดยเป็นฟังก์ชั่นในการคาดการณ์จากค่า X โดย X เป็น n x m เมตริกซ์ของค่าที่ใช้ในการคาดการณ์ และ Y เป็น Vector ของ n คำตอบ (สำหรับ Regression) หรือเป็น Character Array หรือ Cell Array ของ String ซึ่งประกอบด้วย n class name (สำหรับ Classification) หรือในอีกทางหนึ่ง T เป็น Binary Tree ซึ่งแต่ละ non-terminal node นั้นเป็นพื้นฐานของคำตอบที่เป็นค่า X โดยคำสั่ง T = treefit(X,y,'param1',val1,'param2',val2,...) นั้นระบุตัวแปรเพิ่มเติม โดยตัวแปรดังกล่าว ได้แก่

สำหรับ Tree ทั้งต้น

'catidx' เป็น Vector ของ Column ของ X โดย treefit จัดการ Column ดังกล่าวเสมือนเป็น Column ที่ไม่มีการจัดเรียงลำดับของกลุ่มข้อมูล

'method' มีค่าเป็น 'classification' (เป็นค่าปรกติหาก y เป็นตัวอักษร) หรือ 'regression' (เป็นค่าปรกติหาก y เป็นตัวเลข)

'splitmin' เป็นค่าที่บ่งชี้ว่า จะแตก Node ได้ต้องมีคำตอบใน Node นั้นๆ ไม่น้อยกว่าค่าดังกล่าว โดยมีค่าปรกติเป็น 10

'prune' 'on' (เป็นค่าปรกติ) ให้มีการ Prune และ 'off' หากไม่มีค่าการ Prune

Treedisp

เป็นการแสดง Decision Tree ให้เป็นเชิง Graphic

คำสั่ง treedisp(T)

treedisp(T,'param1',val1,'param2',val2,...)

treedisp(T) นั้นเป็นรับ input เป็น Decision tree T และคำนวณโดย treefit function และแสดงอยู่บน Window โดยแต่ละ Tree นั้นถูกตั้งชื่อด้วย Decision Rule และแต่ละ Terminal Node นั้นถูกตั้งชื่อให้เป็นค่าที่เป็นการคาดการณ์

treedisp(T,'param1',val1,'param2',val2,...) นั้นสามารถใส่ตัวแปรเพิ่มเติมเข้าไปได้ โดย

'name' นั้นเป็น Array ของชื่อที่ใช้ในการคาดการณ์ค่าตัวแปรต่างๆ

'prunelevel' เป็นการใช้ค่าเริ่มต้นในการแสดง Pruning Level ที่แสดงผลออกมา

7.2 การทดลอง

การทดลองที่ 1

จุดประสงค์การจำลองข้อมูล

ต้องการ Bias ว่าการอัตราการ Service เป็นปัจจัยเดียวที่มีผลต่อความพึงพอใจ

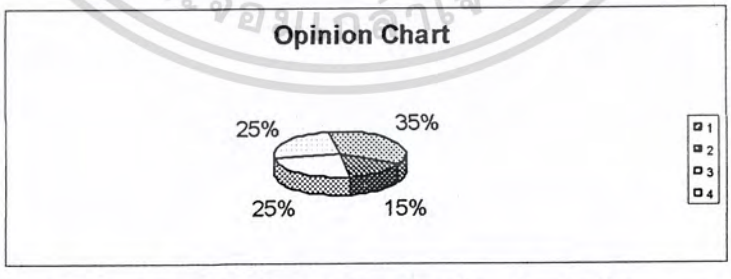
จำนวนตัวแปรที่เกี่ยวข้องกับการ Bias	1
Percentage ที่ถูก Bias โดยตรง (%) :	60
จำนวนข้อมูลตัวอย่าง :	1,000
ภายใน Segment มีคุณสมบัติที่ Random อยู่	No
ทุก Segment มีคุณสมบัติที่ Random อยู่	No
มีส่วนของการตลาดที่ Random อยู่	No

ตารางที่ 7.2-1 แสดงปัจจัยที่เกี่ยวข้องในการสร้างข้อมูลในผลการทดลองที่ 1

สรุปข้อมูลที่ใช้ในการจำลอง

No.	%	รายละเอียด
1*	35	มีอัตราการซ่อมดีมาก ตัวแปรอื่นปานกลาง มีความพึงพอใจสูงมาก ตัวแปรอื่นๆ ปานกลาง
2	15	มีอัตราการซ่อมดี แต่มักซ่อมไม่เสร็จในครั้งเดียว การซ่อมหลังติดตั้งและจำนวนปีไม่กี มีความพึงพอใจสูง
3*	25	มีอัตราการซ่อมไม่ดีมาก ตัวแปรอื่นดีมาก มีความพึงพอใจแย่มาก
4	25	มีอัตราการซ่อมไม่ดีแต่การซ่อมหลังติดตั้งดีมาก มักซ่อมเสร็จในครั้งเดียว ติดตั้งนาน มีความพึงพอใจแย่

ตารางที่ 7.2-2 สรุปข้อมูลที่ใช้ในการจำลองในผลการทดลองที่ 1



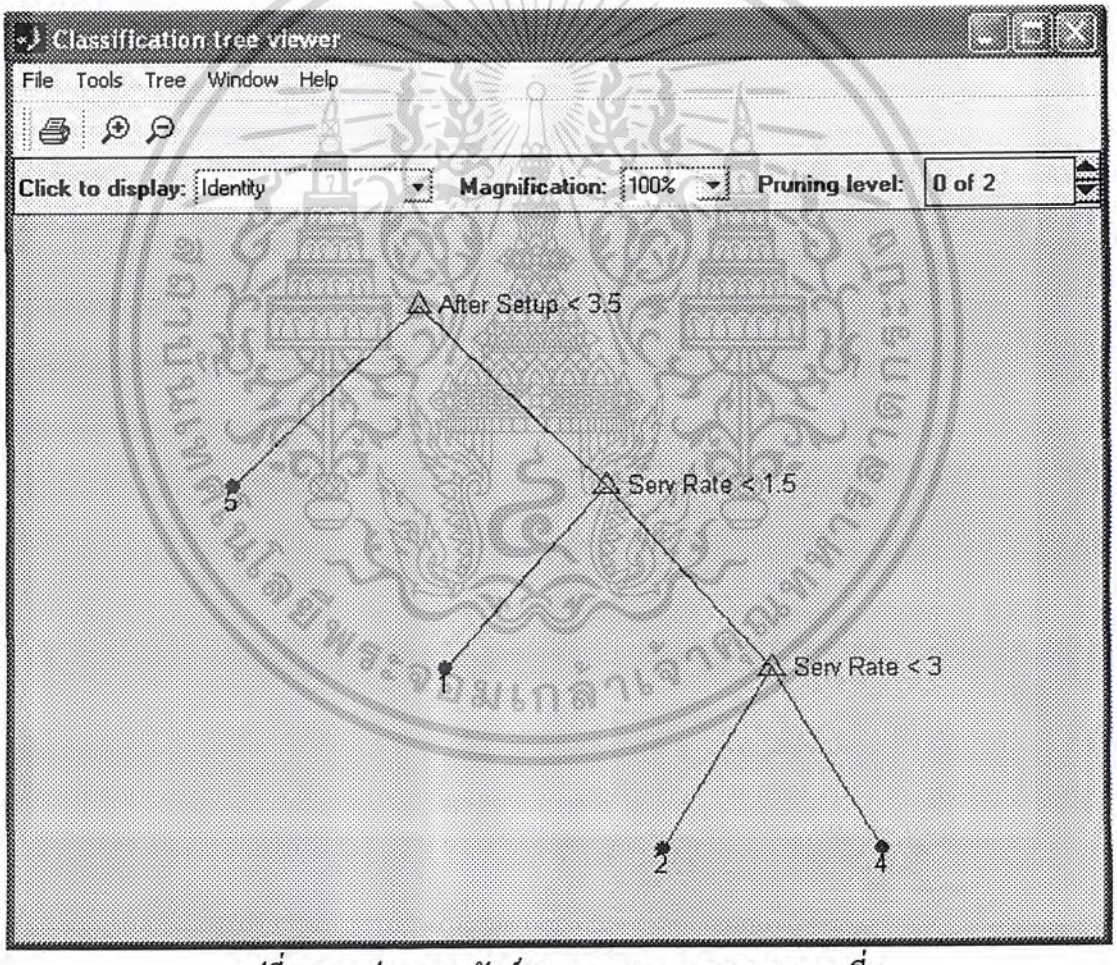
รูปที่ 7.2-1 กราฟแสดงอัตราส่วนในการจำลองข้อมูลในแต่ละส่วนในผลการทดลองที่ 1

โครงสร้างข้อมูลที่ใช้ในการจำลองข้อมูล

No.	%	จำนวน	Service Rate Sat.	After Setup Satisfac.	Not Finish Satisfac.	No. Year	Service Point
1	35	350	5	3	3	3	5
2	15	150	4	4	2	2	4
3	25	250	1	5	5	5	1
4	25	250	2	5	4	4	2

ตารางที่ 7.2-3 โครงสร้างข้อมูลที่ใช้ในการจำลองข้อมูลในการทดลองที่ 1

ผลลัพธ์



รูปที่ 7.2-2 รูปแสดงผลลัพธ์ Decision Tree จากการทดลองที่ 1

สรุปผลการทดลอง

อัตราความถูกต้องของ Decision Tree คือ 100%

การทดลองที่ 2

จุดประสงค์การจำลองข้อมูล

ต้องการ Bias ว่าการอัตราการ Service และจำนวนปีเป็นเพียงสองปัจจัยที่มีผลต่อความพึงพอใจ

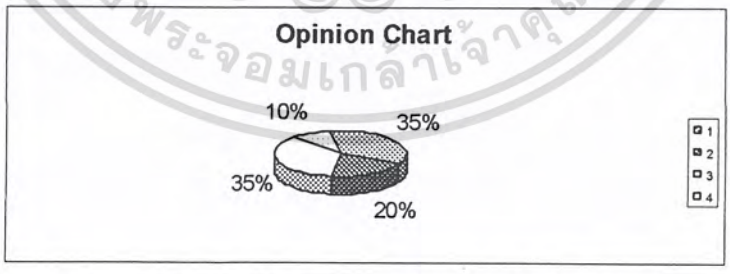
จำนวนตัวแปรที่เกี่ยวข้องกับการ Bias	2
Percentage ที่ถูก Bias โดยตรง (%) :	70
จำนวนข้อมูลตัวอย่าง :	1,000
ภายใน Segment มีคุณสมบัติที่ Random อยู่	Yes
ทุก Segment มีคุณสมบัติที่ Random อยู่	No
มีส่วนของการตลาดที่ Random อยู่	No

ตารางที่ 7.2-4 แสดงปัจจัยที่เกี่ยวข้องในการสร้างข้อมูลในผลการทดลองที่ 2

สรุปข้อมูลที่ใช้ในการจำลอง

No.	%	รายละเอียด
1*	35	มีอัตราการซ่อมดีมาก และระยะเวลาใช้งานดีมาก แต่ตัวแปรอื่นไม่ดีมาก มีความพึงพอใจสูงมาก
2	20	มีอัตราการซ่อมดีมาก และระยะเวลาใช้งานไม่ดี แต่การซ่อมหลังติดตั้งไม่ดี มีความพึงพอใจสูง
3*	35	มีอัตราการซ่อมแย่มาก และระยะเวลาใช้งานไม่ดี แต่ตัวแปรอื่นดีมาก มีความพึงพอใจแย่มาก
4	10	มีอัตราการซ่อมปานกลาง และระยะเวลาใช้งานดีมาก แต่มักซ่อมไม่เสร็จในครั้งเดียว มีความพึงพอใจสูง

ตารางที่ 7.2-5 สรุปข้อมูลที่ใช้ในการจำลองในผลการทดลองที่ 2



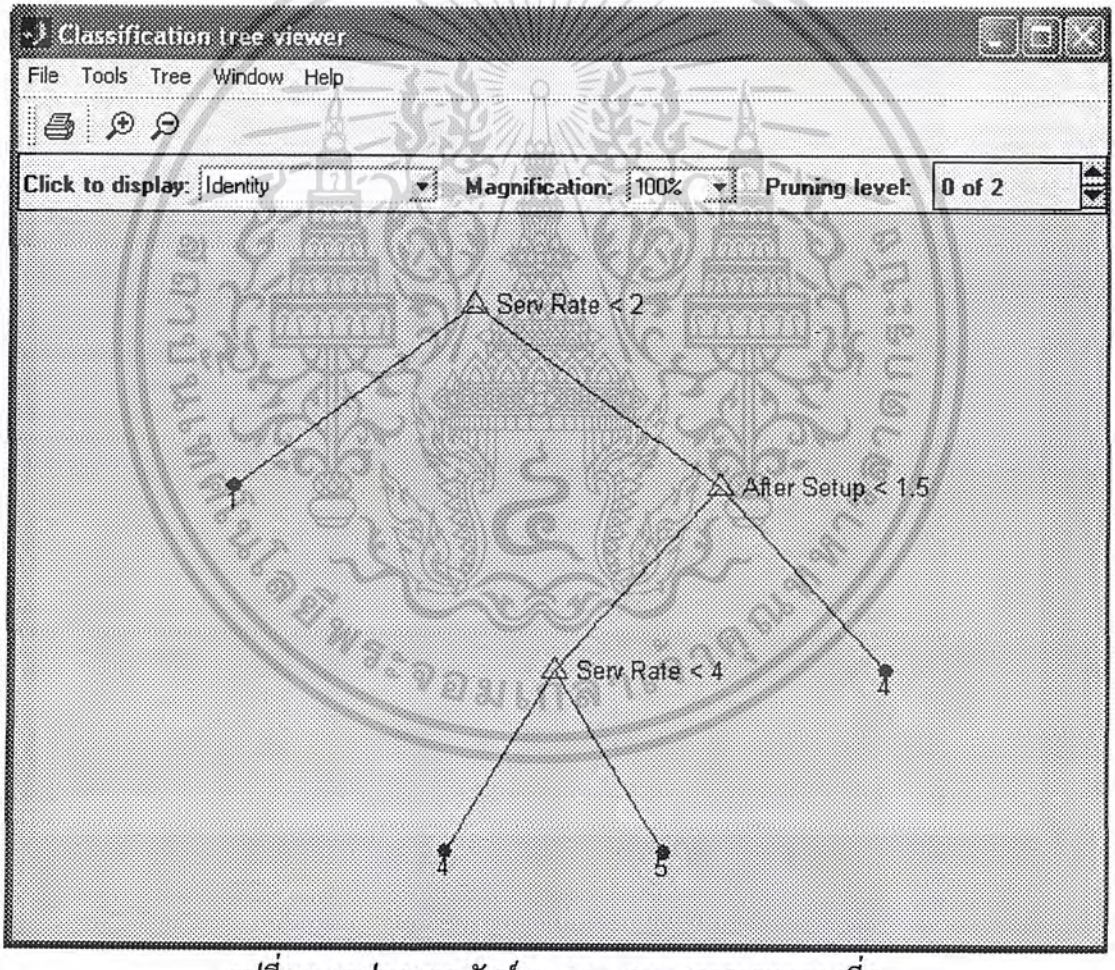
รูปที่ 7.2-3 กราฟแสดงอัตราส่วนในการจำลองข้อมูลในแต่ละส่วนในผลการทดลองที่ 2

โครงสร้างข้อมูลที่ใช้ในการจำลอง

No.	%	Number	Service Rate Sat.	After Setup Satisfac.	Not Finish Satisfac.	No. Year	Service Point
1	35	350	5	1	1	5	5
2	15	150	5	2	Random	2	4
3	25	250	1	5	5	2	1
4	25	250	3	Random	1	5	4

ตารางที่ 7.2-6 โครงสร้างข้อมูลที่ใช้ในการจำลองข้อมูลในการทดลองที่ 2

ผลลัพธ์



รูปที่ 7.2-4 รูปแสดงผลลัพธ์ Decision Tree จากการทดลองที่ 2

สรุปผลการทดลอง

อัตราความถูกต้องของ Decision Tree คือ 100%

การทดลองที่ 3

จุดประสงค์การจำลองข้อมูล

ต้องการ Bias ว่าการซ่อมหลังติดตั้งไม่มีผลต่อความพึงพอใจของลูกค้ามากเท่าไหร่ แต่ปัจจัยอื่นๆ มีผลต่อความพึงพอใจโดยตรง

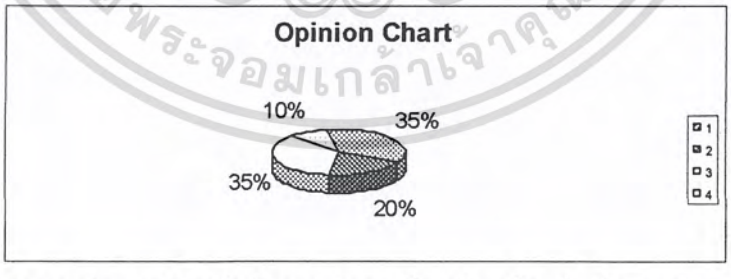
จำนวนตัวแปรที่เกี่ยวข้องกับการ Bias	4
Percentage ที่ถูก Bias โดยตรง (%) :	70
จำนวนข้อมูลตัวอย่าง :	1,000
ภายใน Segment มีคุณสมบัติที่ Random อยู่	No
ทุก Segment มีคุณสมบัติที่ Random อยู่	No
มีส่วนของการตลาดที่ Random อยู่	No

ตารางที่ 7.2-7 แสดงปัจจัยที่เกี่ยวข้องในการสร้างข้อมูลในผลการทดลองที่ 3

สรุปข้อมูลที่ใช้ในการจำลอง

No.	%	รายละเอียด
1*	40	มีอัตราซ่อมดี และอายุการใช้งานดี แต่การซ่อมหลังติดตั้งไม่ดี แต่มีความพึงพอใจดี
2	20	มีอัตราซ่อมปานกลาง และซ่อมเสร็จไม่เสร็จไม่ดีมาก อายุการใช้งานปานกลาง จึงไม่ค่อยพอใจ
3*	30	มีอัตราซ่อมแย่มาก แต่การซ่อมหลังติดตั้งอยู่ในเกณฑ์ต่ำ ใช้งานไม่นาน จึงไม่พอใจมาก
4	10	มีอัตราซ่อมดีมาก และอายุการใช้งานนาน จึงมีความพึงพอใจดีมาก

ตารางที่ 7.2-8 สรุปข้อมูลที่ใช้ในการจำลองในผลการทดลองที่ 3



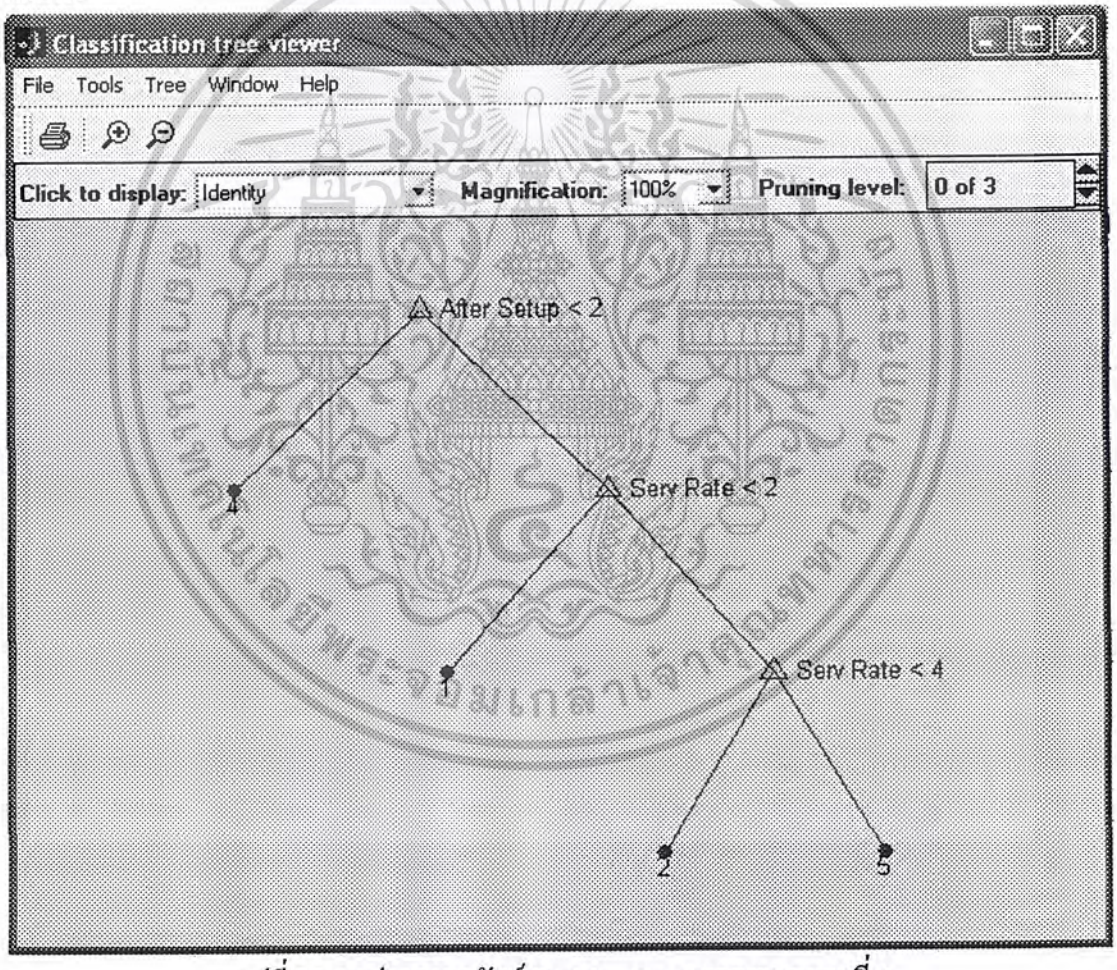
รูปที่ 7.2-5 กราฟแสดงอัตราส่วนในการจำลองข้อมูลในแต่ละส่วนในผลการทดลองที่ 3

โครงสร้างข้อมูลที่ใช้ในการจำลอง

No.	%	Number	Service Rate Sat.	After Setup Satisfac.	Not Finish Satisfac.	No. Year	Service Point
1	40	400	5	1	5	4	4
2	20	200	3	3	1	3	2
3	30	300	1	4	1	2	1
4	10	100	5	5	5	5	5

ตารางที่ 7.2-9 โครงสร้างข้อมูลที่ใช้ในการจำลองข้อมูลในการทดลองที่ 3

ผลลัพธ์



รูปที่ 7.2-6 รูปแสดงผลลัพธ์ Decision Tree จากการทดลองที่ 3

สรุปผลการทดลอง

อัตราความถูกต้องของ Decision Tree คือ 100%

การทดลองที่ 4

จุดประสงค์การจำลองข้อมูล

ต้องการ Bias ว่าอัตราการซ่อม และจำนวนปีมีผลต่อความพึงพอใจสูง แต่มีข้อยกเว้นว่าถ้าอัตราการซ่อมไม่เสร็จในครั้งเดียวมีสูงมาก ความพึงพอใจจะอยู่ในระดับต่ำทันที

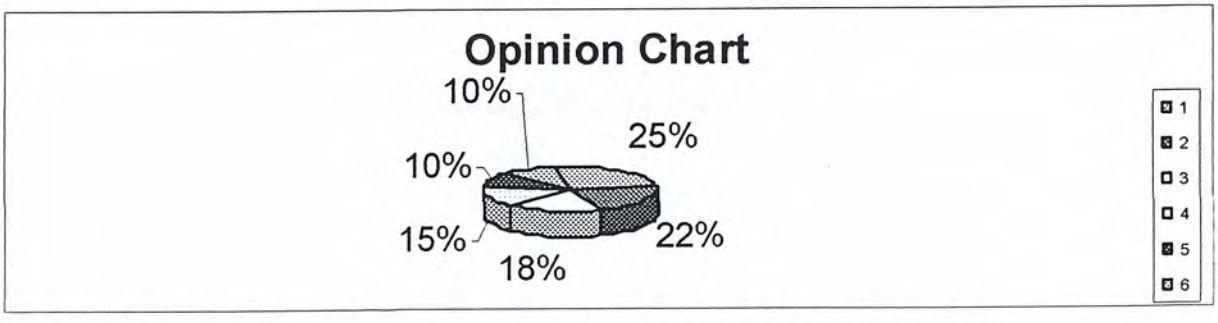
จำนวนตัวแปรที่เกี่ยวข้องกับการ Bias	3
Percentage ที่ถูก Bias โดยตรง (%) :	70
จำนวนข้อมูลตัวอย่าง :	1,000
ภายใน Segment มีคุณสมบัติที่ Random อยู่	Yes
ทุก Segment มีคุณสมบัติที่ Random อยู่	No
มีส่วนของการตลาดที่ Random อยู่	Yes

ตารางที่ 7.2-10 แสดงปัจจัยที่เกี่ยวข้องในการสร้างข้อมูลในผลการทดลองที่ 4

สรุปข้อมูลที่ใช้ในการจำลอง

No.	%	Characteristic
1*	25	มีอัตราการซ่อมดีมาก และระยะเวลาหลังติดตั้งเฉลี่ยดี จึงมีความพอใจสูง
2*	22	มีอัตราการซ่อมเฉลี่ยดี ระยะเวลาหลังติดตั้งค่อนข้างดี แต่อัตราซ่อมไม่เสร็จไม่ดี มีความพอใจเฉลี่ยสูง
3*	18	มีอัตราการซ่อม และระยะเวลาใช้งานเฉลี่ยดี แต่อัตราซ่อมไม่เสร็จดี ความพอใจจึงต่ำมาก อื่นๆ ปานกลาง
4*	15	มีอัตราการซ่อมเฉลี่ยไม่ดี ระยะเวลาหลังติดตั้งเฉลี่ยไม่ดี จึงมีความพอใจต่ำ อื่นๆ เฉลี่ยปานกลาง
5*	10	มีอัตราการซ่อมเฉลี่ยไม่ดี ระยะเวลาหลังติดตั้งเฉลี่ยไม่ดี อื่นๆ ไม่ดี ความพอใจจึงต่ำ
6	10	Random

ตารางที่ 7.2-11 สรุปข้อมูลที่ใช้ในการจำลองในผลการทดลองที่ 4



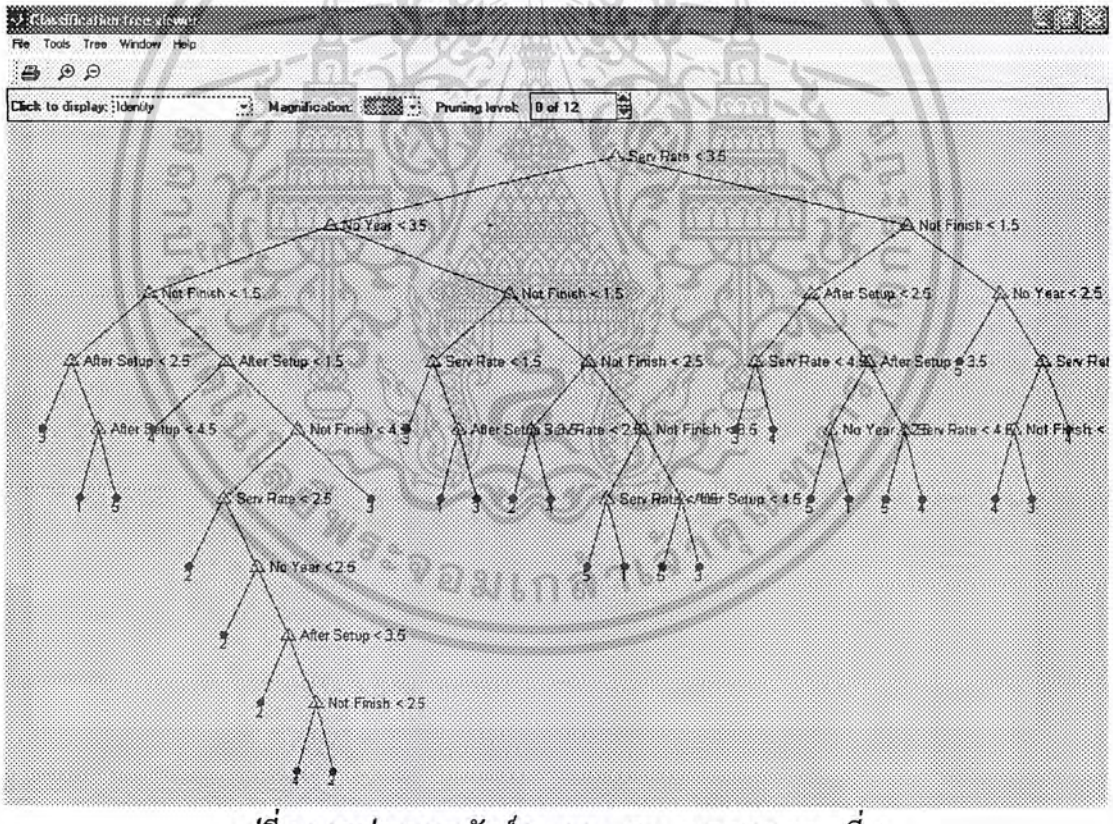
รูปที่ 7.2-7 กราฟแสดงอัตราส่วนในการจำลองข้อมูลในแต่ละส่วนในผลการทดลองที่ 4

โครงสร้างข้อมูลที่ใช้ในการจำลอง

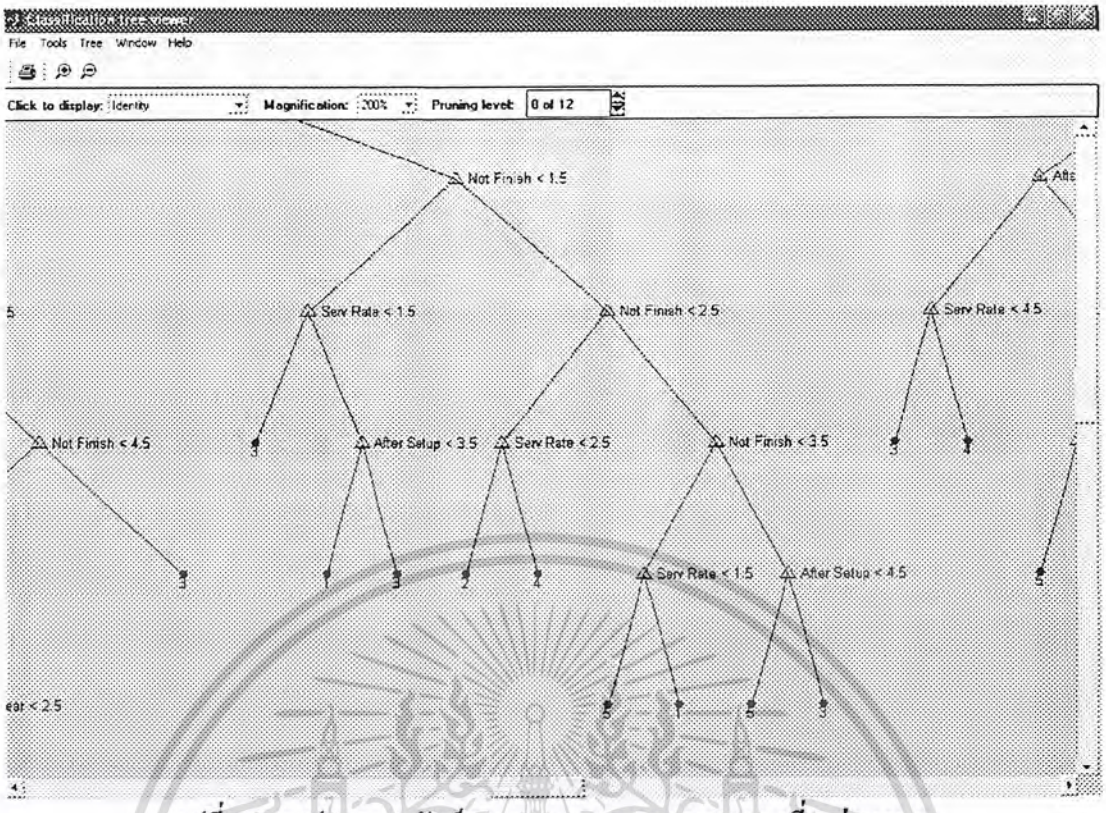
No.	%	Number	Service Rate Sat.	After Setup Satisfac.	Not Finish Satisfac.	No. Year	Service Point
1	25	250	5	Random	Random	3 - 5	4
2	22	220	3 - 5	Random	2	3 - 5	4
3	18	180	3 - 5	3	1	3 - 5	1
4	15	150	1 - 3	2 - 4	2 - 4	1 - 3	2
5	10	100	1 - 3	2	2	1 - 3	2
6	10	100	Random	Random	Random	Random	Random

ตารางที่ 7.2-12 โครงสร้างข้อมูลที่ใช้ในการจำลองข้อมูลในการทดลองที่ 4

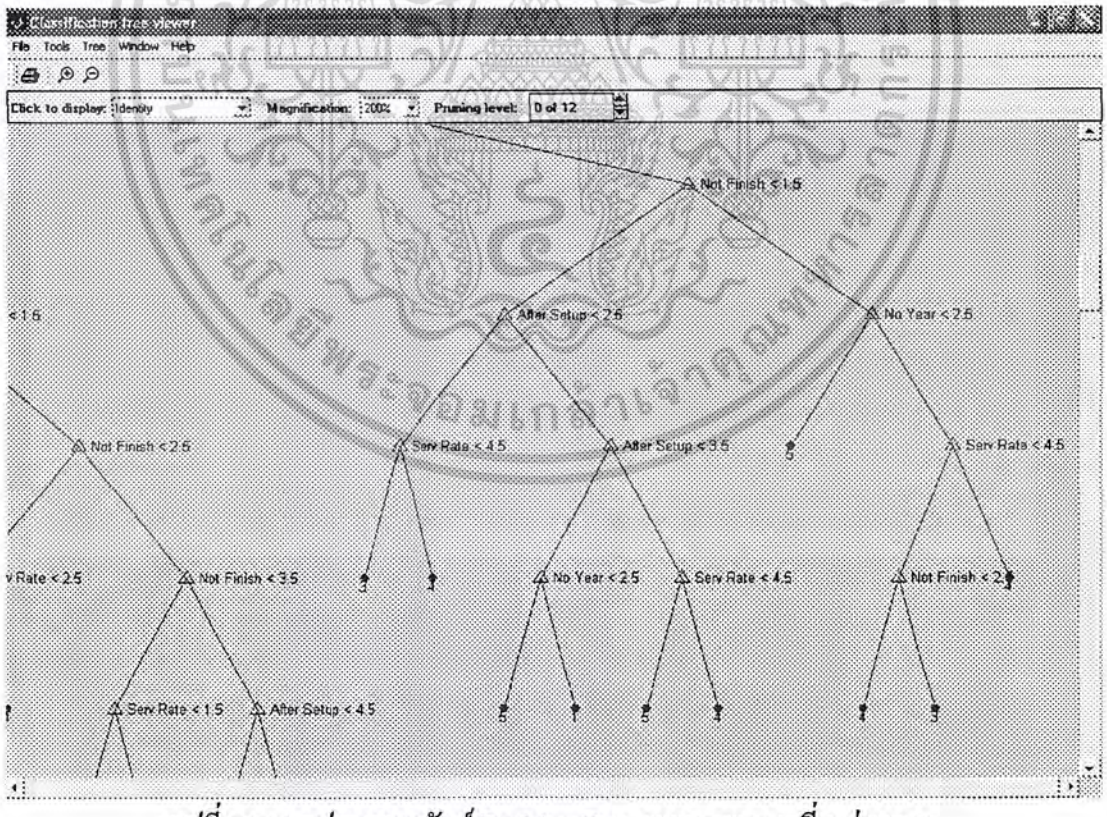
ผลลัพธ์



รูปที่ 7.2-8 รูปแสดงผลลัพธ์ Decision Tree จากการทดลองที่ 4



รูปที่ 7.2-9 รูปแสดงผลลัพธ์ Decision Tree จากการทดลองที่ 4 ส่วนกลาง



รูปที่ 7.2-10 รูปแสดงผลลัพธ์ Decision Tree จากการทดลองที่ 4 ส่วนขวา

สรุปผลการทดลอง

อัตราความถูกต้องของ Decision Tree คือ 92.5%

การทดลองที่ 5

จุดประสงค์การจำลองข้อมูล

ต้องการ Bias ว่าอัตราการซ่อม และจำนวนปีมีผลต่อความพึงพอใจสูง แต่มีข้อยกเว้นว่าถ้าอัตราการซ่อมไม่เสร็จในครั้งเดียว หรืออัตราซ่อมหลังติดตั้งสูงมาก ความพึงพอใจจะอยู่ในระดับต่ำทันที

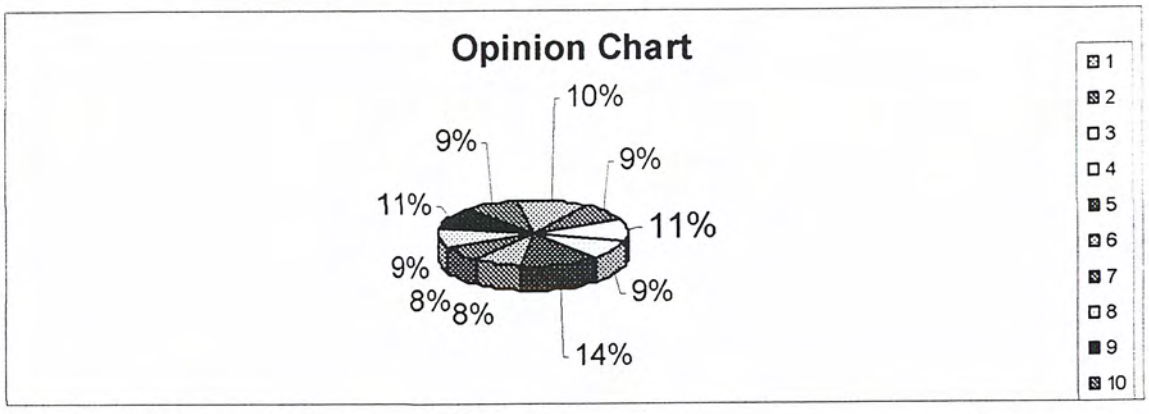
จำนวนตัวแปรที่เกี่ยวข้องกับการ Bias	4
Percentage ที่ถูก Bias โดยตรง (%) :	90
จำนวนข้อมูลตัวอย่าง :	1,000
ภายใน Segment มีคุณสมบัติที่ Random อยู่	Yes
ทุก Segment มีคุณสมบัติที่ Random อยู่	Yes
มีส่วนของการตลาดที่ Random อยู่	Yes

ตารางที่ 7.2-13 แสดงปัจจัยที่เกี่ยวข้องในการสร้างข้อมูลในผลการทดลองที่ 5

สรุปข้อมูลที่ใช้ในการจำลอง

No.	%	Characteristic
1*	12	มีอัตราการซ่อมดีมาก และระยะเวลาหลังติดตั้งเฉลี่ยนาน จึงมีความพอใจสูง
2*	9	มีอัตราการซ่อมเฉลี่ยดี ระยะเวลาหลังติดตั้งค่อนข้างนาน แต่อัตราซ่อมไม่เสร็จต่ำ มีความพอใจเฉลี่ยสูง
3*	11	มีอัตราการซ่อม และระยะเวลาใช้งานเฉลี่ยดี แต่อัตราซ่อมไม่เสร็จไม่ดี ความพอใจจึงต่ำมาก อื่นๆ กลางๆ
4*	9	มีอัตราการซ่อมเฉลี่ยไม่ดี ระยะเวลาหลังติดตั้งเฉลี่ยไม่ดี จึงมีความพอใจต่ำ อื่นๆ เฉลี่ยปานกลาง
5*	14	มีอัตราการซ่อมเฉลี่ยไม่ดี ระยะเวลาหลังติดตั้งเฉลี่ยไม่ดี อื่นๆ ไม่ดี ความพอใจจึงต่ำ
6*	8	มีอัตราการซ่อมดีมาก และระยะเวลาหลังติดตั้งเฉลี่ยนาน แต่อัตราซ่อมหลังติดตั้งไม่ดี มีความพอใจต่ำ
7*	8	มีอัตราการซ่อมดีมาก และระยะเวลาหลังติดตั้งเฉลี่ยนาน แต่ที่เหลือไม่ดี มีความพอใจต่ำมาก
8*	9	มีอัตราการซ่อมดี แต่ที่เหลือเฉลี่ยไม่ดีมาก มีความพอใจต่ำมาก
9*	11	มีอัตราการซ่อมหลังติดตั้ง และไม่เสร็จเฉลี่ยไม่ดี แต่ที่เหลือดีมาก มีความพอใจต่ำมาก
10	9	Random

ตารางที่ 7.2-14 สรุปข้อมูลที่ใช้ในการจำลองในผลการทดลองที่ 5



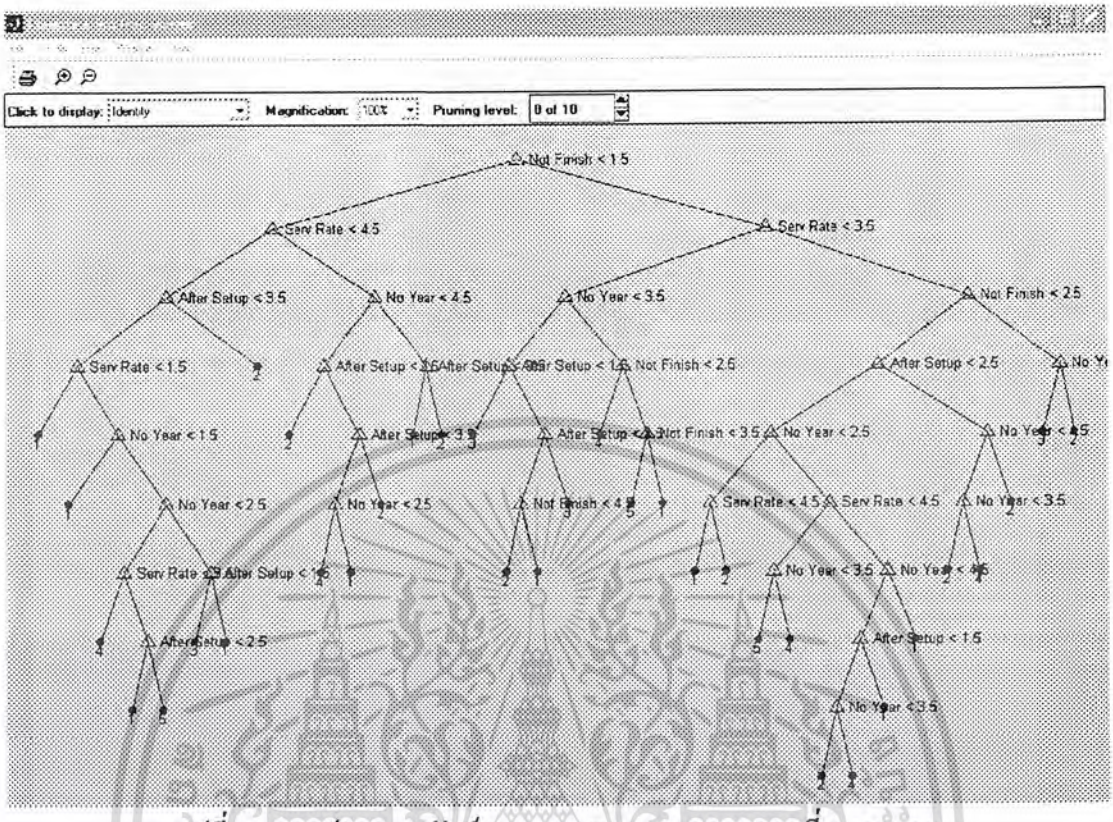
รูปที่ 7.2-11 กราฟแสดงอัตราส่วนในการจำลองข้อมูลในแต่ละส่วนในผลการทดลองที่ 5

โครงสร้างข้อมูลที่ใช้ในการจำลอง

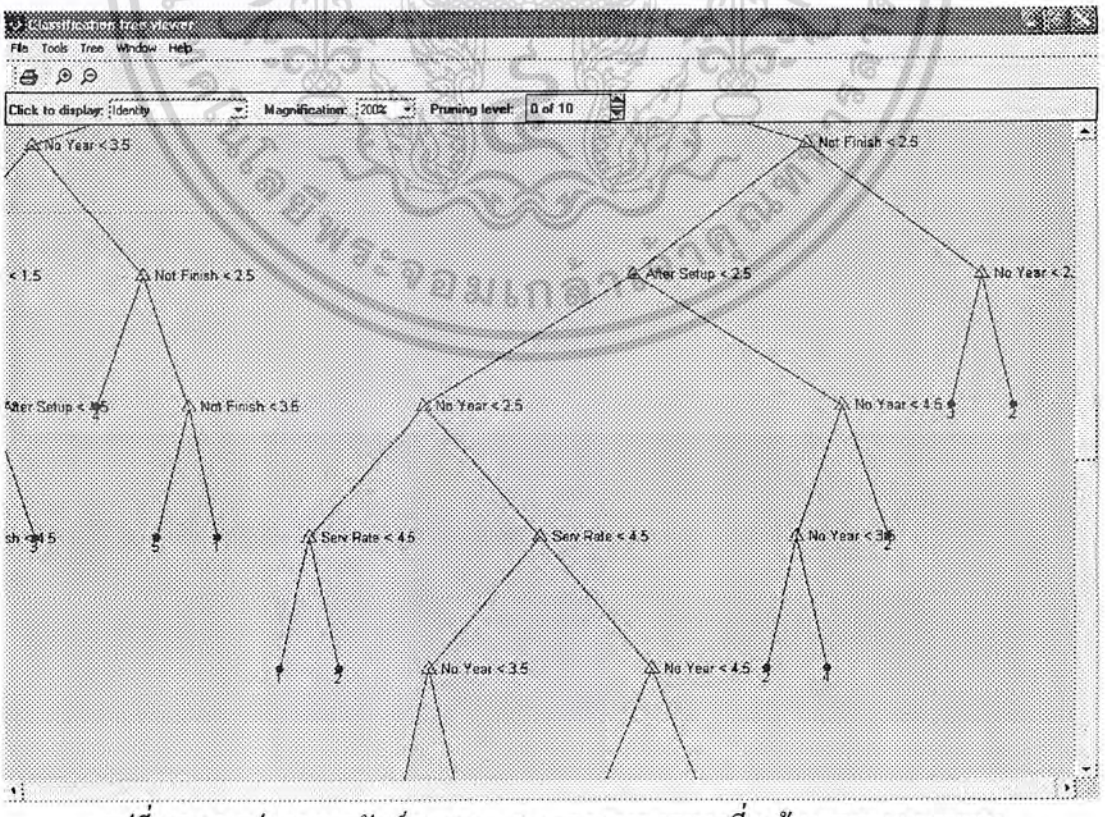
No.	%	Number	Service Rate Sat.	After Setup Satisfac.	Not Finish Satisfac.	No. Year	Service Point
1	12	120	5	Random	Random	3 - 5	4
2	9	90	3 - 5	Random	2	4	4
3	11	110	3 - 5	3	1	3 - 5	1
4	9	90	1 - 3	2 - 4	2 - 4	1 - 3	2
5	14	140	1 - 3	2	2	1 - 3	2
6	8	80	5	2	Random	3 - 5	2
7	8	80	5	2	2	3 - 5	1
8	9	90	4	1 - 2	1 - 2	1 - 2	1
9	11	110	5	1 - 2	1 - 2	5	1
10	9	90	Random	Random	Random	Random	Random

ตารางที่ 7.2-15 โครงสร้างข้อมูลที่ใช้ในการจำลองข้อมูลในการทดลองที่ 5

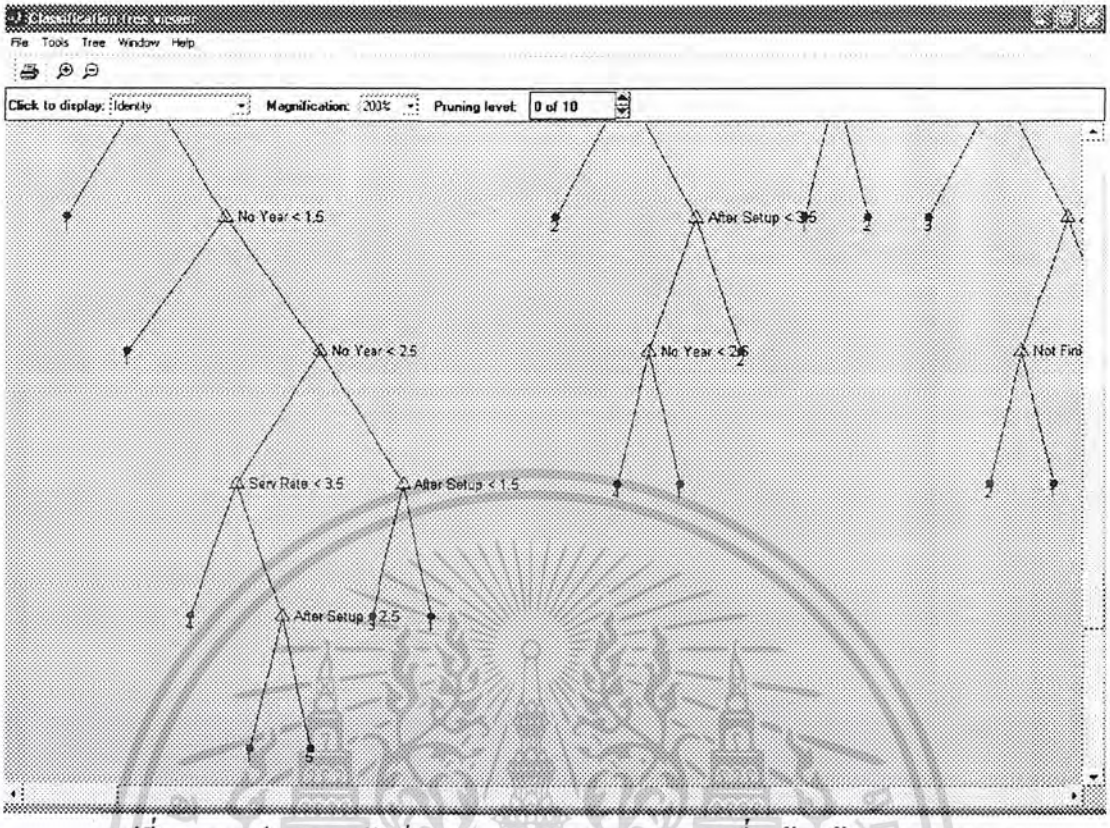
ผลลัพธ์



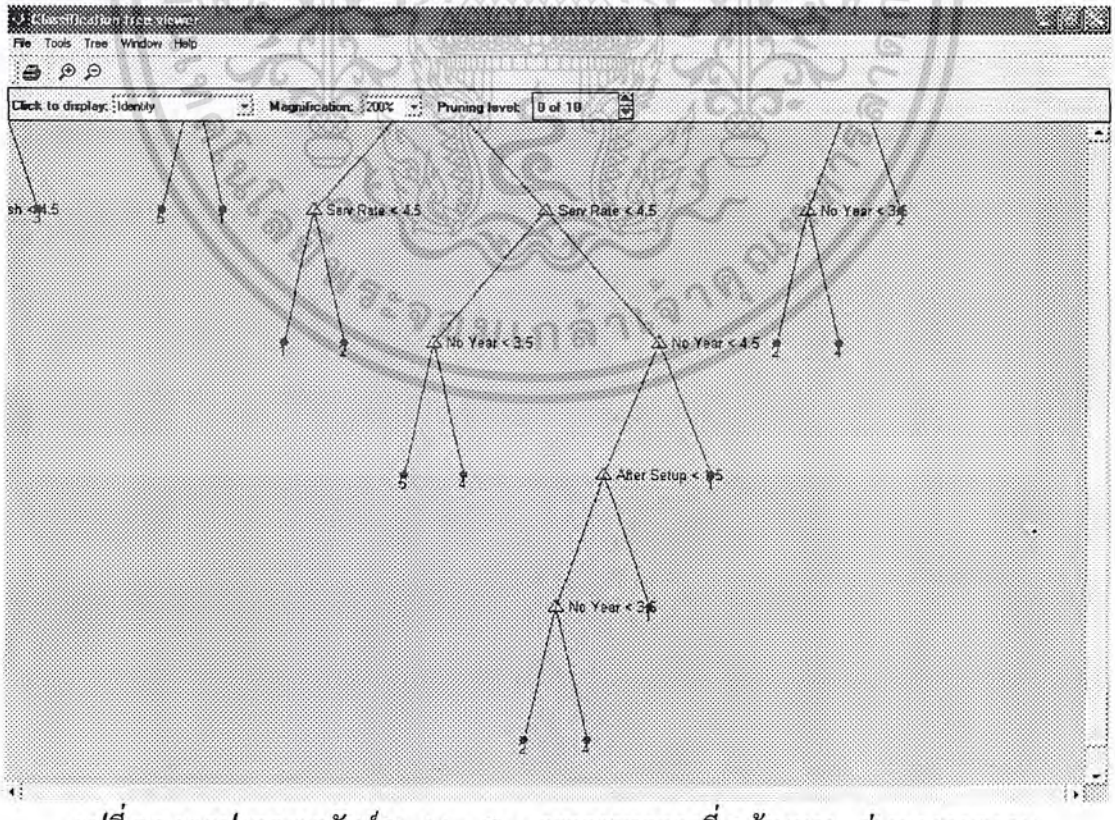
รูปที่ 7.2-12 รูปแสดงผลลัพธ์ Decision Tree จากการทดลองที่ 5 ภาพรวม



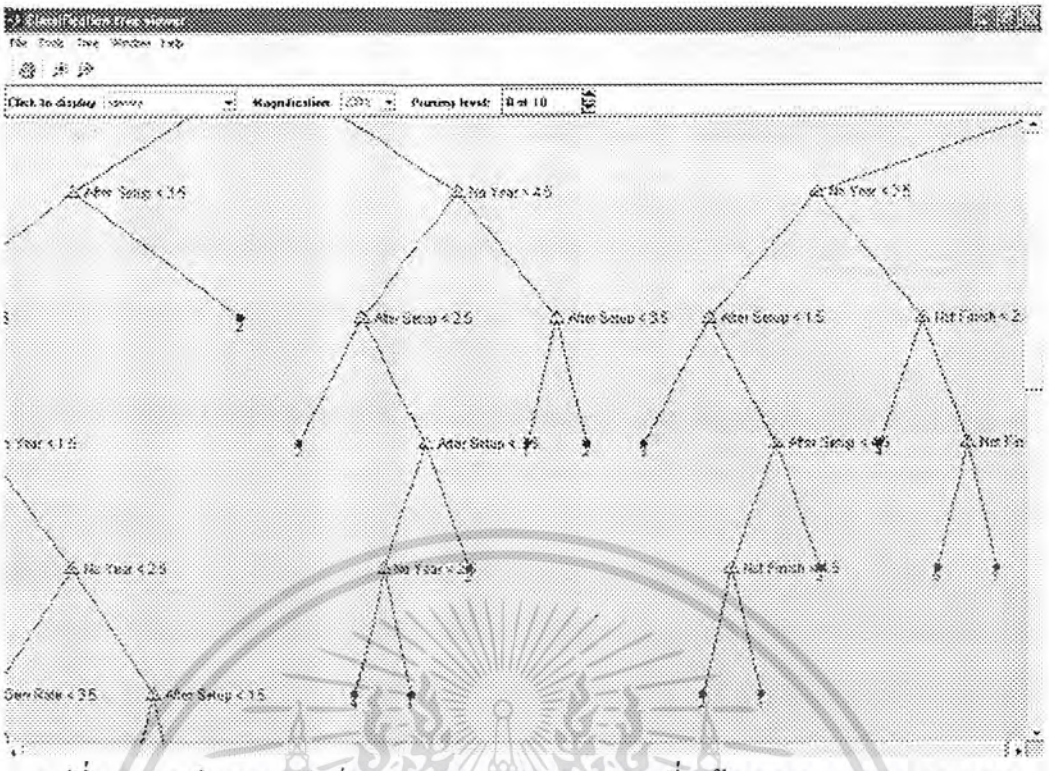
รูปที่ 7.2-13 รูปแสดงผลลัพธ์ Decision Tree จากการทดลองที่ 5 ด้านขวาของภาพรวม



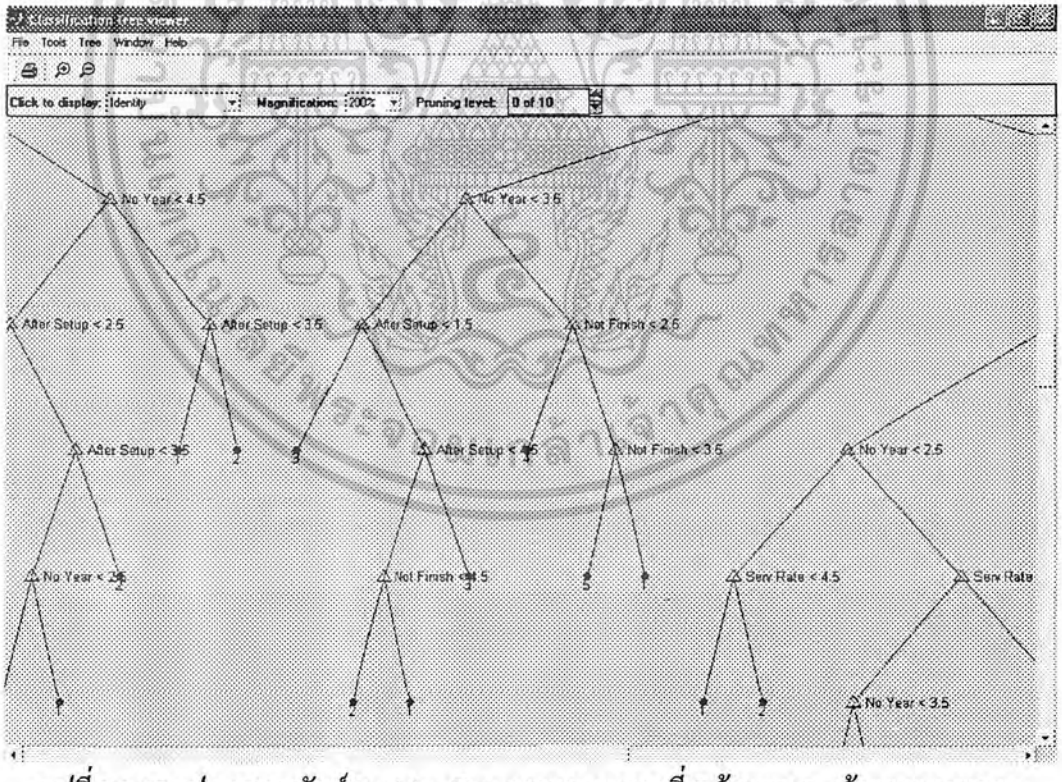
รูปที่ 7.2-14 รูปแสดงผลลัพธ์ Decision Tree จากการทดลองที่ 5 ด้านซ้ายของภาพรวม



รูปที่ 7.2-15 รูปแสดงผลลัพธ์ Decision Tree จากการทดลองที่ 5 ด้านขวา - ล่างของภาพรวม



รูปที่ 7.2-16 รูปแสดงผลลัพธ์ Decision Tree จากการทดลองที่ 5 ด้านกลาง - ขวาของภาพรวม



รูปที่ 7.2-17 รูปแสดงผลลัพธ์ Decision Tree จากการทดลองที่ 5 ด้านกลาง - ซ้ายของภาพรวม

สรุปผลการทดลอง

อัตราความถูกต้องของ Decision Tree คือ 90.3%

บทที่ 8
สรุปการดำเนินงานของโครงงาน

โครงงานนี้ได้ทำการศึกษาเกี่ยวกับ Data Mining และได้นำมาประยุกต์ใช้กับปัญหาทางธุรกิจที่เกี่ยวกับเครื่องปรับอากาศ 3 รูปแบบคือ

1. หาว่าลูกค้าไม่ชอบสินค้า จากการ service มี รูปแบบ อะไรบ้าง โดยได้นำ Decision Tree มาใช้ในการเรียนรู้ รูปแบบ ของการ service ที่ลูกค้าไม่ชอบ โดยอินพุทที่ใช้เป็นข้อมูลที่ได้จำลอง (Simulate) ขึ้นมาเอง ซึ่งผลที่ได้ถูกต้องตามข้อมูลที่ได้ bias เข้าไป

2. หาว่า ลูกค้าซื้อสินค้าเพราะอะไร โดยได้นำการ Clustering มาใช้ในการแบ่งกลุ่มของลูกค้า ว่าแต่ละกลุ่มซื้อสินค้าเนื่องจากปัจจัยใด โดยอินพุทที่ใช้ในปัญหาที่ 2 นี้ แบ่งเป็น 2 ประเภท คือ 1. ข้อมูลจริงที่ได้จากแบบสำรวจความคิดเห็นของประชาชน และ 2. ข้อมูลที่จำลอง (Simulate) ขึ้นมาเอง โดยผลที่ได้จากการ Clustering ข้อมูลจริงนั้นค่อนข้างจะตรงกับความคิดเห็นของลูกค้าส่วนใหญ่ ก็คือ จะเลือกคุณภาพ ประหยัดพลังงาน ราคาและบริการหลังการขายเป็นหลัก ส่วนปัจจัยอื่นๆสนใจค่อนข้างน้อย ส่วนผลที่ได้จากการ Clustering ข้อมูลจำลอง (Simulate) ขึ้นมานั้น ตรงตามชุดข้อมูลที่ได้ bias เข้าไป แต่ยังมีค่าที่ผิดพลาดในกรณีที่มีการ Random ข้อมูลในทุกๆ Cluster อยู่

3. หาความสัมพันธ์ของตัวแปรต่างๆ เพื่อคาดการณ์ยอดขายสินค้าในอนาคต โดยได้นำ Neural Network มาใช้ในการเรียนรู้หาความสัมพันธ์ของปัจจัยแวดล้อมต่างๆ ที่ส่งผลต่อยอดขายสินค้า โดยอินพุทที่ใช้เป็นข้อมูลจริงที่ได้จากกรมอุตุนิยมวิทยา ซึ่งผลที่ได้นั้นยังมีค่าผิดพลาดอยู่พอสมควร เนื่องจากข้อมูลที่ใช้เป็นอินพุทในการทดลองนั้นมีจำนวนที่น้อยเกินไป จึงไม่ได้ทำรายงานเป็นการทดลองออกมา

ซึ่งผลที่ได้จากการนำ Data Mining ไปประยุกต์ใช้กับปัญหาทางธุรกิจที่เกี่ยวกับเครื่องปรับอากาศ 3 รูปแบบข้างต้นนั้นสามารถนำไปวิเคราะห์ในทางธุรกิจต่อไปได้ว่าควรจะปรับปรุงการดำเนินงานและการตลาดอย่างไรให้เหมาะสมกับธุรกิจในปัจจุบัน

แนวทางในการพัฒนาโครงงานนี้ต่อไป คือ เนื่องจากในปัญหาที่ 3 นั้น ข้อมูลที่เป็นอินพุทในการใช้ Neural Network นั้นยังมีจำนวนที่น้อยเกินไป จึงจำเป็นต้องเก็บข้อมูลที่จะใช้เป็นอินพุทให้มากขึ้น เพื่อที่จะได้ผลลัพธ์ที่ถูกต้อง และในส่วนของการทำ Clustering นั้น ข้อมูลจริงควรจะมีการกระจายตัวของข้อมูลที่มากกว่านี้เพื่อให้ได้ผลลัพธ์ที่ถูกต้องแม่นยำมากขึ้น

ภาคผนวก

ภาคผนวก ก
ขั้นตอนการติดตั้งโปรแกรมฐานข้อมูล

ก. ขั้นตอนการติดตั้งโปรแกรมฐานข้อมูล

คอมพิวเตอร์ที่จะติดตั้ง

- PENTIUM 600 MHz หรือมากกว่า
- หน่วยความจำ (Ram) อย่างน้อย 256 MB
- Window 98SE/ME/2000/XP Operating System
- พื้นที่ที่สามารถใช้งานฮาร์ดดิสก์ 100 MB

โดยโปรแกรมฐานข้อมูลที่ต้องติดตั้งก่อนการใช้งานโปรแกรมคาต้าไมนิ่งนั้นประกอบด้วย

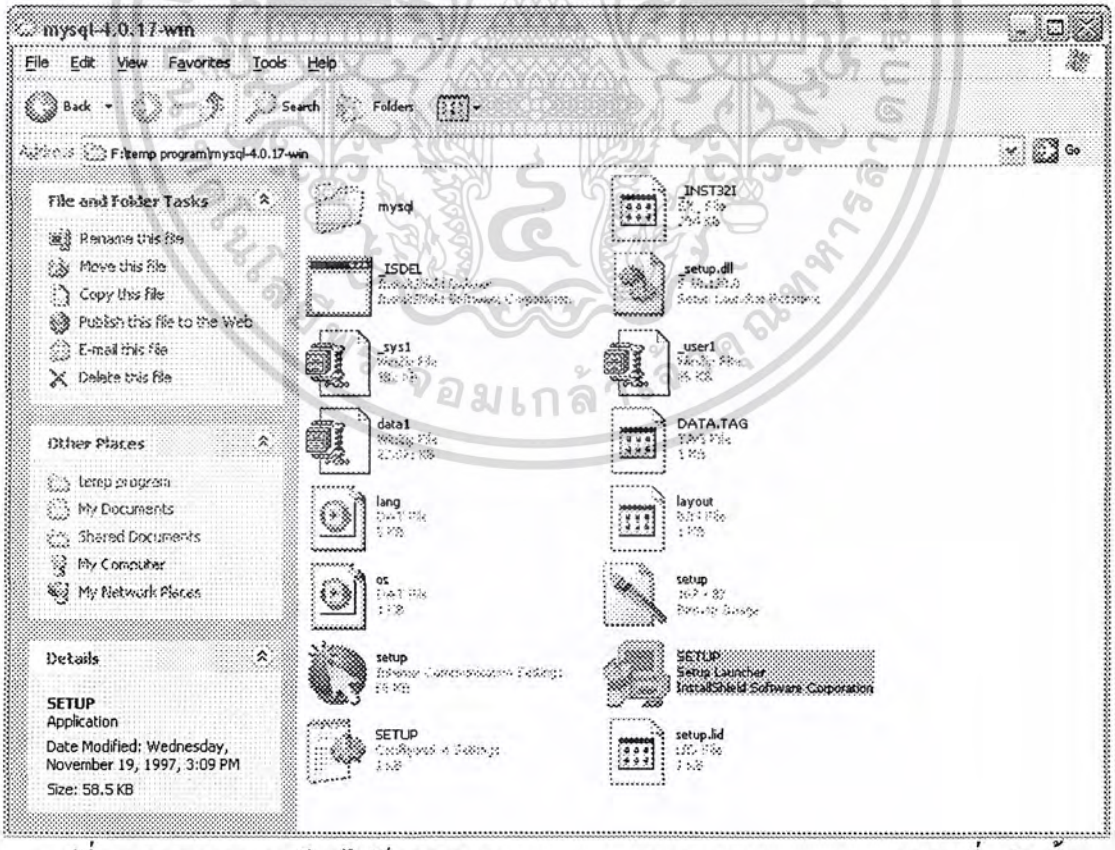
MySQL Servers and Clients window version 4.0.17

MySQL ODBC 3.51.03 Driver

SQLyog version 3.64

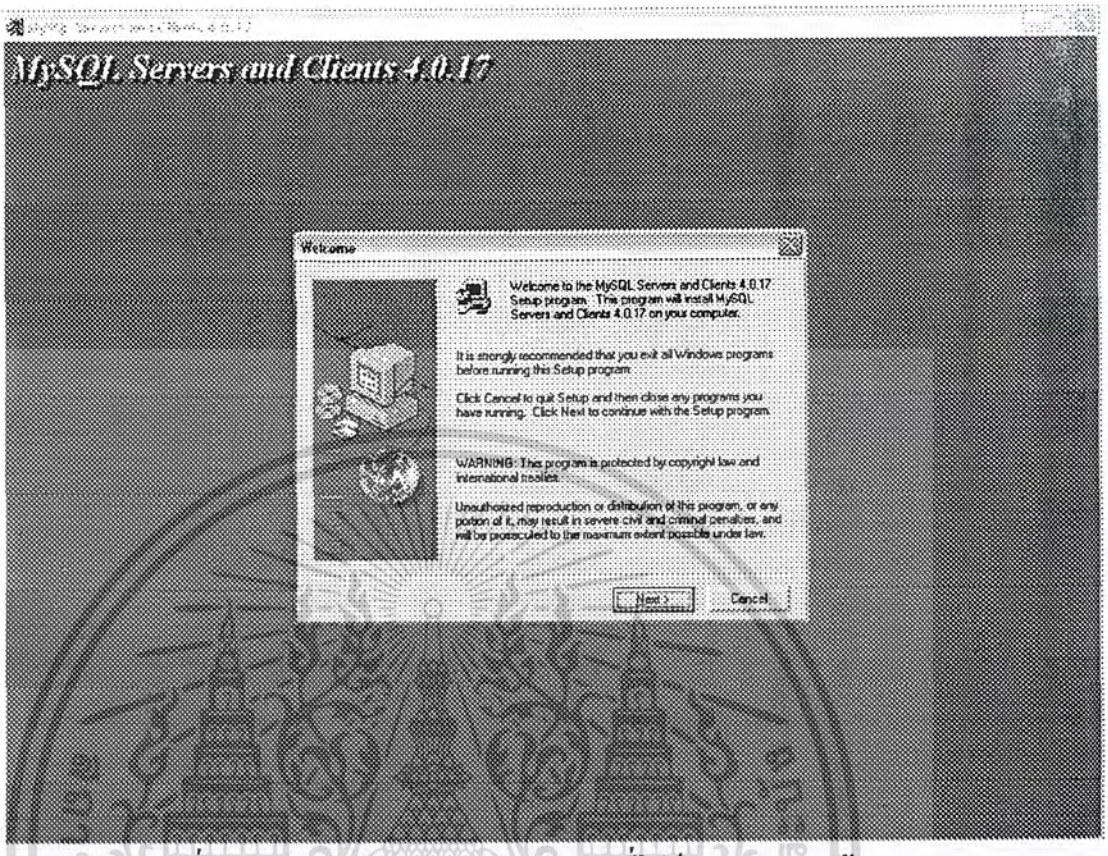
ก.1. ขั้นตอนการติดตั้ง MySQL Servers and Clients window version 4.0.17

1. ดับเบิลคลิกที่ไฟล์ชื่อ SETUP.exe ดังรูปที่ ก.1-1



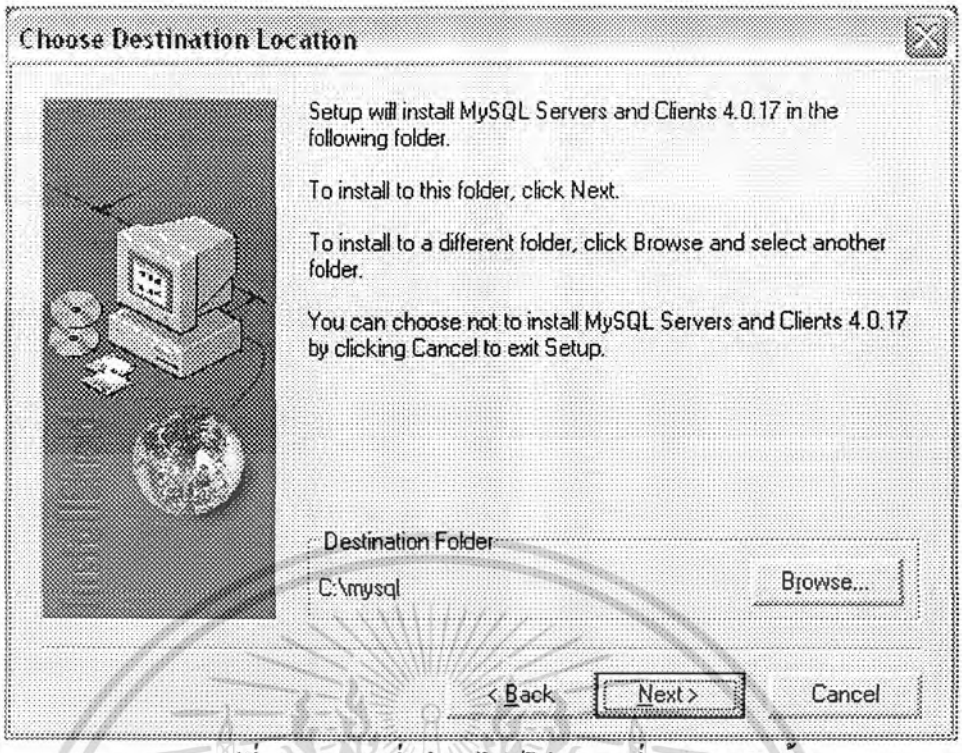
รูปที่ ก.1-1 แสดงรายละเอียดไฟล์ MySQL Servers and Clients window version 4.0.17 ที่จะติดตั้ง

2. เมื่อดับเบิลคลิกแล้วจะขึ้นหน้าจอดังรูปที่ ก.1-2 กดที่ปุ่ม Next เพื่อดำเนินการต่อไป



รูปที่ ก.1-2 แสดงหน้าจอหลังจากดับเบิลคลิกที่ไฟล์ SETUP.exe แล้ว

3. จากนั้นก็กดปุ่ม Next อีกครั้ง จะได้หน้าจอดังรูปที่ ก.1-3 เราสามารถกำหนด Destion Folder ที่จะติดตั้งได้โดยกดที่ปุ่ม Browse.. เพื่อเลือกกำหนด Destion Folder เอง

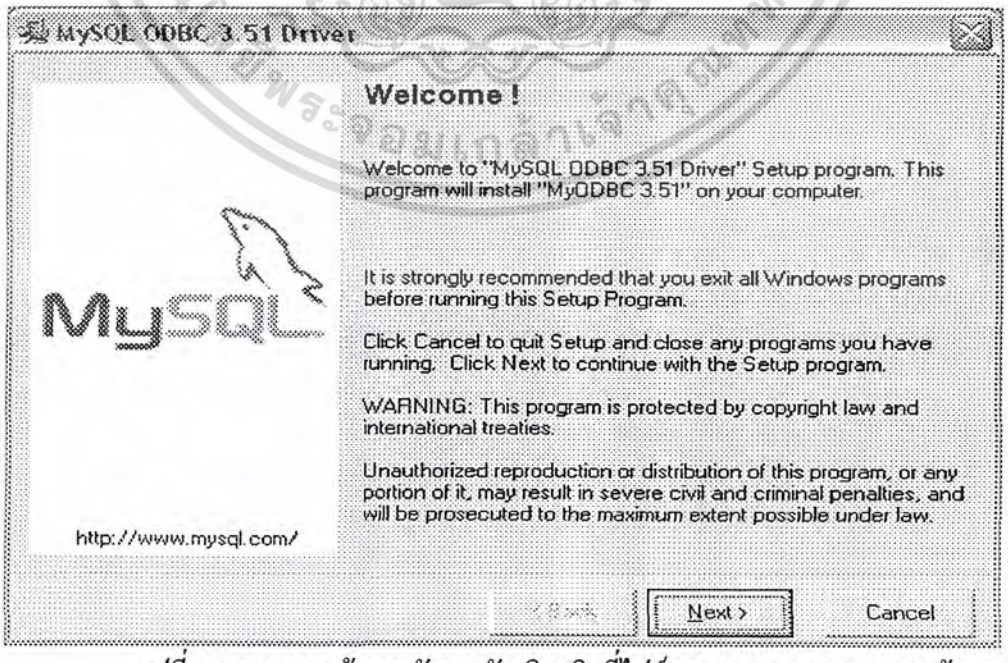


รูปที่ ก.1-3 แสดงที่อยู่ของไฟล์โปรแกรมที่จะทำการติดตั้ง

เมื่อกำหนดเรียบร้อยแล้ว กดปุ่ม Next ไปเรื่อยๆ การติดตั้งก็จะสมบูรณ์

ก.2. ขั้นตอนการติดตั้ง MySQL ODBC 3.51.03 Driver

1. ดับเบิลคลิกที่ไฟล์ชื่อ MyODBC-3.51.03.exe
2. เมื่อดับเบิลคลิกแล้วจะขึ้นหน้าจอดังรูปที่ ก.2-1 กดที่ปุ่ม Next ไปเรื่อยๆ การติดตั้งก็จะสมบูรณ์



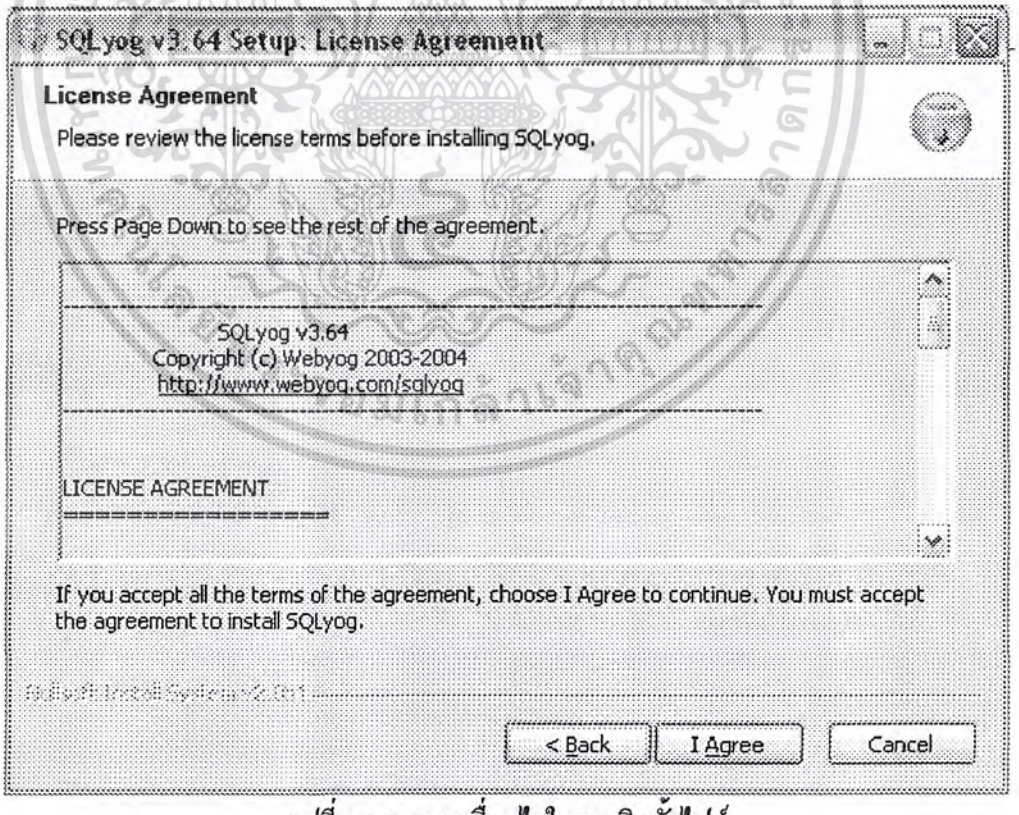
รูปที่ ก.2-1 แสดงหน้าจอหลังจากดับเบิลคลิกที่ไฟล์ MyODBC-3.51.03.exe แล้ว

ก.3. ขั้นตอนการติดตั้ง SQLyog version 3.64

1. ดับเบิลคลิกที่ไฟล์ชื่อ SQLyog364.exe
2. เมื่อดับเบิลคลิกแล้วจะขึ้นหน้าจอดังรูปที่ ก.3-1 กดที่ปุ่ม Next

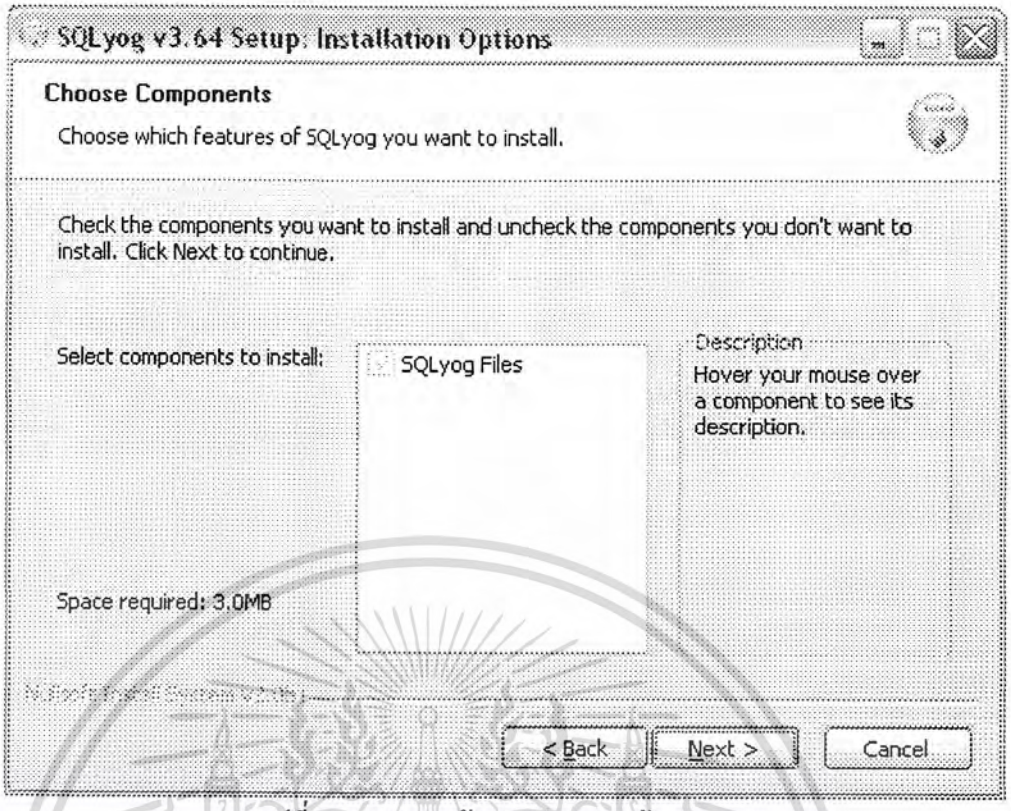
รูปที่ ก.3-1 แสดงหน้าจอหลังจากดับเบิลคลิกที่ไฟล์ SQLyog364.exe แล้ว

3. เมื่อได้หน้าจอดังรูปที่ ก.3-2 แล้ว กดที่ปุ่ม I Agree



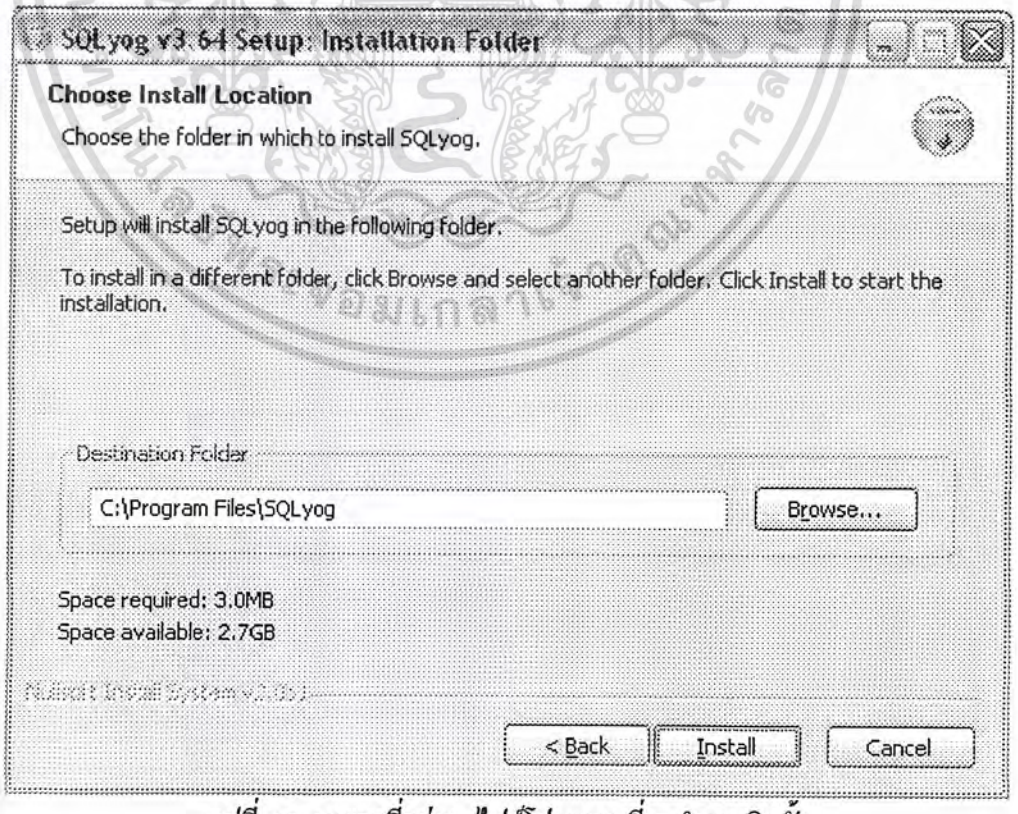
รูปที่ ก.3-2 แสดงเงื่อนไขในการติดตั้งไฟล์

4. เมื่อได้หน้าจอดังรูปที่ ก.3-3 แล้ว กดที่ปุ่ม Next



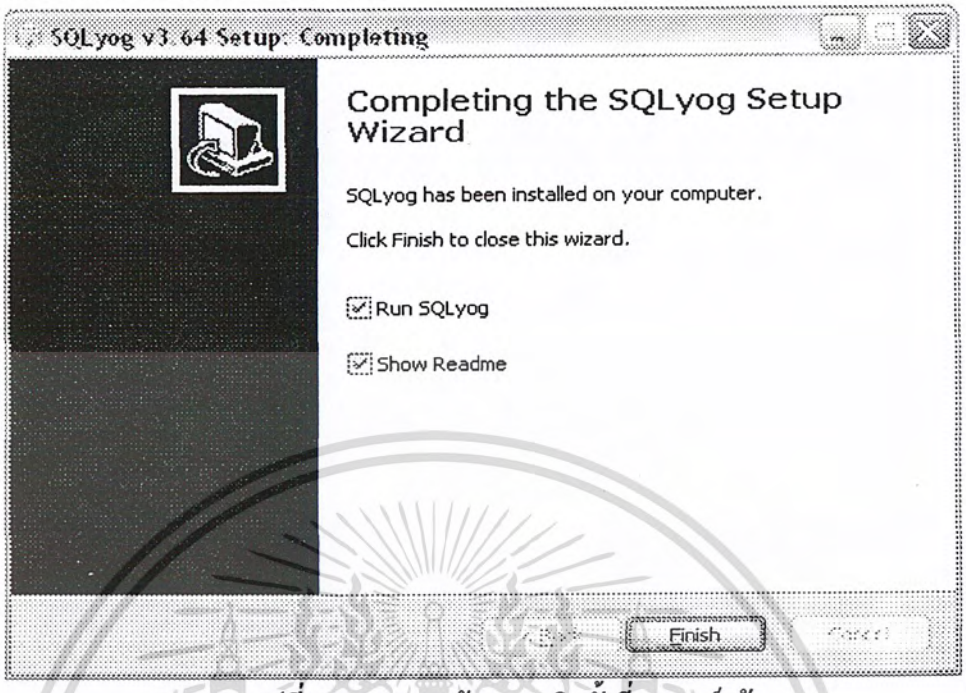
รูปที่ ก.3-3 แสดงหน้าจอหลังจากกดปุ่ม I Agree

5. สามารถกำหนด Destination Folder ที่จะติดตั้งได้โดยกดที่ปุ่ม Browse.. เพื่อเลือกกำหนด Destion Folder เอง



รูปที่ ก.3-4 แสดงที่อยู่ของไฟล์โปรแกรมที่จะทำการติดตั้ง

เมื่อกำหนดเรียบร้อยแล้ว กดปุ่ม Install การติดตั้งก็จะสมบูรณ์ดังรูปที่ ก.3-5



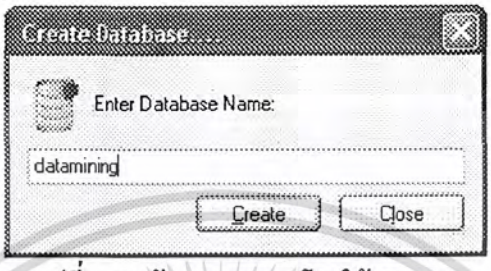
รูปที่ ก.3-5 แสดงหน้าจอการติดตั้งที่สมบูรณ์แล้ว

ภาคผนวก ข
ขั้นตอนการใช้งานโปรแกรม SQLyog

ข. ขั้นตอนการใช้งานโปรแกรม SQLyog

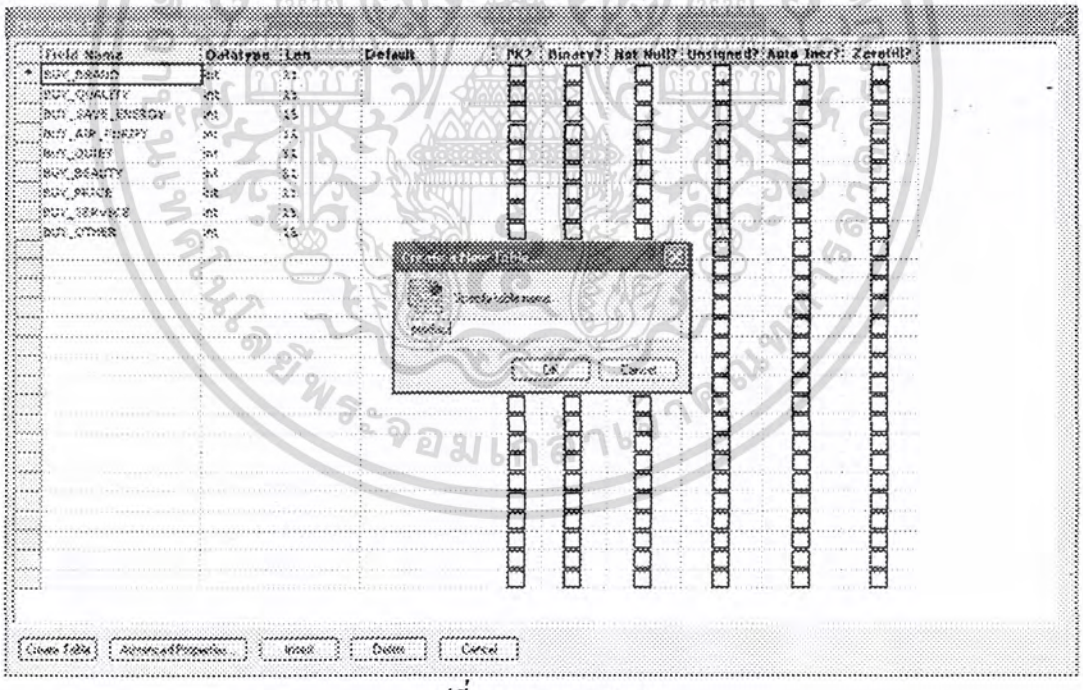
การสร้าง Database โดยใช้ SQLyog

1.เลือก DB ที่ menu bar จากนั้นเลือก Create Database... หรือ กด Ctrl+D จากนั้นให้ใส่ชื่อของ Database ในที่นี้ ต้องการสร้าง ฐานข้อมูลชื่อ datamining



รูปที่ ข-1 สร้าง Database โดยใช้ SQLyog

2.สร้าง Database table โดยการเลือก Database ที่เราต้องการสร้าง table จากนั้น เลือก DB ที่ menu bar และเลือก Create Table In The Database... ใส่ Filed name และ type ตามที่ต้องการแล้วกด ปุ่ม Create แล้วใส่ชื่อ table กดปุ่ม OK ให้สร้าง Database Table ที่จำเป็นต้องใช้งานทั้งหมด



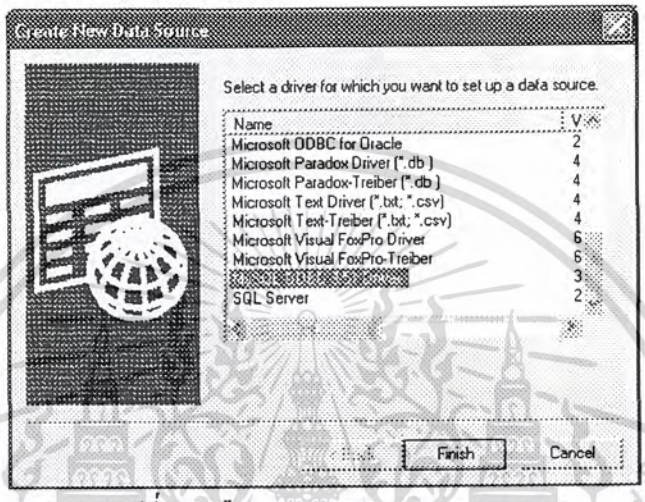
รูปที่ ข-2 Create Table

การติดต่อระหว่าง โปรแกรม Data Generator กับ Database

1.ก่อนอื่นต้องมีการเพิ่ม System Data Source ที่ใช้งานร่วมกันกับ Database ที่สร้างเอาไว้โดยกำหนดที่ Administrative tools 

ใน Control Panel จากนั้นเลือก Data Sources (ODBC) 

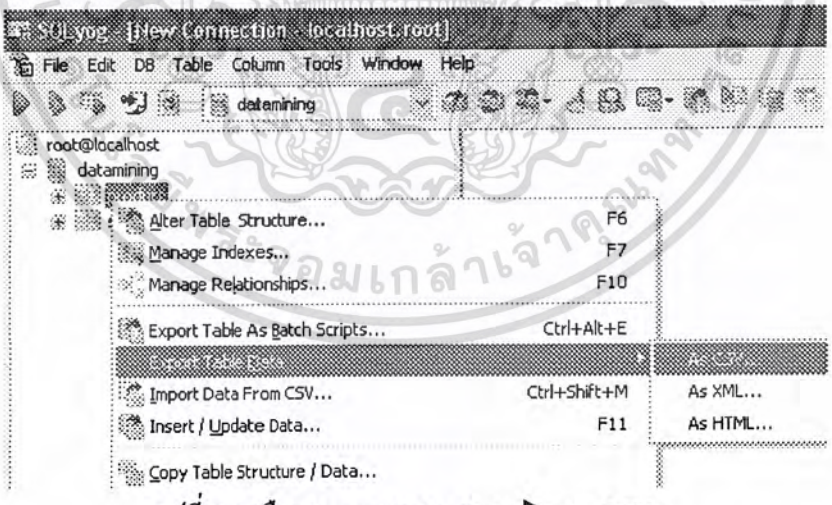
2.Add New System Data Source ของ Database ที่ต้องการใช้งานเข้าไป และเลือก driver ของ database ในที่นี้ ใช้MySQL ให้ใส่ชื่อของ Database ลงไป



รูปที่ ข-3 เลือก Driver ของ Data Source

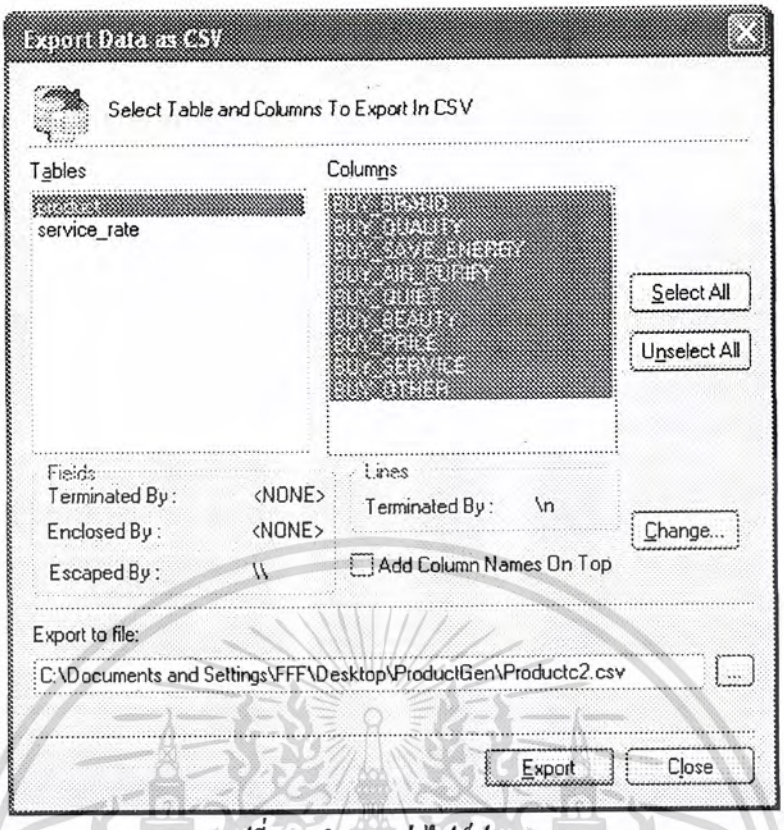
การแปลงข้อมูลใน Database เป็น ไฟล์ .csv มีขั้นตอนดังนี้

1.เลือก ตารางที่ต้องการจะแปลงเป็น .csv คลิ๊กเมาท์ขวา เลือก Export Table Data ▶ As CSV...



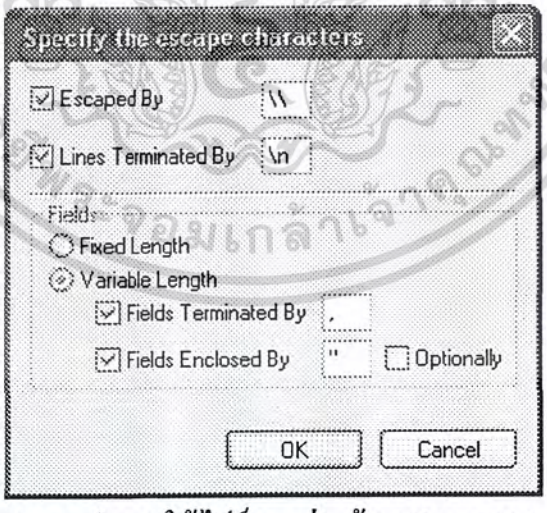
รูปที่ ข-4 เลือก Export Table Data ▶ As CSV...

2.เลือก columns ที่จะแปลง และกำหนด ที่อยู่, ชื่อไฟล์ปลายทาง และกด Export ก็จะแปลงเป็นไฟล์ .csv ให้ทันที



รูปที่ ข-5 กำหนดค่าไฟล์ปลายทาง

หมายเหตุ ถ้าต้องการให้ ไฟล์ .csv อ่านด้วย Microsoft Excel ได้ ให้กด Change... จะมีหน้าต่าง Specify the escape characters ขึ้นมาให้กำหนดค่าตามรูปที่ ข-6



รูปที่ ข-6 การกำหนดให้ ไฟล์.csv อ่านด้วย Microsoft Excel ได้

ภาคผนวก ค
ขั้นตอนการใช้งานโปรแกรมดาตาไมนิ่ง

ค. ขั้นตอนการใช้งานโปรแกรมดาตาไมนิ่ง

โปรแกรมดาต้าไมนิ่ง (Data Mining Program) แบ่งเป็น 3 ส่วนตามการทำงานของโปรแกรม คือ

1. โปรแกรมดาต้าไมนิ่งในส่วนของการจำลองข้อมูล
2. โปรแกรมดาต้าไมนิ่งในส่วนของการคลัสเตอร์ข้อมูล
3. โปรแกรมดาต้าไมนิ่งในส่วนของการทำ Decision Tree

ค.1. การใช้งานโปรแกรมดาตาไมนิ่งในส่วนของการจำลองข้อมูล

ในส่วนการจำลองข้อมูล แบ่งได้เป็น 2 ส่วน คือ

1. ส่วนของการจำลองข้อมูลเพื่อนำไปใช้ในการหาว่าลูกค้าซื้อสินค้าเพราะอะไร
2. ส่วนของการจำลองข้อมูลเพื่อนำไปใช้ในการหาว่าลูกค้าไม่ชอบสินค้า จากการ service มี pattern อะไรบ้าง

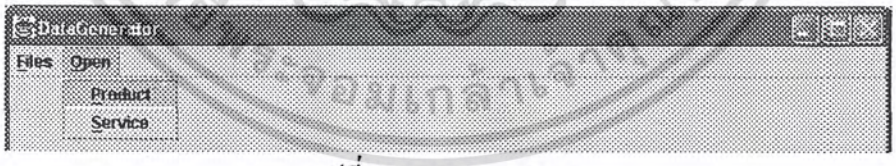
ค.1.1. ขั้นตอนการใช้งานโปรแกรมดาตาไมนิ่งในส่วนของการจำลองข้อมูล

1. ที่หน้าต่างแรกของโปรแกรม ตรง เมนูบาร์ จะมี 2 หัวข้อ ได้แก่

รูปที่ ค.1.1-1 Menu Files

1.1) Files มีตัวเลือก 2ตัวเลือก ได้แก่

- About... : บอกข้อมูลทั่วไปเกี่ยวกับโปรแกรมนี้ เช่น จุดประสงค์ เป็นต้น
- Exit : ปิดโปรแกรม



รูปที่ ค.1.1-2 Menu Open

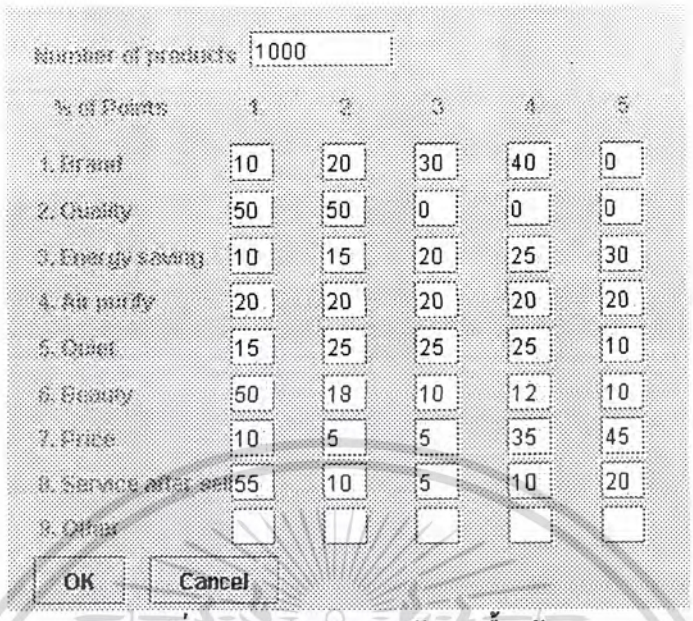
1.2) Open มีตัวเลือก 2 ตัวเลือก ได้แก่

- Product : เปิดหน้าต่าง เพื่อ generate ข้อมูลในตาราง Database Product และใช้งาน โปรแกรม Cluster
- Service: เปิดหน้าต่าง เพื่อ generate ข้อมูลในตาราง Database Service

2. ที่หน้าต่าง Product

รูปที่ ค.1.1-3 หน้าต่าง Product

2.1) ส่วนกำหนดข้อมูลเบื้องต้น



รูปที่ ค.1.1-4 ส่วนกำหนดข้อมูลเบื้องต้น

Number of products : กำหนดจำนวนข้อมูลทั้งหมดที่จะ generate

% of Points : กำหนดเปอร์เซ็นต์ของคะแนนที่จะ generate จากข้อมูลทั้งหมดให้กับแต่ละหัวข้อของแบบสอบถามโดยมี 9 หัวข้อ แต่ละหัวข้อมี คะแนน 5 ประเภท โดยที่ในแต่ละหัวข้อนั้นถ้ามีการใส่ครบทั้ง 1-5 เปอร์เซ็นต์รวมจะต้องได้ 100% พอดี แต่ถ้าใส่ไม่ครบทุกช่อง เปอร์เซ็นต์รวมเฉพาะช่องที่ใส่ต้องไม่เกิน 100% ซึ่งช่องที่ไม่ใส่นั้นจะถือว่าเป็นการ random ซึ่งข้อมูลที่เป็น random นั้นจะนำมากำหนดในส่วนของเงื่อนไขไม่ได้

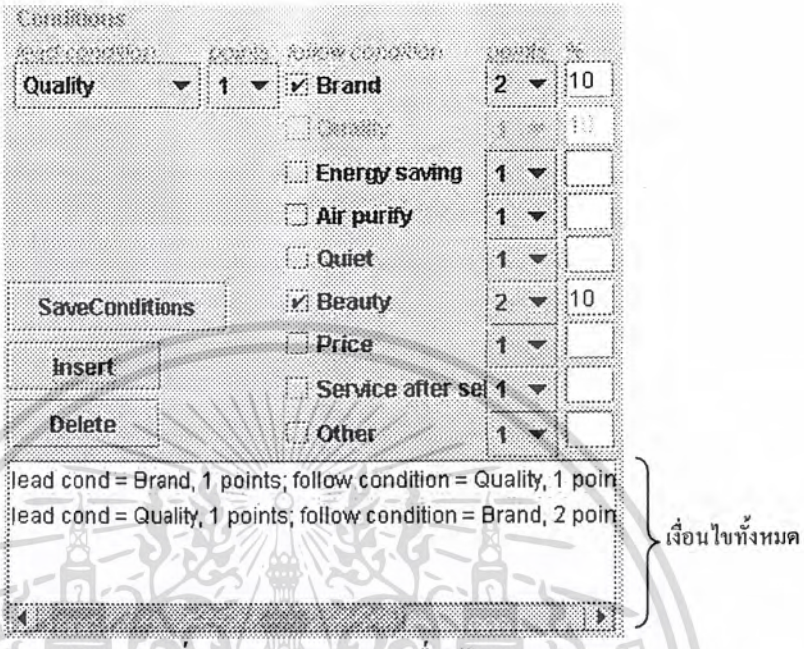
ปุ่ม OK : ยืนยันข้อมูลเบื้องต้น เพื่อใช้ในขั้นตอนต่อไป ซึ่งจะทำให้ตอนกำหนดเงื่อนไขไม่สามารถเปลี่ยนแปลงข้อมูลเบื้องต้นได้

ปุ่ม Cancel : ยกเลิกข้อมูลที่ป้อนเข้ามาในส่วนของข้อมูลเบื้องต้น

ดังตัวอย่างในรูปที่ ค.1.1-4 เป็นการกำหนดว่าจะgenerateข้อมูลใน database Product ทั้งหมด 1000 records ดูจากช่อง Number of products โดยจะมี คะแนน Brand เท่ากับ 1 - 5 อยู่ 10,20,30,40,0 % ตามลำดับ ซึ่งก็คือ column Brand จะมีคะแนน 1 - 5 อยู่ 100,200,300,400,0 records ตามลำดับนั่นเอง ใน column อื่นก็จะเป็นตามที่กำหนดเปอร์เซ็นต์เอาไว้เช่นกันและใน

ส่วนของ Other เว้นเอาไว้ไม่ได้กำหนดลงไปจะหมายถึงทั้งคะแนน 1-5 เป็นการ random ซึ่งจะไม่สามารถนำไปกำหนดลงในส่วนของเงื่อนไขได้

2.2) ส่วนกำหนดเงื่อนไข



รูปที่ ค.1.1-5 ส่วนกำหนดเงื่อนไข

lead condition, points : เลือกหัวข้อ, คะแนนที่เป็นเงื่อนไขนำ เพื่อเป็นตัวแปรในการกำหนดความสัมพันธ์ให้กับเงื่อนไขตาม โดยเลือกหัวข้อ, คะแนน ได้หนึ่งชนิดต่อเงื่อนไขนำหนึ่งเงื่อนไข

follow condition, points,% : เลือกหัวข้อ, คะแนน, เปอร์เซนต์ให้กับเงื่อนไขตามโดยเลือกได้หลายหัวข้อ แต่หัวข้อหนึ่งจะเลือกคะแนน, เปอร์เซ็นต์ ได้เงื่อนไขเดียว โดยจะต้องคำนวณด้วยว่าเปอร์เซ็นต์ที่ใส่ในเงื่อนไขมีจำนวนไม่เกินจากเปอร์เซ็นต์ที่กำหนดในข้อมูลเบื้องต้นในส่วนที่ผ่านมา มิเช่นนั้นจะไม่สามารถ เพิ่ม (Insert) เงื่อนไขลงไปได้

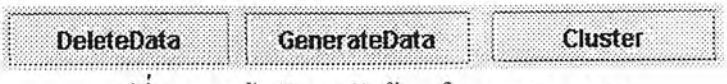
ปุ่ม Insert : ใส่เงื่อนไขเข้าไปใน list เงื่อนไขทั้งหมดโดย โปรแกรมจะเช็คเงื่อนไขที่จะใส่ว่าถูกต้องหรือไม่ ถ้าไม่จะแจ้ง error และยกเลิกการใส่เงื่อนไขนั้น

ปุ่ม Delete : ลบเงื่อนไขที่เลือกใน list เงื่อนไขออก

ปุ่ม SaveConditions : เก็บเงื่อนไขที่เราได้กำหนดไว้ในไฟล์รูปแบบ text

ดังตัวอย่างในรูปที่ ค.1.1-5 เป็นการกำหนดเงื่อนไขให้ record ที่คะแนน Quality เท่ากับ 1 จะมีคะแนน Brand เท่ากับ 2 อยู่ 10% และมี คะแนน Beauty เท่ากับ 2 อยู่ 10% ซึ่งถ้าดูต่อเนื่องจากรูปที่2.2 คะแนนQuality เท่ากับ 1 มี 50% จากทั้งหมด 1000 เพราะฉะนั้น จะมี 500 records ซึ่งใน 500 records นี้จะมี คะแนน Brand เท่ากับ 2 อยู่ 10% ก็คือ มี 50 records เช่นเดียวกัน ใน500 records ก็จะมี คะแนน Beauty 50 records

2.3) ปุ่มจัดการกับข้อมูลในตาราง Product



รูปที่ ค.1.1-6 ปุ่มจัดการกับข้อมูลในตาราง Product

ปุ่ม DeleteData : ลบข้อมูลทั้งหมดในตาราง Product

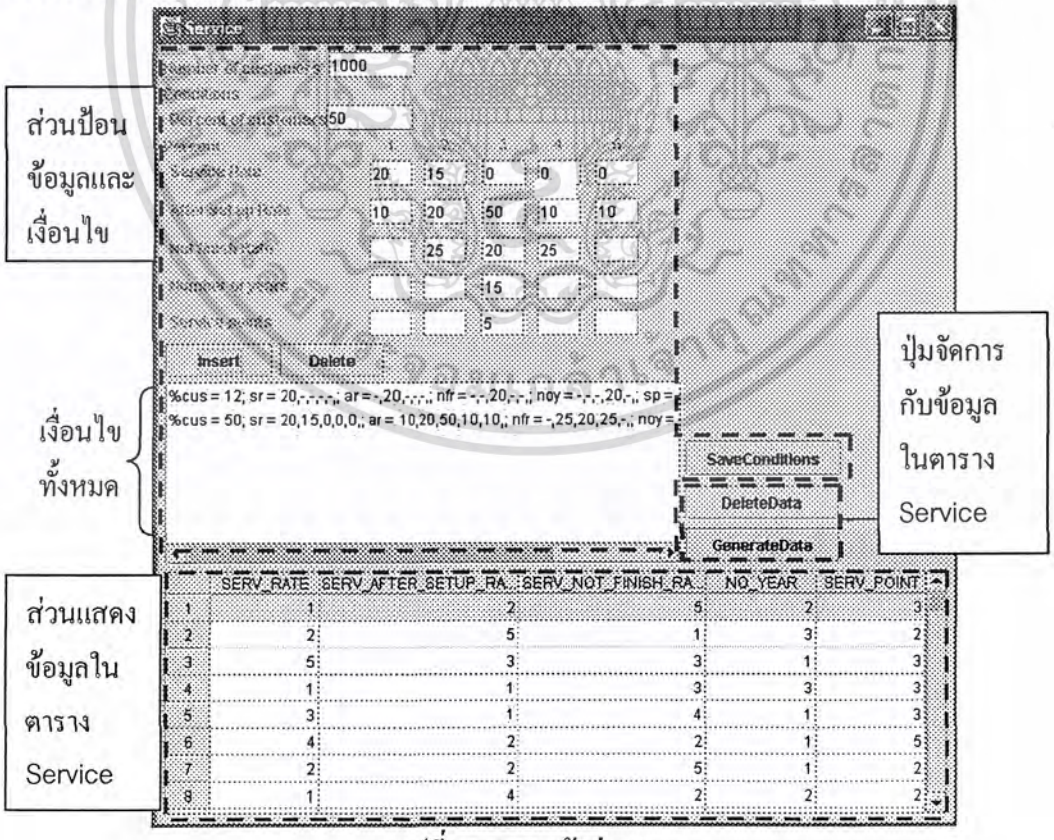
ปุ่ม GenerateData : ทำการ Generate ข้อมูลตามจำนวนและเงื่อนไขที่ได้กำหนดไว้ และเก็บลงในตาราง Product

ปุ่ม Cluster : Cluster ข้อมูลในตาราง Product และ แสดงผลลัพธ์ที่ได้

2.4) ส่วนแสดงข้อมูลในตาราง Product

แสดงข้อมูลทั้งหมดในตาราง Product

3. ที่หน้าต่าง Service



รูปที่ ค.1.1-7 หน้าต่าง Service

3.1) ส่วนป้อนข้อมูลและเงื่อนไข

Number of customers : กำหนดจำนวนทั้งหมดที่จะ generate

Conditions

Percent of customers : กำหนดจำนวนเปอร์เซ็นต์จากข้อมูลทั้งหมด ในเงื่อนไขเหมือนกับ กำหนดจำนวน record ที่จะทำตามเงื่อนไขนี้

Percent

Service Rate : อัตรา service มีคะแนน 1-5 กำหนดเป็นเปอร์เซ็นต์

After set up Rate : อัตรา service หลังติดตั้ง มีคะแนน 1-5 กำหนดเป็นเปอร์เซ็นต์

Not finish Rate : อัตรา service ที่ไม่เสร็จ มีคะแนน 1-5 กำหนดเป็นเปอร์เซ็นต์

Number of years : จำนวนปีตั้งแต่ติดตั้ง มีคะแนน 1-5 กำหนดเป็นเปอร์เซ็นต์

Service points : คะแนนการ service มีคะแนน 1-5 กำหนดเป็นเปอร์เซ็นต์

โดยที่ในแต่ละหัวข้อนั้นถ้ามีการใส่ครบทั้ง 1-5 เปอร์เซ็นต์รวมจะต้องได้ 100 พอดี แต่ถ้าใส่ไม่ครบทุกช่อง เปอร์เซ็นต์รวมเฉพาะช่องที่ใส่ต้องไม่เกิน 100 ซึ่งช่องที่ไม่ใส่นั้นจะถือว่าเป็นการ random

ดังตัวอย่างในรูปที่ ค.1.1-7 เป็นการกำหนดว่าจะ generate ข้อมูลในตาราง Service ทั้งหมด 1000 records ดูจากช่อง Number of customers และกำลังกำหนดเงื่อนไข ดังนี้ ใน 50%จากข้อมูลทั้งหมด จะมีคะแนน Service Rate เท่ากับ 1 อยู่ 20% ซึ่งก็คือ 50%จาก 1000 records เท่ากับ 500 records จะมี คะแนน Service Rate เท่ากับ 1 อยู่ 20% เท่ากับ 100 records นั่นเอง คะแนนอื่นก็จะเป็นตามที่กำหนดเปอร์เซ็นต์เอาไว้เช่นกัน แต่คะแนนเท่ากับ 3 เว้นว่างไว้จะหมายถึง random ซึ่งจะคิดจาก เปอร์เซ็นต์ที่เหลืออยู่

ปุ่ม Insert : ใส่เงื่อนไขเข้าไปใน list เงื่อนไขทั้งหมดโดย โปรแกรมจะเช็คเงื่อนไขที่จะใส่ว่าถูกต้องหรือไม่ ถ้าไม่จะแจ้ง error และยกเลิกการใส่เงื่อนไขนั้น

ปุ่ม Delete : ลบเงื่อนไขที่เลือกใน list เงื่อนไขออก

3.2) ปุ่มจัดการกับข้อมูลในตาราง Service

ปุ่ม DeleteData : ลบข้อมูลทั้งหมดในตาราง Service

ปุ่ม GenerateData : ทำการ Generate ข้อมูลตามจำนวนและเงื่อนไขที่ได้กำหนดไว้ และเก็บลงในตาราง Service

3.3) ส่วนแสดงข้อมูลในตาราง Service

แสดงข้อมูลทั้งหมดในตาราง Service

ค.1.2. ตัวอย่างการใช้งานโปรแกรม

ตัวอย่างที่ 1

สมมติว่าต้องการ generate data ในตาราง Product ดังนี้

ต้องการ generate ทั้งหมด 1000 records โดยมีเงื่อนไขดังนี้

Number	%	Brand	Quality	Energy	Air Purifier	Quiet	Beauty	Price	Service	Others
300	30	Random	5	5	5	5	2	2	Random	Random
200	20	4	5	5	4	5	4	2	4	Random
150	15	5	5	5	4	4	5	5	5	Random
120	12	5	5	5	4	3	2	5	Random	Random
80	8	2	4	4	5	4	2	5	Random	Random
150	15	Random	Random	Random	Random	Random	Random	Random	Random	Random

ตารางที่ ค.1.2-1 ตัวอย่างที่ 1ส่วนเงื่อนไข

หมายเหตุ *Brand คือ ตราสินค้า*

Quality คือ คุณภาพ – อายุการใช้งาน

Energy คือ การประหยัดพลังงาน

Air Purifier คือ การฟอกอากาศ

Quiet คือ ความเงียบ

Beauty คือ ความสวยงาม

Price คือ ราคา

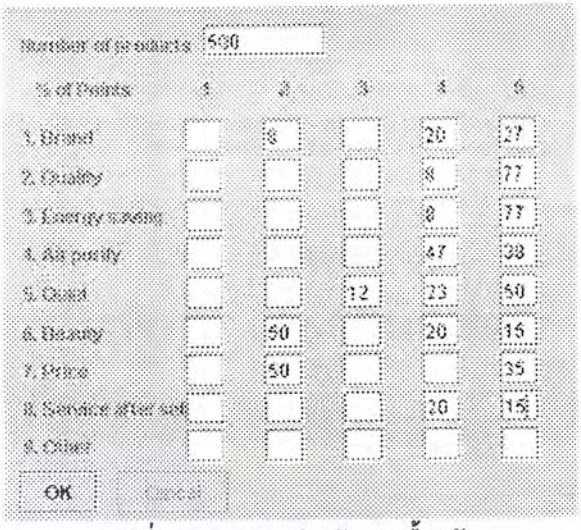
Service คือ การบริการหลังการขาย

Others คือ ปัจจัยอื่นๆ

จากเงื่อนไขทำให้รู้ว่า ข้อมูลของแต่ละหัวข้อต้อง generate กี่ เปอร์เซ็นต์

Brand	คะแนน 2	มี 8%	คะแนน 4	มี 20%	คะแนน 5	มี 27%
Quality	คะแนน 4	มี 8%	คะแนน 5	มี 77%		
Energy	คะแนน 4	มี 8%	คะแนน 5	มี 77%		
Air purify	คะแนน 4	มี 47%	คะแนน 5	มี 38%		
Quiet	คะแนน 3	มี 12%	คะแนน 4	มี 23%	คะแนน 5	มี 50%
Beauty	คะแนน 2	มี 50%	คะแนน 4	มี 20%	คะแนน 5	มี 15%
Price	คะแนน 2	มี 50%	คะแนน 5	มี 35%		
Service	คะแนน 4	มี 20%	คะแนน 5	มี 15%		

Other Random ทั้งหมดดังนั้นจึงกำหนดข้อมูลเบื้องต้นลงไปในโปรแกรมดังรูปที่ ค.1.2-1 และกดปุ่ม OK



รูปที่ ค.1.2-1 ตัวอย่างข้อมูลเบื้องต้น

ต่อไปกำหนดเงื่อนไขจากเงื่อนไขแรก

Number	%	Brand	Quality	Energy	Air Purifier	Quiet	Beauty	Price	Service	Others
300	30	Random	5	5	5	5	2	2	Random	Random

สมมติให้ Quiet คะแนน 5 เป็นเงื่อนไขนำ เปอร์เซ็นต์ ของเงื่อนไขตามหาได้จาก

%ของเงื่อนไข / %ทั้งหมดของเงื่อนไขนำ *100 = 30/50 * 100 = 60%

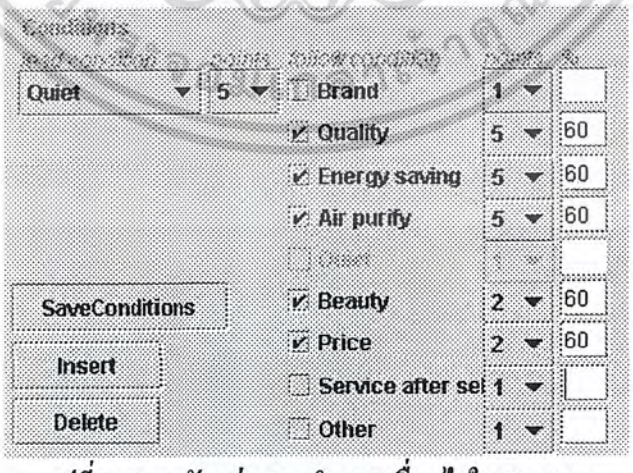
จะกำหนดเงื่อนไขแรกได้ดังนี้ ตามรูปที่ ค.1.2-2

lead condition = Quiet, 5 points;

follow condition = Quality, 5 points; 60% Energy saving, 5 points; 60%

Air purify, 5 points; 60% Beauty, 2 points; 60%

Price, 2 points; 60%



รูปที่ ค.1.2-2 ตัวอย่างการกำหนดเงื่อนไขใน Product

และเช่นเดียวกันเงื่อนไขที่ 2 จะกำหนดได้ดังนี้

lead condition = Brand, 4 points;

follow condition = Quality, 5 points; 100% Energy saving, 5 points; 100%
Air purify, 4 points; 100% Quiet, 5 points; 100%
Beauty, 4 points; 100% Price, 2 points; 100%
Service after sell, 4 points; 100%

เงื่อนไขที่ 3 กำหนดได้ดังนี้

lead condition = Beauty, 5 points;

follow condition = Brand, 5 points; 100% Quality, 5 points; 100%
Energy saving, 5 points; 100% Air purify, 4 points; 100%
Quiet, 4 points; 100% Price, 5 points; 100%
Service after sell, 5 points; 100%

เงื่อนไขที่ 4 กำหนดได้ดังนี้

lead condition = Quiet, 3 points;

follow condition = Brand, 5 points; 100% Quality, 5 points; 100%
Energy saving, 5 points; 100% Air purify, 4 points; 100%
Beauty, 2 points; 100% Price, 5 points; 100%

และเงื่อนไขที่ 5 กำหนดได้ดังนี้

lead condition = Brand, 2 points;

follow condition = Quality, 4 points; 100% Energy saving, 4 points; 100%
Air purify, 5 points; 100% Quiet, 4 points; 100%
Beauty, 2 points; 100% Price, 5 points; 100%

ส่วนเงื่อนไขสุดท้ายที่เป็น Random นั้นไม่ต้องกำหนดอะไรทั้งสิ้น โปรแกรมจะทำการ random ให้เอง ในคะแนนส่วนที่เหลือ

ตัวอย่างที่ 2

สมมติว่าต้องการ generate data ในตาราง Service ดังนี้

ต้องการ generate ทั้งหมด 1000 records โดยมีเงื่อนไขดังนี้

No.	%	Number	Service Rate Sat.	After Setup Satisfac.	Not Finish Satisfac.	No. Year	Service Point
1	25	250	5	Random	Random	3 - 5	4
2	22	220	3 - 5	Random	2	3 - 5	4
3	18	180	3 - 5	3	1	3 - 5	1
4	15	150	1 - 3	2 - 4	2 - 4	1 - 3	2
5	10	100	1 - 3	2	2	1 - 3	2
6	10	100	Random	Random	Random	Random	Random

ตารางที่ ค.1.2-2 ตัวอย่างที่ 2ส่วนเงื่อนไข

หมายเหตุ *Service Rate Sat. คือ อัตราความพึงพอใจของลูกค้าต่อความถี่การบริการ*

After Setup Satisfac. คือ อัตราความพึงพอใจของลูกค้าต่อการบริการหลังการติดตั้ง

Not Finish Satisfac. คือ อัตราความพึงพอใจของลูกค้าต่อการบริการที่ไม่เสร็จในครั้งเดียว

No. Year คือ จำนวนปีหลังการติดตั้ง

Service Point คือ อัตราความพึงพอใจลูกค้า

เงื่อนไขที่ 1 กำหนดได้ดังนี้

ที่Condition % = 25

คะแนน Service Rate มี 5 เท่านั้นจึงให้ ใส่ 100% ลงในช่องคะแนน 5

คะแนน After Setup Rate เป็น Random จึงไม่ต้องกำหนดอะไรทั้งสิ้น

คะแนน Not Finish Rate ก็เป็น Random เช่นกันจึงไม่ต้องกำหนดอะไรทั้งสิ้น

คะแนน No.Year เป็น 3-5 ก็หมายความว่าไม่มีคะแนน1 และ2 นั่นเองให้ใส่ 0%ลงในช่องคะแนน 1และ 2

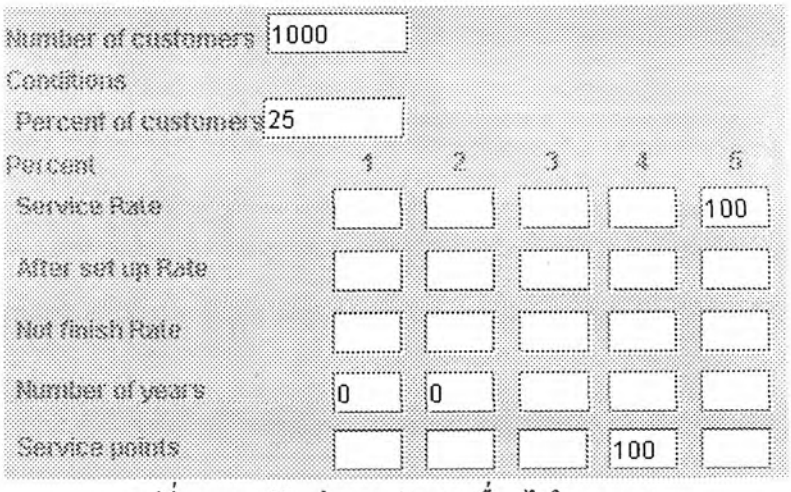
คะแนน Service Point มี 4 เท่านั้นจึงให้ ใส่ 100% ลงในช่องคะแนน 4

ดังนั้นจะกำหนดเงื่อนไขที่ 1 ได้ดังนี้ ตามรูปที่ ค.1.2-3

Percent of customers = 25%

Service Rate = -,-,-,-,100; After Setup Rate= -,-,-,-,-,; Not Finish Rate = -,-,-,-,-,;

No. Year = 0,0,-,-,-,; Service Point = -,-,-,100,-;



รูปที่ ค.1.2-3 ตัวอย่างการกำหนดเงื่อนไขใน Service

และเช่นเดียวกันเงื่อนไขที่ 2 จะกำหนดได้ดังนี้

Percent of customers = 22%

Service Rate = 0,0,-,-,-,; After Setup Rate= -,-,-,-,-,; Not Finish Rate = -,100,-,-,-,;

No. Year = 0,0,-,-,-,; Service Point = -,-,-,100,-;

เงื่อนไขที่ 3 กำหนดได้ดังนี้

Percent of customers = 18%

Service Rate = 0,0,-,-,-,; After Setup Rate = -,-,100,-,-,; Not Finish Rate = 100,-,-,-,-,;

No. Year = 0,0,-,-,-,; Service Point = 100,-,-,-,-;

เงื่อนไขที่ 4 กำหนดได้ดังนี้

Percent of customers = 15%

Service Rate = -,-,-,0,0,; After Setup Rate = 0,-,-,-,0,; Not Finish Rate = 0,-,-,-,0,;

No. Year = -,-,-,0,0,; Service Point = -,100,-,-,-;

เงื่อนไขที่ 5 กำหนดได้ดังนี้

Percent of customers = 10%

Service Rate = -,-,-,0,0,; After Setup Rate = -,100,-,-,-,; Not Finish Rate = -,100,-,-,-,;

No. Year = -,-,-,0,0,; Service Point = -,100,-,-,-;

ส่วนเงื่อนไขที่ 6 เป็น Random จึงไม่ต้องกำหนดอะไรทั้งสิ้น โปรแกรมจะทำการ random ให้เอง ในส่วนที่เหลือ

ค.2. การใช้งานโปรแกรมดาตาไมนิ่งในส่วนของการคลัสเตอร์ข้อมูล

แบ่งเป็น 2 ขั้นตอนหลัก คือ

1. ขั้นตอนในการเตรียมข้อมูลจริงสำหรับการคลัสเตอร์ข้อมูล
2. ขั้นตอนการคลัสเตอร์ข้อมูล

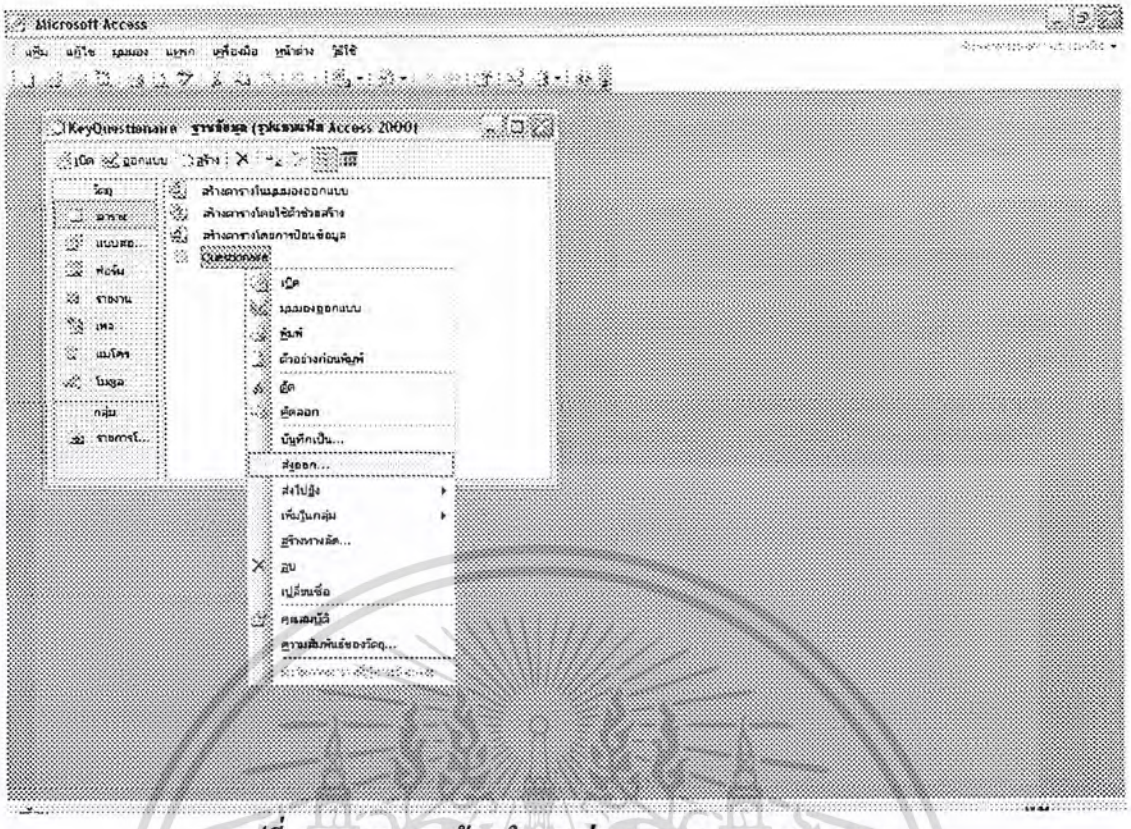
ในส่วนการเตรียมข้อมูลโดยการจำลองข้อมูลสำหรับการคลัสเตอร์ข้อมูล นั้น ได้อธิบายไว้แล้วในตอนต้นในส่วนของการใช้งานโปรแกรมดาตาไมนิ่งในส่วนของการจำลองข้อมูล

ค.2.1. ขั้นตอนในการเตรียมข้อมูลจริงสำหรับการคลัสเตอร์ข้อมูล

จะประกอบด้วยโปรแกรมที่รับอินพุท (Input) ที่ได้จากแบบสำรวจ ซึ่งสร้างจากฟอร์มของ Microsoft Access โดยการคีย์ (key) ค่าต่างๆที่ได้จากแบบสำรวจลงไป เก็บไว้ในฐานข้อมูลของ Microsoft Access ดังรูปที่ ค.2.1-1

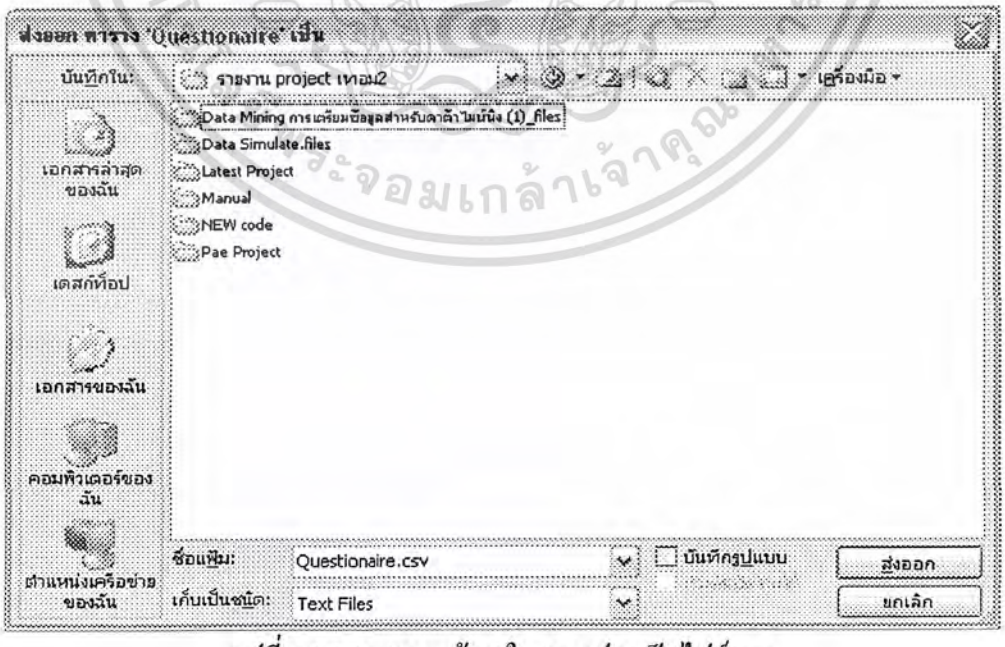
รูปที่ ค.2.1-1 แสดงถึงฟอร์มในการ key ค่าลงในฐานข้อมูล Access

โดยขั้นตอนการใช้ ก็เลือก อาชีพ เพศ ช่วงอายุ รายได้ การพักอาศัย ให้ตรงกับแบบสำรวจที่ได้กรอกมา โดยการคลิกที่ Combobox ของแต่ละตัว จากนั้นก็ทำการ key ค่าต่างๆตามแบบสำรวจที่ได้กรอกมา เมื่อ key ค่าต่างๆครบแล้ว ก็กดปุ่ม  ที่บริเวณด้านล่างทางซ้ายมือ ดังในรูปที่ ค.2.1-1 เพื่อทำการ key ค่าข้อมูลของแบบสำรวจถัดไป ทำเช่นนี้ไปเรื่อยๆ ตามจำนวนของแบบสำรวจ



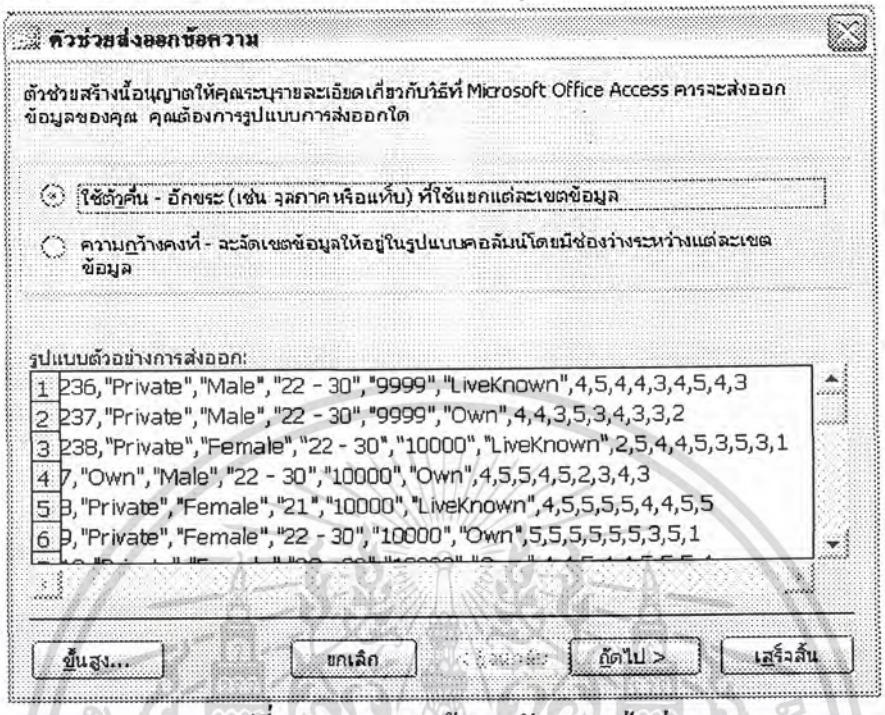
รูปที่ ค.2.1-2 แสดงหน้าจอในการส่งออกตาราง Questionnaire

เมื่อ key ค่าต่างๆเสร็จแล้ว ก็กดปุ่มปิด (รูปกากบาทสีแดงทางขวาบน) จากนั้นคลิกที่ตาราง แล้วคลิกขวาที่ Questionaire เลือกส่งออก ดังรูปที่ ค.2.1-2 เพื่อทำการแปลงฐานข้อมูล Access ให้อยู่ในรูปของ ไฟล์ .csv เพื่อที่จะนำไปใช้ในฐานข้อมูลใน MySQL ต่อไป

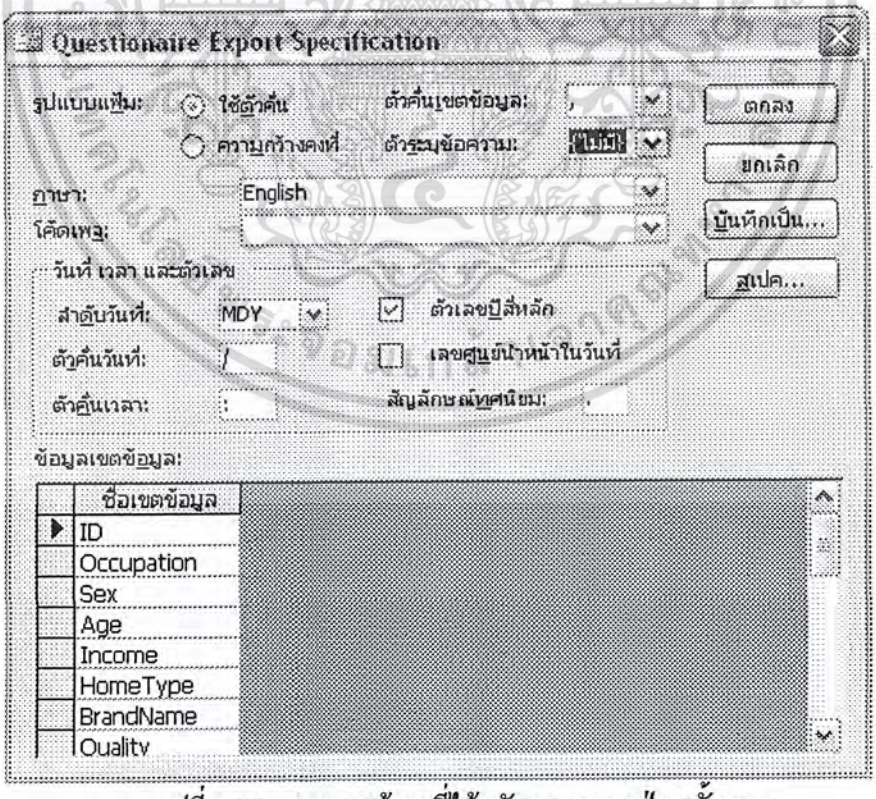


รูปที่ ค.2.1-3 แสดงหน้าจอในการแปลงเป็นไฟล์ .csv

เมื่อเลือกแล้วจะขึ้น Dialog ดังรูปที่ ค.2.1-3 เลือกตำแหน่งที่จะเก็บไฟล์ พร้อมทั้งพิมพ์คำว่า ".csv" ต่อท้ายชื่อไฟล์ตรงบริเวณที่เขียนว่า "ชื่อแฟ้ม" และเลือก Text Files ตรงบริเวณที่เขียนว่า "เก็บเป็นชนิด" จากนั้นกดที่ปุ่ม ส่งออก เพื่อทำขั้นตอนต่อไป ซึ่งจะได้ Dialog ดังรูปที่ ค.2.1-4 กดปุ่มที่เขียนว่า "ขั้นสูง"



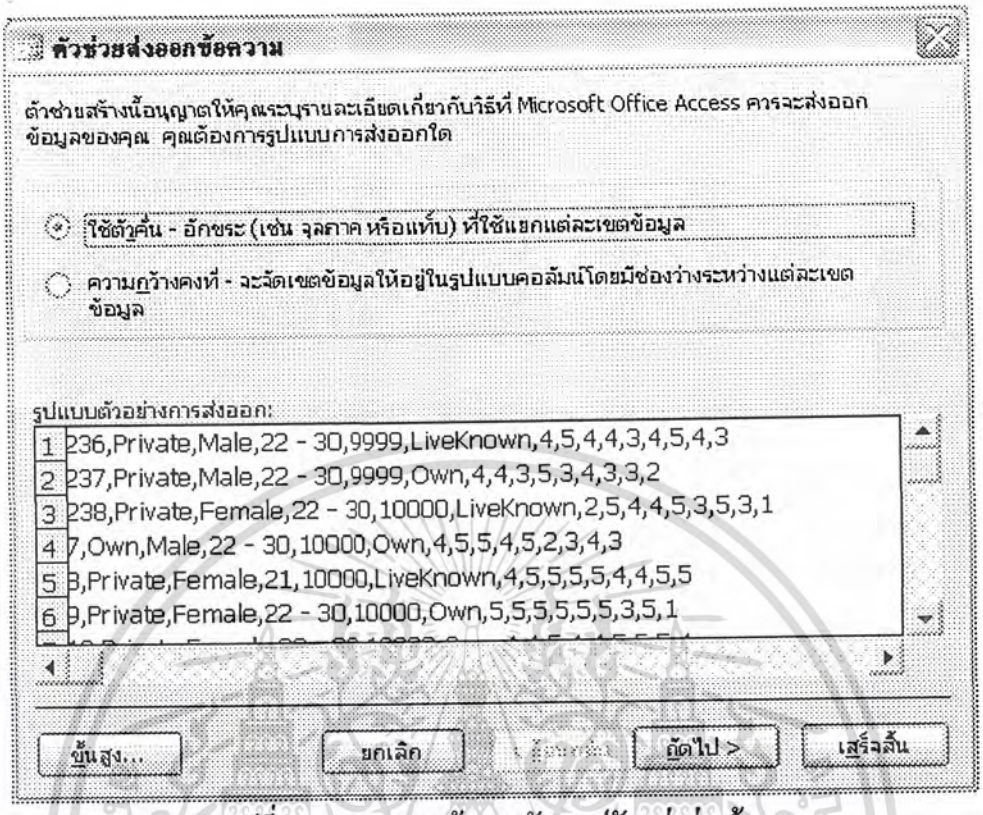
รูปที่ ค.2.1-4 แสดงหน้าจอหลังจากกดปุ่มส่งออก



รูปที่ ค.2.1-5 แสดงหน้าจอที่ได้หลังจากการกดปุ่ม "ขั้นสูง"

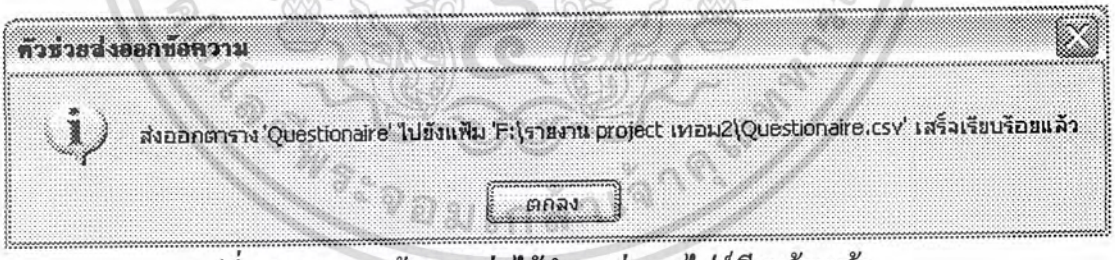
เมื่อกดแล้ว จะได้ Dialog ดังรูปที่ ค.2.1-5 เลือก "ไม่มี" ตรงบริเวณที่เขียนว่า "ตัวระบุข้อความ"

จากนั้นกดที่ปุ่ม ตกลง จะได้ Dialog ดังรูปที่ ค.2.1-6



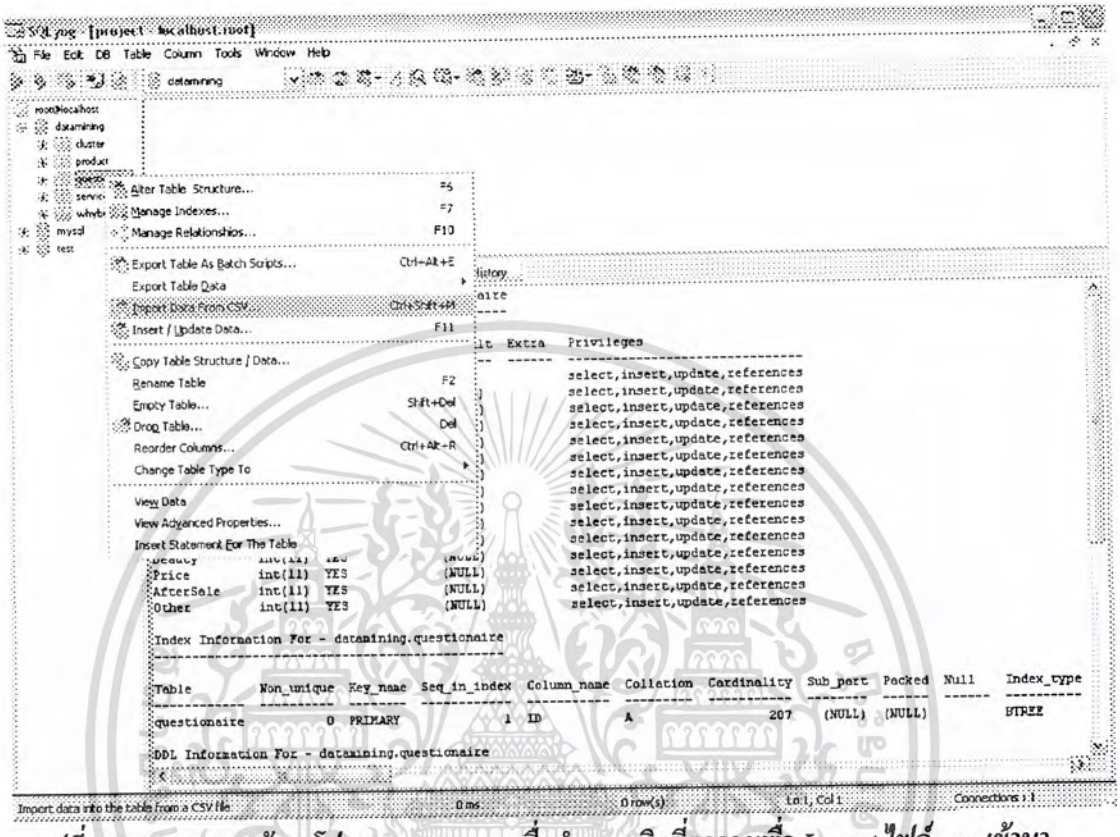
รูปที่ ค.2.1-6 แสดงหน้าจอหลังจากปรับแต่งค่าแล้ว

กดที่ปุ่ม เสร็จสิ้น จะได้ Dialog ดังรูปที่ ค.2.1-7 แสดงข้อความว่าได้ ทำการส่งออก เรียบร้อยแล้ว



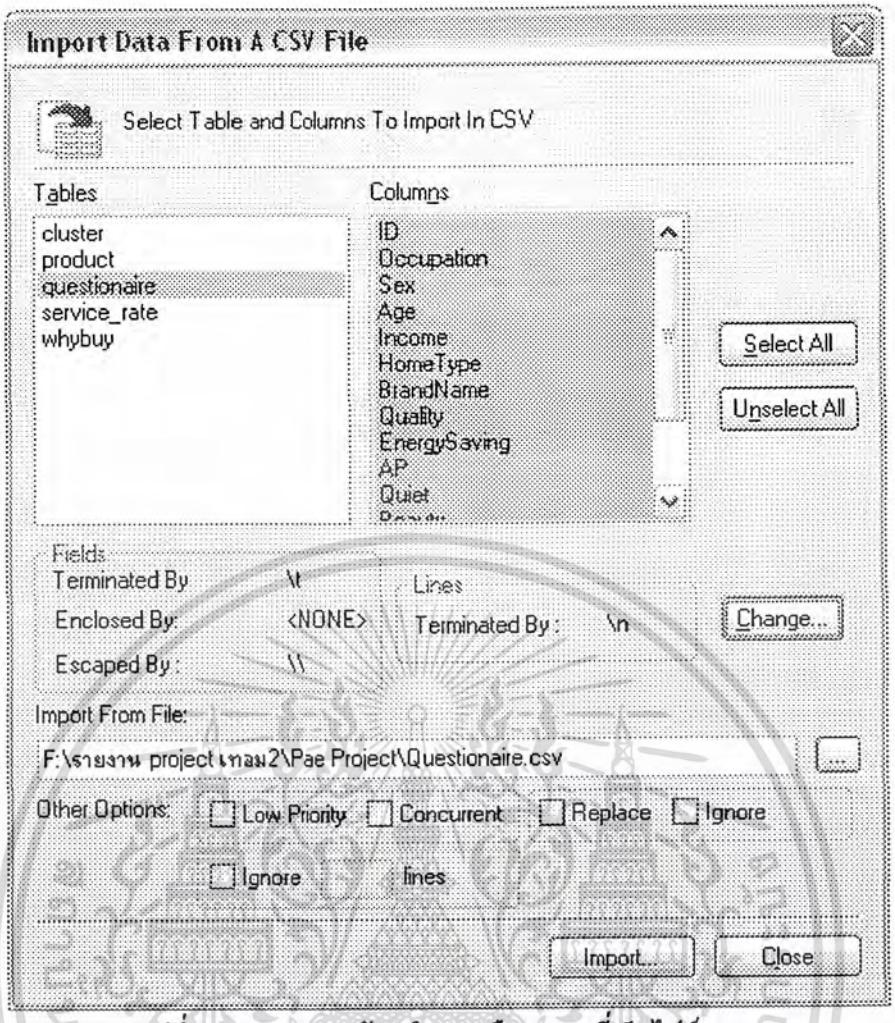
รูปที่ ค.2.1-7 แสดงข้อความว่า ได้ทำการส่งออกไฟล์เรียบร้อยแล้ว

ขั้นตอนต่อไปเปิดโปรแกรม SQLyog ขึ้นมา ทำการเลือกตารางที่ต้องการจะเก็บข้อมูลจากแบบสำรวจ ในที่นี้ จะเลือกตารางที่ชื่อว่า questionnaire โดยการคลิกขวาที่ตาราง เลือก Import Data From CSV ดังรูปที่ ค.2.1-8



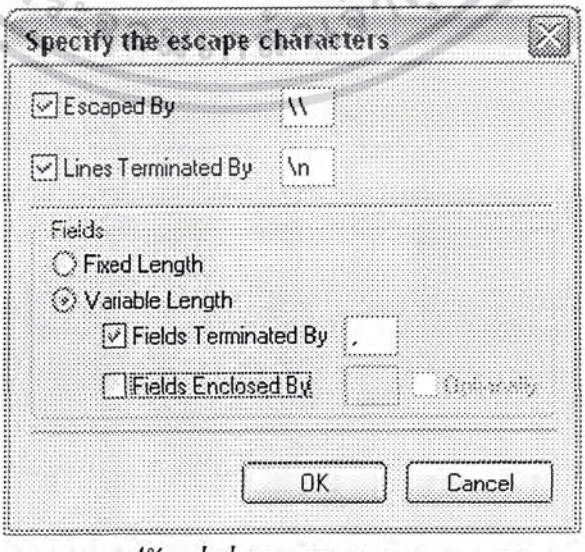
รูปที่ ค.2.1-8 แสดงหน้าจอโปรแกรม SQLyog เมื่อทำการคลิกที่ตารางเพื่อ Import ไฟล์ .csv เข้ามา

เมื่อเลือกแล้ว จะได้ Dialog ดังรูปที่ ค.2.1-9 ทำการเลือกบริเวณที่เก็บไฟล์ที่จะ Import เข้ามา (ไฟล์ questionnaire.csv ที่ได้ทำการส่งออกจาก Microsoft Access ในตอนต้น)โดยกดที่ปุ่ม  เมื่อเลือกไฟล์ที่ต้องการจะ Import ได้แล้ว ก็กดที่ปุ่ม Change เพื่อทำการปรับค่าอีกเล็กน้อย



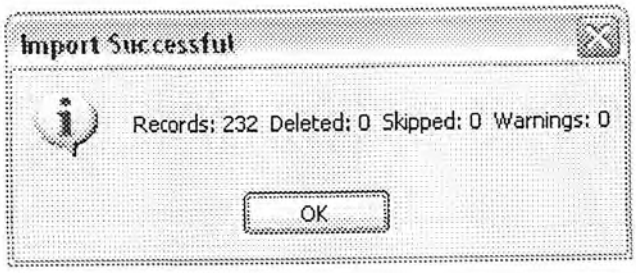
รูปที่ ค.2.1-9 แสดงหน้าจอในการเลือก path ที่เก็บไฟล์ .csv

เมื่อเลือกแล้ว จะได้ Dialog ดังรูปที่ ค.2.1-10 ทำการเปลี่ยนค่า ตรงบริเวณที่เขียนว่า "Fields Terminated By" จากเดิมที่เป็น "\t" ให้เปลี่ยนเป็น " , " แทน จากนั้นกดที่ปุ่ม OK



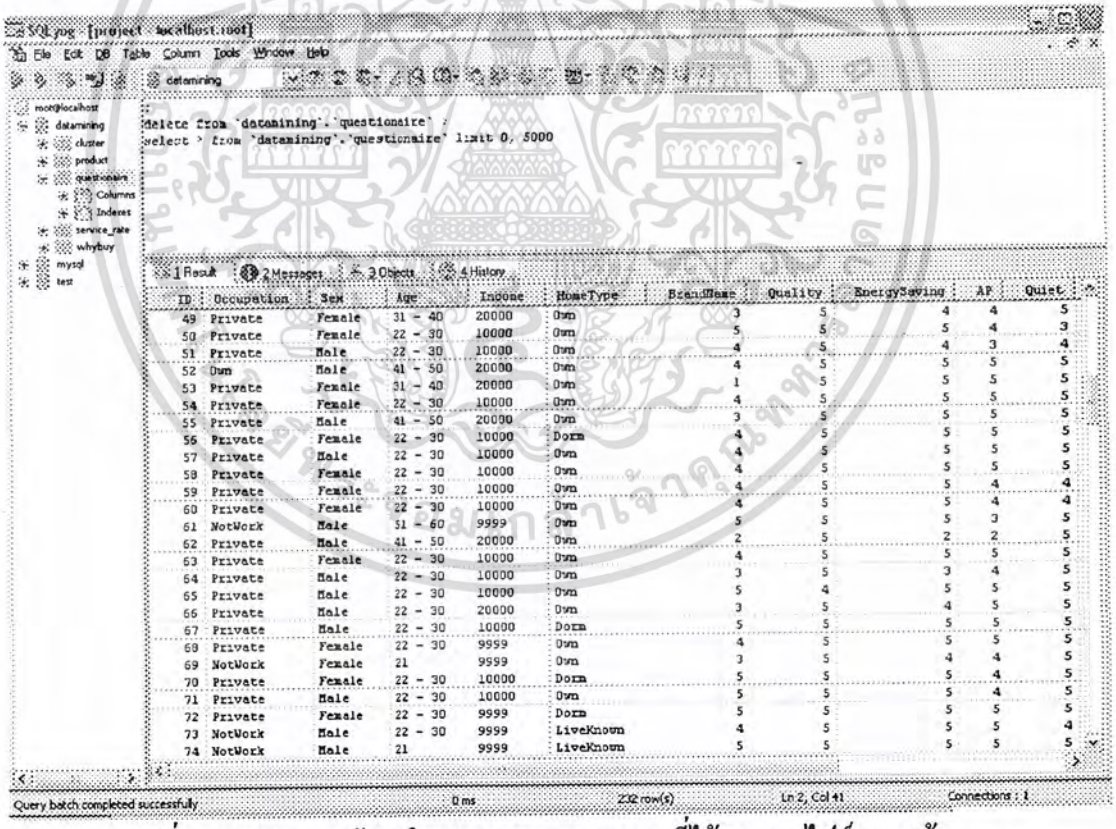
รูปที่ ค.2.1-10 การปรับแต่งค่าตรง "Fields Terminated By"

เมื่อกดแล้ว จะได้ Dialog ดังรูปที่ ค.2.1-9 อีกครั้ง จากนั้น กดที่ปุ่ม Import จะได้ Dialog ดังรูปที่ ค.2.1-11 ซึ่งจะแสดงข้อความว่าได้ Import ไฟล์ .csv เรียบร้อยแล้ว



รูปที่ ค.2.1-11 แสดงข้อความว่าได้ Import ไฟล์ .csv เรียบร้อยแล้ว

จากนั้นลองมาคลิกขวาที่ตาราง questionnaire เลือก View Data จะได้ ดังรูปที่ ค.2.1-12 ซึ่งตาราง questionnaire จะประกอบไปด้วยข้อมูลต่างๆที่ได้จากการ Import ไฟล์ .csv เรียบร้อยแล้ว



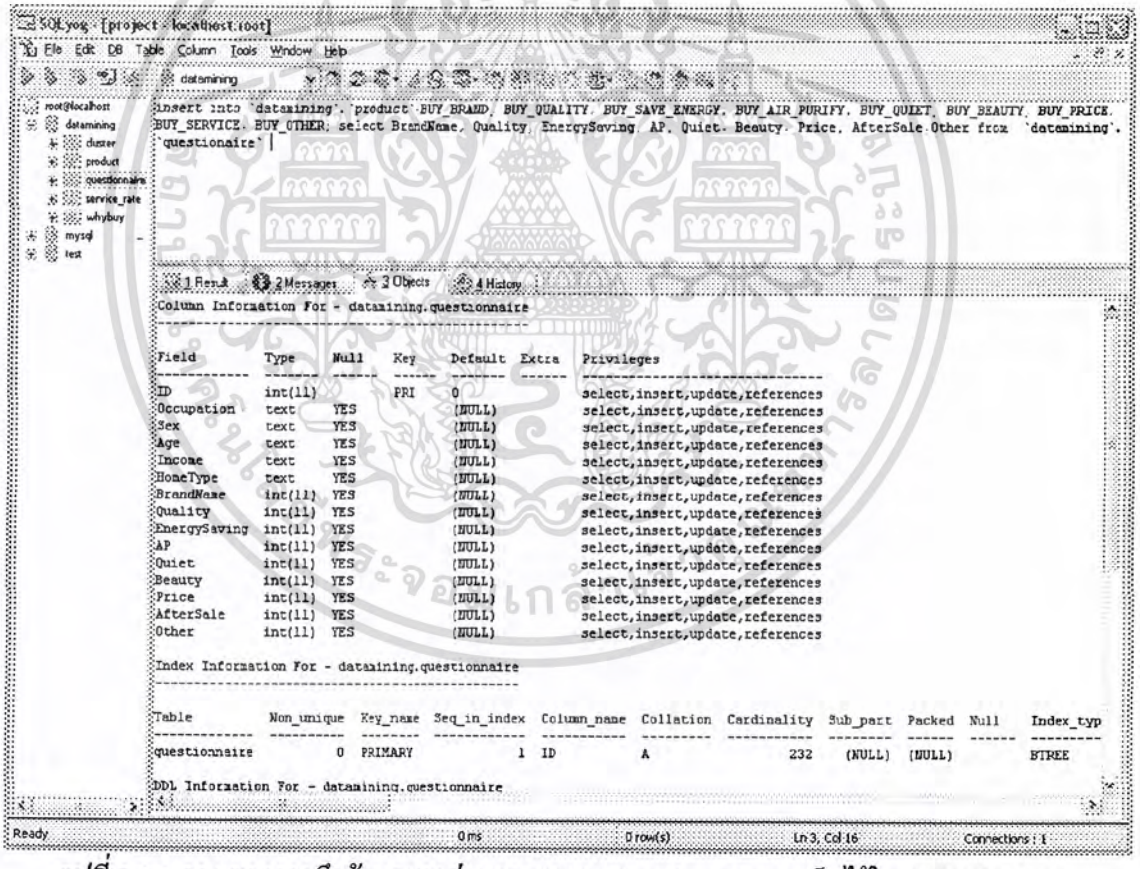
รูปที่ ค.2.1-12 แสดงข้อมูลในตาราง Questionnaire ที่ได้ Import ไฟล์ .csv แล้ว

แต่ตารางที่ใช้ในการ Clustering นั้น จะเป็นตารางที่ชื่อ Product ซึ่งจะต้องทำการพิมพ์คำสั่ง SQL เข้าไปเพื่อดึงข้อมูลจากตาราง questionnaire ไปใส่ในตาราง product อีกที โดยในขั้นตอนนี้ เราสามารถจะเลือก

เอาข้อมูลตามเงื่อนไขที่เราต้องการจะ cluster ได้ เช่น เลือกเอาข้อมูลเฉพาะของผู้ชายอย่างเดียว หรือ เลือกเอาข้อมูลจากรายได้ที่สูงกว่า 10,000 บาทต่อเดือน เป็นต้น ในที่นี้จะใช้ข้อมูลทั้งหมดจากแบบสำรวจ โดยพิมพ์คำสั่ง SQL ดังนี้

insert into `datamining`.`product`(BUY_BRAND, BUY_QUALITY, BUY_SAVE_ENERGY, BUY_AIR_PURIFY, BUY_QUIET, BUY_BEAUTY, BUY_PRICE, BUY_SERVICE, BUY_OTHER) select BrandName, Quality, EnergySaving, AP, Quiet, Beauty, Price, AfterSale,Other from `datamining`.`questionnaire`

ซึ่งคำสั่งดังกล่าว เป็นการดึงข้อมูลจากตาราง questionnaire เฉพาะ Column BrandName, Quality, EnergySaving, AP, Quiet, Beauty, Price, AfterSale และ Other มาเก็บไว้ในตาราง product อีกที ดังรูปที่ ค.2.1-13

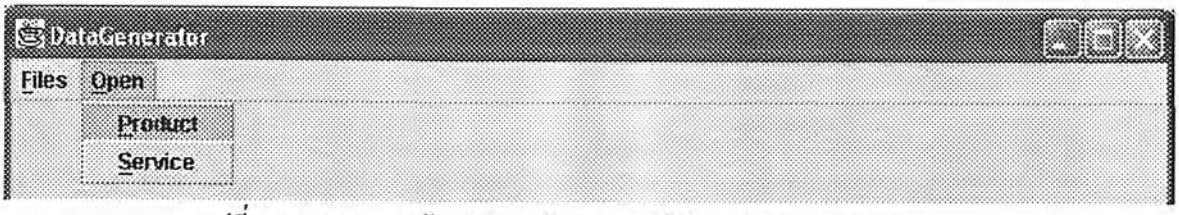


รูปที่ ค.2.1-13 แสดงการดึงข้อมูลบางส่วนจากตาราง Questionnaire มาเก็บไว้ในตาราง Product

เมื่อพิมพ์คำสั่งดังกล่าวแล้ว จากนั้น กดปุ่ม F5 เพื่อทำการประมวลผลคิวรี (Execute query) ก็จะได้ฐานข้อมูลที่จะนำไปใช้ในการ Clustering ต่อไป

ค.2.2. ขั้นตอนการใช้โปรแกรมในการทำ Clustering

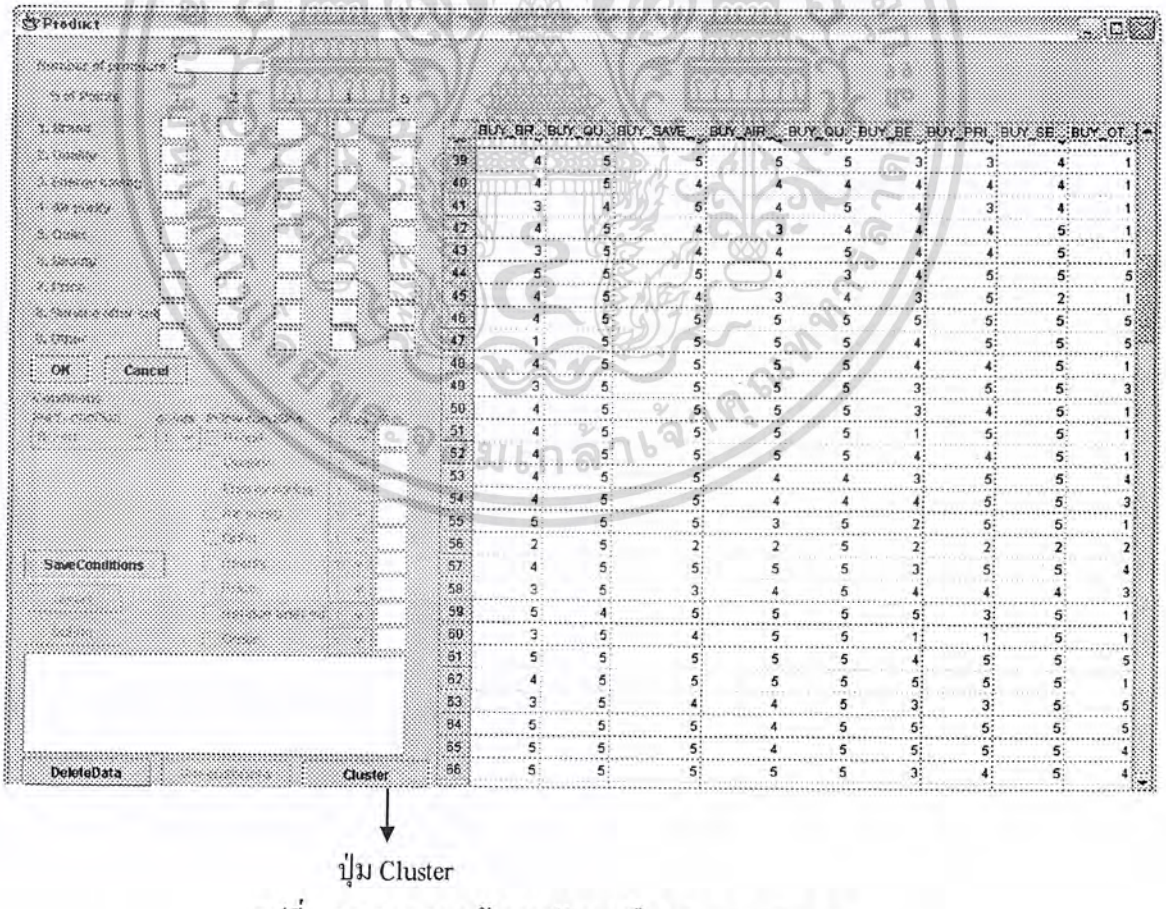
1. ทำการ run โปรแกรม Data Mining ขึ้นมา จากนั้นเลือกที่ Open -> Product ดังรูปที่ 6.3.1



รูปที่ ค.2.2-1 แสดงหน้าจอแรกหลังจาก run โปรแกรม Data Mining

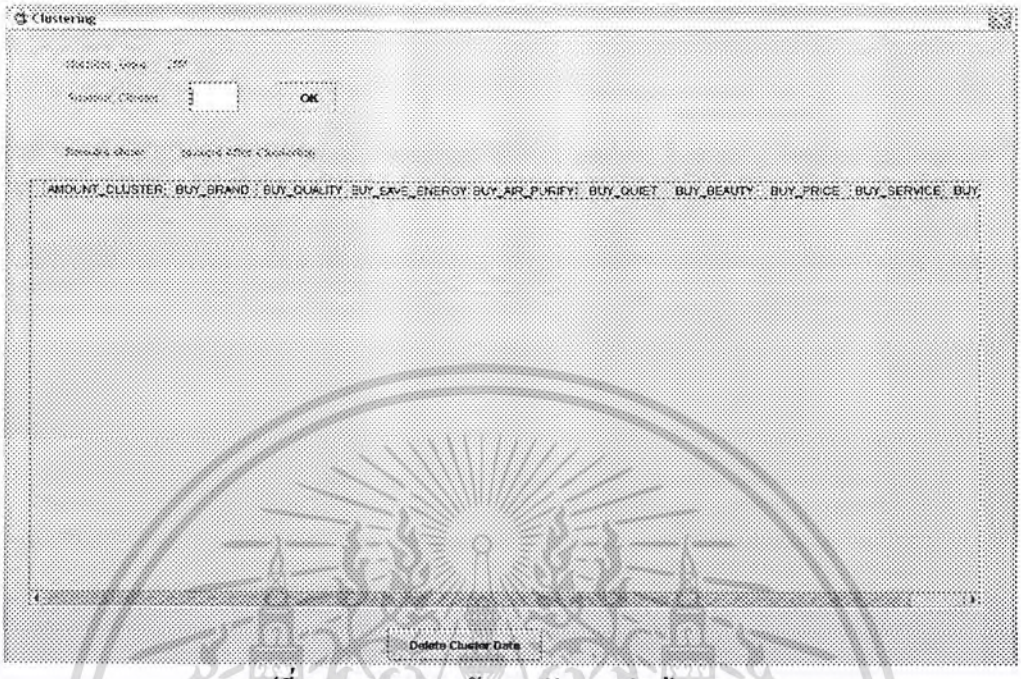
โดยในส่วนของการคลัสเตอร์ข้อมูลจะต้องเริ่มต้นที่ Product ก่อนเสมอ เมื่อเลือก Product แล้ว จะได้ผลลัพธ์ดังรูปที่ ค.2.2-1 หากตารางด้านขวามือยังไม่มีข้อมูลใดๆอยู่เลย จะทำการคลัสเตอร์ข้อมูลไม่ได้ ซึ่งจะต้องทำการจำลองข้อมูลขึ้นมาหรือนำฐานข้อมูลจริงไปเก็บในตาราง Product ในฐานข้อมูล Datamining ใน MySQL ก่อนเสมอ

2. เมื่อได้ฐานข้อมูลแล้ว ขั้นตอนต่อไป คือ การคลัสเตอร์ข้อมูล โดยกดที่ปุ่ม Cluster ดังรูปที่ ค.2.2-2



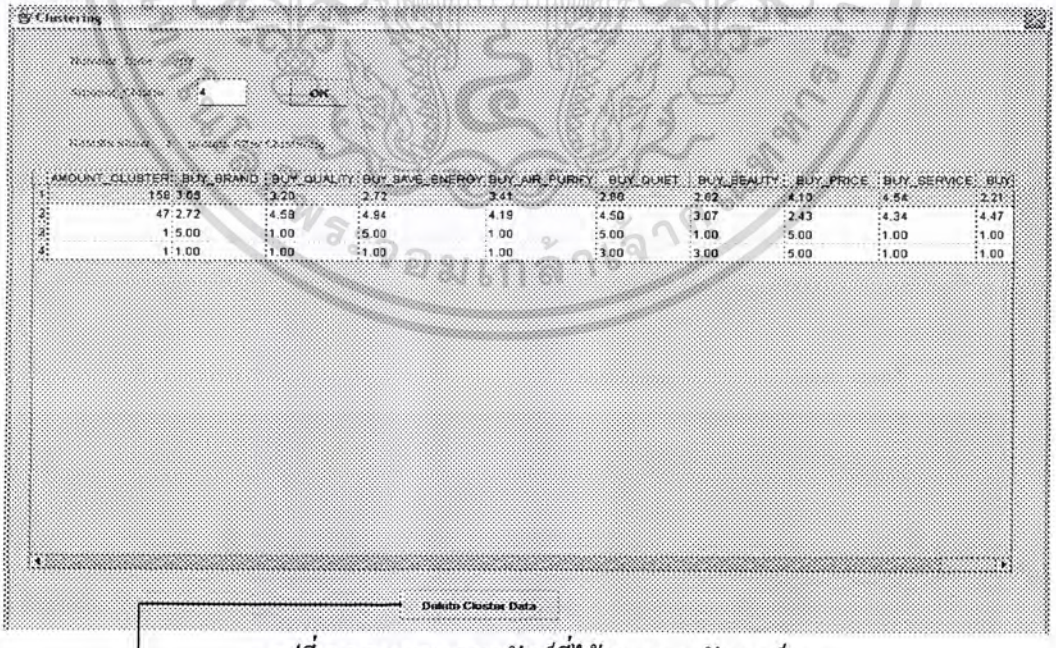
รูปที่ ค.2.2-2 แสดงหน้าจอหลังจากเลือก Open -> Product

ซึ่งจะแสดงผลลัพธ์เป็นหน้าจอ Clustering ดังรูปที่ ค.2.2-3 โดยจะแสดงรายละเอียดต่างๆ ได้แก่ จำนวนข้อมูลทั้งหมดที่จะทำการคลัสเตอร์ , ผลที่ได้จากการคลัสเตอร์ที่อยู่ในรูปแบบของตาราง รวมไปถึงจำนวน Cluster ที่ต้องการ ซึ่งสามารถกำหนดกี่ Cluster ก็ได้ตามที่ต้องการ



รูปที่ ค.2.2-3 แสดงหน้าจอหลังจากคลิกปุ่ม Cluster

3. กรอกจำนวนของ Cluster ที่ต้องการลงไปในช่อง Amount_Cluster จากนั้นกดปุ่ม OK ผลลัพธ์จะได้ดังรูปที่ ค.2.2-4 จากนั้นจึงนำผลลัพธ์ที่ได้ไปทำการวิเคราะห์ทางธุรกิจต่อไป



รูปที่ ค.2.2-4 แสดง ผลลัพธ์ที่ได้จากการคลัสเตอร์

ปุ่ม Delete Cluster Data

4. สามารถลบข้อมูลที่ได้จากการคลัสเตอร์ โดยกดที่ปุ่ม Delete Cluster Data

ค.3. การใช้งานโปรแกรมดาต้าไมนิ่งในส่วนของการทำ Decision Tree

ค.3.1. ขั้นตอนการใช้งานโปรแกรมดาต้าไมนิ่งในส่วนของการทำ Decision Tree

1. ทำการ Load ข้อมูล เข้ามาใน Matlab

ทำการดึงข้อมูลในส่วนของ Input

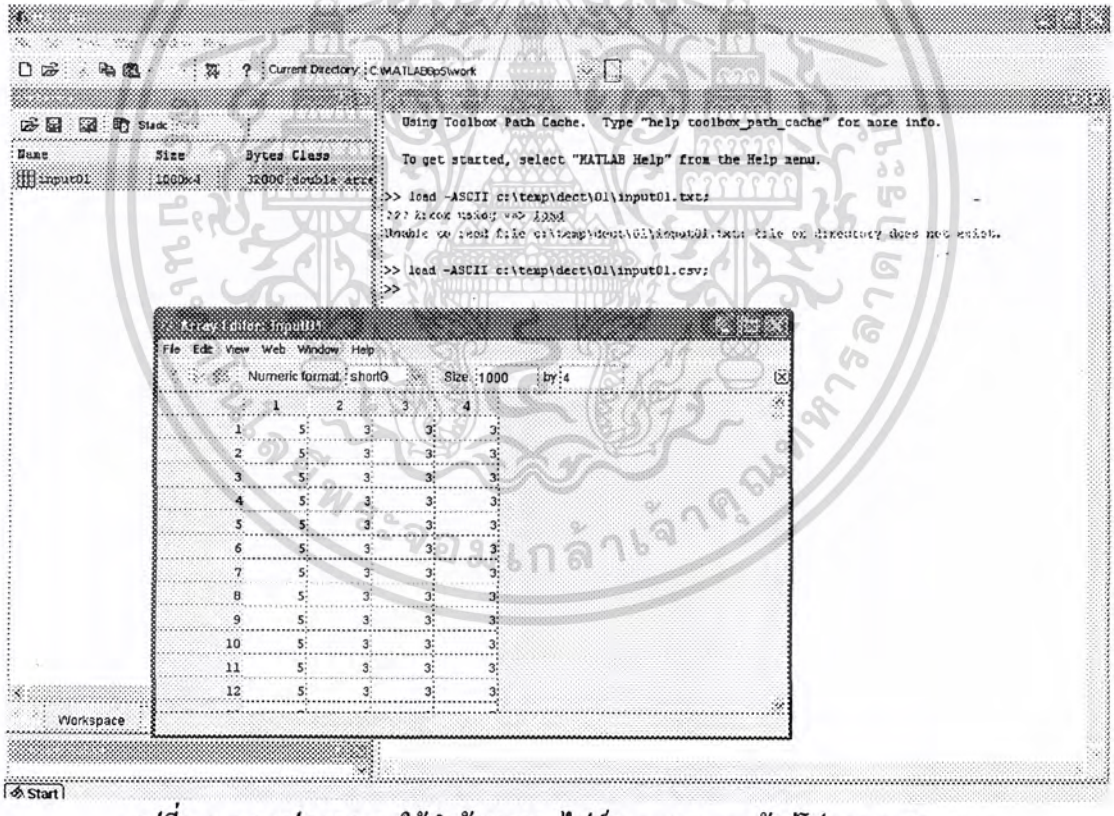
>>load -ASCII c:\temp\dect\01\input01.csv;

โดยคำสั่งดังกล่าวจะดึงข้อมูลจาก File input01.csv ไปเป็นตัวแปรใน Matlab ที่ชื่อ input โดย File Format ของ CSV File ดังกล่าว คือ

ข้อมูลชุดที่ 1:ตัวแปรที่ 1 + Space Bar + ตัวแปรที่ 2 + Space Bar + ตัวแปรที่ 3 + Space Bar + ตัวแปรที่ 4

...

ข้อมูลชุดที่ n:ตัวแปรที่ 1 + Space Bar + ตัวแปรที่ 2 + Space Bar + ตัวแปรที่ 3 + Space Bar + ตัวแปรที่ 4



รูปที่ ค.3.1-1 รูปแสดงการใช้นำข้อมูลจากไฟล์ input01.csv เข้าสู่โปรแกรม Matlab

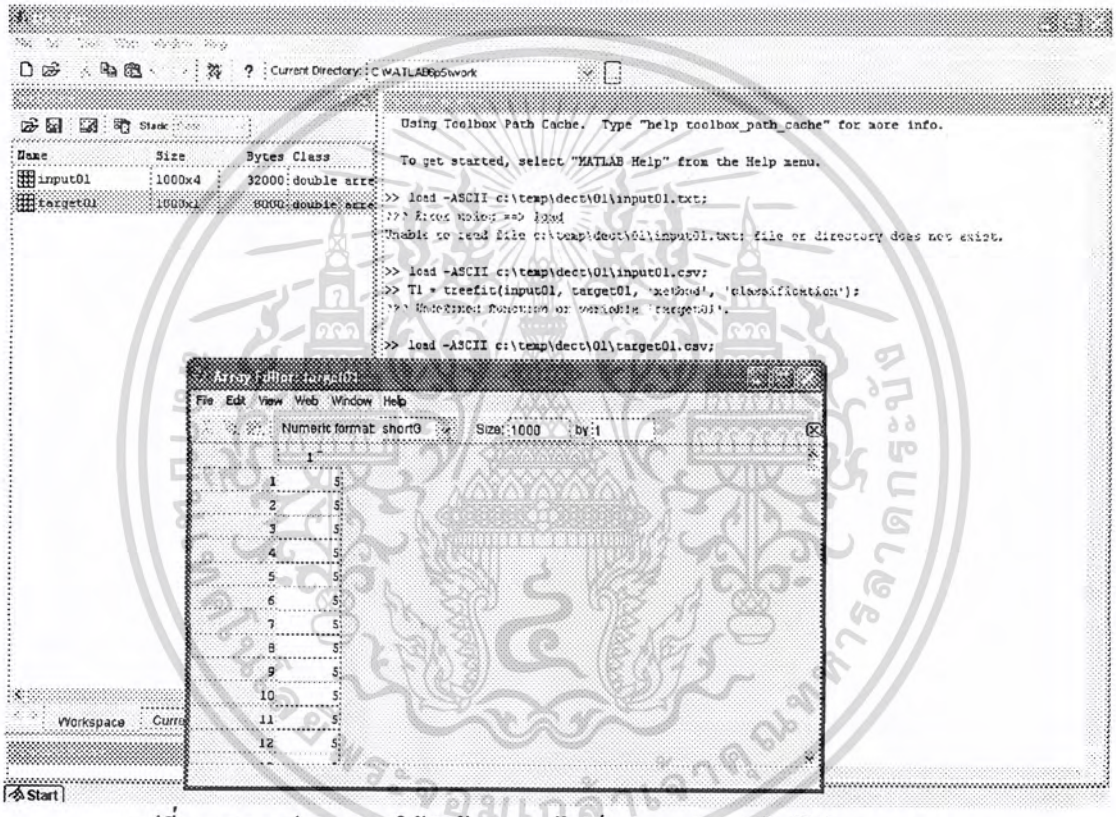
โดยสามารถดูข้อมูลตัวแปร input01 ได้โดย Double Click ที่ตัวแปร input01 ใน Workspace ด้านขวา ซึ่งจะเป็นหน้าจอ Array Editor ดังรูปที่ ค.3.1-1

ทำการดึงข้อมูลในส่วนของ Target

>>load -ASCII c:\temp\dect\01\target01.csv;

โดยคำสั่งดังกล่าวจะดึงข้อมูลจาก File Target01.csv ไปเป็นตัวแปรใน Matlab ที่ชื่อ input โดย File Format ของ CSV File ดังกล่าว คือ

ข้อมูลชุดที่ 1:ตัวแปรที่ 1 ... ข้อมูลชุดที่ n:ตัวแปรที่ 1



รูปที่ ค.3.1-2 รูปแสดงการใช้นำข้อมูลจากไฟล์ target01.csv เข้าสู่โปรแกรม Matlab

โดยสามารถดูข้อมูลตัวแปร Target01 ได้โดย Double Click ที่ตัวแปร Target 01 ใน Workspace ด้านขวา ซึ่งจะเป็นหน้าจอ Array Editor ดังรูปที่ ค.3.1-2

2. สร้าง Decision Tree ด้วยคำสั่ง treefit

>>T1 = treefit(input01, target01, 'method', 'classification');

คำสั่งดังกล่าว เป็นการสร้าง Decision Tree จาก Input ที่มีชื่อว่า Input01 และมีค่าที่ได้จาก Input หลังการประมวลผล คือ Target ซึ่งคือ Target01 นอกจากนั้นยังเป็นการใช้ Decision Tree เพื่อทำ Classification

3. แสดงผล Decision Tree ด้วยคำสั่ง treedisp

>>treedisp(T1,'names',{'Serv Rate' 'After Setup' 'Not Finish' 'No Year'});

เป็นการแสดง Decision Tree โดยใช้ Tree Viewer โดย แสดง Decision Tree ที่ชื่อว่า T1 และมีการแสดงชื่อในตัวแปรต่างๆ โดย

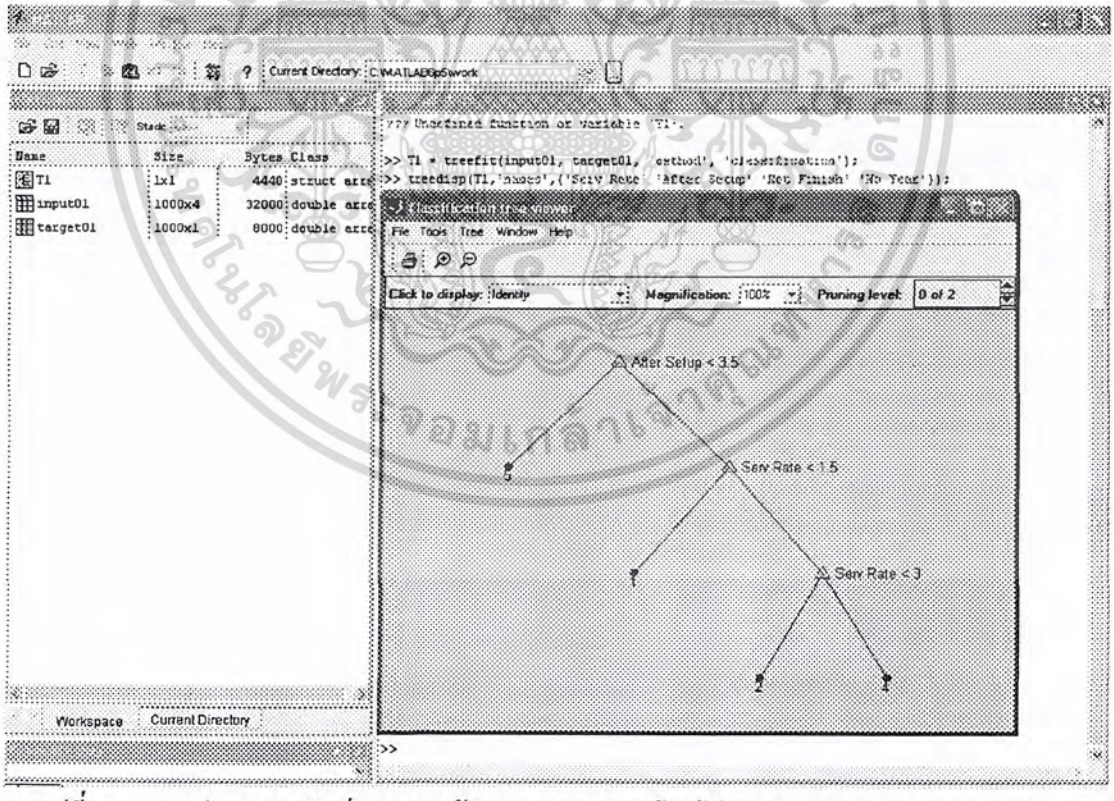
ตัวแปรที่ 1 ชื่อ “Serv Rate”

ตัวแปรที่ 2 ชื่อ “After Setup”

ตัวแปรที่ 3 ชื่อ “Not Finish”

ตัวแปรที่ 4 ชื่อ “No Year”

โดยลำดับของตัวแปรดังกล่าวจะเกี่ยวพันกับ Input01.csv ที่ Import เข้ามาก่อนหน้านี้



รูปที่ ค.3.1-3 รูปแสดงผลลัพธ์ของการสร้าง Decision Tree โดยโปรแกรม Classification tree viewer

ค.3.2. การทำงานของคำสั่งต่างๆ

Treefit

เป็นการสร้าง Tree-Model โดย Regression หรือ Classification

คำสั่ง T = treefit(X,y)

T = treefit(X,y,'param1',val1,'param2',val2,...)

รายละเอียด

T = treefit(X,y) เป็นคำสั่งที่ใช้ในการสร้าง Decision Tree T สำหรับการคาดการณ์ค่า y โดยเป็นฟังก์ชั่นในการคาดการณ์จากค่า X โดย X เป็น n x m เมตริกซ์ของค่าที่ใช้ในการคาดการณ์ และ Y เป็น Vector ของ n คำตอบ (สำหรับ Regression) หรือเป็น Character Array หรือ Cell Array ของ String ซึ่งประกอบด้วย n class name (สำหรับ Classification) หรือในอีกทางหนึ่ง T เป็น Binary Tree ซึ่งแต่ละ non-terminal node นั้นเป็นพื้นฐานของคำตอบที่เป็นค่า X โดยคำสั่ง T = treefit(X,y,'param1',val1,'param2',val2,...) นั้นระบุตัวแปรเพิ่มเติม โดยตัวแปรดังกล่าว ได้แก่

สำหรับ Tree ทั้งต้น

'catidx' เป็น Vector ของ Column ของ X โดย treefit จัดการ Column ดังกล่าวเสมือนเป็น Column ที่ไม่มีการจัดเรียงลำดับของกลุ่มข้อมูล

'method' มีค่าเป็น 'classification' (เป็นค่าปรกติหาก y เป็นตัวอักษร) หรือ 'regression' (เป็นค่าปรกติหาก y เป็นตัวเลข)

'splitmin' เป็นค่าที่บ่งชี้ว่า จะแตก Node ได้ต้องมีคำตอบใน Node นั้นๆ ไม่น้อยกว่าค่าดังกล่าว โดยมีค่าปรกติเป็น 10

'prune' 'on' (เป็นค่าปรกติ) ให้มีการ Prune และ 'off' หากไม่มีค่าการ Prune

Treedisp

เป็นการแสดง Decision Tree ให้เป็นเชิง Graphic

คำสั่ง treedisp(T)

treedisp(T,'param1',val1,'param2',val2,...)

treedisp(T) นั้นเป็นรับ input เป็น Decision tree T และคำนวณโดย treefit function และแสดงอยู่บน Window โดยแต่ละ Tree นั้นถูกตั้งชื่อด้วย Decision Rule และแต่ละ Terminal Node นั้นถูกตั้งชื่อให้เป็นค่าที่เป็นการคาดการณ์

treedisp(T,'param1',val1,'param2',val2,...) นั้นสามารถใส่ตัวแปรเพิ่มเติมเข้าไปได้ โดย

'name' นั้นเป็น Array ของชื่อที่ใช้ในการคาดการณ์ค่าตัวแปรต่างๆ

'prunelevel' เป็นการใช้ค่าเริ่มต้นในการแสดง Pruning Level ที่แสดงผลออกมา

บรรณานุกรม

[1] IBM red book, mining your own business in retail, http://www.ibm.com/redbooks

[2] Genetic Algorithm, http://www.codeproject.com/csharp/btl_GA.asp

[3] Genetic Algorithm, http://cs.felk.cvut.cz/~xobitko/ga/

[4] Genetic Algorithm, http://geneticalgorithms.ai-depot.com/Tutorials.html

[5] Clustering, http://cne.gmu.edu/modules/dau/stat/clustgalgs/clust3_frm.html

[6] Data Mining, http://www.thearling.com

[7] Data Mining & Knowledge Discovery, http://www.kdnuggets.com

[8] Clustering – Introduction, http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/index.html,

[9] Berry, M., and Linoff, G., "Data Mining Techniques for Marketing, Sales, and Customer Support", John Wiley & Sons, 1997.

[10] Neural Network Tutorial with Java Applet, http://diwww.epfl.ch/mantra/tutorial/english/

[11] Artificial Intelligence A Modern Approach, second edition, Stuart Russell, Peter Norvig, Prentice Hall Series in Artificial Intelligence