ระบบวิเคราะห์พฤติกรรมการใช้งานโทรศัพท์ของลูกค้าด้วยดิซิชั่นทรี

Customer Usage Mining For Telephone Company Using Decision Tree

โดย

อัจฉรา ประเสริฐ

รหัสประจำตัว 46066832

อาจารย์ที่ปรึกษา

ผศ.ดร. วรพจน์ กรีสุระเดช

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
ภาคเรียนที่ 1 ปีการศึกษา 2548
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ชื่อหัวข้อ	ระบบวิเคราะห์พฤติกรรมการใช้งานโทรศัพท์ของลูกค้าด้วยดิซิชั่นทรี
นักศึกษา	นางสาวอัจฉรา ประเสริฐ
อาจารย์ที่ปรึกษา	ผศ.ดร.วรพจน์ กรีสุระเดช
ระดับการศึกษา	วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2548

บทคัดย่อ

โครงงานนี้พัฒนาโดยนำหลักการของ ดาต้า ไมนิ่ง โดยผ่านกระบวนการที่เรียกว่า ดิชคิชั่น ทรี โมเดล มาใช้ในการจัดแบ่งกลุ่มหรือส่วนทางการตลาด / ลูกค้า เพื่อที่จะสามารถตอบสนองความต้องการของตลาด/ลูกค้ากลุ่มเป้าหมายและสามารถอธิบายลักษณะหรือพฤติกรรมของลูกค้าว่ามีลักษณะแบบใด ซึ่งจัดได้ว่ามีความสำคัญมากในเชิงธุรกิจ เพราะสามารถนำผลที่ได้จากการวิเคราะห์ไปวางแผนเพื่อเพิ่มกลุ่มเป้าหมาย นั่นหมายถึงขนาดส่วนแบ่งทางการตลาดที่เพิ่มขึ้น

Title	Customer Usage Mining for telephone company using decision tree
Student	Ms. Atchara Prasert
Advisor	Asst. Prof. Dr. Worapoj Kreesuradej
Level of Study	Master of Science in Information Technology
Major	Information Science
Academic Year	2005

ABSTRACT

The project implement by data mining with decision tree we focus on market segmentation operation and classify customer transactions. We propose the characteristic and behavior of customer to predict in market segmentation that is very important for business management. In the results, we could analyze and plan to increase target that means the shared market will also.

กิตติประกาศ

การจัดทำโครงงานพัฒนาระบบนี้สำเร็จล่วงไปด้วยดี เนื่องจากคำแนะนำ และความช่วยเหลือจากบุคคลต่างๆ ดังต่อไปนี้

บิดา และมารดา และครอบครัวที่คอยสนับสนุนและให้กำลังใจในการจัดทำโครงงานมาโดยตลอด และขอขอบคุณ ผศ ดร. วรพจน์ กรีสุระเดช ที่ให้คำแนะนำ และคำปรึกษา แนวทางในการพัฒนาโครงการ อีกทั้งให้กำลังใจส่งผลให้การพัฒนาโครงการ ซึ่งนับเป็นตัวขับเคลื่อนสำคัญทำให้โครงงานสำเร็จลุล่วง ขอบคุณ นายต่อพงษ์ โลหะรังสิกุลและ เพื่อน IS 16.2 ที่คอยเป็นกำลังใจให้กันและกันมาโดยตลอด

ด้วยความเคารพและขอบคุณเป็นอย่างสูง

(นางสาวอัจฉรา ประเสริฐ)

6 กันยายน 2548

สารบัญ

หน้า

สารบัญตาราง

ตารางที่ หน้า

สารบัญภาพ

รูปที่ หน้า

บทที่ 1
บทนำ

1.1 หลักการและเหตุผล

ปัจจุบันจะเห็นได้ว่าธุรกิจต่างๆ มีความเคลื่อนไหวอย่างรวดเร็ว มีการแข่งขันกันสูงขึ้น ปัจจัยหนึ่งที่จะช่วยส่งเสริมการเติบทางธุรกิจหรือเป็นหัวใจหลักของการทำธุรกิจ คือการเข้าถึงความต้องการของลูกค้าให้มากที่สุด และต้องมีการบริหารจัดการองค์กร แผนงานทางธุรกิจให้สอดคล้องกับความต้องการของตลาด พร้อมทั้งสามารถเข้าถึงลูกค้าทุกกลุ่มความต้องการได้อย่างมีประสิทธิภาพ ซึ่งสิ่งนั้นถือได้บริษัท กลุ่มธุรกิจ องค์กรนั้น ประสบความสำเร็จ ในการถือครองใจลูกค้า ซึ่งจะนำไปสู่ความสำเร็จตามความต้องการสูงสุดขององค์กรได้อย่างไม่ยากนัก ซึ่งธุรกิจการสื่อสารและโทรคมนาคม จัดได้ว่ามีการแข่งขันสูงมากในปัจจุบัน ดังนั้นการ ศึกษา เราได้นำดิซิซั่นทรี (Decision Tree) จัดได้ว่าเป็นเครื่องมือในการทำดาต้าไมนิ่ง (Data Mining) เครื่องมือหนึ่ง ที่มีลักษณะการทำงานในรูปแบบ Predictive Model มาใช้ใน การศึกษาถึงพฤติกรรมการใช้บริการของลูกค้า เพื่อสามารถนำข้อมูลที่ได้ จากการทำเซกเมนต์ (Segment) ไปวิเคราะห์ เพื่อเข้าถึงกลุ่มลูกค้าได้อย่างมีประสิทธิภาพมากยิ่งขึ้น

1.2 วัตถุประสงค์

การนำเทคนิคของ ดาต้าไมนิ่ง มาใช้ในการวิเคราะห์ลักษณะการใช้งานโทรศัพท์พื้นฐานของลูกค้าโดย วัตถุประสงค์ที่สำคัญดังนี้ คือ

"ศึกษาถึงลักษณะการใช้งานของลูกค้า โดยวัดจากโปรโมชั่นที่ลูกค้าเลือกใช้ต่อหนึ่งผลิตภัณฑ์ เพื่อคาดเดาความต้องการของลูกค้าในอนาคต"

ซึ่งผลจากการวิเคราะห์จะช่วยให้องค์กรหรือกลุ่มธุรกิจสามารถสร้างโอกาสทางการตลาดได้สูงขึ้น

1.3 ขอบเขตการดำเนินการ

โครงการพัฒนาระบบงานนี้ เป็นการนำข้อมูลการใช้งานของลูกค้าในระบบโทรศัพท์พื้นฐานมาวิเคราะห์โดยผ่านขั้นตอนต่างๆ ของ ดาต้าไมนิ่ง โดยเริ่มจากการคัดเลือกข้อมูล การเตรียมข้อมูลก่อนการประมวลผล และการแปลงข้อมูล เพื่อนำมาสร้างแบบจำลอง โดยอาศัยหลักการทำงาน ของ C4.5 อัลกอริทึม ในรูปแบบของดิซิซั่นทรี (Decision Tree) เพื่อช่วยใน การวิเคราะห์จะช่วยให้องค์กรหรือกลุ่มธุรกิจสามารถสร้างโอกาสทางการตลาดได้สูงขึ้น

1.4 ขั้นตอนการดำเนินการ

ขั้นตอนการดำเนินการแบ่ง ออก เป็น 5 ขั้นตอน

1. ศึกษาและรวบรวมข้อมูล
2. วิเคราะห์กลุ่มข้อมูลที่เก็บรวมรวม เพื่อประโยชน์ในการศึกษา
3. ศึกษาถึงทฤษฎีและขั้นตอน วิธีการทาง ดาต้าไมนิ่ง และ อัลกอริทึม C4.5 เพื่อนำมาเป็นแนวทางในการประยุกต์ใช้กับระบบที่จะพัฒนาขึ้น
4. ออกแบบและพัฒนาระบบเพื่อนำมาใช้ได้ตรงตรมวัตถุประสงค์ที่ตั้งไว้
5. ตรวจสอบและวิเคราะห์ผลการทำงานของระบบ และสรุปผลที่ได้จากการศึกษา

1.5 ประโยชน์ที่คาดว่าจะได้รับ

จากการนำเทคนิคของดาต้าไมนิ่ง มาประยุกต์ใช้การการวิเคราะห์กลุ่มลูกค้าในโครงการนี้ สิงที่คาดว่าจะได้รับคือการเข้าใจถึงหลักการพื้นฐานในการวิเคราะห์ ดาต้าไมนิ่ง โดยใช้ ดิซิซันทรี และยังสามารถนำไปประยุกต์ใช้กับการนำข้อมูลไปวิเคราะห์กับข้อมูลที่จำเป็นในธุรกิจอื่น เพิ่มประสิทธิภาพในการออกแบบและพัฒนาและวิเคราะห์ที่จะเกิดขึ้นต่อไปในอนาคต

บทที่ 2
การแบ่งกลุ่มทางการตลาด ดาต้าไมนิ่ง และทฤษฎีที่เกี่ยวข้อง

2.1 การแบ่งกลุ่มทางการตลาด (Market Segmentation)

การแบ่งกลุ่มทางการตลาด (Market Segmentation) คือการแบ่งตลาดออกเป็นส่วนต่างๆ 100% โดยมีตัวแปรตัวใดอย่างหนึ่ง เป็นเกณฑ์ในการแบ่งตลาดออกเป็นส่วนๆ เพื่อจำได้รู้ว่าส่วนใดของตลาดมีลักษณะอย่างไร และควรลงทุนในส่วนใดของตลาด

2.1.1 ลักษณะของตลาด

ลักษณะของงตลาด สามารถแบ่งได้ ดังนี้(Vriens. 2001)

1. Mass Marketing คือ ผลิตภัณฑ์ 1 ตัว สามารถรองรับตลาดได้ทั้ง 100 % เนื่องจากมีผู้ใช้มาก ข้อดี คือ ต้นทุนในการผลิตต่ำ ผลิตตัวเดียวสามารถขายได้ทั้งตลาด ไม่มีการแบ่งกลุ่มทางการตลาด
2. Product-Varity Marketing คือ ผลิตภัณฑ์ 1 ชิ้น ในสถานการณ์ที่ต่างกันมีความต้องการที่แตกต่างกันออกไป เช่น โค้ก ในเวลาปกติ คนทั่วไปจะเลือกผลิตภัณฑ์ที่บรรจุขวดธรรมดา ในเวลาที่ออกเดินทาง คนทั่วไปมักจะเลือกผลิตภัณฑ์ที่บรรจุกระป๋องแทน
3. Target Marketing คือ การแบ่งตลาดเป็นส่วน และเลือกทำตลาดตามกลุ่มเป้าหมายที่เป็น Target

2.1.2 การแบ่งตลาด

การแบ่งตลาดเป็นส่วน สามารถแบ่งออกได้หลายกลุ่ม หลายประเภท แล้วแต่เกณฑ์ในการจัดแบ่ง (Vriens. 2001)

- แบ่งตามภูมิศาสตร์ (Geographic Segmentation) แบ่งตามขอบเขตของพื้นที่
- แบ่งตามประชากรศาสตร์ (Demographic Segmentation) ซึ่งได้รับความนิยมมากที่สุดในบรรดาตัวแปรที่ใช้ เพราะปกติจะสามารถหาจำนานหรือขนาดของกลุ่มหรือกำลังการซื้อของกลุ่มได้ง่าย เช่น อายุ การศึกษา รายได้ เพศ เป็นต้น

- แบ่งตามจิตวิทยา (Psychographic Segmentation) โดยทั่วไปไม่เป็นที่นิยม เป็นเพียงทฤษฎี ไม่สามารถบอกจำนวนกลุ่มได้แน่ชัด
- แบ่งตามพฤติกรรมที่มีต่อผลิตภัณฑ์ (Behavior Segmentation) ตัวแปรที่ใช้แบ่ง เช่น
    - Occasion คือ โอกาสที่จะซื้อ
    - Benefit
    - Loyalty Status ความซื่อสัตย์ต่อยี่ห้อ มากน้อยเพียงใด
    - User Rate ช่วงเวลาที่ใช้ ปริมาณการใช้ เช่น ปานกลาง น้อย หรือ มาก
    - User Status มีการแบ่งเป็น 4 กลุ่ม
        - Non User ไม่เคยใช้
        - X User เคยใช้แต่เลิกใช้แล้ว
        - Potential User มีโอกาสที่จะใช้
        - User ใช้เป็นปกติ

จะเห็นว่าตัวแปรต่างๆ ที่ใช้แบ่งมีจำนวนมาก ซึ่งไม่สามารถระบุได้ว่าวิธีไหนดีที่สุด ซึ่งวิธีการที่ดูว่าเหมาะกับผลิตภัณฑ์ที่ต้องการผลิตหรือไม่ สามารถวัดได้จาก เกณฑ์ ดังต่อไปนี้ (Vriens. 2001)

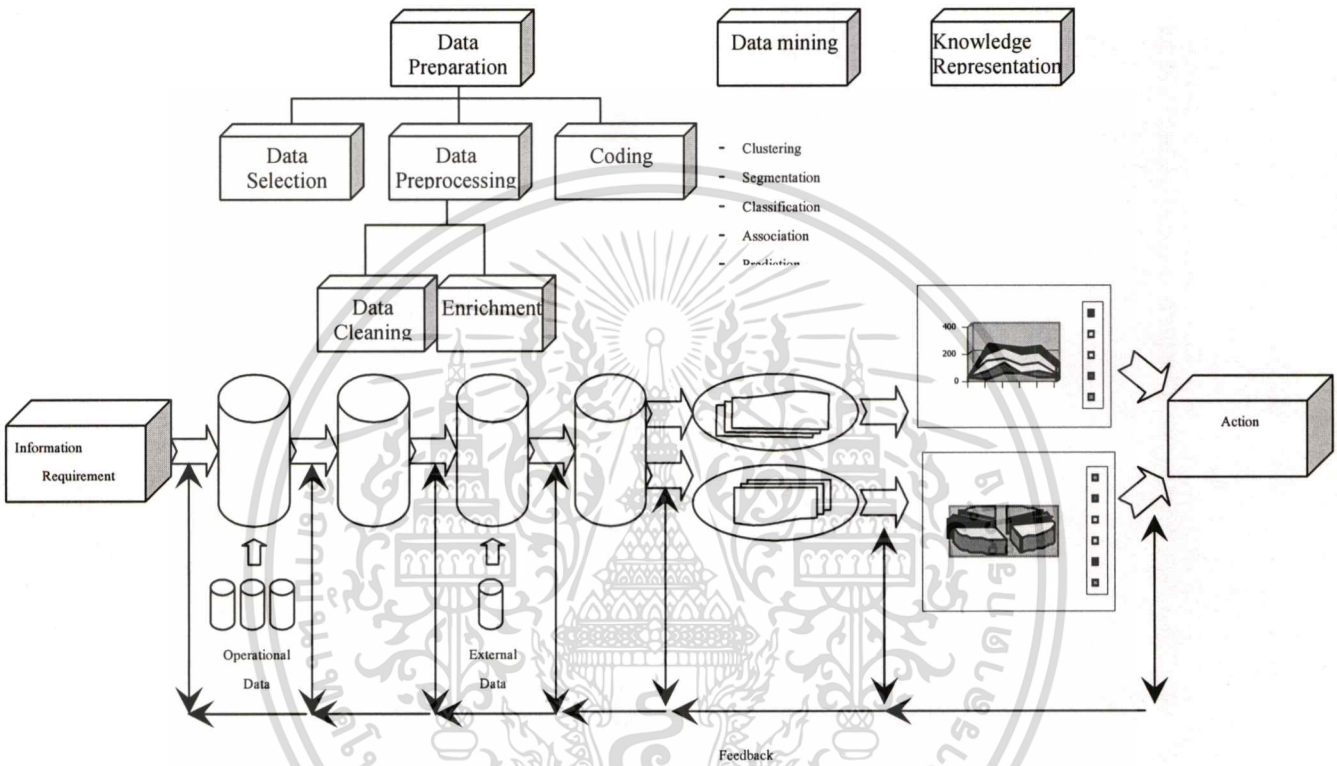
- **Identifiably** คือ สามารถบอกเขตเขตส่วนแบ่งทางการตลาด หรือกลุ่มของลูกค้าได้ และยังสามารถบอก ความแตกต่างระหว่างกลุ่มหรือส่วนแบ่งทางการตลาดได้
- **Substantiality** คือ สามารถประเมินขนาดของกลุ่มตัวอย่างได้ ว่ามีขนาดใหญ่พอที่ทำกำไรได้
- **Accessibility** คือ สามารถบอกได้ถึงขอบเขตความต้องการ ความชอบ ของกลุ่มลูกค้า เพื่อที่จะทำการโฆษณานำเสนอ ได้อย่างถูกที่
- **Responsive** คือ สามารถบอกความต้องการที่แตกต่างกัน ของแต่กลุ่มทางการตลาด
- **Stability** ในที่นี้คือ ระยะเวลาที่คงอยู่ของกลุ่มทางการตลาดที่เราจัดแบ่งขึ้น
- **Action ability** คือ การเข้าถึงความต้องการของกลุ่มเป้าหมาย และสามารถดึงดูดกลุ่มเป้าหมาย จนประสบความสำเร็จ

2.2 ดาต้าไมนิ่ง

Data Mining เป็นวิธีการที่ใช้ในการดึงเอาข้อมูลที่มีความสำคัญ หรือวิเคราะห์ข้อมูลที่เราสนใจหรือให้ความสำคัญจากกลุ่มข้อมูลที่ขนาดใหญ่ เพื่อหาข้อสารสนเทศที่สำคัญออกมา เพื่อนำ

ข้อมูลที่ได้เหล่านั้นไปเป็นองค์ประกอบที่สำคัญที่ช่วยในการตัดสินใจ ซึ่งจะก่อให้เกิดผลดีในเชิงธุรกิจต่อไป (Groth. 1997)

2.2.1 กระบวนการของดาต้าไมนิ่ง



รูปที่ 2.1 Data Mining process

2.2.1.1 การกำหนดวัตถุประสงค์ทางธุรกิจ(Business Objective Determination) ทำความเข้าใจกับข้อมูล และความต้องการทางธุรกิจขององค์กรก่อน เพื่อผลลัพธ์ที่ออกมานั้นสามารถตอบปัญหา หรือความต้องการที่ถูกต้อง และเพื่อให้การทำดาต้าไมนิ่งเกิดประโยชน์สูงสุด

2.2.1.2 การเตรียมข้อมูล (Data Preparation) ซึ่งถือได้ว่าเป็นส่วนสำคัญมาก ในระดับหนึ่งซึ่งต้องใช้เวลาและกระบวนต่างๆ เพื่อให้ได้ข้อมูลมาทำการวิเคราะห์ แบ่งออก เป็นขั้นตอนย่อยได้ถึง 3 ขั้นตอน

2.2.1.2.1 Data selection การระบุหรือเลือกแหล่งข้อมูลที่มีอยู่ การแบ่งข้อมูลออกเป็นกลุ่มย่อยๆ เฉพาะที่มีความจำเป็นต่อการทำ Mining เท่านั้น และยังการพิจารณาไปถึงชนิดรูปแบบของข้อมูล ซึ่งเราอาจจำแนกออกเป็น 2 กลุ่ม ใหญ่ ได้ดังนี้

- ❖ Categorical Data คือ กลุ่มหรือข้อมูลที่สามารถ บอกถึงระดับความสำคัญของสิ่งนั้นได้ชัดเจน แบ่งได้ออกเป็น 2 ประเภทย่อย คือ nominal และ ordinal
  - Nominal Categorical ค่าที่เป็นลำดับไม่มีความสำคัญ เช่น เพศ
  - Ordinal Categorical ค่าที่เป็นลำดับมีความสำคัญ เช่น ดี,ดีมาก,ปานกลาง ถ้าแปลงเป็นตัวเลขความหมายยังเหมือนเดิม
- ❖ Quantitative บอกถึงปริมาณ ค่าความแตกต่างที่สามารถเป็นไปได้
  - Discrete เป็นค่าที่เป็นจำนวนเต็ม เช่นจำนวนพนักงาน
  - Continuous เป็นค่าที่เป็นจำนวนจริง เช่น รายได้

สิ่งสำคัญอีกอย่างคือ อายุการใช้งานของข้อมูลที่เลือก เช่น ข้อมูลมีอัตราการเปลี่ยนแปลงสูง (Variation) ต้องทำการตรวจสอบก่อน ที่จะนำไปใช้งานจริง หลักเกณฑ์ที่ต้องพิจารณาเพิ่มเติมเกี่ยวกับข้อมูล มี 4 ประเด็น

- ❖ ระดับของข้อมูลที่พิจารณา สิ่งที่นำมาช่วยตัดสินใจว่าข้อมูลที่นำมาใช้ควรจะเป็นข้อมูลระดับรายการ(Item) หรือ ข้อมูลที่สรุปแล้ว คือ วัตถุประสงค์ในการทำดาต้าไมนิ่ง
- ❖ ลักษณะของข้อมูลที่จัดเก็บ
- ❖ ข้อมูลที่เป็นข้อความ ข้อมูลที่จัดเก็บแบบไฟล์ อาจก่อนให้เกิดความสับสน เช่น “20040120” กับ “20/042004” ซึ่งอาจทำให้ระบบมองข้อมูลที่แตกต่างกันออกไป ดังนั้นควรนำข้อมูลมาจุดเก็บตามค่าที่ถูกต้องในรูปแบบเดียวกัน
- ❖ ความแตกต่างของแหล่งข้อมูลแต่ละแหล่ง

2.2.1.2.2 Data Preprocessing การทำ Data cleaning คือ การตรวจสอบข้อมูลที่มีการคัดเลือกมาว่า เป็นข้อมูลที่มีความเหมาะสมหรือไม่ โดยต้องแปลงข้อมูลที่มีให้มีค่าทางสถิติ เช่น พวกข้อมูล แบบ categorical เป็นต้น หรือ อาจทำการวัดการกระจายของข้อมูล เพื่อให้เข้าใจข้อมูลที่มีอยู่มากยิ่งขึ้น ทำให้สามารถหาแนวโน้มของข้อมูลที่อาจเกิดขึ้นได้ ส่วนข้อมูลที่เป็นแบบ Quantitative อาจข้อมูลที่ได้จากการวิเคราะห์ข้อมูลโดยการหา ค่าสูงสุด ต่ำสุด ค่าเฉลี่ยได้เป็นต้น ซึ่งสิ่งที่อธิบายในส่วนนี้สามารถมาแก้ปัญหาเหล่านี้

- ❖ Noisy data คือ ข้อมูลมีความแตกต่างจากกลุ่มออกไป อาจจะเกิดจากการป้อนข้อมูลผิดพลาด วิธีการแก้ปัญหา อาจจะใช้ วิธีการหาค่าเฉลี่ย ที่ได้มาใช้แทน เรียกว่าการทำ Binding Method

- ❖ Missing data คือ ข้อมูลที่ค่านั้นหายไป กรณีที่ค่าน้อยมากอาจจะตัดทิ้ง หรืออาจใช้ค่าเฉลี่ยแทนค่าที่หายไป
- ❖ Outlier คือ ข้อมูลที่มีอาจจะผิดพลาดที่เกิดจากการบันทึกหรือ จากการจัดเก็บข้อมูลไม่ดี พิจารณาและแก้ไขให้ถูกต้อง

หลังจากนั้นนำข้อมูลที่ได้มาทำ Data Integration or Enrichment คือ ข้อมูลอาจมีการซ้ำซ้อนของข้อมูลเนื่องมาจากถูกส่งมาหลายแหล่ง เช่น รูปแบบของการเก็บข้อมูลไม่เหมือนกัน, มีการพิมพ์ไม่เหมือนกันแต่เป็นข้อมูลเดียวกัน เป็นต้น

2.2.1.2.3 Data Transformation (Coding) การแปลงข้อมูลนี้มี วัตถุประสงค์ เพื่อปรับรูปแบบข้อมูลให้เหมาะสม ตามแบบของ Algorithm ของ Data Mining ที่เลือกใช้ โดยมีเทคนิคการแปลงข้อมูลได้หลายรูปแบบ ได้แก่

- ❖ Discrimination คือ การแปลงข้อมูลที่เป็นแบบ Quantitative ให้เป็นแบบ Categorical
- ❖ One of N coding คือ การแปลงข้อมูลที่เป็นแบบ Categorical ให้อยู่ในรูปตัวเลข ซึ่งจะใช้ในกรณีที่ ข้อมูลมีลำดับของข้อมูลไม่มีความสำคัญ เช่น ชาย/หญิง แปลเป็น 0/1 เป็นต้น

2.2.1.3 Data Mining จัดได้เป็นกระบวนการสำคัญ ในการเลือกเทคนิคการสร้าง Model และนำมาประยุกต์ใช้ ซึ่งกระบวนการนี้ที่อาจจะต้องกลับมาทำซ้ำในขั้นนตอนที่ทำไปแล้ว (Iterative Process) จะต้องเลือก Operation ที่เหมาะสม เพื่อจะได้นำมาใช้ให้เกิดประโยชน์สูงสุดต่อองค์กรที่นำไปใช้ คาต้าไมนิ่ง มี 2 ประเภท คือ การหาลักษณะของข้อมูล (Description) และการทำนายค่าในอนาคต (Prediction) โดย Description จะหารูปแบบ(Pattern) เพื่ออธิบายข้อมูล ในรูปแบบที่บุคคลากรเข้าใจได้ง่าย แต่ Prediction จะเป็นการทำนายค่าในอนาคตหรือค่าที่เราไม่รู้ของ Attribute ที่สนใจ โดยใช้ข้อมูลที่เรามีอยู่ Data Mining มีเครื่องมือที่ใช้ในการค้นหาข้อมูลในการทำ Market Segmentation 2 model ใหญ่ๆ ด้วยกันคือ

- ❖ Descriptive Model เป็นการนำเอาข้อมูลที่ถูกจัดเก็บในฐานข้อมูล มาหาความสัมพันธ์เพื่อหาข้อสารสนเทศที่สำคัญออกมา โดยการทำ Clustering ซึ่ง Data mining มีเครื่องมือ(Method) และ Algorithm จำนวนมาก ยกตัวอย่าง เช่น

- K-means จัดได้ว่าเป็น Methods ที่มีความสำคัญมากอย่างหนึ่งในการทำ Clustering เนื่องจากง่ายต่อการใช้งาน โดย k จะเป็นพารามิเตอร์ที่ส่งเข้าไปเป็นข้อมูล Input และ เซตของข้อมูล n กลุ่มที่ถูกจัดแบ่งใน k cluster ลักษณะผลลัพธ์ที่ได้ แบ่งออกเป็น 2 ลักษณะดังนี้ กลุ่มข้อมูลที่ได้ Intra cluster ได้ผลดีมีประสิทธิภาพ ส่วนกลุ่มข้อมูล Inter cluster ได้ผลที่มีความเหมือนกันค่อนข้างน้อย

❖ Predictive Model เป็นเครื่องมือหนึ่ง ใน Data Mining โดยลักษณะพิเศษของ Predictive Model คือเรารู้ว่าเราต้องการอะไร เช่น ต้องการแบ่งกลุ่มข้อมูลแบบไหน เช่นเราต้องการ แบ่งกลุ่มข้อมูลออกเป็น 3 กลุ่ม คือ ทำกำไรมากกว่า 50 เปอร์เซ็นต์ ทำกำไรมากกว่า 10 เปอร์เซ็นต์ แต่ไม่มากกว่า 50 เปอร์เซ็นต์ และกลุ่มสุดท้ายคือทำกำไร ต่ำกว่า 10 เปอร์เซ็นต์ โดยเรานำค่าเหล่านี้ไปกำหนดใน Predictive variable เพื่อที่จัดให้ Methods นั้นทำการจัดแบ่งกลุ่มข้อมูลออกมาตามที่เราต้องการ จากนั้นจึงส่งข้อมูลที่มีอยู่เข้าไป Predictive Model จะนำข้อมูลเหล่านั้นมามาประมวลผล และแสดงผลลัพธ์ต่างตามที่ข้อกำหนดที่เราต้องการออกมา ซึ่งการประมวลผลที่ว่าเหล่านี้ มีเครื่องมือหรือ Algorithm ที่น่าสนใจ หลายกลุ่มด้วยกัน ซึ่งจะอธิบาย Method ที่น่าสนใจดังต่อไปนี้

- Neural networks จัดได้ว่าเป็น Predictive Model แบบหนึ่ง ที่จัดได้ว่ามีประสิทธิภาพมากกับข้อมูลที่มีขนาดใหญ่ และสามารถกำหนดค่า Predictive variables ได้เป็นจำนวนมากกว่า ร้อยค่า และยังสามารถคาดเดา กรณีใหม่หรือแนวทางทางด้านข้อมูลใหม่ได้อย่างรวดเร็ว แต่อย่างไรก็ตาม ข้อมูลที่นำเข้ามาประมวลผลเป็นตัวเลขประเภท Numerical ที่มีค่าระหว่าง 0 กับ 1 ได้ดีมากกว่ากลุ่มหรือประเภทข้อมูลที่แบ่งข้อมูลตามลำดับ เช่น หญิงชาย, โสดหรือแต่งงาน เป็นต้น
- Decision Trees เป็นเทคนิคหนึ่งของการใช้ Classification เพื่อทำการคาดเดาข้อมูลที่จะเกิดขึ้นในอนาคต และที่เราเรียกว่า Decision Tree เนื่องจากการทำงานมีลักษณะ คล้ายกับ Tree Structure โดยเราจะเริ่มมองจาก root node แล้วท่องไปใน Internal Node ไปยัง

Leaf node เพื่อหาข้อสรุปและผลลัพธ์ที่ควรจะได้ ซึ่งนับจาก root Node ถึง Leaf node จะมี decision อยู่ภายใน เทคนิคนี้จะใช้ดีกับข้อมูลที่แบ่งระดับ หรือวัดระดับความพึงพอใจที่เรียกว่า Categorical แต่ใช้ไม่ได้กับพวกข้อมูล Numerical

2.2.1.4 Knowledge Representation เป็นขั้นตอนที่ นักวิเคราะห์ข้อมูล และนักวิเคราะห์นำผลที่ได้จาก 2.2.1.4 มาแปลความหมาย และประเมินค่าผลที่ได้ เพื่อนำสารสนเทศที่ถูกต้อง มาผสมผสานกับประสบการณ์ของนักวิเคราะห์เพื่อหาวิธี ในการสร้างมูลค่าของข้อมูลเพื่อตอบสนองต่อความต้องการหรือวัตถุประสงค์ที่องค์กรต้องการค้นหา

2.3 Predictive Model กับการทำ Market Segmentation

Predictive Model มีความสามารถในการศึกษาและทำนายลักษณะของสิ่งที่กำลังศึกษาได้ และเมื่อนำมาใช้เพื่อการแบ่งกลุ่มทางการตลาด Predictive Model จึงมีส่วนช่วยในการทำนายลักษณะของกลุ่ม Segment นั้น ทำให้เข้าใจหรือคาดเดาลักษณะของกลุ่มลูกค้าได้อย่างมีประสิทธิภาพมากยิ่งขึ้น(Vriens. 2001)

บทที่ 3
การจัดกลุ่ม(Classification)

การจัดหมวดหมู่ของข้อมูล เป็นการแบ่งประเภทของข้อมูลแต่ละหน่วยออกตาม Class ที่ได้กำหนดไว้ การทำ Classification คือ การวิเคราะห์ Training data เพื่อสร้างเป็น Model ของแต่ละ Class ขึ้นมา เพื่อแสดงถึงลักษณะของข้อมูล และข้อมูลที่จะนำมาใช้ในอนาคต จะถูกแบ่งตาม Model ของ Class ที่มีไว้

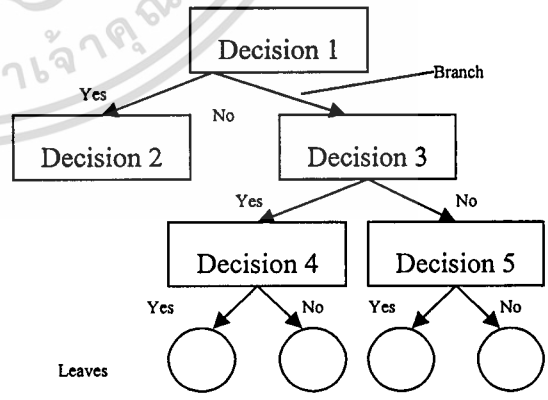
โดยทั่วไปการจัดหมวดหมู่สามารถแบ่งได้เป็น 2 แบบ คือ Tree Induction และ Neural Induction ซึ่งในที่นี้จะเลือกใช้วิธีที่เลือกว่า Decision Tree ซึ่งเป็นรูปแบบหนึ่งการการทำ Tree Induction

3.1 Decision Tree

Decision Tree เป็นรูปแบบการคัดเลือกจัดแบ่งกลุ่มข้อมูลออกเป็นกลุ่มย่อยๆ โดยใช้เงื่อนไขหรือข้อกำหนดที่เราต้องการจะวิเคราะห์ มาเป็นหลักเกณฑ์ช่วยในการตัดสินใจในการแบ่งกลุ่มข้อมูล (ลักษณะการมองย้อนกลับแบบ Agglomerative) เราสามารถจำแนกส่วนประกอบของ Decision Tree ออกเป็น 3 ส่วนหลัก ๆ ด้วยกัน ดังนี้ (Witten. 1999)

- Root Node ซึ่งเป็น Node บนสุดของ Tree
- Child Node ซึ่งเป็น Node ลูกของ Tree หรือในส่วนที่เรียกกันว่า Branch จะบอกถึงผลการของการทดสอบ

Leaf Node ซึ่งเป็น Node ที่อยู่ระดับล่างสุด ซึ่งจะแสดงถึง Class ที่ได้ออกมา



รูปที่ 3.1 Decision Tree

3.2 ขั้นตอนพื้นฐานในการสร้าง Tree

- หา Attribute ที่สำคัญที่สุดมาแบ่งข้อมูลโดย Attribute นี้จะถูกสร้างเป็น Root Node
- นำค่าที่เป็นไปได้ใน Attribute ที่ถูกเลือกแตกออกมาเป็นกลุ่ม
- แบ่งข้อมูลทั้งหมดตามกลุ่มที่แตกออกจาก Root Node
- นำข้อมูลแต่ละกลุ่มมาทำซ้ำขั้นตอนแลก คือ หา Attribute ที่สำคัญที่สุด

สำหรับ Algorithm ที่ใช้สร้าง Decision Tree มีหลาย Algorithm อาทิเช่น Inferring rudimentary rules, Statistical modeling เป็นต้น แต่ในการศึกษาครั้งนี้ขอกล่าวถึง C4.5 Algorithm (Witten. 1999)

3.3 C4.5 Algorithm

C4.5 Algorithm เป็น Algorithm ที่พัฒนามาจาก ซึ่งพัฒนาจาก ID3 และ Divide and conquer Algorithm เนื่องจาก ID3 ไม่สามารถจัดการกับค่าที่มีชนิด Continue ได้ C4.5 Algorithm ใช้เป็นเทคนิคหนึ่งที่เรียกว่า Tree induction (top-down induction of decision trees) เป็นการแบ่ง Segment ที่จะนำข้อมูลมาแตกแบบบนลงล่าง และยังมีการจัดหมวดหมู่หรือ Segment อีกแบบ ที่เรียกว่า Neural Induction แต่จะไม่ได้กล่าวถึงในที่นี้ (Witten. 1999)

C4.5 Algorithm มีการสร้าง Tree ตาม Algorithm ดังนี้

- Tree เริ่มต้นด้วย Node หนึ่งแสดงถึง ข้อมูลที่ใช้ Train ในที่นี้ขอเรียกว่า T1
- หากพบว่า T1 เป็น class เดียวกับ Node ให้กำหนด Node นี้เป็น Leaf Node โดยให้มีชื่อตามชื่อ Class
- หากไม่พบ T1 ใน Class ให้ทำการแตก T1 ออกเป็น Class ซึ่ง Attribute จะกลายเป็น Attribute ที่ใช้ในการทดสอบหรือใช้เพื่อการตัดสินใจที่ Node ในแต่ละกิ่ง (Branch) ของ Tree ซึ่ง T1 จะถูกแบ่งไปตามค่า Attribute และจะมีการทำซ้ำ ณ กระบวนเดิม และจะหยุดก็ต่อเมื่อ
  - T1 ทั้งหมดอยู่ใน Class เดียวกัน
  - T1 ไม่มี Attribute เหลืออยู่อีก
  - ไม่มี T1 เหลืออยู่แล้ว

ใช้การวัดแบบ Entropy-based หรือ ที่รู้จักกันดีคือ Information gain ซึ่งจะเป็นการเลือก Attribute ที่เหมาะที่สุดและถือว่าเป็นสิ่งที่สำคัญที่สุดเพื่อใช้ในการแบ่งข้อมูล โดยเราจะกำหนดให้

T แทน Training Set

S แทน Set ของข้อมูลใด ๆ

p แทน ความน่าจะเป็นที่ T1 นั้นจะเป็นของ Class C_i

m แทน จำนวน Class

สูตร Info(S)

$$\text{Info}(S) = \left|-\sum_{I=0}^{m} p_i \log(p_i)\right|$$

และเมื่อนำสูตรมาประยุกต์ใช้กับการ Training Set จะได้ Info(T), $\text{Info}_x(T)$ เป็นการวัดค่าของ Information เพื่อแบ่ง T โดยคิดจากค่าที่เป็นไปได้ของ Attribute X

สูตร Info(T)

$$\text{Info}_x(T) = \sum_{J=1}^{n} \frac{|T_i|}{|T|} * \text{Info}(T_i)$$

Gain(X) เป็นการวัดค่าของ Information ที่ได้รับเลือก Attribute X

สูตร Gain(X)

$$\text{Gain}(X) = \text{Info}(X) - \text{Info}_x(T)$$

❖ กรณีที่ Attribute มีค่าเป็น unique คือ การแบ่งข้อมูลโดยใช้ Attribute แล้วเกิดพบว่า มี Subset จำนวนมาก แต่ละ Subset มีข้อมูลเพียง 1 record เท่านั้น ทำให้ $\text{Info}_x(T)=0$ ส่งผลให้ค่า Information Gain ของ Attribute มีค่าสูงมาก และการแบ่งข้อมูลโดยใช้ Attribute นี้ไม่ก่อให้เกิดประโยชน์ใด ๆ นั้น สามารถแก้ไขได้โดยการวัดค่า Gain ratio ซึ่งเป็นค่าในการวัดว่าการแบ่งข้อมูลโดยใช้ Attribute นั้น ๆ ก่อให้เกิดประโยชน์ต่อการทำนายหรือไม่ ซึ่ง Gain ratio criterion จะทำให้ Tree ได้มีขนาดเล็กกว่าการใช้ Gain Criterion ซึ่งค่า Gain ratio(X) สามารถคำนวณได้จากการนำ gain (X) และ Split info(X) (เป็นค่า Information ที่ได้จากการแบ่ง T ออกเป็น n Subset)

สูตร Split info(x)

$$\text{Split Info}(X) = \sum_{J=1}^{n} \frac{|T_i|}{|T|} \log_2 \frac{|T_i|}{|T|}$$

สูตร gain ratio(X)

$$\text{gain ratio}(X) = \text{gain}(X)/\text{split info}(X)$$

ตัวอย่าง 1 การข้อมูลรายละเอียดของอากาศที่มีผลต่อการเล่นกอล์ฟ โดย Attribute Play ถือว่าเป็น categorical เฉพาะเกี่ยวกับการตัดสินใจ เล่นหรือไม่ ส่วนข้อมูล non-categorical attributes มีดังนี้ (Witten. 1999)

ตารางที่ 3.1 non-categorical attributes

ATTRIBUTE	POSSIBLE VALUES
outlook	sunny, overcast, rainy
temperature	continuous
humidity	continuous
windy	true, false

ตารางที่ 3.2 Weather data with some numeric attributes

OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
sunny	85	85	false	Don't Play
sunny	80	90	true	Don't Play
overcast	83	78	false	Play
rain	70	96	false	Play
rain	68	80	false	Play
rain	65	70	true	Don't Play
overcast	64	65	true	Play
sunny	72	95	false	Don't Play
sunny	69	70	false	Play
rain	75	80	false	Play
sunny	75	70	true	Play
overcast	72	90	true	Play
overcast	81	75	false	Play
rain	71	80	true	Don't Play

สามารถสรุปออกมาเป็นตารางได้ดังนี้

ตารางที่ 3.3 ตารางการสรุปการเล่นหรือไม่เล่น โดยใช้ outlook เป็นเกณฑ์

outlook	Play	Don't Play	Total
sunny	2	3	5

Overcast	4	0	4
Rain	3	2	5
Total	9	5	14

หาค่า information ต่างๆ ดังนี้

Sunny info([2,3]) $=-((\log_2(2/5)*2/5)+(\log_2(3/5)*3/5))$

$=0.971$

Overcast info([4,0]) $=-(\log_2(4/4)*4/4)$

$=0$

Rain info([3,2]) $=-((\log_2(3/5)*3/5)+(\log_2(2/5)*2/5))$

$=0.971$

$\text{Info}_x(T)$ $= \text{info}([2,3],[4,0],[3,2])$

$= (5/14)* 0.971+(4/14)*0+(5/14)* 0.971$

$=0.693$

info([9,5]) $=-((\log (9/14)*9/14)+(\log_2(5/14)*5/14))$

$=0.940$

Gain(outlook) $=\text{info}([9,5])- \text{Info}(T)$

$= 0.940-0.693$

$=0.247$

หาค่า Split Info

Split Info (Outlook,T) $= -5/14*\log(5/14) - 4/14*\log(4/14) - 5/14*\log(5/14)$

$= 1.577$

GainRatio of Outlook $= 0.246/1.577$

$= 0.156$

SplitInfo(Windy,T) $= -6/14*\log(6/14) - 8/14*\log(8/14)$

$= 6/14*0.1.222 + 8/14*0.807 = 0.985$

GainRatio of Windy $= 0.048/0.985$

$= 0.049$

จะพบว่าค่า Gain ที่จะได้จากการแบ่ง Training set โดยใช้ Attributes Outlook มากว่า Windy ดังนั้น ควรใช้ Attributes Outlook ในการแบ่ง Training set แล้วนำ Sample ในแต่ละกิ่งของ Attribute ที่ใช้ทดสอบมาทำซ้ำขั้นตอนตั้งแต่ต้น คือหา Attribute ที่มีค่าสูงสุดมาแบ่งข้อมูลต่อไป

การหา Attribute เพื่อใช้แบ่งข้อมูล ในกรณีที่ (Unknown attribute value)

- ค่า info(T) และ $info_x(T)$ โดยพิจารณาเฉพาะข้อมูลที่รู้ค่าของ A
- หาค่า gain(X) โดย

สูตร gain (X) (Lan H.Witten. 1999)

Gain(X)=probability A is Know *(info(T)- $info_x(T)$)

❖ หาค่า split info(x) โดยพิจารณาจากกลุ่มข้อมูลที่ไม่รู้ค่า A เป็นอีก 1 Subset คือ ถ้า Attribute ที่จะนำมาทดสอบมีค่าที่เป็นไปได้ n ค่า split info(X) จะถูกคำนวณโดยการแบ่งข้อมูลออกเป็น n+1 subsets

การแบ่ง Training Set สมมุติ Attribute ที่เลือกจากขั้นตอนแรกมีค่าเป็นไปได้ คือ $O_1, O_2, \ldots\ldots, O_n$ เมื่อข้อมูล 1 record ใน T ซึ่งมีค่า O_i ถูกกำหนดให้ Subset T_i ค่าความน่าจะเป็นที่ข้อมูลนี้อยู่ใน Subset T_i เท่ากับ 1 และความน่าจะเป็นที่ข้อมูลอยู่ใน Subset อื่น ๆ เท่ากับ 0 แต่ค่าใน Attribute ไม่ทราบค่า จะมีความน่าจะเป็นน้อยลง สำหรับข้อมูลในแต่ละ record ในแต่ละ Subset T_i weight จะเท่ากับความน่าจะเป็นของ O_i ที่จุดนั้น ๆ ทำให้ $|T_i|$ เป็นผลรวมของค่า weight w ซึ่งค่าใน Attribute ไม่ทราบค่าจะถูกกำหนดให้แต่ละ Subset T_i ด้วย weight

สูตร

W * probability of Outcome O_i

ความน่าจะเป็น คือ ผลรวมของ Weight ของข้อมูลทั้งหมดใน T ซึ่งค่า Oi หารด้วยผลรวมของ Weight ของข้อมูลทั้งหมดใน T ซึ่งค่าใน Attribute เป็นค่าที่ทราบค่า

การใช้ Decision Tree มาทำนายกลุ่มของข้อมูล ในกรณีที่ค่าใน Attribute ที่จะทดสอบที่ decision node เป็นค่าที่ไม่ทราบค่า ทำให้ไม่สามารถแบ่งข้อมูลได้ โดยจะสำรวจความเป็นไปได้ของเส้นที่อาจจะเกิดขึ้น และผลรวมที่ได้จากการจัดกลุ่ม (Classification๗) ด้วยวิธีการทางคณิตศาสตร์ ผลที่เกิดจากเส้นทาง จากรูท (root) ของทรี (tree) หรือซับทรี (sub tree) มายังลีฟโหนด (leaf node) และ class ที่ได้จากการทำนายความน่าจะเป็นสูงสุด

ตัวอย่าง 2 ใช้ข้อมูลจากตัวอย่างที่ 1 โดยกำหนดให้ค่า outlook บางอย่างเป็นค่าที่ไม่ทราบ โดยแบ่งเป็น 3 ขั้นตอนดังนี้

ตารางที่ 3.4 ใช้ข้อมูลจากตัวอย่างที่ 1 โดยกำหนดให้ค่า outlook บางอย่างเป็นค่าที่ไม่ทราบ

OUTLOOK	TEMPERATURE	HUMIDITY	W INDY	PLAY	HEIGHT
sunny	85	85	false	Don't Play	1
sunny	80	90	true	Don't Play	1
overcast	83	78	false	Play	1

OUTLOOK	TEMPERATURE	HUMIDITY	W INDY	PLAY	HEIGHT
rain	65	70	true	Don't Play	1
overcast	64	65	true	Play	1
sunny	72	95	false	Don't Play	1
sunny	69	70	false	Play	1
rain	75	80	false	Play	1
sunny	75	70	true	Play	1
X	72	90	true	Play	5/13
overcast	81	75	false	Play	1
rain	71	80	true	Don't Play	1

1. การหา Attribute เพื่อใช้แบ่งข้อมูล สมมุติค่าใน Attribute outlook ที่ record ที่ 12 เป็นค่าที่ไม่ทราบค่า ซึ่งแทนด้วย X ซึ่งเราจะพิจารณาเฉพาะข้อมูล 13 record ที่เหลือจะได้ความถี่ดังแสดงในตารางที่ 4 ทำการคำนวณค่าต่างๆ โดยพิจารณา Attribute outlook ดังนี้

Sunny info([2,3]) $=-((\log_2(8/13)*8/13)+(\log_2(5/13)*5/13))$

$=0.9691$

Info $_x$(T) $=$ info([2,3],[3,0],[3,2])

$=(5/13)*0.9691+(3/13)*0+])+(5/13)*0.9691$

$=0.747$

Gain(X) $=13/14*(0.9691-0.747)$

$=0.199$

ตารางที่ 3.5 ความถี่ของข้อมูล

outlook	Play	Don't Play	Total
sunny	2	3	5
Overcast	2	0	2
Rain	3	2	5
Total	9	5	13

จะพบว่าค่า gain ที่ได้จะลดลงเล็กน้อย จากเดิม 0.247 เป็น 0.199 bits ส่วนค่า split information จะพิจารณาจากข้อมูลใน training set ทั้งหมด จึงทำให้ค่าที่ได้เพิ่มขึ้นจาก 1.577 เป็น 1.809 ดังนี้

$-5/14 * \log_2(5/14)$ (for sunny)

$-3/14 * \log_2(3/14)$ (for overcast)

$-5/14 * \log_2(5/14)$ (for rain)

$-1/14 * \log_2(1/14)$ (for X)

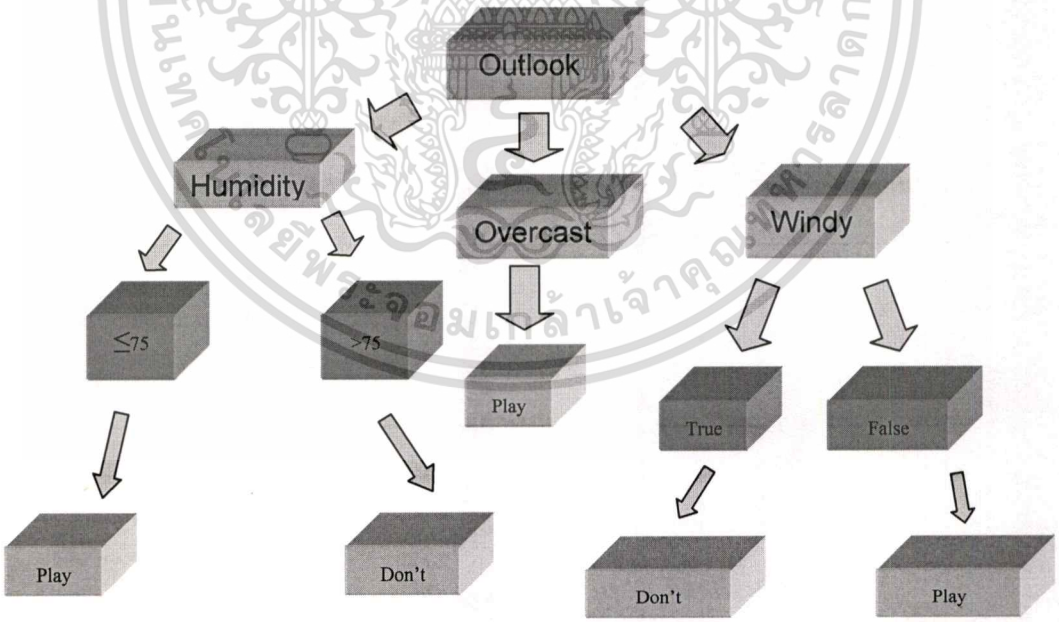
และค่า gain ratio ลดลงจาก 0.156 เป็น 0.111

2. การแบ่ง Training Set เมื่อข้อมูลใน Training set ทั้ง record ถูกแบ่งออกโดยใช้ค่าใน Attribute outlook record เป็นเกณฑ์ ซึ่งมี Attribute outlook เป็นค่าที่ไม่ทราบค่า จะถูกกำหนดให้ในทุก subset คือ sunny ,overcast และ rain ด้วยค่า weight เท่ากับ 5/13 , 3/13 และ 5/13 ตามลำดับ พิจารณา subset ดังต่อปนี้

   a. Humidity □ 75 2 class Play,0 class don't play

   b. Humidity > 75 5/13 class Play,3 class don't play

Decision Tree ที่ได้จะมีโครงสร้างดังนี้



รูปที่ 3.2 โครงสร้าง Decision Tree

โดยค่าของตัวเลขที่ leaf node จะอยู่ในรูป (N) หรือ (N/E) โดยที่ N หมายถึง จำนวนข้อมูลทั้งหมดที่มาถึง leaf node นั้นๆ และ E เป็นจำนวนข้อมูลที่อยู่ใน class ที่ระบุไว้ เช่น Don't Play

(3.4/0.4) หมายความว่า จำนวนข้อมูลที่มาถึงที่ leaf node นี้ เท่ากับ 3.4 และ 0.4 ในจำนวนนี้ไม่อยู่ใน class Don't Play

3. การใช้ Decision Tree ที่ได้ มาทำนายกลุ่มของข้อมูล สมมุติข้อมูล คือ Sunny , Outlook temperature 70°, unknown humidity ,windy false

จากค่าใน Outlook พบว่าต้อง move ไปยัง subset แรกแต่เนื่องจากไม่สามารถตรวจสอบค่า humidity ได้ จึงทำการพิจารณา ดังนี้

- ถ้า humidity □ 75% จะได้ class play มีค่าน่าจะเป็นเท่ากับ0.4/3.4 หรือ 12 %
- ถ้า humidity > 75% จำได้ class Don't Play มีค่าน่าจะเป็นเท่ากับ 3/3.4 หรือ 88%

จะเห็นได้ว่าการกระจายของ Class สุดท้ายสำหรับข้อมูลนี้ เท่ากับ

Play : 2.0/5.4*100% + 3.4/5.4 * 12% =44%

Don't Play : 3.4/5.4 * 88% =56%

Continue attribute values

การที่เราจะสรุปหรือวิเคราะห์ว่า Attribute นี้เป็น continuous value หรือไม่นั้น เราสามารถตรวจสอบได้จาก การแบ่งกลุ่มของข้อมูลออกกลุ่มย่อย อาทิเช่น สมมติให้ A เป็น Attribute ชนิด Continue value เราจะแบ่ง A ออกเป็น A <= X และ A> X โดยการเปรียบเทียบค่า ของ A กับค่า Threshold value X โดยการหาค่า Threshold ที่เหมาะสมมีขั้นตอน ดังนี้

- เรียงลำดับ Training Set ด้วยค่าใน Attribute A จากน้อยไปมาก และจะเลือกเฉพาะค่าไม่ซ้ำ กันมาพิจารณาจะได้ $\{v_1, v_2, \ldots\ldots, v_n\}$
- การหาค่า Threshold ใดๆ จะอยู่ระหว่าง v_i และ v_i+1 โดยการคำนวณเพื่อหา Mid Point ของแต่ละช่วงดังนี้ $(v_i + v_i + 1)/2$ และจะเลือกค่าที่มากที่สุดใน Attribute A แต่ต้องไม่เกินค่า Midpoint นั้นๆ จาก Training Set เป็นค่า Threshold ของแต่ละช่วง เพื่อค่า Threshold ทั้งหมดที่อยู่ใน Tree หรือ Rule จะเป็นค่าที่เกิดขึ้นจริงในข้อมูล
- เลือกค่า Threshold ที่เหมาะสม โดยสามารถพิจารณาได้จากค่า Threshold ที่มีค่า Information gain สูงสุด

Pruning decision tree

การเป็นการแบ่งข้อมูล Training set เพื่อสร้าง Decision จะทำไปเรื่อย ๆ จนกระทั่งข้อมูลในแต่ละ Subset อยู่ใน Class เดียวกัน ซึ่งผลลัพธ์ที่ได้อาจจะทำให้ Tree เกิดความซับซ้อนมากเกินไป ทำให้เกิด Over-fits the data ปัญหานี้สามารถแก้ไขได้โดยการทำ Pruning ซึ่งทำให้แต่

ละ Leaf node ที่ได้นั้นอาจจะไม่มีความจะเป็นที่ต้องประกอบด้วยข้อมูลที่อยู่ใน Class เดียวกันทั้งหมด โดยแต่ละ Leaf node จะมีการกระจายของข้อมูลแต่ละ Class ไว้ ซึ่งสิ่งนี้จะบอกถึงความน่าจะเป็นที่ข้อมูลอยู่ใน Class นั้นๆ ซึ่ง C4.5 Algorithm ทำ pruning โดยการตัด sub tree บางกลุ่มออกเมื่อพบว่า Sub tree ทำให้เกิดข้อผิดพลาดในการทำนายออกมา แล้วทำการแทนที่ sub tree นั้นด้วย Leaf node (ใช้เทคนิคนี้กับข้อมูลใน Training set ที่ใช้ในการสร้าง Tree เท่านั้น) และการคำนวณความผิดพลาดที่เกิดจากการทำนาย ของแต่ละ Leaf node และ Sub tree จะทำโดยการสมมติว่าจะทำการแบ่งกลุ่ม set ของข้อมูลที่ไม่เคยพบมาก่อนที่มีขนาดเท่ากับ Training set โดยการคำนวณจะใช้ Function ทางสถิติ ซึ่งอยู่บนพื้นฐานของการกระจายแบบ Binomial จำนวน Error ที่เกิดขึ้นเมื่อข้อมูลมีขนาดเท่ากับ N

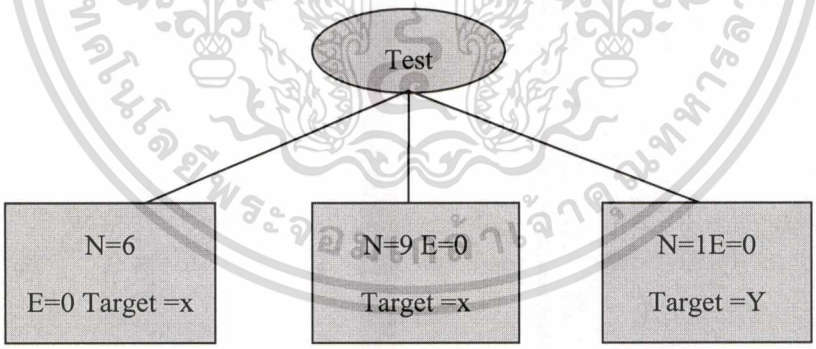
$$= N * U_{CF}(E,N)$$

โดย N แทน ขนาดของข้อมูลที่ Leaf node ใดๆ

E แทน จำนวนของ Error ที่เกิดขึ้นใน set ของข้อมูลที่ node ใดๆ

$U_{CF}(E,N)$ แทน ความน่าจะเป็นสูงสุดที่จะเกิด Error และ C4.5 ใช้ Confidence level เท่ากับ 0.25 หรือ 25%

ต่อไปจะอธิบายการ Pruning โดยพิจารณา sub tree ดังรูป



รูปที่ 3.3 Sub tree

บทที่ 4
การวิเคราะห์และออกแบบระบบดาต้าไมนิ่ง

การออกแบบพัฒนาระบบดาต้าไมนิ่งเพื่อวิเคราะห์พฤติกรรมการใช้งานของลูกค้ากลุ่มบริษัทเทเลคอมมิวนิเคชั่นแห่งหนึ่ง ซึ่งอยู่ในธุรกิจในกลุ่มการสื่อสารที่มีการเข่งขันการสูงมากนั้น การเข้าถึงลักษณะของลูกค้าย่อมมีผลต่อการได้เปรียบในการดำเนินธุรกิจไปสู่ความสำเร็จได้สูงยิ่งขึ้น

4.1 สถาปัตยกรรมระบบ

ในการออกแบบจะใช้สถาปัตยกรรมแบบ Client-Server based แบบ 3-Tiers โดยการทำงานแบบ Application ส่วนประกอบของแต่ละ Tier มีดังนี้

1. **Database Server** เป็นส่วนที่เก็บฐานข้อมูลของระบบ โดยมี Software ที่จัดการเกี่ยวกับระบบฐานข้อมูล โดยมีการติดต่อกับ ODBC เพื่อที่จะนำข้อมูลมาประมวลผล
2. **Application Server** เป็นส่วนที่รองรับงานทางด้าน Application โดยจะมีการติดต่อกับ Database Server ผ่านทาง ODBC ดึงข้อมูลจาก ฐานข้อมูลเพื่อใช้ในการวิเคราะห์ และผลข้อมูลที่จะนำข้อมูลที่ได้ไปวิเคราะห์เพื่อออกโปรโมชั่น ตอบสนองลูกค้าต่อไป
3. **Client** ผู้ใช้ ซึ่งจะเป็นส่วนของฝ่ายการตลาด เพื่อนำข้อมูลที่ได้ไปวิเคราะห์เพื่อออกโปรโมชั่น ตอบสนองลูกค้าต่อไป

4.2 Problems and Benefits

Problem

- ต้องการนำข้อมูลการใช้งานของลูกค้าไปวิเคราะห์เพื่อสร้างผลกำไรทางการตลาดมากยิ่งขึ้น

Benefits

- ระบบสามารถรองรับข้อมูลที่มีขนาดใหญ่และนำข้อมูลที่ได้ไปวิเคราะห์ และประมวลผลออกมาเพื่อให้ง่ายต่อการเข้าใจ เพื่อให้ผู้ใช้ได้นำข้อมูลเหล่านี้ไปใช้ในการเพื่อมูลค่าทางการตลาดมากยิ่งขึ้น

4.3 Functional Requirement

Accessibility Of System

ความสามารถและกระบวนการวิเคราะห์ข้อมูลของระบบ

ตารางที่ 4.1 Accessibility of System

Requirement	Marketing Plan/Manager
Process 1 Data Preparation	
Process 1.1 Connect Database	☒
Process 1.2 Selection Attribute for Process	☒
Process 1.3 Manage Missing Value	☒
Process 1.4 Create Rule for Coding	☒
Process 1.5 View Frequency	☒
Process 1.6 View Data	☒
Process 2 Build Tree by C4.5 Algorithm	
Process 2.1 Build tree	☒
Process 2.2 Create Testing Data	☒
Process 2.3 Pruning	☒

- **รายละเอียด functional requirements**

  Process 1 Data Preparation

  - การติดต่อกับฐานข้อมูล(Process 1.1 Connect Database)

    หน้าที่ของ Function : เพื่อติดต่อกับฐานข้อมูลที่ต้องการจะใช้ในการวิเคราะห์ข้อมูล

    - **Class of Input**

ตารางที่ 4.2 Class of Input(Process 1.1 Connect Database)

Input	Valid	Invalid	Error Message	Description
Database Driver	เป็นข้อความ		General Error	ชื่อ Driver database
Database URL	เป็นข้อความ		General Error	ชื่อ Database URL

User Name	เป็นตัวอักษร		General Error	อาจจะมีหรือไม่ตามแต่กำหนดตอนติดตั้ง database
Password	เป็นตัวอักษร		General Error	อาจจะมีหรือไม่ตามแต่กำหนดตอนติดตั้ง database

- **Class of Output**

  แสดงหน้าจอ Sql Command

- การเลือกข้อมูล (Process 1.2 Selection Attribute for Process)

  หน้าที่ของ Function : ดึงข้อมูลมาจาก Database

- **Class of Input**

ตารางที่ 4.3 Class of Input(Process 1.2 Selection Attribute for Process)

Input	Valid	Invalid	Error Message	Description
Sql command	Sql command		[Microsoft][ODBC Microsoft Access Driver]Too few parameter	ตัวแรกเป็น Target value และตัวต่อไปที่ค่าที่เราต้องการจำทำมาศึกษาร่วมด้วย

- **Class of Output**

  แสดงข้อมูล Data Typeที่ทำการเลือก

- การจัดการกับข้อมูลที่ไม่สมบูรณ์ (Process 1.3 Manage Missing Value)
  - ➢ การลบข้อมูลที่ไม่สมบูรณ์ออก
  - ➢ การแทนที่ข้อมูลที่ไม่สมบูรณ์ โดยให้ป้อนค่าตามชนิดและข้อมูลที่จะแทนที่

- **Class of Output**

  แสดงข้อมูล Data ที่เหลือที่มีสภาพสมบูรณ์

- การกำหนดRule (Process 1.4 Create Rule for Coding)

  หน้าที่ของ Function : แปลงข้อมูลที่มีค่าเป็นตัวเลขให้อยู่ในรูป String เพื่อเพิ่มประสิทธิภาพในการประมวลผลของโปรแกรม

- **Class of Input**

ตารางที่ 4.4 Class of Input(Process 1.4 Create Rule for Coding)

Input	Valid	Invalid	Error Message	Description
ค่าที่ต้องการจะแปลงข้อมูล	ค่าตัวเลข			แปลงค่าเป็น String

- **Class of Output**

  ข้อมูลที่จะเตรียมส่งไปประมวลผล

- การดูกลุ่มข้อมูล (Process 1.5 View Frequency)

  - **Class of Output**

    ข้อมูลโดยแสดงตามความถี่ของข้อมูล

- การดูข้อมูล(Process 1.6 View Data)

  หน้าที่ของ Function :แสดงข้อมูลที่เลือกจากฐานข้อมูล

  - **Class of Output**

    ข้อมูลที่เลือกจากหน้าจอ Sql Command

Process 2 Build Tree by C4.5 Algorithm

- การสร้างต้นไม้ (Process 2.1 Build tree)

  หน้าที่ของ Function :แสดงข้อมูลที่เลือกจากฐานข้อมูล ที่ผ่านกระบวนการ Data Preparation มาสร้างเป็นกราฟต้นไม้

  - **Class of Output**

    กราฟรูปต้นไม้

- การสร้างตัวทดสอบข้อมูล(Process 2.2 Create Testing Data)

  หน้าที่ของ Function : นำข้อมูลที่เลือกจากฐานข้อมูล ที่ผ่านกระบวนการ Data Preparation มาแบ่งเป็น Case ทดสอบเพื่อตรวจสอบความถูกต้องที่ได้จากการประมวลของ c4.5 algorithm

  - **Class of Output**

    ข้อมูลที่ได้จากการสุ่ม(Random) จาก data ที่ส่งเข้ามาประมวลผล

- การทำ Pruning (Process 2.3 Pruning)

  หน้าที่ของ Function :คือส่วนลด error rate ที่เกิดจากการประมวลด้วย Algorithm

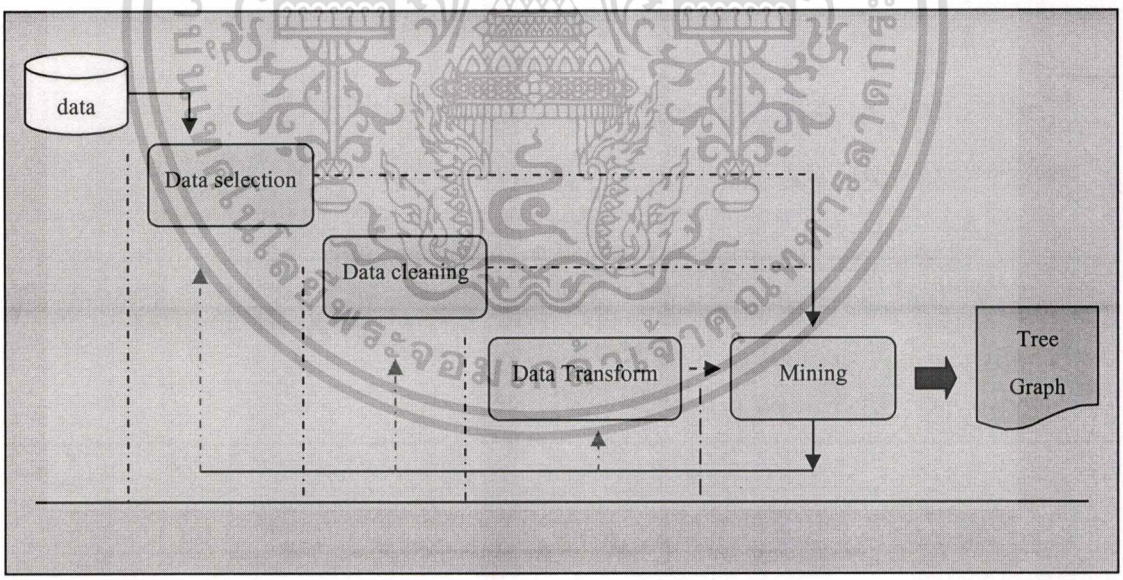
  - **Class of Output**

    กราฟรูปต้นไม้ที่มีระดับชั้นลดลงลดจำนวนลง

4.4 สรุปขั้นตอนการทำงานของระบบ

ระบบจะประกอบด้วย 5 ส่วนหลัก ด้วยกัน คือ

1. ส่วนในการรับข้อมูลดิบ ซึ่งผู้ใช้จะสามารถรับได้ระบบฐานข้อมูล
2. เมื่อรับข้อมูลเข้าสู่ระบบ ระบบให้ผู้ใช้เลือกข้อมูลที่ต้องการจะวิเคราะห์ ซึ่งอยู่ในส่วน Data Selection
3. ผู้ใช้เลือกข้อมูลที่จะทำการวิเคราะห์ระบบจะตรวจสอบ หากพบข้อมูลที่ต้องทำการ Cleaning อยู่ในส่วนของการทำ Data Cleaning
4. เมื่อเสร็จขั้นตอนข้างต้น ผู้ใช้อาจจะต้องการทำการ Transform ข้อมูล ถ้ามีข้อมูลประเภท Quantitative ระบบจะอนุญาติให้ผู้ใช้ ทำการ Transform
5. เสร็จสิ้นทุกขั้นตอนก็จะเข้าสู่การทำ Data mining ซึ่งอาจเกิดการวิเคราะห์ที่ผิดพลาด สามารถย้อนกลับไปทำขั้นตอนข้างต้น โดยผลจะแสดงออกมาในรูปของกราฟต้นไม้



รูปที่ 4.1 ขั้นตอนการทำงานของระบบ

บทที่ 5
ระบบการวิเคราะห์การจัดแบ่งกลุ่มลูกค้าที่ใช้โทรศัพท์พื้นฐาน

ในบทนี้จะเป็นรายละเอียดทั้งหมดของระบบการวิเคราะห์การจัดแบ่งกลุ่มลูกค้าที่ใช้โทรศัพท์พื้นฐาน โดยใช้ดิซิซั่นทรี โดยระบบจะรับข้อมูลที่จะใช้ในการสร้างโมเดลเป็นอินพุตท์ แชะระบบจะประมวลผลข้อมูลด้วยการกระบวนการทางดาต้าไมนิ่ง เพื่อให้ได้เอาพุตท์ออกมาเป็นโมเดลในแบบดิซิซั่นทรี ซึ่งในบทนี้ จะกล่าวถึงการขั้นตอนทำงานของระบบ

5.1 เครื่องมือที่ใช้ในการพัฒนาระบบ

ระบบการวิเคราะห์การจัดแบ่งกลุ่มลูกค้าที่พัฒนาขึ้นโดยใช้ JBuilder ในการพัฒนาระบบเป็นเครื่องมือการพัฒนาระบบด้วยภาษาจาวา (Java Language) เหตุผลที่เลือกใช้ภาษาจาวาด้วย JBuilder ในการพัฒนาระบบนี้ เนื่องจากภาษาจาวาเป็นภาษาเชิงวัตถุ (Java Programming Language) ซึ่งให้ความยืดหยุ่น, ความเป็นมาตรฐาน, ความชัดเจน และกลไกซึ่งส่งเสริมการนำโปรแกรมที่สร้างไว้แล้วมาใช้งานใหม่ได้ (Reusability) ทำให้สามารถนำบาง method และ class ที่มีส่วนเกี่ยวข้องกับระบบซึ่งมีอยู่แล้ว และมาใช้ในการพัฒนาระบบได้

5.2 แนวทางในการพัฒนาระบบ

ระบบพัฒนาให้อยู่ในรูปแบบของแอปพลิเคชั่นบนวินโดวส์ ซึ่งเป็นระบบปฏิบัติการที่นิยมใช้กันอย่างแพร่หลาย เพื่อให้ผู้ใช้สามารถเรียนรู้การใช้งานระบบได้ง่าย เนื่องจากผู้ใช้ส่วนใหญ่จะมีความเคยชินในการใช้โปรแกรมแอปพลิเคชั่นอยู่แล้ว ระบบยังมีการออกแบบให้ผู้ใช้ได้รับทราบหากมีข้อมูลที่ขาดหายไป โดยระบบจะให้ลูกค้ากำหนด หรือลบข้อมูลเหล่านั้นทิ้งไป ส่งผลให้ข้อมูลมีความสมบูรณ์มากขึ้น เพื่อนำไปสร้างโมเดลที่มีประสิทธิภาพต่อไปได้ และในรูปแบบการแสดงผลการวิเคราะห์เราจะให้ผู้ใช้สามารถเห็น รูปแบบของข้อมูลในลักษณะต้นไม้ได้

5.3 โครงสร้างการทำงานของระบบ

ระบบประกอบด้วยส่วนการรับข้อมูลการเตรียมข้อมูล การสร้างโมเดลและการแสดงผล

การรับข้อมูล

ระบบจะรับข้อมูลโดยการเลือกติดต่อกับฐานข้อมูลที่ต้องการ โดยระบุ URL และ Driver ของฐานข้อมูล รวมทั้งUser และ Password (ถ้ามี) หลังจากติดต่อกับฐานข้อมูลเป็นที่เรียบร้อย เราสามารถเลือกข้อมูลในฐานข้อมูลโดยใช้คำสั่ง SQL ระบุแอตทริบิวต์ และตารางที่ต้องการ โดยหลักการในการเลือกข้อมูลนั้นความเลือกข้อมูลที่เป็นข้อมูล target เก็บข้อมูลแรก และเลือกข้อมูลที่มีความเกี่ยวข้องในการสร้างโมเดล และเหมาะสมกับการทำนานดังที่กล่าวไว้ เพื่อให้ผลการสร้างโมเดลมีความถูกต้อง และยอมรับได้

5.4 การเตรียมข้อมูล (Data Preparation)

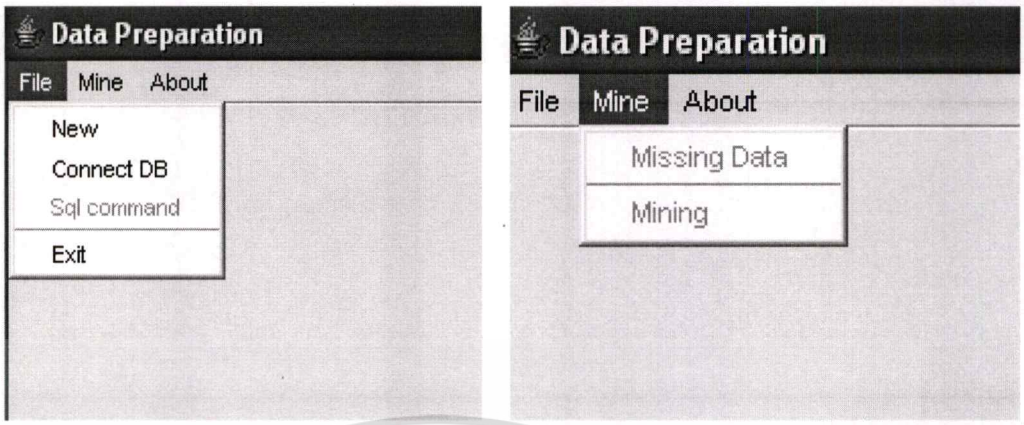
เมื่อติดต่อ และเลือกข้อมูลจากฐานข้อมูลเป็นที่เรียบร้อยแล้ว ระบบจะแสดงรายการแอตทริบิวต์ที่มีทั้งหมด และแสดงสถานะว่าข้อมูลที่เลือกเข้ามามีความสมบูรณ์หรือไม่ หากไม่สมบูรณ์ก็จะให้ผู้ใช้ปรับเปลี่ยนหรือ เลือกลบข้อมูลเหล่านั้น จากมากดูค่าข้อมูลว่าแต่กลุ่มมีจำนวนมากน้อยเท่าไหร่

5.5 การสร้างโมเดล

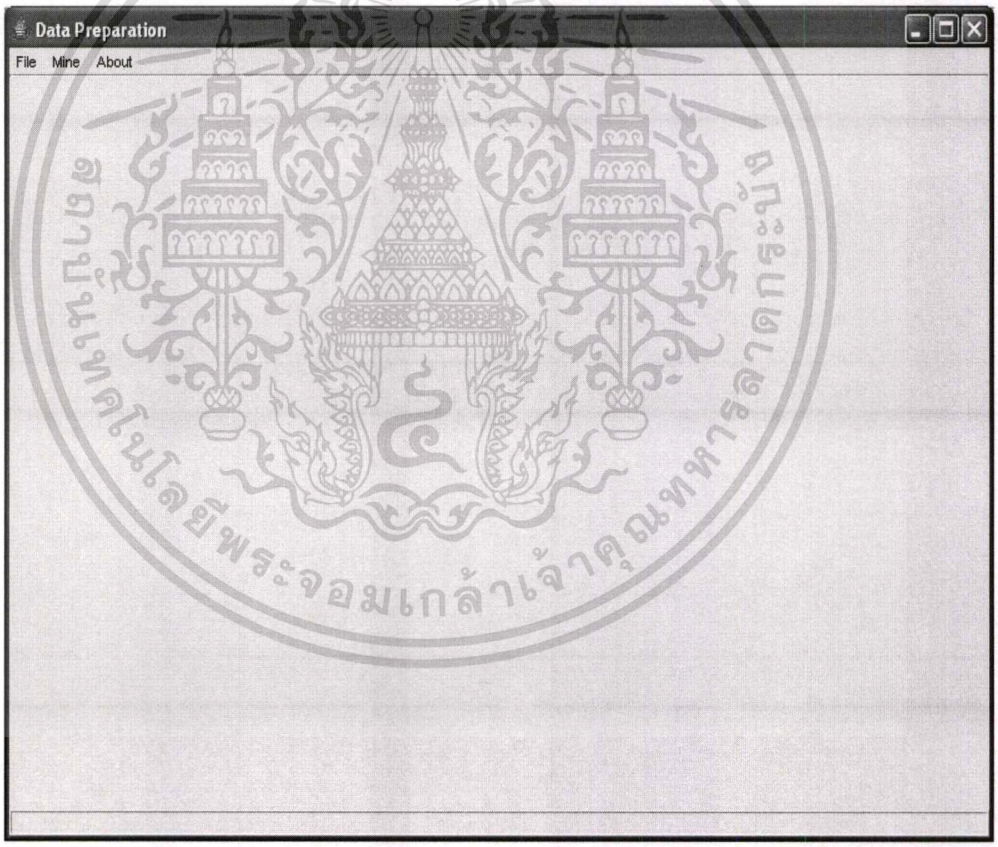
หลังจากการเตรียมข้อมูลเป็นที่เรียบร้อย ข้อมูลที่เราเลือกเป็นอันดับแรกจะจัดเป็น Target value หรือ แอตทริบิวต์ที่มีความสนใจ หรือ แอตทริบิวต์ที่มีค่าเป็นลาเบลที่การแบ่งกลุ่มไปล่วงหน้าแล้ว ตาม Target ที่เราตั้งเป้าหมายไว้

5.6 รายละเอียดของหน้าจอการทำงาน

หน้าจอหลัก แสดงดังรูปที่ 5.2 คือมีการสร้างโมเดลใหม่ ดูเลือกเมนู File กดปุ่ม New หรือเลือก เมนู Mine เลือก Connect DB

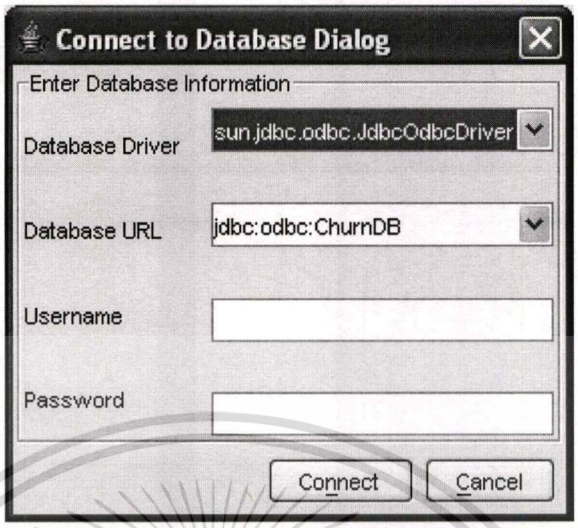


รูปที่ 5.1 เมนูหน้าจอหลัก



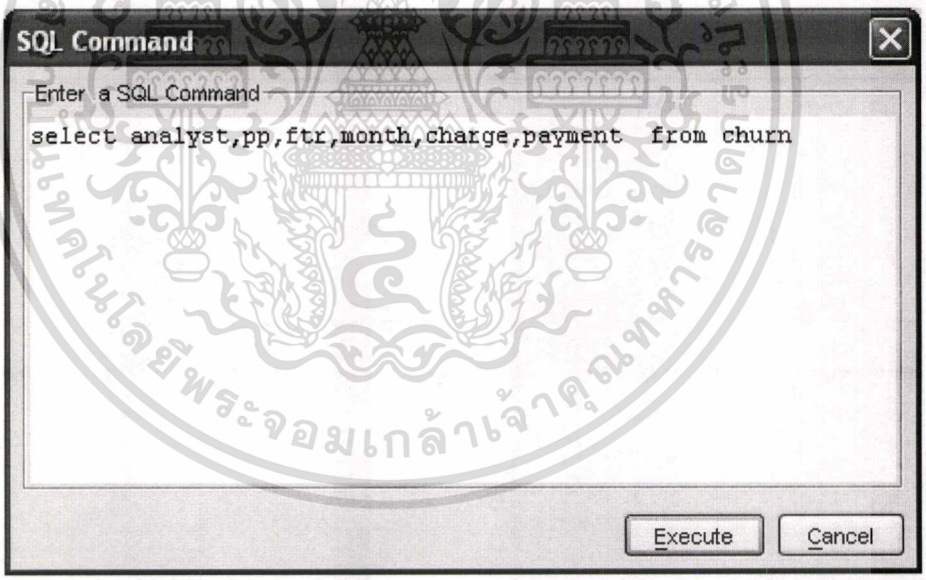
รูปที่ 5.2 หน้าจอหลัก

○ เมื่อมีการเลือก New หรือ Connect DB จะแสดงหน้าจอให้เลือกว่าจะ Connect กับ Database ตัวใด ใส่ User and Password กด Connect



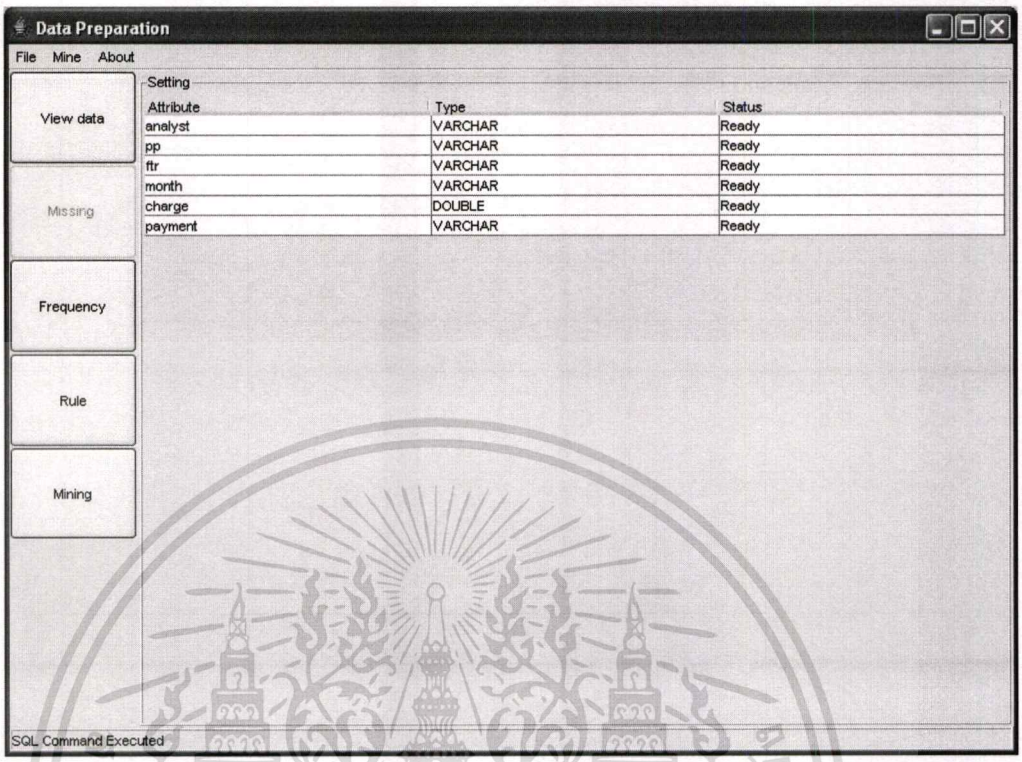
รูปที่ 5.3 Connect Dialog

○ หลังจาก กด Connect จะมีหน้าจอให้ป้อนคำสั่ง SQL ซึ่งต้องป้อน ตัวแรกเป็นตัว Target value ก่อน



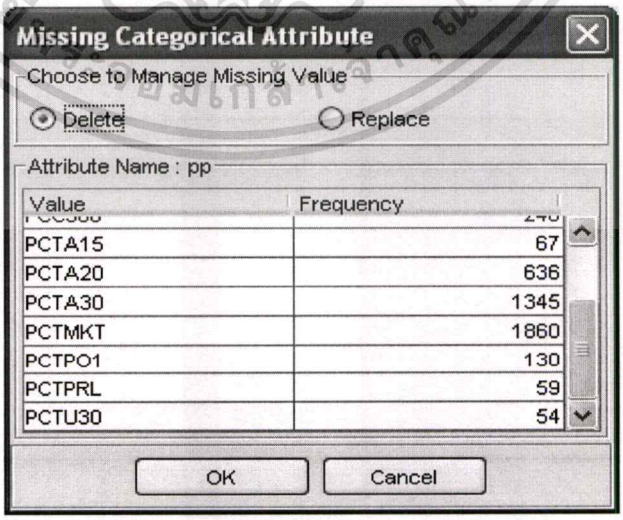
รูปที่ 5.4 SQL Dialog

○ หลังจากกด Execute จะแสดง หน้าจอ Typeของ ข้อมูลที่เลือก ขึ้นมา เพื่อให้เราทำงานต่อไป



รูปที่ 5.5 แสดง type ของข้อมูลที่เลือก

○ หากมีข้อมูลที่ สูญหายเป็นบาง Field จะแสดงข้อมูลเพื่อให้มีการแก้ ไข และปุ่ม Missing Value จะสามารถทำงานได้ เมื่อมีการกดปุ่ม Missing Value จะ แสดงหน้าจอตามลักษณะของข้อมูล ซึ่ง ถ้าเป็น Categorical จะแสดงข้อมูลที่มีทั้งหมด ว่ามี ค่าอะไรบ้าง จำนวนเท่าไหร่ เพื่อประกอบการตัดสินใจ

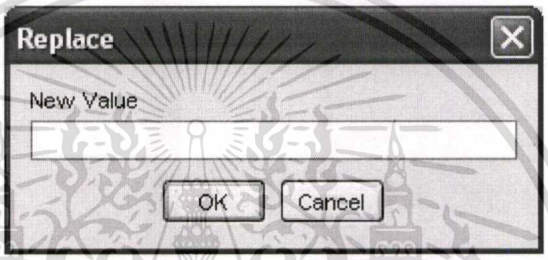


รูปที่ 5.6 Missing Value Dialog of Categorical

หาก เลือก Delete แล้วกด OK จะ แสดง Dialog เพื่อ มีการยืนยันดังรูป

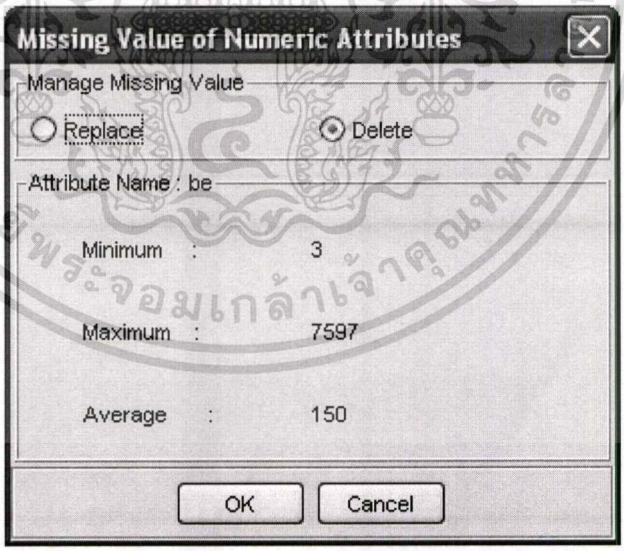
รูปที่ 5.7 Confirm Dialog for Delete

หากต้องการแทนที่ด้วยค่าใดค่าหนึ่งให้กด เลือก Replace กด OK จะมีหน้าจอให้ป้อนค่าที่ต้องการ ดังรูป



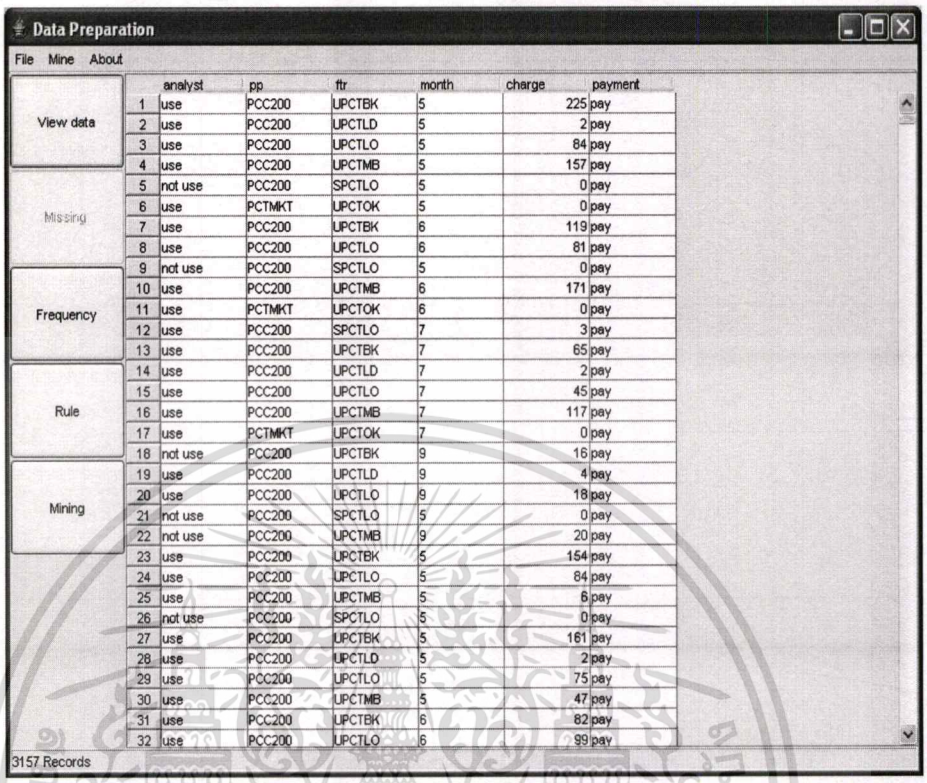
รูปที่ 5.8 Replace new value Dialog

หากข้อมูลเป็น Numeric จะแสดงหน้าจอ



รูปที่ 5.9 Missing Value of Numeric Attribute

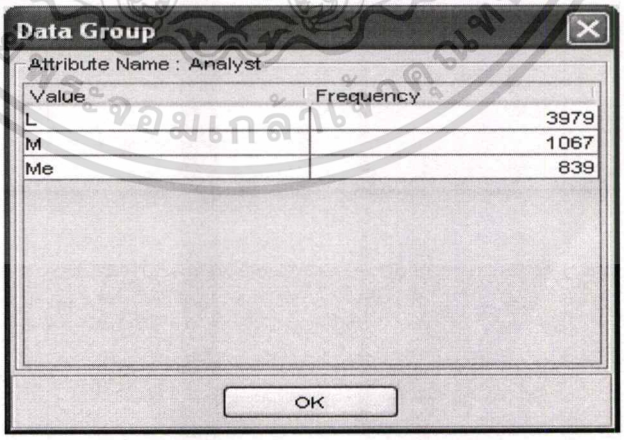
○ ถ้ากดปุ่ม View data จะแสดงข้อมูลทั้งหมด ที่เลือก ออกมา ดังรูปที่



	analyst	pp	ftr	month	charge	payment
1	use	PCC200	UPCTBK	5	225	pay
2	use	PCC200	UPCTLD	5	2	pay
3	use	PCC200	UPCTLO	5	84	pay
4	use	PCC200	UPCTMB	5	157	pay
5	not use	PCC200	SPCTLO	5	0	pay
6	use	PCTMKT	UPCTOK	5	0	pay
7	use	PCC200	UPCTBK	6	119	pay
8	use	PCC200	UPCTLO	6	81	pay
9	not use	PCC200	SPCTLO	5	0	pay
10	use	PCC200	UPCTMB	6	171	pay
11	use	PCTMKT	UPCTOK	6	0	pay
12	use	PCC200	SPCTLO	7	3	pay
13	use	PCC200	UPCTBK	7	65	pay
14	use	PCC200	UPCTLD	7	2	pay
15	use	PCC200	UPCTLO	7	45	pay
16	use	PCC200	UPCTMB	7	117	pay
17	use	PCTMKT	UPCTOK	7	0	pay
18	not use	PCC200	UPCTBK	9	16	pay
19	use	PCC200	UPCTLD	9	4	pay
20	use	PCC200	UPCTLO	9	18	pay
21	not use	PCC200	SPCTLO	5	0	pay
22	not use	PCC200	UPCTMB	9	20	pay
23	use	PCC200	UPCTBK	5	154	pay
24	use	PCC200	UPCTLO	5	84	pay
25	use	PCC200	UPCTMB	5	6	pay
26	not use	PCC200	SPCTLO	5	0	pay
27	use	PCC200	UPCTBK	5	161	pay
28	use	PCC200	UPCTLD	5	2	pay
29	use	PCC200	UPCTLO	5	75	pay
30	use	PCC200	UPCTMB	5	47	pay
31	use	PCC200	UPCTBK	6	82	pay
32	use	PCC200	UPCTLO	6	99	pay

รูปที่ 5.10 Show Data (View data)

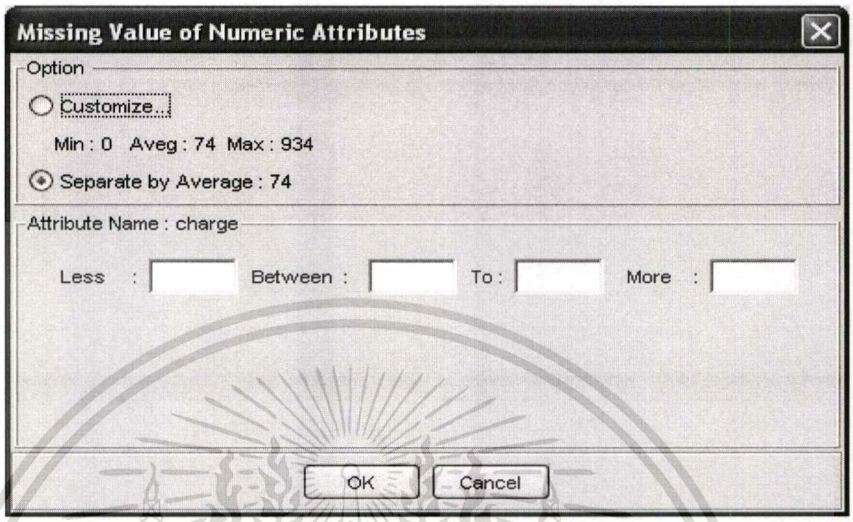
○ หากไม่มี ข้อมูลที่ สูญหาย จะสามารถเลือก ปุ่ม Group และ Mining จะสามารถทำงานได้ ซึ่งหาก กดปุ่ม Group จะแสดงข้อมูลความถี่ข้อมูลที่จะใช้ในการทำ Mining ออกมา โดยจะแสดงที่ละกลุ่มไปจนครบเพื่อให้ ผู้ใช้ไม่สับสน ในการดูข้อมูล



รูปที่ 5.11 Group of data Dialog

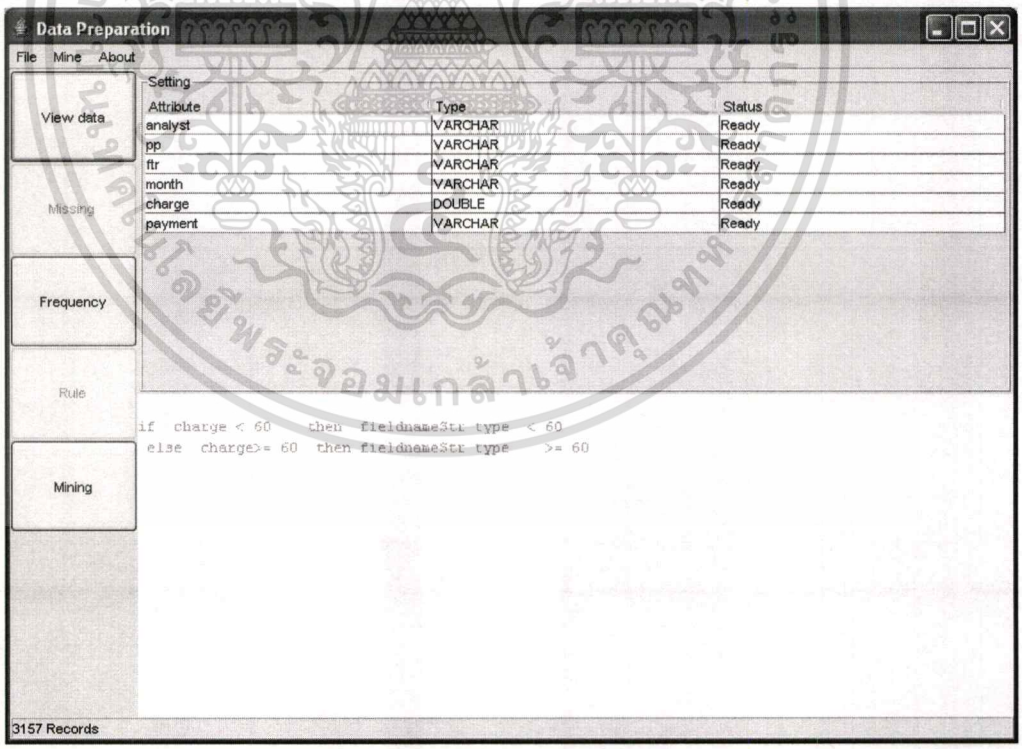
○ หากมีข้อมูลที่มีลักษณะเป็น Quantitative อาจจะมาเป็น Categorical ก่อน เพื่อให้การประมวลผลเป็นไปได้อย่างรวดเร็วยิ่งขึ้น โดยการคิดปุ่ม Rule โดยจะแสดงให้

เลือกกำหนด ช่วงของข้อมูลว่าจะกำหนดเอง หรือ แบ่งตามค่าเฉลี่ย หลังจากเลือกเป็นที่เรียบร้อย



รูปที่ 5.12 Rule Dialog

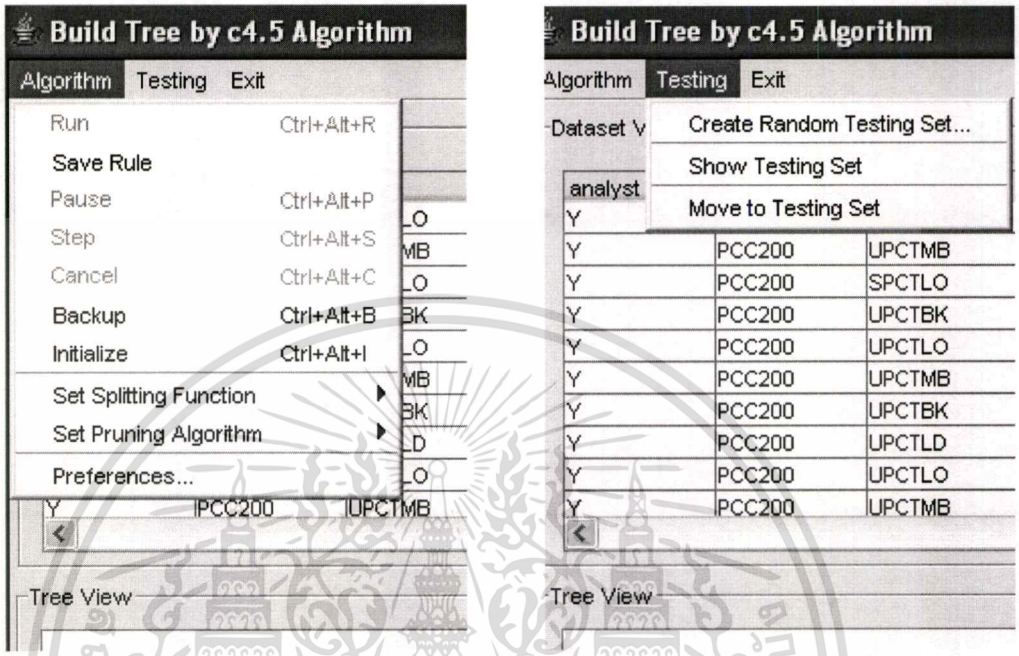
หลังจากเลือกเป็นที่เรียบร้อย จะแสดง Rule ดังหน้าจอต่อไปนี้



รูปที่ 5.13 Main Dialog

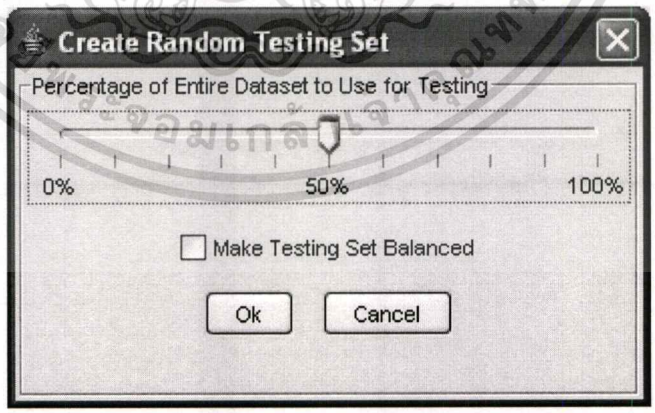
○ หากกดปุ่ม Mining จะแสดงหน้าจอ Build Tree by c4.5 Algorithn แสดงความพร้อมในการสร้าง Tree จะถูกแบ่งเป็น 3 ส่วน คือ ส่วนของ Menu การทำงาน ส่วนของ

ข้อมูลที่จะมาสร้าง Tree และส่วนของ Tree ที่จะมีการสร้างขึ้น โดยส่วน Tree หรือ ส่วน Tree View จะ การ แสดง 3 ส่วนด้วยกัน คือ



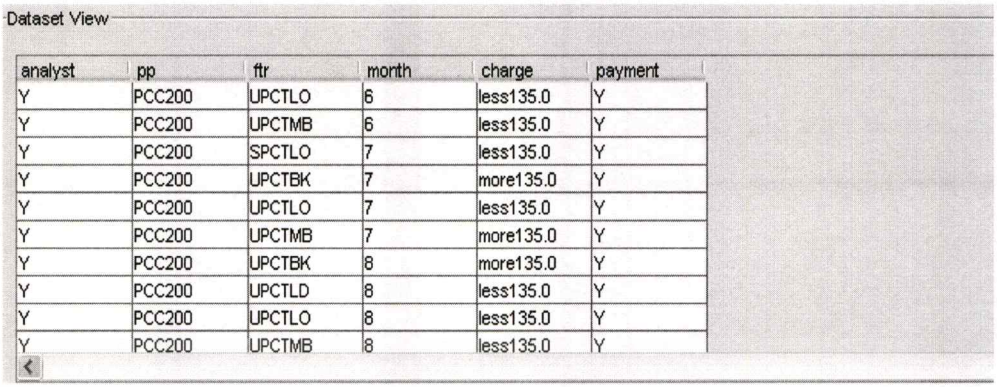
รูปที่ 5.14 Menu Bar of Mining Dialog

การสร้าง Test Set คือการแบ่งจำนวนข้อมูลที่ใช้ในการ Training set ว่าจะนำไปใช้ในส่วนของ test set จำนานเท่าไหร่ โดยผ่านหน้าจอ



รูปที่ 5.15 Menu Bar of Mining Dialog

ซึ่งจะ random ข้อมูลที่จะนำไปใช้ใน Test set

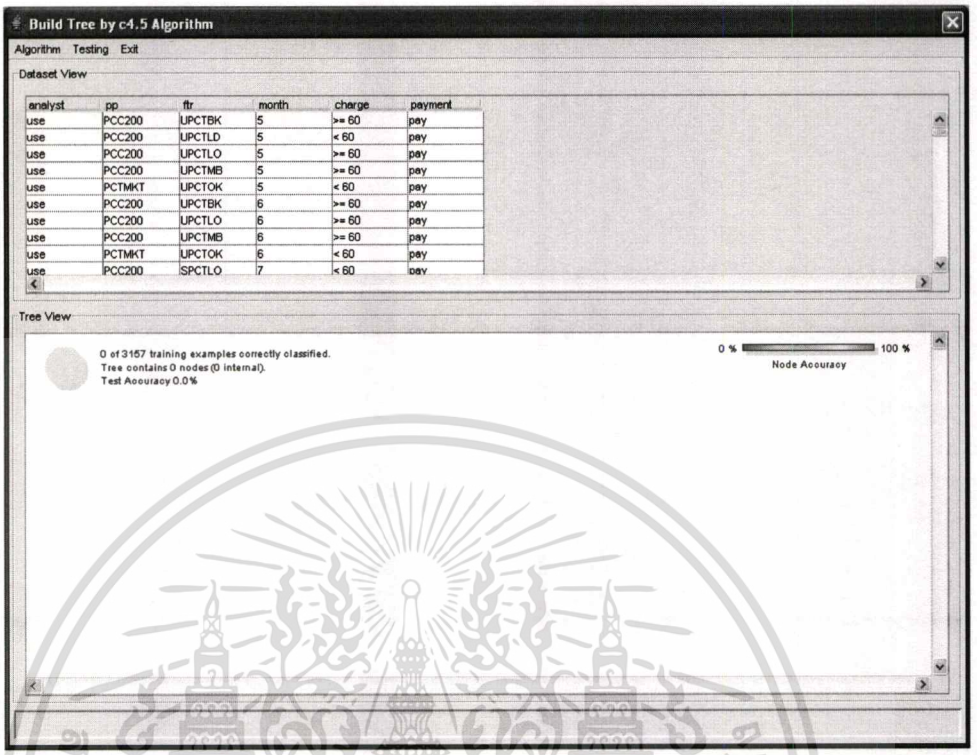


รูปที่ 5.16 ส่วนของข้อมูลที่จะมาสร้าง Tree

- Pie chart หรือ กราฟวงกลม จะ แสดงข้อมูลว่ามีข้อมูลจำนวนเท่าไหร่ ถ้ามีการสั่งให้ประมวลผลก็จะแสดงผลข้อมูลที่ใช้ในการ training และยังบอกถึง leaf node ว่ามีจำนวนเท่าไหร่ที่จะเกิดขึ้นการการไมนิ่ง Mining และมี internal จำนวนเท่าไหร่
- Gradient Bar เพื่อเป็นตัวบอกถึงสีของ Leaf node ว่ามีความถูกต้องประมาณเท่าไหร่
- Tree ซึ่งจะแสดงหลังจากมีการประมาลผล

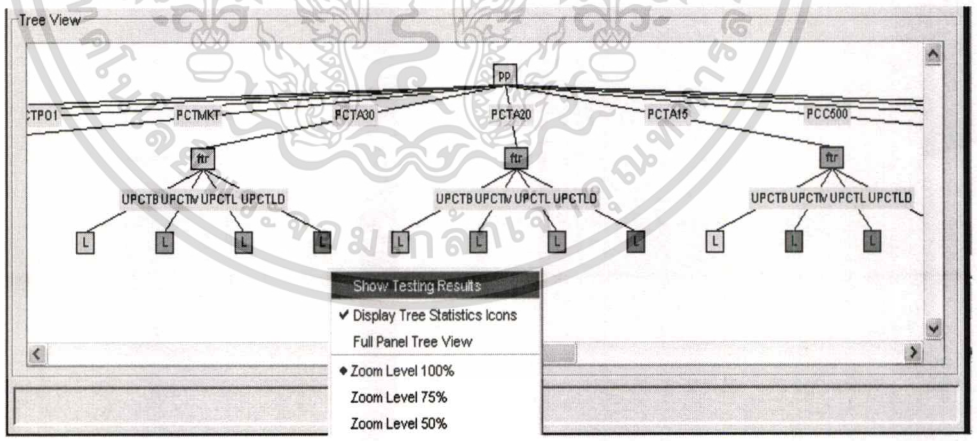
รูปที่ 5.17 แสดงถึงส่วนประกอบ ของTree View

ซึ่งหากเราคลิกบริเวณ Panel จะแสดงจำนวนข้อมูล ที่เกิดขึ้นในกลุ่มของ Target Value

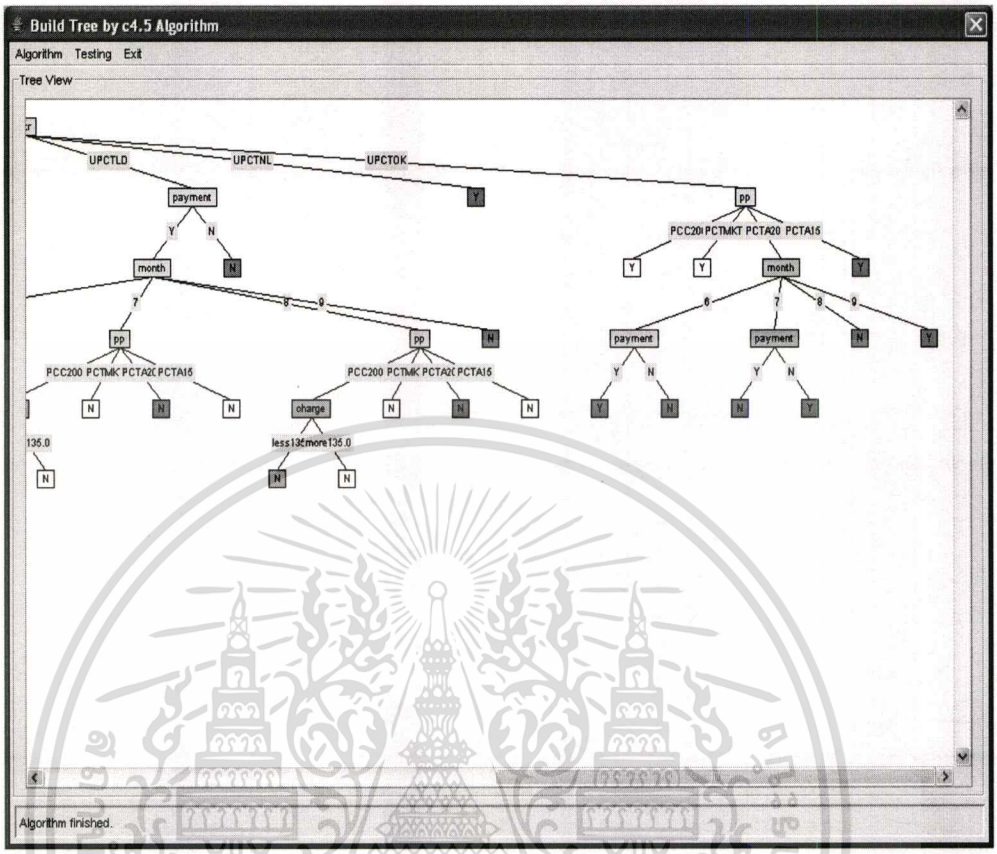


รูปที่ 5.18 Mining Dialog

ซึ่งหากเราคลิกซ้าย เราสามารถ ขยายหน้า จอ เพื่อดู Tree ได้



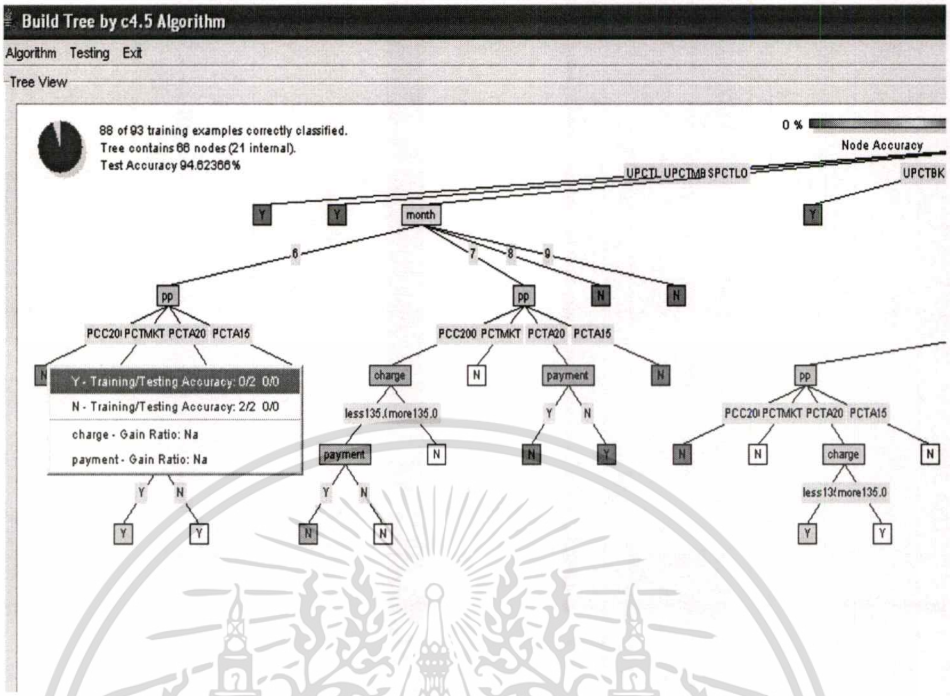
รูปที่ 5.19 Tree View



รูปที่ 5.20 Full Panel Tree View

ซึ่งหน้าจอจะมี Menu เพื่อทำการ Mining หาก เรากด Run จะแสดง Tree ออกมา

ซึ่งเมื่อไมนิ่งเป็นที่เรียบร้อย จะแสดง กราฟทรีออกมา ซึ่ง Leaf node จะมี สี ที่แตกต่างกัน ซึงสามารถสังเกตได้จาก Gradient Bar หาก มีสี แดง จะมีความถูกต้อง สูง หรือผู้ใช้อาจคลิก ที่ leaf node ที่ต้องการศึกษา แสดงข้อมูลการ Training ออกมา ซึ่งเราสามารถบอกถึงความถูกต้อง ในแต่ละ Leaf Node ได้ ซึงนับว่าเป็นผลดี ต่อการ วิเคราะห์อย่างมาก



รูปที่ 5.21 Full Panel Tree View

5.7 ขั้นตอนการทำงานของระบบ

1) ดับเบิ้ลคลิกๆไอคอน จะแสดงหน้าจอรูปที่ 5.2

2) จากนั้นไป Menu File เลือก Connect จะแสดงหน้าจอรูปที่ 9 เมื่อ Connect สำเร็จ จะแสดงหน้าจอที่ 5.3

3) หลังจากนั้นให้ใส่คำสั่ง Sql โดยตัวแรกของการ Select เป็นตัว Target Values เสมอ

4) หลังจากกดปุ่ม OK หากคำสั่งที่ส่งเข้าไปถูกต้อง จะแสดง Filed ที่เลือกออกมา หากมีข้อมูลที่ Missing อยู่ระบบจะให้ผู้ใช้ทำการแก้ไข โดยกดปุ่ม Missing ในรูปที่ 16 จะแสดงหน้าจอรูปที่ 5.4 ขึ้นมา ให้เลือก หากเราต้องการลบถึง ระบบจะแสดงหน้าจอรูปที่ 5.5 เพื่อให้ผู้ใช้ยืนยัน หรือ หากผู้ใช้เลือกที่แทนที่ค่าลงไป ระบบจะแสดงหน้าจอรูปที่ 14 เพื่อให้ผู้ใช้ ใส่ค่าที่ต้องการ

5) หากไม่มีแต่ข้อมูลที่เป็นค่า Number หรือ Double ระบบจะยอมให้ผู้ใช้กำหนด Rule โดยกดปุ่ม Rule ในรูปที่ 5.10 จะแสดงหน้าจอรูปที่ 5.12 เพื่อให้กำหนดค่า ซึ่งการทำเช่นนี้จะช่วยลดเวลาในการทำ Mining ลงได้อย่างมากทีเดียว

6) หลังจากเสร็จกระบวนการต่างๆ ผู้ใช้สามารถเลือก Mining ซึ่งจะแสดงหน้าจอที่ 5.25

7) เลือก Menu Algorithm เลือก Initialize ระบบจะแสดงหน้าจอ ไม้ ดังที่เห็นในรูปที่ 5.17

8) เลือก Menu Algorithm เลือก run ระบบจะสร้างกราฟต้นไม้ ดังที่เห็นในรูปที่ 5.21

บทที่ 6
การประยุกต์ใช้ดาต้าไมนิ่งเพื่อการจัดแบ่งกลุ่มลูกค้าที่ใช้โทรศัพท์พื้นฐาน

การออกแบบพัฒนาระบบดาต้าไมนิ่งเพื่อวิเคราะห์พฤติกรรมการใช้งานของลูกค้ากลุ่มบริษัทเทเลคอมมิวนิวเคชั่นแห่งหนึ่ง ซึ่งอยู่ในธุรกิจในกลุ่มการสื่อสารที่มีการเข่งขันการสูงมากนั้น การเข้าถึงลักษณะของลูกค้าย่อมมีผลต่อการได้เปรียบในการดำเนินธุรกิจไปสู่ความสำเร็จได้สูงยิ่งขึ้น

6.1 กำหนดวัตถุประสงค์

1. ศึกษาถึงการเทคนิคของ ดาต้าไมนิ่ง โดยใช้ C4.5 Algorithms มาใช้ในการวิเคราะห์ได้
2. ศึกษาข้อมูลลักษณะการใช้งานของลูกค้า โดยวัดจากโปรโมชั่นที่ลูกค้าเลือกใช้ต่อหนึ่งผลิตภัณฑ์ เพื่อคาดเดาความต้องการของลูกค้าในอนาคต ซึ่งผลจากการวิเคราะห์จะช่วยให้องค์กรหรือกลุ่มธุรกิจสามารถสร้างโอกาสทางการตลาดได้สูงขึ้น

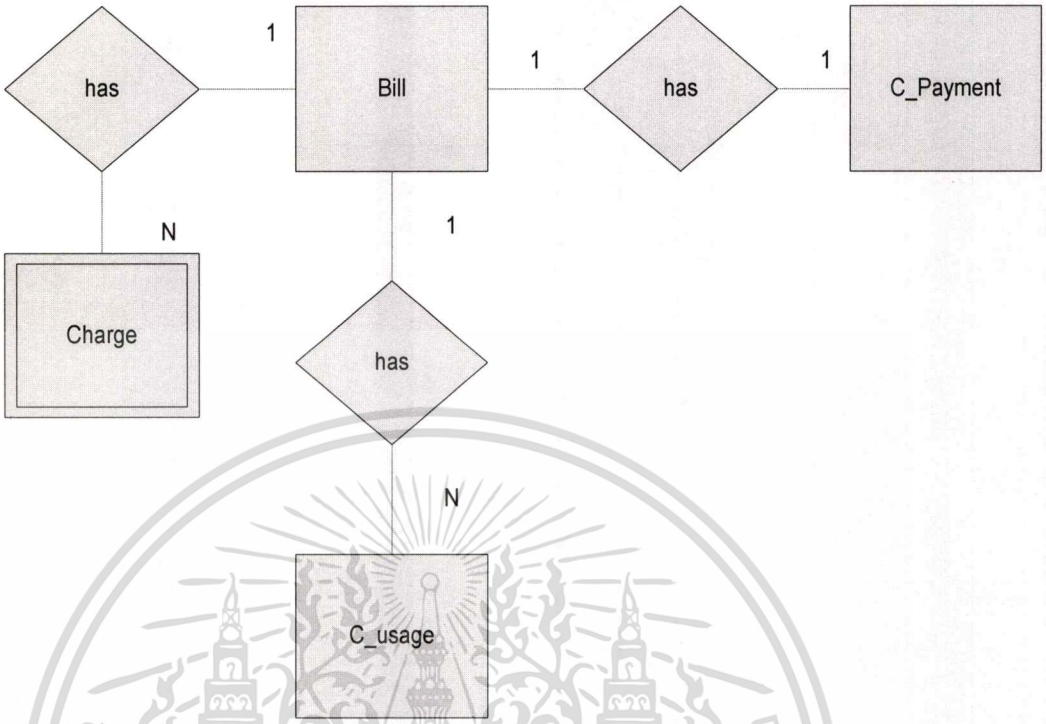
6.2 การคัดเลือกข้อมูล

ขั้นตอนในการคัดเลือก Attribute ของข้อมูล โดย Attribute ที่เป็นปัจจัยในการหลักที่มีส่งผลต่อการวิเคราะห์ตามวัตถุประสงค์ที่ตั้งไว้ การเลือกข้อมูลการใช้งานลูกค้าลูกค้าตามโปรโมชั่นต่างๆ มาวิเคาระห์

โดยเลือกข้อมูลจาก C_payment (ตารางที่ 6.1) และ C_usagc (ตารางที่ 6.2) ซึ่งถือได้ว่าเป็นข้อมูลดิบและมีจำนวนมาก เรานำข้อมูลนั้น มาทำการ รวม ข้อมูลให้ มองเห็นลักษณะการใช้งานของลูกค้าในแต่ละเดือน โดย เรานำ ยอดการใช้แต่ละเดือนตามโปรโมชั่น (PP_CODE) และลักษณะการโทรขอลูกค้า เช่น โทรทางไกล, มือถือ หรือ โทรในกรุงเทพ และรวมทั้งการรวมกันของข้อมูลการ ชำระเงินค่าบริการของลูกค้า ดังตารางที่ 6.3 Summary

ER-Diagram

ก่อนอื่นต้องให้ข้อมูลเกี่ยวกับการเลือกดาต้าว่า ข้อมูลที่จะมาทำ Data Mining ไม่จำเป็นต้องเหมือนกันเสมอไป การวิเคราะห์อาจจะต้องทำกลับไปกลับมาหลายครั้ง ดังนั้นข้อมูลที่จะมาทำการวิเคราะห์อาจจะไม่ใช้ข้อมูลดังที่แสดงก็ได้ ขึ้นอยู่กับความจำเป็นและสิ่งที่ต้องทำวิเคราะห์และดุลพินิจของการเลือกข้อมูลของผู้เลือก แต่เราจะแสดงส่วนหนึ่งของข้อมูลที่เลือกมาใช้



รูปที่ 6.1 ER-Diagram

- **Data Dictionary**

○ C_payment จะเก็บข้อมูลการชำระค่าใช้บริการของลูกค้า ซึ่งมีรายละเอียด field ที่สำคัญดังต่อไปนี้

ตารางที่ 6.1 C_payment

Field_name	Field_type	Description	Constaint
CODE	INT(12)	รหัสของลูกค้า	Primary Key
PRODUCT_ID	CHAR(16)	เบอร์โทรศัพท์	Primary Key
ENT_SEQ_NO	CHAR(1)	เลขที่การเรียกเก็บเงิน	Primary Key
MONTH	CHAR(1)	เดือนที่ การชำระค่าบริการ	
PAYMENT_IND	CHAR(1)	การชำระค่าบริการ	

○ C_usage จะเก็บข้อมูลการใช้งานของลูกค้าในแต่ละวันเพื่อนำข้อมูลเหล่านี้มาวิเคราะห์ลักษณะการใช้งานของลูกค้าที่เกิดขึ้น ซึ่งมีรายละเอียด field ที่สำคัญดังต่อไปนี้

ตารางที่ 6.2 C_Usage

Field_name	Field_type	Description	Field_name
CODE	INT(12)	รหัสของลูกค้า	Primary Key
PRODUCT_ID	CHAR(16)	เบอร์โทรศัพท์	Primary Key
PRODUCT_TYPE	CHAR(1)	Type ของ product	Primary Key
MONTH	CHAR(1)	เดือนที่ ทำการประมวลผล	
CALL_TYPE	CHAR(2)	ชนิดของการโทร	Foreign Key
DIALED_TN	VARCHAR2(18)	เบอร์ที่โทรออก	
PP_CODE	CHAR(6)	รหัส PP	Foreign Key
FTR_CODE	CHAR(6)	รหัส feature	Foreign Key
PROD_STATUS	CHAR(1),	status ของ product	Foreign Key
QTY_OF_IU	INT(9),	ปริมาณการใช้	
CHARGE_AMT	FLOAT(9,3),	ค่าใช้จ่ายที่ใช้ไป	

หลังจากการออกแบบ เพื่อจะนำไปพัฒนาต่อไป ให้เป็นที่เรียบร้อยเพื่อความถูกต้อง จึงมีโจทย์ที่ระบบต้องตอบให้ได้คือ

1) ลูกค้ามีลักษณะการใช้ Promotion เป็นอย่างไรจงยกตัวอย่างมาสัก 3 โปรโมชั่น
2) การชำระหรือไม่ชำระค่าบริการขึ้นกับปัจจัยใดบ้าง

ซึ่งระบบที่เราพัฒนาขึ้น ผู้ใช้ต้องมีการ วิเคราะห์ข้อมูลลูกค้า ต้องมีการตรวจสอบสรุปการใช้งานของลูกค้า ก่อน แต่ในพัฒนาเรานำข้อมูลเราได้มีการเตรียมข้อมูลอย่างดีอย่างที่กล่าวมาข้างต้น แล้ว

ตารางที่ 6.3 Summary

Field_name	Field_type	Description
PP	CHAR(6)	โปรโมชั่นที่ลูกค้าเลือก
FEATURE_CD	CHAR(6)	Code ลูกค้าใช้ในการโทร
USE_MIN	INTEGER(10)	จำนวนนาทีที่ใช้ต่อเดือน
USE_FIXED	INTEGER(10)	จำนวนครั้งที่ใช้โทรศัพท์
MONTH	CHAR(1)	เดือน
PAYMENT_IND	CHAR(1)	การชำระค่าบริการ

6.3 การเตรียมข้อมูล

○ **Data Cleaning**

กระบวนการ Data cleaning จัดได้ว่าเป็นอีกหนึ่งกระบวนการที่มีความสำคัญ ใช้ในการตรวจสอบเพื่อระบุถึงความผิดพลาด หรือความไม่สมบูรณ์ต่างๆที่เกิดขึ้นภายในตัวข้อมูลที่ได้จากการคัดเลือกในเบื้องต้น พร้อมทั้งปรับค่าที่ผิดพลาดเหล่านั้นให้อยู่ในรูปแบบที่เหมาะสม และถูกต้อง ซึ่งในการศึกษา ครั้งนี้ เราพบข้อมูลที่อาจจะเกิดความผิดพลาดขึ้นได้ ดังนั้น เราจะมีการให้ผู้ใช้ได้เลือกว่าหากมีข้อมูลหายไป จะทำอย่างไร เช่น ลบหรือแทนค่าข้อมูลที่ขาดหายไป

○ **Data Reduction**

ขั้นตอนหรือกระบวนการที่จะช่วยลดขนาดของข้อมูลให้มีจำนวนข้อมูลที่น้อยลง โดยเจาะจงเฉพาะกลุ่มที่ในสนใจ เท่านั้น สำหรับกรณีศึกษานี้เราสนใจในส่วน 2 จุดที่สำคัญคือ

1. ข้อมูลลูกค้าที่ใช้โทรศัพท์พื้นฐาน และโทรศัพท์เคลื่อนที่ เราไม่สนใจกลุ่มลูกค้าที่ใช้เพื่อต่อ Internet หรือใช้ ADSL ดังนั้นเราจึงกรองข้อมูลที่จะนำไปใช้ โดยเลือก Call Type ที่เป็นไม่ใช่อักษรภาษาอังกฤษเนื่องจากเป็นกลุ่มการใช้งานโทรศัพท์เพื่อต่อ Internet ออกไป ซึ่ง ขั้นตอนนี้ผู้ใช้ต้องทำการคัดเลือกข้อมูลมาก่อน

เมื่อลดจำนวนข้อมูลกลุ่มคงเหลือเฉพาะกลุ่มที่ให้สนใจเป็นที่เรียบร้อย เราจะเข้าสู่กระบวนการต่อไป คือการทำ Data Transformation ซึ่งจะนำข้อมูลหรือ field ที่คัดเลือกทำการเปลี่ยนแปลงแก้ไขเพื่อเอาเฉพาะข้อมูลที่มีต้องการวิเคราะห์ต่อไป

○ **Data Transformation (Coding)**

ขั้นตอนการแปลงข้อมูลนี้มี วัตถุประสงค์ เพื่อปรับรูปแบบข้อมูลให้เหมาะสม เพื่อให้สอดคล้องกับความต้องการของ Algorithm ที่แต่จริง โดย C4.5 Algorithm จะประมวลผลได้ดีกับข้อมูลที่เป็น Categorical ดังนั้นหากต้องการการประมวลให้ดียิ่งขึ้น ระบบอาจให้ช่วยในการเปลี่ยนข้อมูลบางข้อมูลที่มีลักษณะเป็น Quantitative มาเป็น Categorical ก่อนเพื่อให้การประมวลผลเป็นไปได้อย่างรวดเร็วยิ่งขึ้น

6.4 การ Mining

คือกระบวนการการสร้างทรี (Tree) โดยใช้ C4.5 Algorithm ในการแบ่ง Tree โดย

- ❖ Tree เริ่มต้นด้วย Node หนึ่งแสดงถึง ข้อมูลที่ใช้ Train ในการศึกษาครั้งนี้ คือ การ field Analyst หรือ T1 ในที่นี้จะขอเรียก T1
- ❖ หากพบว่า T1 เป็น class เดียวกับ Node ให้กำหนด Node นี้เป็น Leaf Node
- ❖ หากไม่พบ T1 ใน Class ให้ทำการแตก T1 ออกเป็น Class ซึ่ง Attribute จะกลายเป็น Attribute ที่ใช้ในการทดสอบหรือใช้เพื่อการตัดสินใจที่ Node ในแต่ละกิ่ง (Branch) ของ Tree ซึ่ง T1 จะถูกแบ่งไปตามค่า Attribute และจะมีการทำซ้ำ ณ กระบวนเดิม และจะหยุดก็ต่อเมื่อ
  - T1 ทั้งหมดอยู่ใน Class เดียวกัน
  - T1 ไม่มี Attribute เหลืออยู่อีก
  - ไม่มี T1 เหลืออยู่แล้ว

โดยใช้การวัดแบบ Entropy-based และหาค่า Gain ratio ซึ่ง ได้กล่าวไปแล้วในบทที่ 3 ซึ่งมีสูตรโดยสังเขปดังนี้

สูตร Info(S)

$$\text{Info}(S) = \left| -\sum_{I=0}^{m} p_i \log(p_i) \right|$$

และเมื่อนำสูตรมาประยุกต์ใช้กับการ Training Set จะได้ Info(T), $\text{Info}_x(T)$ เป็นการวัดค่าของ Information เพื่อแบ่ง T โดยคิดจากค่าที่เป็นไปได้ของ Attribute X

สูตร Info(T)

$$\text{Info}_x(T) = \sum_{J=1}^{n} \frac{|T_i|}{|T|} * \text{Info}(T_i)$$

Gain(X) เป็นการวัดค่าของ Information ที่ได้รับเลือก Attribute X

สูตร Gain(X)

$$\text{Gain}(X) = \text{Info}(X) - \text{Info}_x(T)$$

สูตร Split info(x)

$$\text{Split Info}(X) = \sum_{J=1}^{n} \frac{|T_i|}{|T|} \log_2 \frac{|T_i|}{|T|}$$

สูตร gain ratio(X)

$$\text{gain ratio}(X) = \text{gain}(X)/\text{split info}(X)$$

ข้อมูลที่ทำการศึกษา เช่น

ตารางที่ 6.4 ตัวอย่างข้อมูลที่ทำการศึกษา

Month	PP	Feature_cd	charge	payment
7	APCC200	UPCABK	0	pay
7	APCC200	UPCALD	0	pay
7	APCC200	UPCALO	93	Not pay
7	APCC200	UPCAMB	0	pay
9	APCC200	UPCABK	108	pay
7	APCAMKT	UPCANL	198	pay
6	APCAA20	UPCABK	326	pay
6	APCAA20	UPCALD	0	pay
6	APCAA20	UPCALO	0	Not pay
6	APCAA20	UPCAMB	21	pay
6	APCAA20	UPCAOK	0	pay
7	APCAA20	UPCABK	197	pay
6	APCAA20	UPCABK	145	pay

PP หมายถึง โปรโมชั่นที่ลูกค้าได้รับ โดยโปรโมชั่นต่างๆ มีความหมายดังนี้

APCC200 คือ เสียค่าบริการรายเดือน จ่าย 200 บาท

-โทรมือถือ นาทีละ 3 บาท

- โทรในกรุงเทพและพื้นที่ ครั้ง 3 บาท

- โทรนอกพื้นที่ นาทีละ 3 บาท

- โทรในเครือข่ายเดียวกันฟรี

APCAMKT คือ เสียค่าบริการรายเดือน 100 บาท

-โทรในมือถือ นาทีละ 3 บาท

- โทรในกรุงเทพและพื้นที่ ครั้งละ 3 บาท

- โทรนอกพื้นที่ นาทีละ 3 บาท

- โทรในเครือข่ายเดียวกันครั้งละ 3 บาท

APCAA20 คือ เสียค่าบริการรายเดือน จ่าย 200 บาท โทรฟรี 150 บาท

-โทรในมือถือ ครั้งละ 3 บาท

- โทรในกรุงเทพและพื้นที่ ครั้งละ 3 บาท

- โทรนอกพื้นที่ ครั้งละ 3 บาท

-โทรในเครือข่ายเดียวกันฟรี

APCAU30 คือ เสียค่าบริการรายเดือน จ่าย 300 บาท โทรฟรี 300 บาท

-โทรในมือถือ ครั้งละ 3 บาท

- โทรในกรุงเทพและพื้นที่ ครั้งละ 3 บาท

- โทรนอกพื้นที่ ครั้งละ 3 บาท

-โทรในเครือข่ายเดียวกันฟรี

Feature_cd คือ การใช้งานของลูกค้า ซึ่งความหมายดังต่อไปนี้

- UPCABK โทรในกรุงเทพ
- UPCALD โทรไปต่างจังหวัด
- UPCALO โทรในเขตปริมณฑล
- UPCAMB โทรเข้าเครือข่ายโทรศัทพ์เคลื่อนที่
- UPCAOK โทรในเครือข่ายเดียวกัน

ผลที่ได้จากระบบเป็นโมเดลการทำนายที่เป็น Decision Tree ซึ่งเป็น Model ที่สามารถเข้าใช้ได้ง่าย โดยเฉพาะการแสดงผลทำให้เราเห็นมุมมองที่ชัดเจนยิ่งขึ้น

1) ลูกค้ามีลักษณะการใช้ Promotion เป็นอย่างไรจงยกตัวอย่างมาสัก 3 โปรโมชั่น

2) การชำระหรือไม่ชำระค่าบริการขึ้นกับปัจจัยใดบ้าง

ผลที่ได้รับ (1) จากข้อมูลที่มีขนาด 2931 ราย ที่เป็นลูกค้าบริษัทโทรคมนาคมแห่งหนึ่ง ใน เดือน 5,6,7,8,9 ของ ปีพุทธศักราช 2548 พบว่า 100% ของลูกค้าใช้บริการโทรศัพท์พื้นฐานแบบพกพา โดย

- ลูกค้าที่เลือก promotion APCA200 ทุกเดือนประมาณ มากกว่า 50%
  - มีการ โทรออกไปยังโทรศัพท์ เคลื่อนที่
  - มีการ โทรออกไปโทรศัพท์พื้นฐานแบบพกพา
  - มีการ โทรออกไปโทรภายในกรุงเทพ
  - โดยมีค่าใช้จ่ายโดยรวมมากกว่า 0 บาท
  - และมีการจ่ายเงินตรงตามกำหนดทุกครั้ง

- ลูกค้าที่เลือก promotion APCAU30 ทุกเดือนประมาณ มากกว่า 60%
    - มีการ โทรออกไปยังโทรศัพท์ เคลื่อนที่
    - มีการ โทรออกไปโทรศัพท์พื้นฐานแบบพกพา
    - มีการ โทรออกไปโทรภายในกรุงเทพ
    - โดยมีค่าใช้จ่ายโดยรวมมากกว่า 0 บาท
    - แต่มีการจ่ายเงินตรงไม่ตามกำหนดทุกครั้ง
- ลูกค้าที่เลือก promotion APCAMKT ทุกเดือนประมาณ มากกว่า 60%
    - มีการ โทรออกไปยังโทรศัพท์ เคลื่อนที่ จำนวนน้อยมาก
    - มีการ โทรออกไปโทรศัพท์พื้นฐานแบบพกพา
    - มีการ โทรออกไปโทรภายในกรุงเทพ
    - โดยมีค่าใช้จ่ายโดยรวมมากกว่า 0 บาท
    - และมีการจ่ายเงินตรงตามกำหนดทุกครั้ง

ผลที่ได้รับ (2) จากข้อมูลที่มีขนาด 2931 ราย ที่เป็นลูกค้าบริษัทโทรคมนาคมแห่งหนึ่ง ใน เดือน 5,6,7,8,9 ของ ปีพุทธศักราช 2548 พบว่า 100% ของลูกค้าใช้บริการโทรศัพท์พื้นฐานแบบพกพา โดย

- ลูกค้าที่ทุกคนชำระเงินค่าบริการ ทั้งหมด แต่ APCAU30 ถ้ามีค่าใช้มากกว่า 25 บาทจะจ่ายช้าลง

บทที่ 7
การสรุปผลการศึกษาและข้อเสนอแนะ

โครงการพัฒนาระบบนี้ทำการพัฒนาระบบเพื่อประโยชน์ในการสนับสนุนวิเคราะห์ทำนายลักษณะการใช้งานของกลุ่มลูกค้า โดยในบทนี้จะสรุปผลการศึกษา และข้อเสนอแนะ ในการพัฒนาระบบให้มีประสิทธิภาพ และตอบสนองต่อความต้องการในการใช้ระบบให้มากที่สุด

7.1 สรุปผลการดำเนินงาน

ระบบการวิเคราะห์การจัดแบ่งกลุ่มลูกค้า เป็นระบบที่อยู่ในรูปแบบแอปพลิเคชั่น บนวินโดวส์ โดยระบบจะทำการสร้างโมเดล เริ่มจากการติดต่อกับฐานข้อมูล เพื่อเลือกข้อมูลที่ทำการวิเคราะห์ผ่านคำสั่ง SQL ซึ่งสามารถรองรับข้อมูลที่เป็นทั้ง categorical และ numerical และยังสามารถจัดการับค่าที่สูญหาย ตามชนิดของข้อมูลได้ และยังนำข้อมูลมาสร้างโมเดลเพื่อวิเคราะห์ได้

ระบบสร้างโมเดลโดยให้ C4.5 อัลกอริทึม ซึ่งใช้เป็นเทคนิคหนึ่งการแบ่งกลุ่ม (Classification) แบบดิซิซั่นทรีและได้นำมาใช้ในการแบ่งกลุ่มการใช้งานของลูกค้า คือ การใช้งานปริมาณมาก น้อย หรือปานกลาง ซึ่ง C4.5 อัลกอริทึม มีความสามารถในการรองรับกลุ่มของข้อมูลที่มีขนาดใหญ่ได้เป็นอย่างดี

7.2 ข้อเสนอแนะ

ดังที่กล่าวมาแล้ว โครงการัฒนาระบบนี้เพื่อเข้าใจถึงการใช้งานของลูกค้าให้มากขึ้น ดังนั้นระบบนี้สามารถนำไปพัฒนาให้มีประสิทธิภาพและตรงกับความต้องการมากขึ้น โดยผู้พัฒนามีข้อเสนอแนะ ให้พัฒนาระบบนี้ ดังนี้

- ❖ ข้อมูลที่ได้จากการวิเคราะห์ ควรจะเก็บในรูปแบบรูปภาพ เพื่อป้องการแก้ไขและไม่ต้อง process ใหม่ทุกครั้ง
- ❖ พัฒนาให้รองรับระบบไฟล์ข้อมูล เพื่อการประมวลได้หลายรูปแบบมากขึ้น

นอกเหนือ จากการพัฒนาระบบ ให้ประสิทธิภาพมากขึ้นแล้วระบบนี้มีความยืดหยุ่นมาก สามารถวิเคราะห์ข้อมูลที่หลากหลาย ดังนั้นระบบนี้สามารถนำไปศึกษา เพื่อทำนาย ลักษณะการใช้งานของลูกค้า หรือ กรณีอื่นๆ ต่อไป

บรรณานุกรม

Geary ,David. 2001. **Java Mastering the JFC.** Sun Microsystems.

Walton,Edward. 2001. **Data Generaion For Machine Learning Techniques.** University of Bristol.

DBMS, Data Mining Solutions Supplement. 1998. [Online]. Available: http://www.dbmsmag.com .

Witten. 1999. **Data Mining Practical Machine Learning Tools and Techniques with Java Implementations** . England : Morgran Kaufmann Publishers.

Vriens ,Marco. 2001. **Market Segmentation Analytical Developments and Application Guidelines.** [Online]. Available: http://www.citoseer.com/THD=1141545465.

Groth ,Robert. 1997. **Data Mining** . New Jersey :A Simoon &Company.

Yang ,Yinghui (Catherine) and Padmanabhan ,Balaji. January 11, 2004. **Data Mining for Customer Segmentation: A Behavioral Pattern-Based Approach1.** [Online]. Available: www.empower.com/market-segmentation.htm .

ประวัติผู้เขียน

ชื่อ นามสกุล นางสาว อัจฉรา ประเสริฐ

วัน เดือน ปี 6 สิงหาคม พ.ศ.2523

สถานที่เกิด จังหวัดกรุงเทพมหานคร

วุฒิการศึกษา ปริญญาตรี วิทยาศาสตร์บัณฑิต สาขาวิทยาการคอมพิวเตอร์ประยุกต์ คณะวิทยาศาสตร์ มหาวิทยาลัย ศรีนครินทรวิโรฒ