ตัวแยกแยะ SLIQ เพื่อการทำดาต้าไมนิ่ง :
กรณีศึกษาการวิเคราะห์ข้อมูลการใช้งานโทรศัพท์เคลื่อนที่

SLIQ Classifier for Data Mining :
A Case Study of Mobile Phone Usage Analysis

โดย

วุฒิไกร มะลิลา

รหัส 46066339

อาจารย์ที่ปรึกษา

ผศ.ดร. ภัทรชัย ลลิตโรจน์วงศ์

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน

หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

ภาคเรียนที่ 2 ปีการศึกษา 2547

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ชื่อหัวข้อ	ตัวแยกแยะ SLIQ เพื่อการทำดาต้าไมนิ่ง : กรณีศึกษาการวิเคราะห์ข้อมูลการใช้งานโทรศัพท์เคลื่อนที่
นักศึกษา	นายวุฒิไกร มะลิลา
อาจารย์ที่ปรึกษา	ผศ.ดร. ภัทรชัย ลลิตโรจน์วงศ์
ระดับการศึกษา	วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2547

บทคัดย่อ

ตัวแยกแยะ SLIQ เพื่อการทำดาต้าไมนิ่ง เป็นเครื่องมือที่ช่วยให้ธุรกิจทางด้านโทรคมนาคม โดยเฉพาะผู้ให้บริการโทรศัพท์เคลื่อนที่ สามารถวิเคราะห์ข้อมูลการใช้งานโทรศัพท์เคลื่อนที่ของลูกค้าที่ใช้บริการอยู่ในปัจจุบัน และลูกค้าที่ยกเลิกการใช้บริการไปแล้ว มาศึกษาและเรียนรู้พฤติกรรมการใช้งานโทรศัพท์ของลูกค้าในกลุ่มต่างๆ เพื่อนำเสนอแคมเปญที่เหมาะสมกับลูกค้าในแต่ละกลุ่มได้อย่างถูกต้อง ซึ่งจะเป็นการรักษาฐานลูกค้าเก่า และขยายฐานกลุ่มลูกค้าใหม่ให้เพิ่มมากขึ้นได้อย่างถูกกลุ่มเป้าหมาย

การวิเคราะห์ข้อมูลการใช้งานโทรศัพท์เคลื่อนที่ ใช้ข้อมูลที่เกี่ยวข้องกับโปรโมชั่น อายุของลูกค้า ค่าใช้จ่ายต่อเดือน ระยะเวลาในการใช้โทรศัพท์ และจำนวนครั้งในการโทรศัพท์ เป็นต้น โดยใช้หลักการของต้นไม้การตัดสินใจ และอัลกอริทึม SLIQ การพัฒนาระบบได้นำวิชวลเบสิค เวอร์ชั่น 6 มาเป็นภาษาในการพัฒนา และใช้ Microsoft SQL Server 2000 เป็นฐานข้อมูล

Title	SLIQ Classifier for Data Mining : A Case Study of Mobile Phone Usage Analysis
Student	Mr. Wuttikrai Malila
Advisor	Asst.Prof.Dr. Pattarachai Lalitrojwong
Level of Study	Master of Science in Information Technology
Major	Information Science
Academic Year	2004

ABSTRACT

The SLIQ classifier for data mining is the tools to analyze the telecommunication information especially mobile phone service or Call Detail Record (CDR). This tools use to analysis customer usage and churn customer to study customer usage behavior in each customer segment. The result of this process can lead to initiate the proper campaign. The new proper campaign can keep the existing customers and gain the new customers.

The usage information analyzing use information such as customer promotion, age of customer, fix payment amount, usage period per call and number of call to analyze. Decision tree and SLIQ algorithm is used to develop the system. The language to develop this system is Visual Basic version 6 and database use Microsoft SQL Server 2000

กิตติกรรมประกาศ

การพัฒนาระบบงานในครั้งนี้ ผู้จัดทำขอขอบพระคุณ ผศ.ดร. ภัทรชัย ลลิตโรจน์วงศ์ เป็นอย่างสูง ที่ได้ให้คำปรึกษาให้ความรู้ และให้คำแนะนำในการพัฒนาระบบงานเป็นอย่างดียิ่ง ขอขอบพระคุณ ดร.ธนารัตน์ ชลิดาพงศ์ ที่ได้สอนให้ผู้จัดทำสามารถนำวิชา Information System Development มาใช้กับการพัฒนาระบบได้อย่างถูกต้องและสนุกสนาน ขอบขอบคุณเพื่อนๆ ห้อง IS 15.2 ทุกคนที่คอยเป็นแรงกระตุ้น และแรงใจให้ผู้จัดทำมีความกระตือรือร้นในการพัฒนาระบบอยู่เสมอ ขอขอบคุณพี่ๆ เพื่อนๆ ทุกคนที่บริษัท โทเทิ่ล แอ็คเซ็ส คอมมูนิเคชั่น จำกัด (มหาชน) ที่ให้คำแนะนำ ให้ความช่วยเหลือ และให้คำติชมต่างๆ ท้ายที่สุดนี้ขอขอบพระคุณ คุณพ่อ คุณแม่ ที่เป็นแรงใจและพลังใจที่ดีที่สุด ที่ทำให้ผู้จัดทำมีวันนี้ได้อย่างภาคภูมิใจ

วุฒิไกร มะลิลา

กุมภาพันธ์ 2548

สารบัญ

สารบัญ(ต่อ)

หน้า

สารบัญตาราง

หน้า

ตารางที่

สารบัญรูป

หน้า

รูปที่

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันธุรกิจการสื่อสารโทรคมนาคม โดยเฉพาะธุรกิจผู้ให้บริการโทรศัพท์เคลื่อนที่ มีสภาวะการแข่งขันที่สูงมาก เพื่อให้ประสบความสำเร็จทางธุรกิจผู้ให้บริการรายต่างๆ จึงจำเป็นต้องมีกลยุทธ์ที่เชื่อมั่นได้ว่าจะช่วยเพิ่มรายได้ สร้างความพึงพอใจให้กับลูกค้า และลดความเสี่ยงของธุรกิจ ซึ่งกลยุทธ์หรือวิธีการต่างๆ จำเป็นต้องมีฐานความรู้เพื่อใช้ในการสร้างกรอบการทำงาน ที่ช่วยตอบสนองกับกลยุทธ์ทางธุรกิจ การที่จะได้มาซึ่งฐานความรู้และกรอบการทำงานที่มีประโยชน์จำเป็นต้องมีเทคโนโลยีสารสนเทศที่สามารถวิเคราะห์ข้อมูลทางธุรกิจ สามารถกลั่นกรองแยกแยะข้อมูลทางธุรกิจที่มีปริมาณมหาศาล เพื่อให้ได้ข้อมูลที่มีประโยชน์ในการพัฒนา แก้ไขและปรับปรุงกลยุทธ์ ดังนั้นในขณะนี้ดาต้าไมนิ่ง จึงเป็นเทคโนโลยีสารสนเทศ ที่ได้รับการกล่าวถึงมากที่สุด นั่นก็เพราะว่าดาต้าไมนิ่งเป็นเทคโนโลยีสารสนเทศ ที่สามารถกลั่นกรอง แยกแยะ วิเคราะห์ข้อมูล ที่มีปริมาณมหาศาล เพื่อให้ได้ข้อมูลที่มีประโยชน์หรือได้ข้อมูลที่ซ่อนเร้น และนำข้อมูลที่มีประโยชน์นั้น มาใช้เป็นฐานความรู้ เพื่อช่วยบริหารงาน เช่น การบริหารความสัมพันธ์ลูกค้า

การจัดการแคมเปญ (Campaign Management) คือ การนำเสนอรายการส่งเสริมการขายหรือการออกโปรโมชั่นใหม่ให้กับผู้ใช้บริการโทรศัพท์เคลื่อนที่ เพื่อให้แคมเปญที่นำเสนอออกมานั้นมีประสิทธิภาพสูงสุด สามารถประยุกต์ใช้ได้อย่างทันสถานการณ์ และสามารถช่วยเพิ่มยอดขายได้เป็นอย่างดี การจัดการในเรื่องนี้จึงจำเป็นที่จะต้องมีข้อมูลของลูกค้า และเทคนิคที่จะนำมาใช้ในการวิเคราะห์ ซึ่งดาต้าไมนิ่งเป็นวิธีการหนึ่งที่เหมาะสมในการวิเคราะห์ในเรื่องนี้

โครงงานนี้จะกล่าวถึงตัวแยกแยะ SLIQ เพื่อการทำดาต้าไมนิ่ง โดยจะยกกรณีศึกษาในเรื่องการวิเคราะห์ข้อมูลการใช้งานโทรศัพท์เคลื่อนที่ เพื่อนำสารสนเทศที่ซ่อนเร้นอยู่มาศึกษา แล้วเสนอรายการส่งเสริมการขาย หรือการออกโปรโมชั่นใหม่ที่เหมาะกับผู้ใช้บริการโทรศัพท์เคลื่อนที่ในช่วงเวลาต่างๆ โดยใช้ต้นไม้การตัดสินใจและนำหลักการของอัลกอริทึม SLIQ มาเป็นแนวทางในการพัฒนาระบบ ซึ่งระบบนี้จะช่วยสนับสนุนข้อมูลในด้านการตัดสินใจให้กับผู้บริหาร เพื่อใช้ในการวิเคราะห์การออกโปรโมชั่น โดยผลลัพธ์ที่ได้จากการวิเคราะห์จะแสดงในรูปของแบบจำลองต้นไม้

1.2 วัตถุประสงค์ของโครงการ

1. เพื่อศึกษาและทำความเข้าใจเกี่ยวกับแนวคิดและขั้นตอนการทำดาต้าไมนิ่ง สำหรับข้อมูลทางธุรกิจโทรคมนาคม
2. เพื่อศึกษาและทำความเข้าใจในหลักการตลาดเบื้องต้น เพื่อเป็นความรู้ในการทำงานและเปิดวิสัยทัศน์ให้กว้างไกลยิ่งขึ้น
3. เพื่อศึกษาและทำความเข้าใจเกี่ยวกับการใช้แนวคิดของการแยกแยะ โดยนำหลักการของต้นไม้การตัดสินใจและใช้อัลกอริทึม SLIQ (Supervised Learning in Quest)
4. เพื่อศึกษาถึงแนวทางและความเป็นไปได้ ในการนำแนวคิดของต้นไม้การตัดสินใจมาใช้ในการวิเคราะห์ข้อมูลการใช้งานโทรศัพท์เคลื่อนที่ เพื่อนำเสนอรายการส่งเสริมการขายหรือการออกโปรโมชั่น ว่าสามารถนำมาวิเคราะห์ข้อมูลได้จริงและมีประสิทธิภาพหรือไม่
5. เพื่อเป็นแนวทางในการประยุกต์ใช้ข้อมูลจากการทำดาต้าไมนิ่ง มาสนับสนุนหรือประกอบการตัดสินใจให้ผู้บริหาร ในการวางแผนกลยุทธ์ทางด้านการตลาด

1.3 ขอบเขตของการพัฒนาแอพพลิเคชัน

1. หน้าที่การทำงานของโปรแกรม

โปรแกรมวิเคราะห์ข้อมูลการใช้งานโทรศัพท์เคลื่อนที่ เพื่อนำเสนอรายการส่งเสริมการขาย โดยใช้ต้นไม้การตัดสินใจมีขอบเขตการทำงานหลักดังต่อไปนี้

- ระบบสามารถให้ผู้ใช้นำข้อมูลที่จะนำมาวิเคราะห์ได้ 2 ทาง คือ จากฐานข้อมูลที่ผู้ใช้ระบบมีอยู่แล้ว ซึ่งในระบบนี้จะกำหนดให้ใช้ได้เฉพาะฐานข้อมูลที่มาจาก SQL Server 2000 เท่านั้น และจากการโหลดข้อมูลจากไฟล์ข้อมูล ซึ่งมีรูปแบบตามที่ผู้พัฒนาระบบกำหนดไว้เท่านั้น
- ระบบจะวิเคราะห์ข้อมูลโดยใช้แนวคิดการแยกแยะ โดยนำหลักการของต้นไม้การตัดสินใจและใช้อัลกอริทึม SLIQ การแสดงผลลัพธ์จะแสดงในรูปของแบบจำลองต้นไม้เท่านั้น
- ในส่วนของการทดสอบผลของการวิเคราะห์แบบจำลองต้นไม้ ใช้แนวคิดดังนี้ คือ แบ่งข้อมูลออกเป็น 2 ส่วน ข้อมูลส่วนแรกเรียกว่าข้อมูลสำหรับฝึกอบรมใช้ในการสร้างแบบจำลองต้นไม้หรือแบจำลอง และข้อมูลส่วนที่สองเรียกว่าข้อมูลสำหรับการทดสอบ ใช้สำหรับการทดสอบแบบจำลองต้นไม้ที่ระบบสร้างขึ้น

2. ข้อมูลที่นำมาวิเคราะห์

ข้อมูลที่นำมาใช้การพัฒนาระบบในวิเคราะห์ข้อมูลในโครงการนี้ ยกกรณีศึกษาการวิเคราะห์ข้อมูลการใช้งานโทรศัพท์เคลื่อนที่ เพื่อนำเสนอรายการส่งเสริมการขายโดยใช้ต้นไม้การตัดสินใจ เป็นข้อมูลที่ได้จากหน่วยงานของฝ่ายบริหารความสัมพันธ์ลูกค้า ซึ่งข้อมูลเหล่านี้ได้มาจากคลังอีกทีหนึ่ง ข้อมูลชุดนี้ได้ที่ผ่านการทำความสะอาด และพร้อมใช้อยู่เสมอ มีรายละเอียดดังนี้

- ข้อมูลที่ใช้เป็นกรณีศึกษาการวิเคราะห์ข้อมูลการใช้งานโทรศัพท์เคลื่อนที่ เป็นข้อมูลจริงของลูกค้าในองค์กร ซึ่งข้อมูลที่นำมาวิเคราะห์เป็นข้อมูลของลูกค้าที่เปิดบริการในระบบแบบใช้ก่อนจ่ายเงินทีหลัง ระหว่างวันที่ 1 กรกฎาคม ถึง 30 กันยายน พ.ศ. 2547 จำนวน 1,500 เรคอร์ด
- ข้อมูลที่นำมาใช้ในการวิเคราะห์นั้น เป็นข้อมูลที่ผ่านการเตรียมข้อมูลมาแล้วตามหลักการดาต้าไมนิ่ง

1.4 ขั้นตอนการพัฒนาระบบ

1. การกำหนดระยะเวลาในการดำเนินงาน

- วางแผนในการดำเนินงาน โดยกำหนดระยะเวลาในการพัฒนาระบบ ตามเป็นความจริงที่สามารถทำได้ตามแผนงานที่วางไว้ดังนี้

ตารางที่ 1.1 ตารางการทำงาน

	Task Name	Duration	Start	Finish
1	**Mobile Phone Usage Mining Project**	**92 days**	**Wed 11/10/04**	**Thu 3/17/05**
2	Send proposal	2 days	Wed 11/10/04	Thu 11/11/04
3	Analysis and design program	5 days	Mon 11/15/04	Fri 11/19/04
4	Design interface	5 days	Mon 11/22/04	Fri 11/26/04
5	Design database	5 days	Mon 11/29/04	Fri 12/3/04
6	Prepare environment	6 days	Mon 12/6/04	Mon 12/13/04
7	Coding phase load data	10 days	Wed 12/8/04	Tue 12/21/04
8	Develop progress document (2 copies)	5 days	Wed 12/22/04	Tue 12/28/04
9	Code phase algorithm and testing	30 days	Fri 12/31/04	Thu 2/10/05
10	Develop original document (4 copies)	5 days	Fri 2/11/05	Thu 2/17/05
11	Presentation	4 days	Mon 3/7/05	Thu 3/10/05
12	Correct complete document	5 days	Fri 3/11/05	Thu 3/17/05

2. **การออกแบบระบบ**
   - ศึกษาทฤษฎีและความต้องการของระบบ โดยอ้างอิงจากทฤษฎีดาต้าไมนิ่ง แนวคิดการแยกแยะ โดยนำหลักการของต้นไม้การตัดสินใจและใช้อัลกอริทึม SLIQ เพื่อวิเคราะห์หน้าที่การทำงานหลักที่เกี่ยวข้องกับการวิเคราะห์ข้อมูล และหาขอบเขตของระบบที่จะพัฒนาขึ้น
   - กำหนดเครื่องมือและทรัพยากรที่ใช้ในการพัฒนาระบบ
   - ออกแบบหน้าจอของระบบ
   - ออกแบบฐานข้อมูล
   - ออกแบบโครงสร้างของโปรแกรม ซึ่งแบ่งออกเป็น 3 ส่วนดังนี้ ส่วนแรกเป็นการนำเข้าข้อมูลระบบ ส่วนที่สองเป็นการวิเคราะห์ข้อมูล และส่วนที่สามเป็นการทดสอบผลการวิเคราะห์
3. **การเขียนโปรแกรมและทดสอบระบบ**
   - เขียนโปรแกรมตามที่ได้ออกแบบไว้
   - ทดสอบการทำงานของระบบที่ได้พัฒนาขึ้น โดยแบ่งการทดสอบออกเป็น 3 ส่วนดังนี้
     - ทดสอบการทำงานส่วนย่อยของโปรแกรม ว่าสามารถทำงานได้ถูกต้องตามที่ออกแบบไว้หรือไม่
     - ทดสอบการทำงานร่วมกันของแต่ละฟังก์ชั่น ว่าสามารถทำงานร่วมกันได้สอดคล้องถูกต้องหรือไม่
     - ทดสอบภาพรวมของระบบ ว่าสามารถทำงานได้ครบถ้วนถูกต้องตรงกับความต้องการหรือไม่
4. **การทำเอกสารประกอบระบบ**
   - ทำเอกสารประกอบการออกแบบระบบ
   - ทำเอกสารคู่มือการใช้งานระบบ
5. **การติดตั้งระบบ**
   - ติดตั้งระบบ
   - บำรุงรักษาและดูแลระบบ
   - ปรับปรุงแก้ไขระบบตามกำหนดเวลา

1.5 เครื่องมือที่ใช้ในการพัฒนาระบบ

1. เครื่องมือที่ใช้ในการเขียนโปรแกรม คือ Microsoft Visual Basic Version 6
2. ระบบฐานข้อมูลคือ Microsoft SQL Server 2000
3. เครื่องคอมพิวเตอร์ที่นำมาใช้ในการพัฒนา คือ Microsoft Windows XP Professional 2002 Pentium III, CPU 1.8 GHz, RAM 256 MB

1.6 ประโยชน์ที่คาดว่าจะได้รับ

1. ผู้พัฒนาระบบมีความรู้ความเข้าใจในหลักการของดาต้าไมนิ่งแบบการแยกแยะ โดยใช้ต้นไม้การตัดสินใจและอัลกอริทึม SLIQ มากยิ่งขึ้น
2. ผู้ใช้ระบบสามารถนำผลการวิเคราะห์ไปใช้ในการวางกลยุทธ์ทางด้านการตลาดได้
3. สามารถนำระบบที่จัดทำขึ้น มาช่วยเพิ่มประสิทธิภาพในการทำงานขององค์กร
4. เป็นกรณีศึกษาเพื่อเป็นแนวทางในการพัฒนาระบบงานอื่นๆ ต่อไป

บทที่ 2
ดาต้าไมนิ่ง

เทคนิคการแยกแยะประเภทข้อมูล เป็นเทคนิคหนึ่งที่สำคัญของการสืบค้นความรู้เกี่ยวกับลูกค้า หรือข้อมูลที่น่าสนใจบนฐานข้อมูลขนาดใหญ่ หรือดาต้าไมนิ่ง (Berry and Linoff. 2000; Berson et al. 1999) เทคนิคการแยกแยะประเภทของข้อมูล เป็นกระบวนการสร้างโมเดลสำหรับจัดการข้อมูล ให้อยู่ในกลุ่มที่กำหนดมาให้ จากกลุ่มข้อมูลตัวอย่าง ที่เรียกว่าข้อมูลสอนระบบ ที่แต่ละแถวของข้อมูลประกอบด้วยแอตทริบิวต์ หรือฟิลด์จำนวนมาก แอตทริบิวต์นี้จะอาจเป็นค่าต่อเนื่องหรือค่ากลุ่มก็ได้ โดยจะมีแอตทริบิวต์ที่ใช้แบ่งข้อมูล ซึ่งเป็นตัวบ่งชี้คลาสของข้อมูล จุดประสงค์ของการแยกแยะประเภทข้อมูล คือ การสร้างโมเดลของการแยกแอตทริบิวต์หนึ่ง โดยขึ้นอยู่กับแอตทริบิวต์อื่น โมเดลที่ได้จากการแยกแยะประเภทข้อมูลนี้จะทำให้สามารถพิจารณาคลาสจากข้อมูลที่ยังมิได้แบ่งกลุ่มในอนาคตได้ เทคนิคการแยกแยะประเภทข้อมูล สามารถนำไปประยุกต์ใช้ในงานหลายด้าน เช่น การป้องกันการฉ้อโกง การจัดกลุ่มลูกค้าทางการตลาด เป็นต้น

2.1 ความหมายของดาต้าไมนิ่ง

ดาต้าไมนิ่ง คือ กระบวนการทำงาน ที่ค้นหาความรู้ที่น่าสนใจ ไม่ว่าจะเป็นรูปแบบความสัมพันธ์ การเปลี่ยนแปลง ความผิดปกติ และสารสนเทศที่มีนัยสำคัญจากฐานข้อมูลที่มีปริมาณข้อมูลขนาดใหญ่ โดยการสกัดข้อมูลจากฐานข้อมูล เพื่อให้ได้สารสนเทศที่ยังไม่รู้ โดยเป็นสารสนเทศที่มีเหตุผล และสามารถนำไปใช้งานได้ เพื่อเป็นข้อมูลที่ช่วยในการตัดสินใจ ดาต้าไมนิ่งอาจเรียกอีกอย่างหนึ่งได้ว่า Knowledge Discovery in Databases (KDD) (Piatetsky-Shapiro and Frawley. 2001)

2.2 ขั้นตอนการทำดาต้าไมนิ่ง

การทำดาต้าไมนิ่ง เป็นกระบวนการสร้างแบบจำลองของกลุ่มข้อมูล เพื่อให้ได้ข้อมูลที่สามารถนำไปใช้งานได้จริงและเป็นประโยชน์ มี 5 ขั้นตอนดังนี้ (Han and Kamber. 2001)

2.2.1 การกำหนดวัตถุประสงค์ทางธุรกิจ

การกำหนดวัตถุประสงค์ทางธุรกิจ (Business Objective Determination) คือ ขั้นตอนที่มีความสำคัญที่สุด เนื่องจากเป็นการกำหนดขอบเขตและเป้าหมาย ซึ่งจะมีผลต่อทุกๆ ขั้นตอนของ

การทำดาต้าไมนิ่ง โดยนักวิเคราะห์ธุรกิจจะต้องระบุความต้องการ หรือปัญหาที่เกิดขึ้นในการทำธุรกิจให้ครอบคลุมและชัดเจน รวมทั้งวัตถุประสงค์ของการทำดาต้าไมนิ่งด้วย

2.2.2 การเตรียมข้อมูล

การเตรียมข้อมูล (Data Preparation) คือ ขั้นตอนที่ใช้เวลานานที่สุด เนื่องจากต้องพิจารณาเกี่ยวกับข้อมูล และวัตถุประสงค์ในการวิเคราะห์ ชนิด ประเภท จำนวนและอายุของข้อมูล หน้าที่ของขั้นตอนนี้ คือ การจัดการข้อมูลให้สามารถนำเข้าสู่อัลกอริทึมของดาต้าไมนิ่งได้ ซึ่งต้องคัดเลือกข้อมูลที่เหมาะสม อาจมีการปรับเปลี่ยนรูปแบบหรือแปลงข้อมูลเพื่อให้เหมาะสมกับอัลกอริทึมที่ใช้และอยู่ในประเด็นที่ต้องการด้วย เช่น การทำความสะอาดข้อมูล การปรับข้อมูล และการลดขนาดข้อมูล ขั้นตอนในการเตรียมข้อมูลเพื่อที่จะทำให้ข้อมูลมีคุณภาพ มีขั้นตอนย่อย 4 ขั้นตอนดังนี้

1. การคัดเลือกข้อมูล

การคัดเลือกข้อมูล (Data Selection) เป็นการคัดเลือกชุดของข้อมูลที่ต้องการใช้จากฐานข้อมูลที่สัมพันธ์กับวัตถุประสงค์ในการวิเคราะห์ ซึ่งข้อมูลนี้จะต้องเป็นข้อมูลที่มีความสมบูรณ์ และถูกต้องมากที่สุด โดยกำหนดรูปแบบข้อมูลที่ต้องการ ระบุลักษณะข้อมูล และนำข้อมูลที่ไม่ต้องการออกไป การเลือกจะต้องคำนึงถึงอายุของข้อมูลด้วย ซึ่งข้อมูลอาจไม่ได้อยู่บนแหล่งข้อมูลเดียวกัน จำเป็นต้องดึงข้อมูลจากแหล่งอื่นมาประกอบใช้ร่วมกัน สามารถแบ่งประเภทของข้อมูลได้เป็น 2 ประเภท คือ

1. ข้อมูลตัวเลข (Quantitative Data) แบ่งเป็น 2 ประเภท
   - เลขไม่ต่อเนื่อง (Discrete) คือ ตัวเลขจำนวนเต็ม เช่น รหัสนักศึกษา รหัสบัตรประจำตัวประชาชน หรือรหัสพนักงาน
   - เลขต่อเนื่อง (Continuous) คือ ตัวเลขจำนวนจริง เช่น รายได้ เกรดเฉลี่ย หรืออัตราภาษี
2. ข้อมูลที่ไม่ใช่ตัวเลข (Categorical Data) แบ่งเป็น 2 ประเภท
   - ข้อมูลที่มีลำดับความสำคัญ (Ordinal) เช่น เกรด (A, B, C, D, F)
   - ข้อมูลที่ไม่มีลำดับความสำคัญ (Nominal) เช่น ชื่อ-นามสกุล เพศ และอาชีพ

2. การประมวลผลก่อนข้อมูล

การประมวลผลก่อนข้อมูล (Data Preprocessing) เป็นการทำให้ข้อมูลครบถ้วนสมบูรณ์ และเลือกข้อมูลที่สำคัญที่คาดว่าจะนำมาใช้ประโยชน์ได้ เป็นการกลั่นกรองข้อมูลให้เหมาะสมก่อนที่จะนำไปทำดาต้าไมนิ่ง ซึ่งข้อมูลที่เลือกมานั้นอาจมีข้อมูลที่มีความผิดพลาด หรือค่าของข้อมูลขาดหายไป จึงต้องใช้หลักการทางสถิติมาช่วยเพิ่มความถูกต้องให้กับข้อมูล เช่น ข้อมูลที่

ไม่ใช่ตัวเลข จะใช้การจัดการกระจายของข้อมูล หรือการนำข้อมูลมาสร้างกราฟ เพื่อช่วยให้เห็นความโน้มเอียงของข้อมูล และข้อมูลที่ผิดปกติได้ ส่วนข้อมูลประเภทที่เป็นตัวเลข สามารถวิเคราะห์ได้โดยการหาค่าสูงสุด ค่าต่ำสุด ค่าเฉลี่ย หรือค่าที่ปรากฏบ่อย เป็นต้น ซึ่งสิ่งที่จะแสดงให้เห็น คือ

1. ค่าของข้อมูลที่ผิดปกติ (Noisy Data) คือ ค่าของข้อมูลที่มีลักษณะแตกต่างไปจากค่าข้อมูลที่เป็นไปได้ ซึ่งอาจจะเกิดจากเป็นค่าที่เกิดขึ้นจริงๆ หรือเป็นความผิดพลาดในการบันทึกข้อมูล
2. ข้อมูลขาดหายไป (Miss Value) คือ มีข้อมูลบางส่วนขาดหายไป แก้ไขได้โดยการตัดข้อมูลรายการนั้นทิ้งไปทั้งรายการ หรือแทนส่วนที่หายไปด้วยค่าเฉลี่ย หรือค่าที่ปรากฏบ่อย หรือบันทึกเป็น “Unknown”

3. การแปลงข้อมูล

การแปลงข้อมูล (Data Transformation) ให้อยู่ในรูปแบบที่เหมาะสมกับโมเดล ที่จะใช้ในการทำดาต้าไมนิ่ง คือ ขั้นตอนที่ข้อมูลจะถูกเปลี่ยนรูป และรวบรวมให้อยู่ในรูปแบบที่เหมาะสมแก่การทำดาต้าไมนิ่งซึ่งมีแนวคิดการทำงานดังนี้

1. Smoothing คือ การเอาข้อมูลที่ผิดไปจากค่าที่ควรจะเป็นออกจากข้อมูล ซึ่งวิธีที่ใช้คือ การจัดกลุ่ม, Binning, Regression
2. Aggregation คือ วิธีในการรวมหรือสรุปข้อมูล เช่น ข้อมูลการขายประจำวัน อาจจะทำสรุปข้อมูลเป็นรายปี หรือรายเดือน
3. Generalization คือ การแปลงข้อมูลดิบหรือข้อมูลที่อยู่ในระดับต่ำให้อยู่ในระดับที่สูงกว่าตามลำดับขั้น เช่น ฟิลด์บ้านเลขที่ เราอาจจะนำข้อมูลไปรวมอยู่ในฟิลด์ที่อยู่ และฟิลด์เกรด เราอาจจะแบ่งฟิลด์เกรดออกเป็น เกรดต่ำ เกรดปานกลาง และเกรดสูง
4. Normalization คือ การกำหนดช่วงของค่าให้แคบลง เช่น กำหนดให้ข้อมูลอยู่ในช่วง 0.0 – 1.0
5. Attribute Construction คือ การสร้างแอตทริบิวต์ใหม่เพื่อทำให้ข้อมูลมีความถูกต้องและมีความเข้าใจมากยิ่งขึ้น เช่น การสร้างแอตทริบิวต์ขึ้นมาใหม่ เช่น Age ได้มาจากการนำแอตทริบิวต์ Now Date ลบกับแอตทริบิวต์ Birth Date เป็นต้น

2.2.3 การทำดาต้าไมนิ่ง

การทำดาต้าไมนิ่ง (Data Mining) เป็นขั้นตอนการทำไมนิ่งข้อมูล โดยสามารถดำเนินการทำดาต้าไมนิ่งได้หลายรูปแบบ เช่น การแบ่งส่วนฐานข้อมูล (Database Segmentation),การจำลองแบบพยากรณ์ (Predictive Modeling), การวิเคราะห์ความเชื่อมโยง (Link Analysis) เป็นต้น แต่ละวิธีดำเนินการของดาต้าไมนิ่ง จะมีอัลกอริทึมให้เลือกใช้ เช่น ถ้าเป็นการจำลองแบบพยากรณ์ อาจใช้

SLIQ (Supervised Learning in Quest), CART (Classification And Regression Tree) หรืออาจใช้เครือข่ายประสาทเทียมที่มีการเรียน (Supervised Learning Neural Network) เช่น Backpropagation Neural Net

2.2.4 การวิเคราะห์ผลลัพธ์

การวิเคราะห์ผลลัพธ์ (Analysis of Results) ผลลัพธ์ที่ได้เป็นเพียงข้อมูลที่ช่วยให้มองเห็นรูปแบบของข้อมูลที่มีอยู่ว่าเป็นอย่างไร ขั้นตอนนี้จะทำการเก็บผลลัพธ์ของดาต้าไมนิ่ง และสรุปความหมายของผลลัพธ์ที่ได้ ทำการประเมินคำตอบ และวิเคราะห์ผลว่าได้ตามต้องการหรือผิดพลาดอะไรบ้าง ซึ่งจะเป็นข้อมูลความรู้นำไปเป็นสารสนเทศที่ช่วยในการตัดสินใจ

2.2.5 การนำเสนอความรู้

การนำเสนอความรู้ (Assimilation of Knowledge) เป็นการนำความรู้ใหม่ๆ ที่ได้มาใช้ให้เป็นประโยชน์ต่อการดำเนินธุรกิจ หรือช่วยในการตัดสินใจ การนำข้อมูลที่ได้ไปใช้ประโยชน์ โดยรวบรวมจากความเข้าใจทางธุรกิจ และผลจากการทำดาต้าไมนิ่ง ซึ่งเกี่ยวข้องกับวัตถุประสงค์ของการดำเนินธุรกิจ จะต้องพิจารณาหลักสำคัญ 2 ประการ คือ

1. นำเสนอถึงแนวคิดใหม่ทางธุรกิจที่ค้นพบ
2. หาแนวทางในการใช้ความรู้ใหม่ที่ค้นพบ เพื่อก่อให้เกิดประโยชน์สูงสุด

2.3 เทคนิคของดาต้าไมนิ่ง

ดาต้าไมนิ่ง มีเทคนิคและอัลกอริทึมที่สามารถนำมาใช้งานให้ประสบความสำเร็จได้ มีอยู่หลายประเภทด้วยกัน ขึ้นอยู่กับรูปแบบของแอพพลิเคชั่น ที่ต้องการนำมาใช้งาน ซึ่งสามารถแบ่งออกเป็นรูปแบบต่างๆ ดังตารางที่ 2.1 (Han and kamber. 2001)

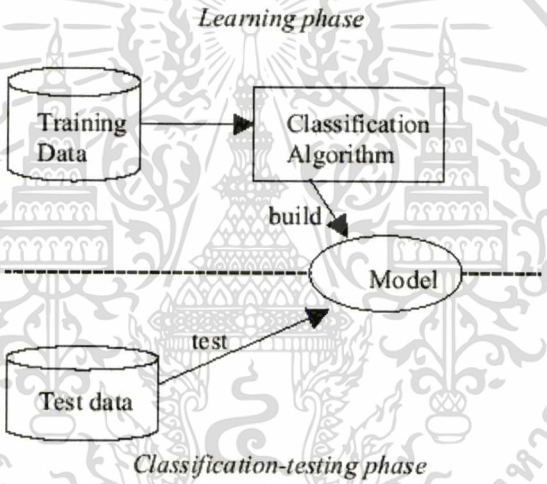
2.3.1 การจำลองแบบการทำนาย

การจำลองแบบการทำนาย (Predictive Modeling) เป็นการคาดคะเน ทำนายถึงความเป็นไปได้ โดยใช้การสังเกต จากรูปแบบของข้อมูลที่มีอยู่ โดยมีลักษณะเป็นการเรียนรู้จากกลุ่มข้อมูล ที่ได้กำหนดไว้ แล้วจึงนำไปวิเคราะห์กลุ่มข้อมูลที่ต้องการ การจำลองแบบนี้สามารถแบ่งออกได้เป็น 2 ระยะ (Cabena et al. 1997) ดังรูปที่ 2.1 คือ

1. ระยะการฝึกอบรม (Training Phase) คือ ขั้นตอนที่สร้างแบบจำลองขึ้นมาใหม่โดยใช้ข้อมูลในอดีต โดยใช้ข้อมูลประมาณ 80% ของข้อมูลทั้งหมด
2. ระยะการทดสอบ (Testing Phase) คือ ขั้นตอนการทดสอบแบบจำลองที่ได้สร้างขึ้น โดยนำข้อมูลในส่วนที่เหลือ 20% จากส่วนการฝึกอบรม

ตารางที่ 2.1 เทคนิคของดาต้าไมนิ่ง

Predictive Modeling	**Link Analysis**
Classification	Association discovery
Value prediction	Sequential pattern discovery
Database Segmentation	**Deviation Detection**
Demographic clustering	Visualization
Neural clustering	Statistics



รูปที่ 2.1 กระบวนการของการแยกแยะประเภทข้อมูล

1. การแยกแยะ

การแยกแยะ (Classification) หรือการแบ่งกลุ่มของข้อมูล ตามชนิดของกลุ่มข้อมูล เป็นกระบวนการสร้างแบบจำลอง โดยจัดข้อมูลให้อยู่ตามกลุ่มที่กำหนดไว้แล้ว สามารถแบ่งกลุ่มของข้อมูลได้อย่างชัดเจน เป็นการสร้างแบบจำลองเพื่อทำนายกลุ่มของข้อมูลที่เราสนใจ ซึ่งกลุ่มต่างๆ จะมีการกำหนดไว้ล่วงหน้าแล้ว เช่น การทำนายการปล่อยสินเชื่อของธนาคารแก่ลูกค้าว่ามีความเสี่ยงหรือปลอดภัย ซึ่งแนวคิดของการแยกแยะมีหลายรูปแบบดังนี้ ต้นไม้การตัดสินใจ (Decision Tree), การแยกแยะแบบเบยส์ (Bayesian Classification), เครือข่ายความเชื่อแบบเบยส์ (Bayesian Belief Network), เครือข่ายประสาทเทียม (Neural Network), จีนีติกอัลกอริทึม (Genetic Algorithms), รัฟเซต (Rough Set) และ ตรรกศาสตร์คลุมเครือ (Fuzzy Logic)

2. การทำนายค่าความต่อเนื่องของข้อมูล

การทำนายค่าความต่อเนื่องของข้อมูล (Value Prediction) เป็นการประมาณการข้อมูลทางตัวเลขในฐานข้อมูล โดยใช้เทคนิคทางสถิติ หรือเรียกอีกอย่างหนึ่งว่าการทำ Scoring มีเทคนิคที่นิยมใช้ คือ การถดถอยเชิงเส้น (Linear Regression) และการถดถอยไม่เชิงเส้น (Nonlinear Regression) ซึ่งเป็นตัวเลขที่บอกถึงความเป็นไปได้ของข้อมูล เช่น ธนาคาร ต้องการประเมินการอนุมัติการสินเชื่อกู้เงินแก่ลูกค้า ซึ่งต้องนำตัวแปรต่างๆ มาประกอบการพิจารณา เช่น อายุ ประวัติการใช้เงิน การชำระเงิน เงินเดือน และอาชีพ เป็นต้น

2.3.2 การแบ่งกลุ่มของข้อมูล

การแบ่งกลุ่มของข้อมูล (Database Segmentation หรือ Data Clustering) เป็นวิธีแบ่งกลุ่มของข้อมูล โดยจัดข้อมูลที่มีความคล้ายคลึงกันอยู่ด้วยกัน ส่วนข้อมูลที่มีความแตกต่างกันก็อยู่คนละกลุ่ม โดยสามารถรู้ล่วงหน้าได้ว่าจะมีข้อมูลเป็นจำนวนกี่กลุ่ม และเป็นกลุ่มใดบ้าง

2.3.3 การค้นหาความสัมพันธ์

การค้นหาความสัมพันธ์ (Link Analysis) คือ แนวคิดในการค้นหาความสัมพันธ์ระหว่างข้อมูลแต่ละเรคอร์ด หรือกลุ่มของเรคอร์ดในฐานข้อมูล ซึ่งความสัมพันธ์นั้นจะเรียกว่า Association วิธีการค้นหาความสัมพันธ์เหมาะในการวิเคราะห์ความสัมพันธ์ระหว่างสินค้า หรือบริการที่ลูกค้ามีแนวโน้มว่าจะใช้ร่วมกัน เช่น การขายสินค้าข้ามประเภท (Cross Selling), การหาตลาดเป้าหมาย (Target Marketing) และการเคลื่อนไหวของราคาหุ้น (Stock Price Movement) วิธีที่ใช้ในการค้นหาความสัมพันธ์นั้นมี 3 วิธีการดังนี้ การค้นพบความสัมพันธ์ (Association Discovery), ลำดับของการค้นพบรูปแบบ (Sequential Pattern Discovery) และการค้นพบลำดับเวลาที่คล้ายคลึงกัน (Similar Time Sequence Discovery)

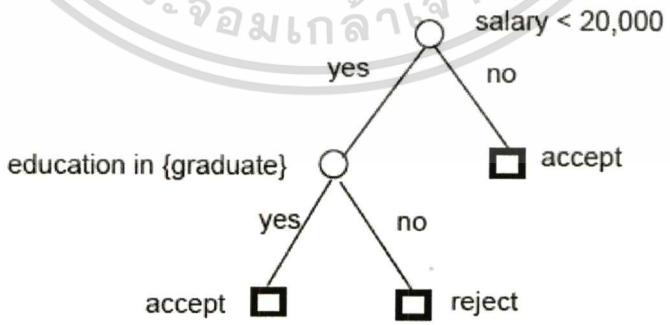
2.3.4 การตรวจหาความผิดปกติของข้อมูลจากข้อมูลปกติ

การตรวจหาความผิดปกติของข้อมูลจากข้อมูลปกติ (Deviation Detection) เป็นแนวคิดใหม่ซึ่งมีความสำคัญ การตรวจหาความผิดปกติของข้อมูลจากข้อมูลปกติมีอยู่ 2 วิธี คือ สถิติ (Statistics) กับเทคนิคการสร้างภาพนามธรรม (Visualization Techniques) ที่จะใช้ตรวจหาความผิดปกติจากข้อมูลปกติหรือ ข้อมูลที่คาดว่าควรจะเป็น ซึ่งในปัจจุบันเทคนิคการทำให้เห็นภาพสามารถทำได้ง่ายยิ่งขึ้น สามารถแสดงผลการวิเคราะห์ในลักษณะเป็นกราฟิก ส่วนมากจะอาศัยการเขียนกราฟ แล้วดูการกระจายของจุด ทำให้สามารถตรวจจับความผิดปกติได้ง่าย ตัวอย่างของการตรวจความผิดปกติของข้อมูลจากข้อมูลปกติ เช่น การตรวจสอบการปลอมลายเซ็น การอ้างสิทธิ์การประกัน การโกงบัตรโทรศัพท์ การใช้บัตรเครดิต การตรวจสอบคุณภาพ หรือการตรวจจุดบกพร่องของชิ้นงานในโรงงานอุตสาหกรรม

บทที่ 3
รายละเอียดและขั้นตอนการทำต้นไม้การตัดสินใจ

ต้นไม้การตัดสินใจเป็นรูปแบบการตัดสินใจในแบบมองจากบนลงล่าง ซึ่งโครงสร้างของต้นไม้การตัดสินใจจะประกอบไปด้วยรูทโหนด ซึ่งจะเป็นโหนดบนสุด และแตกออกไปเป็นโหนดลูก ซึ่งแต่ละโหนดอาจมีลูกมากกว่า 2 โหนดก็ได้ ส่วนโหนดที่อยู่ระดับล่างสุดเราเรียกว่า ลีฟโหนด ส่วนใหญ่จะอยู่ในรูปแบบของไบนารีต้นไม้ คือเป็นต้นไม้ที่มีทางเลือกเป็น 2 ทางเลือก มีวิธีการทำงานโดยการกำหนดตัวแปร ที่มีค่าความสำคัญสูงที่สุด ที่มีผลกระทบในการกำหนดการจัดกลุ่มเพื่อนำมาเป็นรูทโหนด จากนั้นทางเลือกจากรูทโหนดก็จะเป็นการกำหนดจากโหนดต่อๆ ไป สำหรับค่าที่เป็นไปได้ในการเลือกทางเลือกต่อไป ดังรูปที่ 3.1 (Cebena et al. 1997)

salary	*education*	label
10,000	high-school	reject
40,000	under-graduate	accept
15,000	under-graduate	reject
75,000	graduate	accept
18,000	graduate	accept



รูปที่ 3.1 ต้นไม้การตัดสินใจวิเคราะห์การอนุมัติเครดิต

จากรูปที่ 3.1 เป็นลักษณะของต้นไม้การตัดสินใจ ในการวิเคราะห์การอนุมัติเครดิตให้กับลูกค้า ซึ่งลักษณะของต้นไม้นี้บอกได้ว่า ลูกค้าที่มีเงินเดือนมากกว่า 20,000 บาทจะได้รับอนุมัติ

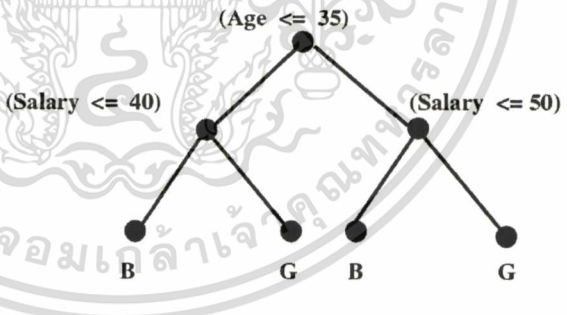
เครดิตทันที หรือลูกค้าที่มีเงินเดือนน้อยกว่าหรือเท่ากับ 20,000 บาทและมีการศึกษาเป็น Graduate จะอนุมัติเครดติเช่นกัน แต่ถ้าเป็นลูกค้าที่มีเงินเดือนน้อยกว่าหรือเท่ากับ 20,000 บาทและมีการศึกษาไม่ใช่ Graduate ก็จะถูกปฏิเสธการอนุมัติเครดิต

3.1 การแยกแยะประเภทข้อมูล

เทคนิคดาต้าไมนิ่งที่สำคัญเทคนิคหนึ่ง คือ การแยกแยะข้อมูล (Data Classification) เป็นกระบวนการสร้างโมเดลจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ โดยการสร้างกฎเพื่อช่วยในการตัดสินใจจากข้อมูลที่มีอยู่ เพื่อใช้ทำนายแนวโน้มการเกิดขึ้นของข้อมูลที่ยังไม่เกิดขึ้น โดยนำเสนอกฎที่ได้จากเทคนิคการจำแนกประเภทข้อมูล นิยมนำเสนอในรูปแบบของแผนภูมิต้นไม้ ซึ่งเรียกว่าต้นไม้ตัดการสินใจ

ต้นไม้การตัดสินใจ (Decision Tree) เป็นโครงสร้างต้นไม้ที่ใช้แสดงกฎ ที่ได้จากเทคนิคการแยกแยะประเภทข้อมูล (Mehta et al. 2001) โดยต้นไม้การตัดสินใจจะมีลักษณะคล้ายโครงสร้างต้นไม้ ที่แต่ละโหนดแสดงคุณลักษณะหรือแอตทริบิวต์ แต่ละกิ่งแสดงเงื่อนไขในการทดสอบ และลีฟโหนด (Leaf Node) แสดงกลุ่มที่กำหนดไว้ดังรูปที่ 3.2 (Han and Kamber. 2001)

Age	Salary	Class
30	65	G
23	15	B
40	75	G
55	40	B
55	100	G
45	60	G



รูปที่ 3.2 ตัวอย่างการใช้ต้นไม้การตัดสินใจ

อัลกอริทึมพื้นฐานของการสร้างต้นไม้การตัดสินใจ จะสร้างต้นไม้จากบนลงล่างแบบวนซ้ำ ด้วยวิธีการแบ่งปัญหาใหญ่เป็นปัญหาย่อย โดยการแล้วสร้างต้นไม้จากรูทโหนดให้แตกกิ่งย่อยออกไปตามลีฟโหนด ต้นไม้การตัดสินใจสร้างขึ้นโดยการเรียนรู้จากข้อมูลสอนระบบเป็นหลัก มีอัลกอริทึมพื้นฐานในการหาต้นไม้การตัดสินใจ ดังนี้ (Han and Kamber. 2001)

(1) create a node *N*;

(2) if *samples* are all of the same class *C* then

(3) return *N* as a leaf node labeled with the class *C*;

(4) if *attribute-list* is empty then

(5) return *N* as a leaf node labeled with the most

common class in *samples*; // majority voting

(6) select *test-attribute*, the attribute among *attribute-list*

with the highest information gain;

(7) label node *N* with *test-attribute*;

(8) for each known value a_i of *test-attribute*

(9) grow a branch from node *N* for the condition *test-attribute*= a_i;

(10) let s_i be the set of samples in *samples* for which *test-attribute*= a_i;

(11) if s_i is empty then

(12) attach a leaf labeled with the most common class in *samples*;

(13) else attach the node returned by Generate decision tree(s_i, *attribute-list, test-attribute)*;

1. ต้นไม้เริ่มสร้างโหนดแรก จากข้อมูลสอนระบบ (บรรทัดที่ 1)
2. ถ้าข้อมูลตัวอย่างเป็นคลาสเดียวกันทั้งหมดแล้ว โหนดนั้นก็จะกลายเป็นลีฟโหนด หรือคลาสปลายทาง (บรรทัดที่ 2 และ 3)
3. ถ้าอัลกอริทึมใช้ค่าของเอ็นโทรปี หรือค่า Information Gain เป็นค่าฮิวริสติกแล้ว ในการเลือกว่าแอตทริบิวต์ใด ที่จะเป็นตัวแยกข้อมูลไปสู่คลาสได้ดีที่สุด (บรรทัดที่ 6) แอตทริบิวต์ที่ได้เป็นแอตทริบิวต์ทดสอบ หรือตัดสินใจ (บรรทัดที่ 7) แอตทริบิวต์ทั้งหมดในอัลกอริทึมนี้เป็นค่าไม่ต่อเนื่อง ดังนั้น ต้องแปลงแอตทริบิวต์ที่มีค่าเป็นตัวเลขต่อเนื่อง ให้เป็นค่าที่ไม่ต่อเนื่องก่อน
4. กิ่งก็จะถูกสร้างขึ้นสำหรับแต่ละค่าในแอตทริบิวต์ทดสอบ และข้อมูลทดสอบจะถูกแบ่งส่วนตามลำดับ (บรรทัดที่ 8-10)
5. อัลกอริทึมใช้กระบวนการเดิมทำวนซ้ำ เพื่อที่จะสร้างต้นไม้สำหรับข้อมูลที่แต่ละจุดแบ่งข้อมูล (บรรทัดที่ 13)
6. ทำวนซ้ำไปจนกว่าจะไม่สามารถแยกความแตกต่างของคลาสภายในข้อมูลได้อีกแล้ว

การเลือกแอทริบิวต์เพื่อใช้ในการแบ่งแยกข้อมูล จะเลือกแอทริบิวท์ที่มีค่า Information Gain สูงสุด (หรือค่าเอ็นโทรปีดีสุด) จากสูตร $Gain(A) = I(s_1, s_2, \ldots, s_m) - E(A)$ โดยที่ A เป็นแอทริบิวท์ที่พิจารณา และ s_i เป็นจำนวนข้อมูล S ในคลาส C_i

อัลกอริทึมจะคำนวณค่า Information Gain สำหรับแต่ละแอทริบิวท์ แอทริบิวท์ที่มีค่า Information gain สูงสุดจะถูกเลือกให้เป็นแอทริบิวต์ทดสอบในเซต S จากนั้นจะสร้างโหนด และกิ่งที่แสดงค่าในแอทริบิวต์ และแบ่งข้อมูลต่อไปตามลำดับ

3.2 อัลกอริทึม SLIQ

ปัญหาที่สำคัญประการหนึ่งของการทำดาต้าไมนิ่ง โดยใช้อัลกอริทึมสำหรับการแยกแยะ คือ อัลกอริทึมส่วนใหญ่จะถูกออกแบบให้ใช้พื้นที่ในหน่วยความจำขนาดใหญ่ ซึ่งทำให้จำนวนชุดของข้อมูลที่ใช้ในการวิเคราะห์จะถูกจำกัด ด้วยขนาดของหน่วยความจำบนดิสก์ โครงการนี้เลือกใช้อัลกอริทึม SLIQ (Supervised Learning in Quest) (Han and Kamber. 2001; Hong. 2004) ซึ่งเป็นอัลกอริทึมในการทำแยกแยะ โดยใช้แนวความคิดต้นไม้สำหรับการตัดสินใจตัว SLIQ นั้นเหมาะกับข้อมูลที่มีขนาดใหญ่ สามารถจัดการกับข้อมูลได้ทั้งแบบตัวเลข และข้อมูลที่จัดเป็นหมวดหมู่ โดยใช้เทคนิคใหม่ ในการเรียงลำดับของข้อมูลก่อนในช่วงการสร้างต้นไม้ ที่เรียกว่า Breadth-First Growing เพื่อที่จะย้ายข้อมูลบางส่วนลงไปเก็บในดิสก์แทน และในการทำ Tree-Pruning ก็นำเทคนิค MDL (Minimum Description Length Principle) มาใช้ ซึ่งจะมีผลทำให้เพิ่มความถูกต้องของต้นไม้มากยิ่งขึ้น ซึ่งทั้งหมดนี้จะทำให้ SLIQ สามารถจัดการกับการแบ่งกลุ่มข้อมูลได้โดยไม่ต้องคำนึงถึงขนาด หรือชนิดของข้อมูลอีกต่อไป ซึ่งจะทำให้สามารถสร้างเครื่องมือในการทำดาต้าไมนิ่งได้ดีขึ้น

ข้อดีของ SLIQ

1. สามารถใช้งานกับข้อมูลจำนวนมากได้ โดยยังคงความถูกต้องในการทำงานค่อนข้างสูง ทำให้สามารถลองรับการขยายตัวของระบบได้
2. ใช้เทคนิคการเรียงลำดับก่อน ในการจัดการกับข้อมูลแบบตัวเลข ทำให้ช่วยลดเวลาในการประมวลผล เนื่องจากไม่ต้องมีการจัดการเรียงข้อมูลใหม่ทุกครั้ง
3. โครงสร้างการจัดการข้อมูล มีการนำเทคนิคการแบ่งข้อมูลออกเป็นคลาสลิสต์และแอตทริบิวต์ลิสต์ แล้วใช้ดัชนี เข้ามาช่วยในการอ้างอิงถึงกัน ทำให้ช่วยลดขนาดของข้อมูลที่ใช้ในการประมวลผล

3.3 การทำงานของ SLIQ

การทำงานโดยใช้อัลกอริทึม SLIQ แบ่งออกได้เป็น 2 ขั้นตอน (Hong. 2004) คือ

1. การสร้างต้นไม้ (Tree Building)
2. การปรับแต่งต้นไม้ (Tree Pruning)

ซึ่งในแต่ละขั้นตอนนั้นสามารถอธิบายวิธีการทำงานได้ดังต่อไปนี้

3.3.1 การสร้างต้นไม้

การสร้างต้นไม้ เป็นขั้นตอนแรกในการทำต้นไม้การตัดสินใจ โดยแบ่งกลุ่มจากข้อมูลชุดฝึกอบรม ข้อมูลชุดนี้จะถูกแบ่งออกเป็นสองส่วน หรือมากกว่า 2 ส่วนก็ได้ตามแอตทริบิวต์ที่ได้กำหนดไว้ ขั้นตอนนี้จะเป็นการทำงานแบบรีเคอร์ซีฟ คือ ทำวนซ้ำเป็นรอบๆ จนกระทั่งแต่ละตัวนั้นอยู่ในคลาสเดียว ซึ่งมีขั้นตอนการทำงานดังนี้

```
Make Tree(Trainging Data T)
    Partition(T)
Partition(Data S)
    If (All points in S are in the same class) then return;
    Evaluate splits for each attribute A
    Use best split found to partition S into S1 and S2;
    Partition(S1);
    Partition(S2);
```

ในการสร้างต้นไม้ นั้นมี 2 ขั้นตอนที่สำคัญคือ

1. คัดเลือกแอตทริบิวต์ที่มีความเหมาะสมในการวิเคราะห์ข้อมูล เพื่อหาจุดที่ดีที่สุด (Best Split) ในการเลือกและแบ่งแอตทริบิวต์ มีสูตรคำนวณหาค่าgini ดังนี้

$$Gini(T) = 1 - \sum p^2_j$$

T เป็นชุดของข้อมูลที่เก็บตัวอย่างจาก N คลาส

p_j เป็นความถี่สัมพันธ์ของคลาส j ใน T

ถ้าชุดของข้อมูล T ได้แบ่งเป็น 2 เซตย่อย ด้วยขนาด N1 และ N2 คลาส ค่า $Gini_{split}$ จากจำนวน N คลาสจะเป็น ดังนี้

$$Gini_{split}(T) = N_1/N\ gini(T_1) + N_2/N\ gini(T_2)$$

2. สร้างจุดแบ่งโดยใช้ค่าของ Gini split ที่ต่ำที่สุดมาเป็นตัวกำหนดจุดแบ่ง การกำหนดจุดแบ่งนั้นจะขึ้นอยู่กับว่าแอตทริบิวต์นั้นเป็นแอตทริบิวต์ประเภทไหน ดังนี้

1. การแบ่งข้อมูลสำหรับแอตทริบิวต์ที่เป็นตัวเลข (Numeric) เป็นการแบ่งข้อมูลแบบแบ่งเป็น 2 ทาง (Binary Split) จากรูปแบบ $A \leq v$ โดยที่ v เป็นตัวเลขจำนวนจริงของแอตทริบิวต์ A ให้เรียงข้อมูลที่นำมาใช้ทดสอบซึ่งจะขึ้นอยู่กับค่าของแอตทริบิวต์ A ที่พิจารณา เช่น เรียงค่า v1, v2,…,vn และเมื่อหาค่า Best Split ได้ว่าเป็นเรคอร์ดที่ v_i ก็ให้นำค่ากลางระหว่าง $v_i + v_{i+1}$ มาเป็น Best Split
2. การแบ่งข้อมูลสำหรับแอตทริบิวต์ ที่เป็นหมวดหมู่ ถ้าให้ S(A) เป็นชุดของค่าที่เป็นไปได้ในแอตทริบิวต์ A ดังนั้น การแบ่งข้อมูลจะอยู่ในรูปของ $A \in S'$, ซึ่ง $S' \subset S$ ดังนั้น จำนวนของเซตย่อยที่เป็นไปได้ n จำนวนจะมีค่าเท่ากับ 2^n สำหรับแอตทริบิวต์ที่มีค่าเป็นไปได้ n ค่า โดยจะต้องมีการกำหนดค่า MAXSETSIZE คือจำนวน n สูงสุดที่ทำให้ค่า Best Split ที่ดีที่สุด

ขั้นตอนการสร้างต้นไม้ใน SLIQ

1. Pre-Sorting และ Breadth-First Growth

การเรียงลำดับข้อมูลถือเป็นสิ่งที่จำเป็นที่สุด ดังนั้นเทคนิคแรกของการทำ SLIQ คือ การนำข้อมูลเป็นตัวเลขมาเรียงลำดับ (การเรียงลำดับนั้นให้เรียงเพียงครั้งแรกครั้งเดียว) ดังรูปที่ 3.3 ในการทำงานของการเรียงลำดับก่อน (Pre-Sorting) นั้นต้องจัดโครงสร้างข้อมูลให้เป็นดังนี้

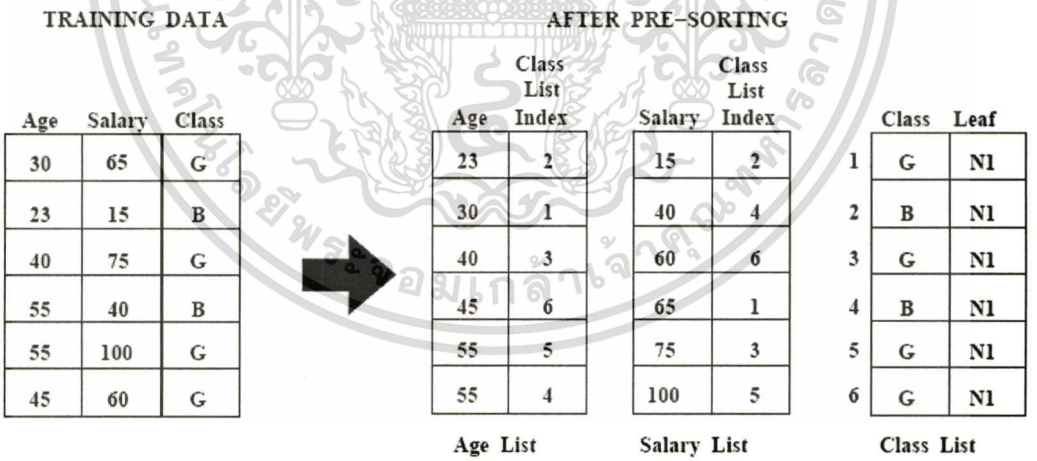
- กำหนดคลาสลิสต์ (Class List) ขึ้นมาก่อนในคลาสลิสต์นั้นจะประกอบด้วย 2 ฟิลด์ ดังนี้คือ คลาสเลเบล (Class Label) และลีฟโหนด ซึ่งคลาสเลเบล คือ ค่าของแอตทริบิวต์ที่ใช้ในการทำนาย เช่น G = Good, B = Bad
- สร้างแอตทริบิวต์ลิสต์ขึ้นมา ซึ่งในตัวแอตทริบิวต์ลิสต์นั้นจะประกอบไปด้วยแอตทริบิวต์และค่าดัชนี ลำดับที่ i ในคลาสลิสต์จะมีผลต่อลำดับที่ i ของข้อมูลแต่ละลีฟโหนดของต้นไม้การตัดสินใจ จะแสดงผลของการแบ่งข้อมูลใน Training Set (T)
- การสร้างต้นไม้ ครั้งแรกจะกำหนดให้ ลีฟโหนดในคลาสลิสต์ชี้ไปที่รูท

2. Node Splits

คำนวณเพื่อแบ่งโหนดจากลีฟทั้งหมด ซึ่งจะทำและประเมินไปพร้อมกัน ทุกจุดแบ่ง

ที่ดีที่สุดจะถูกเก็บไว้ที่แต่ละลีฟโหนด ขั้นตอนการแบ่งโหนดมีอัลกอริทึมดังนี้

```
EvaluateSplits()
    For each attribute list of A DO
        Traverse attributed list of A
        For each value v in the attribute list DO
            Find the corresponding entry in the class list, and
                hence the corresponding class and the leaf node (say l)
            Update the class histogram in the leaf l
            IF A is a numeric attribute then
                Compute splitting index for test (A ≤ v) for leaf l
        IF A is a categorical attribute then
            For each leaf of the tree do
                Find subset of A with best split
```



รูปที่ 3.3 การเรียงลำดับก่อน

การคำนวณหาจุดในการแบ่งต้นไม้ จากค่าของ Gini Index และ $Gini_{split}$ จากทุกๆ เรคอร์ด และจะใช้ค่า $Gini_{split}$ ที่มีค่าน้อยที่สุดมาหาจุดแบ่งของต้นไม้ ยกตัวอย่างการคำนวณข้อมูลจากดัชนีที่ 1 พบว่าค่า $Gini_{split}$ ที่มีค่าน้อยที่สุด จึงใช้ดัชนีที่ 1 เป็นจุดแบ่งต้นไม้ ดังนั้นค่ากลางระหว่าง 30 และ 40 ซึ่งก็คือ (30+40)/2 = 35 จากสูตร $v = \nu_i + \nu_{i+1}/2$ เป็นค่าแบ่งที่ดีที่สุดดังตารางที่ 3.1 และได้

ลักษณะของต้นไม้ที่เกิดจากการทำ Bread-Firs Growth ดังรูปที่ 3.4

$$Gini(S1) = 1-[(1/2)^2 +(1/2)^2] = 0.5$$

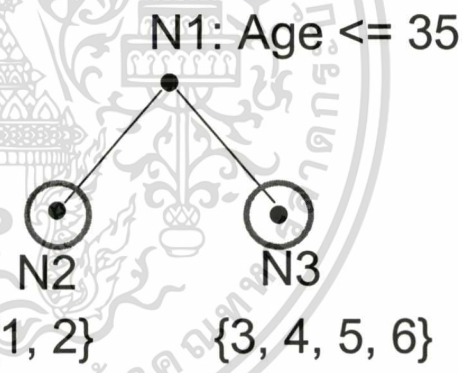
$$Gini(S2) = 1-[(3/4)^2 +(1/4)^2] = 0.375$$

$$Ginisplit = 2/6(0.5) +4/6(0.375) = 0.417$$

ตารางที่ 3.1 ฮิสโตรแกรมของนัมเบอร์แอตทริบิวต์

	G	B	Sum
C above	1	1	2
C below	3	1	4

Age	Salary	Class
30	65	G
23	15	B
40	75	G
55	40	B
55	100	G
45	60	G

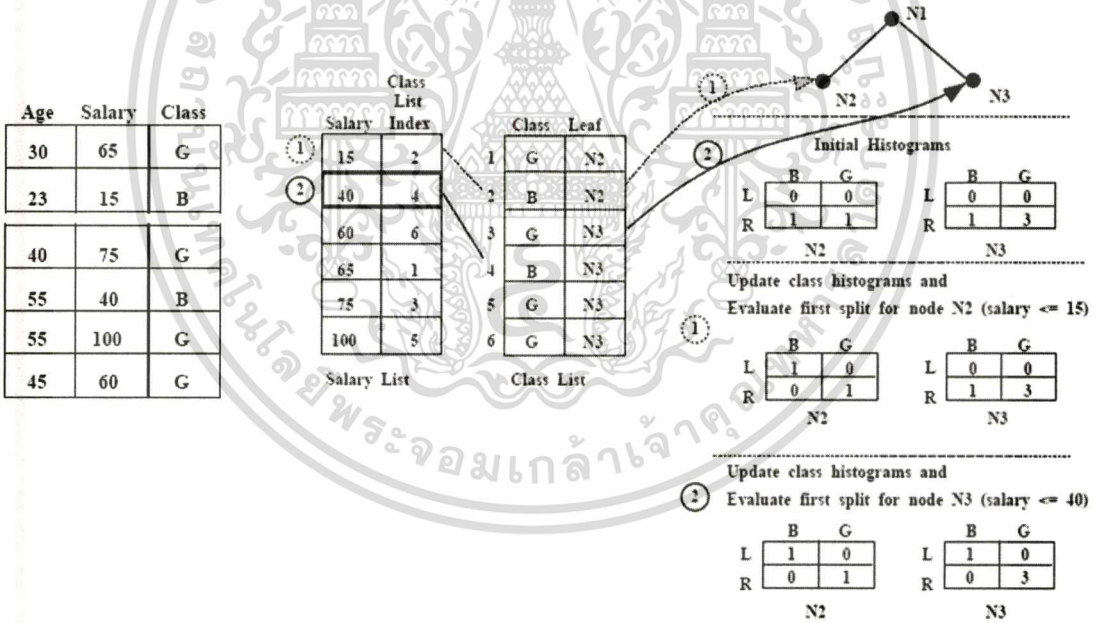


รูปที่ 3.4 ลักษณะของต้นไม้ที่เกิดจากการทำ Bread-First Growth

การคำนวณหา Gini Split Index สำหรับการแบ่งโหนดนั้นจะนำเทคนิค Gini เข้ามาช่วยซึ่งในการคำนวณสำหรับแต่ละแอตทริบิวต์นั้น คลาสฮิสโตรแกรมที่ติดมากับแต่ละลีฟโหนด จะใช้สะสมความถี่ของค่าในคลาส โดยข้อมูลจะต้องตรงกับโหนดนั้นๆ ซึ่งก็คือ การทำฮิสโตรแกรม ถ้าแอตทริบิวต์ที่เป็นตัวเลขฮิสโตรแกรมจะมีรูปแบบดังนี้ (คลาส, ความถี่) แต่ถ้าค่าของแอตทริบิวต์เป็นหมวดหมู่จะแตกต่างกัน โดยตัวฮิสโตรแกรมจะมีรูปแบบดังนี้ (ค่าแอตทริบิวต์, คลาส, ความถี่) ในรูปที่ 3.5 แสดงการหาค่าจุดแบ่งจากแอตทริบิวต์เงินเดือน ในการแบ่งครั้งแรกจากแอตทริบิวต์อายุที่มีค่าน้อยกว่าหรือเท่ากับ 35 แต่ละคลาสฮิสโตรแกรมแสดงถึงการกระจายของลีฟโหนด ค่า L

หมายถึง การกระจายที่น่าพอใจ ส่วนค่า R หมายถึง ผลการทดสอบที่ไม่น่าพอใจ ค่าแรกใน Salary List จะเป็น N2 ดังนั้นการแบ่งครั้งแรกจึงถูกกำหนดไว้ที่เงินเดือนน้อยกว่าหรือเท่ากับ 15 สำหรับ N2 และหลังจากการแบ่งจะได้เป็น (เงินเดือน 15, ดัชนีตัวที่ 2) ค่าที่ได้จะถูกส่งไปที่กิ่งทางด้านซ้าย ส่วนค่าที่เหลือจะถูกส่งไปที่กิ่งทางด้านขวาทั้งหมด คลาสฮิสโตรแกรมของโหนด N2 จะบันทึกจาก N1 เปลี่ยนเป็น N2 ต่อมาเป็นการแบ่งเงินเดือนที่น้อยกว่าหรือเท่ากับ 40 จากข้อมูลชุดหลัง จะถูกกำหนดไว้สำหรับโหนด N3 โดยหลังจากการแบ่งจะได้ (เงินเดือน 40 , ดัชนีตัวที่ 4) อยู่ที่ฝั่งทางซ้ายของกิ่ง และคลาสฮิสโตรแกรมของโหนด N3 บันทึกการเปลี่ยนแปลงที่เกิดขึ้น

การแบ่งข้อมูลที่มีลักษณะเป็นหมวดหมู่ (Categorical Attribute) ใช้วิธีการกำหนดเซตย่อยของ S และค่าของแอตทริบิวต์ โดยที่จำนวนสมาชิกใน S ต้องน้อยกว่า Threshold ไม่เช่นนั้นสมาชิกของ S ที่ได้จากการแบ่งที่ดีที่สุด จะถูกเพิ่มเข้าไปที่เซตเดิมของ S ซึ่งว่างเปล่า กระบวนการดังกล่าวจะต้องทำวนซ้ำไปเรื่อยๆ จนไม่มีการเปลี่ยนแปลงของการแบ่งเลย



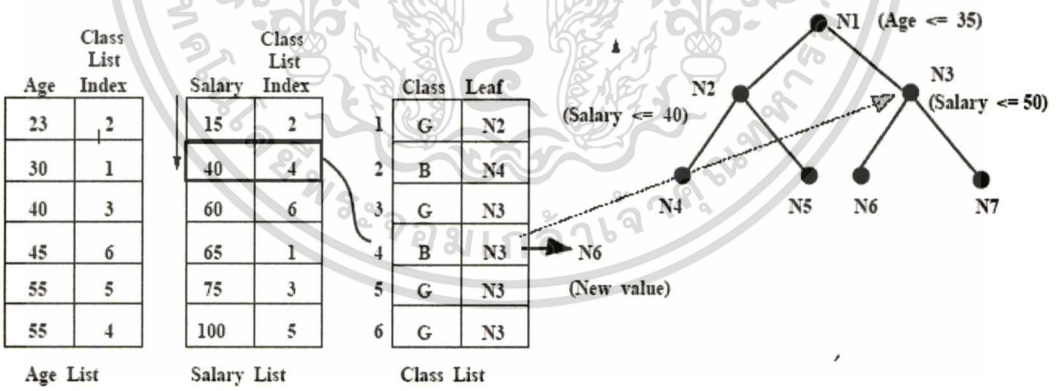
รูปที่ 3.5 การแตกกิ่งของต้นไม้

3. การปรุบปรุงคลาสลิสต์

เมื่อหาจุดแบ่งที่ดีที่สุดได้แล้วจะสร้างต้นไม้ไปตามกิ่งที่เกิดขึ้น ซึ่งมักแตกกิ่งออกไปทั้งซ้ายและขวา จากนั้นปรับปรุงค่าคลาสลิสต์ให้เป็นค่าใหม่ที่ถูกต้องให้เป็นไปตามค่าที่คลาสนั้นเปลี่ยนแปลงไป ซึ่งใช้อัลกอรึทึมดังนี้

```
UpdateLabels()
    For each attribute A used in a spit do
        Traverse attribute list of A
        For each value v in the attribute list do
            find the corresponding entry in the class list (say e)
            find the new class c to which v belongs by applying
                the splitting test at node referenced from e
            update the class label for e to c
            update node referenced in e to the child corresponding to class c
```

พบว่าคลาสลิสต์จากรูปที่ 3.6 จะถูกปรับปรุงหลังจากที่โหนด N2 และโหนด N3 แบ่งออกด้วยแอตทริบิวต์เงินเดือน ซึ่งจะเดินทางไปที่ดัชนีที่ 4 ซึ่งมีค่าของเงินเดือนเท่ากับ 40 จะถูกปรับปรุงค่าลีฟโหนด ขั้นตอนแรกลีฟโหนดของดัชนีที่ 4 ของคลาสลิสต์จะถูกใช้ในการค้นหาโหนด เช่น N3 หลังจากนั้นการแบ่งที่ถูกเลือกโดย N3 จะนำมาใช้ค้นหาลูกตัวใหม่นั่น คือ N6 ฟิลด์ลีฟโหนดของดัชนีตัวที่ 4 ในคลาสลิสต์ก็จะถูกปรับปรุงเพื่อให้สอดคล้องกับค่าใหม่ที่เกิดขึ้น



รูปที่ 3.6 การปรับปรุงคลาสลิสต์

3.3.2 การปรับแต่งต้นไม้

การปรับแต่งต้นไม้ (Tree Pruning) เพื่อทำให้ต้นไม้นั้นมีขนาดเล็กลง และช่วยลดข้อมูลที่ไม่เกี่ยวข้องกับการวิเคราะห์ออกไป ขั้นตอนนี้จะต้องตรวจสอบต้นไม้ที่ได้สร้างไป และเลือกต้นไม้ย่อย ที่มีข้อผิดพลาดน้อยที่สุด ซึ่งมีวิธีหาค่าผิดพลาด อยู่ 2 วิธี คือ

1. ใช้ข้อมูลสอนระบบเดิมวิธีนี้เรียกว่า Cross-Validation โดยนำกลุ่มข้อมูลสอนระบบมาแบ่งออกเป็นหลายๆ ตัวอย่าง แล้วสร้างต้นไม้จากตัวอย่างเหล่านี้ จากนั้นก็ใช้ต้นไม้นี้ประเมินหาข้อผิดพลาดกับต้นไม้ที่ได้สร้างขึ้น ซึ่งถึงแม้ว่าวิธีการนี้จะทำให้ได้ต้นไม้ที่ขนาดเล็กลงและมีความถูกต้องสูง แต่ว่าต้นทุน (Cost) ในการสร้างต้นไม้ค่อนข้างสูง ซึ่งทำให้วิธีนี้ไม่ค่อยเหมาะสมกับกลุ่มข้อมูลสอนระบบขนาดใหญ่

2. ใช้ข้อมูลที่เป็นข้อมูลสอนระบบใหม่ วิธีการนี้จะแบ่งข้อมูลสอนระบบออกเป็น 2 ส่วนด้วยกัน ส่วนหนึ่งใช้ในการสร้างต้นไม้ อีกส่วนหนึ่งจะใช้ในการปรับแต่งต้นไม้ ข้อมูลที่ใช้ในการปรับแต่งต้นไม้ ก็ควรจะเลือกข้อมูลที่มีการกระจายตามความเป็นจริง เพราะถ้าเลือกข้อมูลขนาดไม่เหมาะสม หรือเลือกข้อมูลไม่ถูก จะมีผลทำให้ไปลดขนาด และความถูกต้องของกลุ่มข้อมูลสอนระบบ ในการสร้างต้นไม้เพื่อป้องกันการสร้างต้นไม้ ที่เกินพอดีจึงได้นำหลักการของ MDL มาใช้ คือ เป็นการปรับแต่งต้นไม้ในระหว่างการสร้างต้นไม้ และหาต้นไม้ย่อยที่มีการใช้จำนวนบิตในการเข้ารหัสน้อยที่สุด ซึ่งต้นทุนของการเข้ารหัส โดยที่ให้เซต S ประกอบไปด้วย n รายการ ซึ่งแต่ละรายการอยู่ในคลาส k โดยที่ ni เท่ากับจำนวนรายการในคลาส i มีสูตรหนึ่งในการคำนวณ คือ

$$C(S)=\sum_i n_i \log\frac{n}{n_i}+\frac{k-1}{2}\log\frac{n}{2}+\log\frac{\pi^{k/2}}{\tau(k/2)}$$

ในเทอมแรก คือ n * E(S) ซึ่ง E(S) คือ เอ็นโทรปีของเซต S ต้นทุน (Cost) ในการเข้ารหัสต้นไม้ ประกอบด้วย 3 ต้นทุนดังนี้

1. ต้นทุนการเข้ารหัสโครงสร้างของต้นไม้
2. ต้นทุนการเข้ารหัสแต่ละจุดแบ่ง โดยพิจารณาจากประเภทของแอตทริบิวต์และค่าของจุดแบ่ง
3. ต้นทุนการเข้ารหัสคลาสของข้อมูลในแต่ละลีฟโหนดของต้นไม้

บทที่ 4
การพัฒนาโปรแกรม

ในการศึกษาโครงการนี้ จะเป็นการนำทฤษฎีดาต้าไมนิ่งแนวคิดต้นไม้การตัดสินใจโดยใช้อัลกอริทึมที่ชื่อว่า SLIQ มาพัฒนาเป็นโปรแกรมประยุกต์ เพื่อใช้ในวิเคราะห์หาสาเหตุการเปลี่ยนผู้ให้บริการโทรศัพท์เคลื่อนที่ของลูกค้า โดยแบ่งดำเนินการศึกษาเป็น 4 ขั้นตอน คือ

4.1 การศึกษาทฤษฎีที่เกี่ยวข้อง

ในการพัฒนาระบบงานนี้มีทฤษฎีที่ผู้พัฒนาต้องทราบดังนี้

1. แนวคิดการจัดการลูกค้าสัมพันธ์ (Customer Relationship Management) หรือ (CRM) เบื้องต้น โดยเฉพาะในส่วนของการจัดการเคมเปญเพื่อใช้ในการหาข้อมูล
2. ทฤษฎีดาต้าไมนิ่ง โดยเฉพาะแนวคิดการแยกแยะ ในส่วนของต้นไม้การตัดสินใจ
3. อัลกอริทึม SLIQ ที่ใช้ในการสร้างแบบจำลองต้นไม้
4. ความรู้ความเข้าใจในด้านการพัฒนาโปรแกรมซึ่งโดยใช้ภาษาวิชวลเบสิค เวอร์ชั่น 6 และ Microsoft SQL Server 2000

4.2 การรวบรวมข้อมูลที่เกี่ยวข้อง

ในการรวบรวมข้อมูลที่จะนำมาใช้วิเคราะห์นั้น ทางผู้จัดทำได้ทำหนังสือขอข้อมูลไปถึงแผนกการตลาดขององค์กรที่ให้บริการด้านโทรศัพท์เคลื่อนที่แห่งหนึ่ง ซึ่งข้อมูลที่ได้เป็นข้อมูลของลูกค้าในระบบแบบใช้ก่อนจ่ายทีหลัง (Post-Paid) ที่ปิดบริการตั้งแต่วันที่ 1 กรกฎาคม ถึง 30 กันยายน พ.ศ. 2547 จำนวน 1,500 เรคอร์ด ซึ่งเป็นข้อมูลที่เพิ่งได้นำเข้ามาสู่ระบบดาต้าแวร์เฮ้าส์ รายละเอียดของข้อมูลจะกล่าวในบทที่ 5 เรื่องการเตรียมข้อมูล

4.3 การศึกษาความต้องการของระบบ

จากการศึกษาความต้องการของระบบนั้น พบว่าระบบนั้นต้องมีหน้าที่หลักในการทำงาน 7 หน้าที่ดังต่อไปนี้

1. ระบบสามารถนำข้อมูลเข้าได้ 2 ทาง ดังนี้

- แฟ้มข้อความ จะต้องมีรูปแบบตามที่กำหนดดังนี้ ชื่อตัวแปร จะอยู่ในแถวแรก และข้อมูลแต่ละตัวจะคั่นด้วยเครื่องหมายไปป์ (|)
- ข้อมูลจากฐานข้อมูล ซึ่งข้อมูลจะต้องมาจากฐานข้อมูลที่เป็น Microsoft SQL Server 2000 เท่านั้น

2. ผู้ใช้ระบบสามารถนำตัวแปรเข้ามาวิเคราะห์ได้ไม่เกิน 10 ตัวแปร (รวมตัวแปร ที่เป็นคลาสเลเบล) และสามารถกำหนดได้ว่าจะนำตัวแปรใดเข้ามาวิเคราะห์ ซึ่งต้องสามารถเพิ่มหรือลดจำนวนตัวแปรที่นำมาวิเคราะห์ได้ในภายหลัง
3. ระบบต้องสามารถแบ่งข้อมูลออกเป็น 2 ส่วนจากแหล่งข้อมูลเดียวกัน
   - ส่วนที่ 1 คือ ส่วนของข้อมูลที่นำมาใช้ในการสร้างแบบจำลองต้นไม้
   - ส่วนที่ 2 คือ ส่วนของข้อมูลที่นำมาใช้ทดสอบแบบจำลองต้นไม้ว่าที่สร้างขึ้น
4. ผู้ใช้ระบบสามารถกำหนดเงื่อนไขในการสร้างแบบจำลองต้นไม้ได้ดังนี้
   - ผู้ใช้สามารถกำหนดระดับของแบบจำลองต้นไม้ที่จะแตกได้
   - ผู้ใช้สามารถกำหนดจำนวนข้อมูลที่น้อยที่สุดในแต่ละโหนดเพื่อวิเคราะห์ได้
5. ระบบสามารถคำนวณผลการวิเคราะห์ได้ถูกต้องน่าเชื่อถือ ตามแนวคิดการแยกแยะโดยนำหลักการต้นไม้การตัดสินใจและอัลกอริทึม SLIQ เข้ามาใช้
6. ระบบแสดงค่าความเชื่อมั่นเพื่อให้ผู้ใช้ระบบใช้ประกอบการตัดสินใจ โดยค่าความเชื่อมั่นมีวิธีในการคิดดังนี้

$$\text{ค่าความเชื่อมั่น} = \frac{\text{จำนวนข้อมูลที่อยู่ในคลาส}}{\text{จำนวนข้อมูลในโหนดทั้งหมด}} \times 100\%$$

7. ระบบแสดงผลลัพธ์ให้ผู้ใช้ระบบสามารถเข้าใจได้ง่าย ซึ่งในที่นี้จะแสดงในรูปแบบจำลองต้นไม้

4.4 การวิเคราะห์และออกแบบโปรแกรม

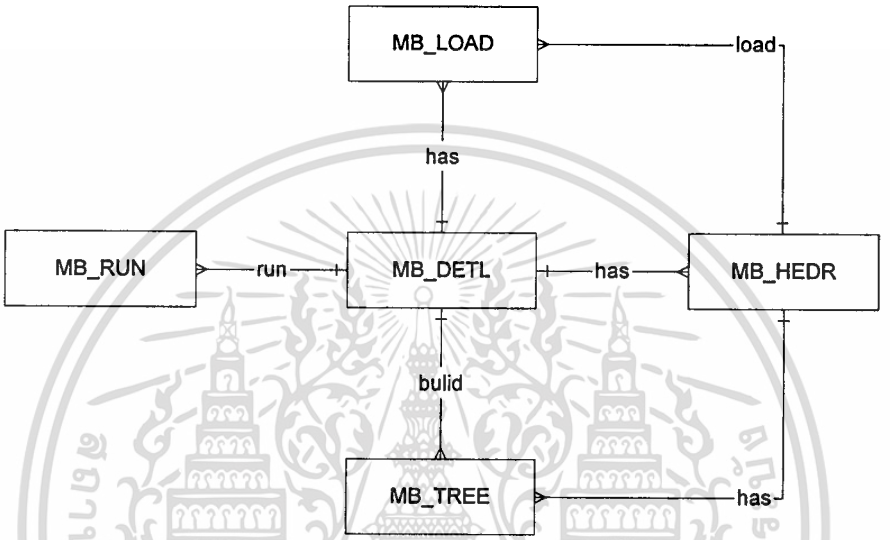
ในการวิเคราะห์และออกแบบโปรแกรมนั้นแบ่งเป็น 2 ส่วนดังนี้

1. การวิเคราะห์และออกแบบส่วนของฐานข้อมูล
2. การวิเคราะห์และออกแบบส่วนของหน้าจอ

4.4.1 การวิเคราะห์และออกแบบส่วนของฐานข้อมูล

1. ฐานข้อมูลระบบที่ใช้ในการวิเคราะห์

ใช้ Microsoft SQL Server 2000 เป็นฐานข้อมูลซึ่งจะประกอบด้วย 5 ตาราง ดังรูปที่ 4.1 คือ



รูปที่ 4.1 แผนภาพอีอาร์ของการพัฒนาระบบ

1. MB_TREE เป็นตารางที่ใช้เก็บข้อมูลผลของแบบจำลองที่ผ่านการวิเคราะห์ โดยรายละเอียดแสดงดังตารางที่ 4.1
2. MB_LOAD เป็นตารางที่ใช้เก็บข้อมูลที่จะนำมาวิเคราะห์ ข้อมูลที่อยู่ในตารางนี้จะเป็นข้อมูลดิบ โดยรายละเอียดแสดงดังตารางที่ 4.2
3. MB_RUN เป็นตารางที่ใช้เก็บข้อมูลที่เป็นรายละเอียดของข้อมูลที่ใช้ในการวิเคราะห์ และเก็บค่าลีฟโหนดที่ข้อมูลแต่ละตัวโดยรายละเอียดแสดงดังตารางที่ 4.3
4. MB_HEDR เป็นตารางที่ใช้เก็บข้อมูลชื่อตัวแปร และประเภทของข้อมูลที่จะวิเคราะห์ โดยรายละเอียดแสดงดังตารางที่ 4.4
5. MB_DETL เป็นตารางที่ใช้เก็บข้อมูลชุดของข้อมูล และค่าพารามิเตอร์ที่ใช้ในการวิเคราะห์ โดยรายละเอียดแสดงดังตารางที่ 4.5

ตารางที่ 4.1 รายละเอียดฐานข้อมูลของตาราง MB_TREE

ชื่อฟิลด์	ประเภท	รายละเอียด	ชนิดของคีย์	ตารางที่อ้างถึง
INPT_CODE	INT(4)	รหัสของชุดข้อมูลนำเข้า	PK,FK	MB_DETL
VAR_CODE	INT(4)	รหัสของตัวแปร	PK,FK	MB_HEDR
NODE_NUMB	CHAR(10)	หมายเลขโหนดหรือค่าลีฟโหนด	PK	
BCAT_SPLT	CHAR(50)	ค่าแบ่งแยกที่ดีที่สุดของข้อมูล ประเภท Category		
BNUM_SPLT	REAL(4)	ค่าแบ่งแยกที่ดีที่สุดของข้อมูล ประเภท Numeric		
LEAF	CHAR(50)	แสดงค่าหมายเลขโหนดที่มีค่าลีฟโหนด [- = มี 2 ค่ารวมกัน] [X = มี 1 ค่าจะแสดงค่าตาม X]		
CLSS1_TOTL	INT(4)	จำนวนเรคอร์ดของคลาสที่ 1		
CLSS2_TOTL	INT(4)	จำนวนเรคอร์ดของคลาสที่ 2		

ตารางที่ 4.2 รายละเอียดฐานข้อมูลของตาราง MB_LOAD

ชื่อฟิลด์	ประเภท	รายละเอียด	ชนิดของคีย์	ตารางที่อ้างถึง
INPT_CODE	INT(4)	รหัสของชุดข้อมูลนำเข้า	PK,FK	MB_DETL
VAR_CODE	INT(4)	รหัสของตัวแปร	PK,FK	MB_HEDR
RECD_SEQN	INT(4)	ลำดับของข้อมูล	PK	
VAR_VLUE	VARCHAR(50)	ค่าของข้อมูล		

ตารางที่ 4.3 รายละเอียดฐานข้อมูลของตาราง MB_RUN

ชื่อฟิลด์	ประเภท	รายละเอียด	ชนิดของคีย์	ตารางที่อ้างถึง
INPT_CODE	INT(4)	รหัสของชุดข้อมูลนำเข้า	PK,FK	MB_DETL
RECD_SEQN	INT(4)	ลำดับของข้อมูล	PK,FK	MB_LOAD
NODE_NUMB	INT(4)	หมายเลขโหนดหรือค่าลีฟโหนด		
CLSS	CHAR(50)	ค่าคลาสเลเบล		
VAR1	CHAR(50)	ค่าของตัวแปรตัวที่ 1		
VAR2	CHAR(50)	ค่าของตัวแปรตัวที่ 2		
VAR3	CHAR(50)	ค่าของตัวแปรตัวที่ 3		
VAR4	CHAR(50)	ค่าของตัวแปรตัวที่ 4		
VAR5	CHAR(50)	ค่าของตัวแปรตัวที่ 5		
VAR6	CHAR(50)	ค่าของตัวแปรตัวที่ 6		
VAR7	CHAR(50)	ค่าของตัวแปรตัวที่ 7		
VAR8	CHAR(50)	ค่าของตัวแปรตัวที่ 8		
VAR9	CHAR(50)	ค่าของตัวแปรตัวที่ 9		
VAR10	CHAR(50)	ค่าของตัวแปรตัวที่ 10		

ตารางที่ 4.4 รายละเอียดฐานข้อมูลของตาราง MB_HEDR

ชื่อฟิลด์	ประเภท	รายละเอียด	ชนิดของคีย์	ตารางที่อ้างถึง
INPT_CODE	INT(4)	รหัสของชุดข้อมูลนำเข้า	PK,FK	MB_DETL
VAR_CODE	INT(4)	รหัสของตัวแปร	PK	
VAR_NAME	VARCHAR(50)	ชื่อของตัวแปร		
VAR_TYPE	INT(4)	ประเภทของตัวแปร [0 = Numeric] [1 = Category]		
SELT_TYPE	CHAR(1)	ประเภทของตัวแปรที่ถูกเลือก [0 = ถูกเลือก] [1 = ไม่ถูกเลือก] [2 = ถูกเลือกเป็นคลาสเลเบล]		

ตารางที่ 4.5 รายละเอียดฐานข้อมูลของตาราง MB_DETL

ชื่อฟิลด์	ประเภท	รายละเอียด	ชนิดของคีย์	ตารางที่อ้างถึง
INPT_CODE	INT(4)	รหัสของชุดข้อมูลนำเข้า	PK	
LOAD_ORGN	INT(4)	เก็บ INPT_CODE เดิมของข้อมูลเพื่อใช้ใหม่หากมีการแก้ไขโดยไม่ต้องสร้างตารางใหม่		
DB_FLAG	CHAR(1)	ประเภทของข้อมูลนำเข้า [Y = ฐานข้อมูล] [N = แฟ้มข้อความ]		
DATA_PATH	VARCHAR(80)	ที่เก็บชุดข้อมูล		
VAR_NUMB	INT(4)	จำนวนตัวแปร		
CLSS_VLUE1	CHAR(10)	ค่าของคลาสเลเบลที่ 1		
CLSS_VLUE2	CHAR(10)	ค่าของคลาสเลเบลที่ 2		
TEST_TYPE	CHAR(1)	แบ่งข้อมูลไว้ทดสอบ [Y = แบ่งข้อมูล] [N = ไม่แบ่งข้อมูล]		
TEST_VLUE	INT(4)	เปอร์เซ็นต์ของข้อมูลที่ใช้ทดสอบ		
MAX_LEVL	INT(4)	จำนวนระดับสูงสุดที่ใช้แตกทรี		
MIN_VLUE	INT(4)	จำนวนข้อมูลน้อยสุดในแต่ละโหนด		
TOTL_TRIN	INT(4)	จำนวนข้อมูลสอนระบบทั้งหมด		
TOTL_ANLY	INT(4)	จำนวนข้อมูลที่ใช้ในการวิเคราะห์ทั้งหมด		
STRT_DTTM	DATETIME(4)	วันที่และเวลาที่นำเข้าเข้าระบบ		

บทที่ 5
การทำดาต้าไมนิ่งวิเคราะห์ข้อมูลการใช้งานโทรศัพท์เคลื่อนที่

5.1 วัตถุประสงค์ทางธุรกิจ

กรณีศึกษาการวิเคราะห์ข้อมูลการใช้งานโทรศัพท์เคลื่อนที่ มีวัตถุประสงค์ดังนี้

1. ต้องการทราบว่าลูกค้าที่ยกเลิกไปแล้วเกิดมีพฤติกรรมอย่างไรบ้าง
2. เพื่อหาแนวทางในการเสนอเคมเปญที่ตรงใจลูกค้า
3. ค่าใช้จ่ายต่อเดือนมีผลต่อการเปลี่ยนใจไปใช้บริการจากค่ายอื่นหรือไม่
4. ลูกค้าใช้ Promotion เหมาะสมกับตัวเองหรือไม่

5.2 การเตรียมข้อมูล

ได้รับข้อมูลของผู้ใช้บริการจำนวน 1,500 เรคอร์ด เพื่อใช้ในการวิเคราะห์ข้อมูลการใช้งานโทรศัพท์เคลื่อนที่ โดยรับความอนุเคราะห์ข้อมูลนี้จากฝ่ายการจัดการการเปลี่ยนแปลงผู้ให้บริการและการรักษาผู้ใช้บริการ แผนกวิเคราะห์และวิจัยข้อมูลลูกค้า เป็นข้อมูลเรียกใช้ผ่านงานโปรแกรม Business Object จากคลังข้อมูล สามารถนำมาใช้งานได้ในทันที เนื่องจากได้มีเฉพาะฟิลด์ที่ต้องการเท่านั้น และผ่านขั้นตอนของเตรียมข้อมูลมาอย่างดีแล้วจากกระบวนการของคลังข้อมูล

5.3 การใช้โปรแกรมดาต้าไมนิ่งกับข้อมูล

ผู้ใช้งานโปรแกรมวิเคราะห์ข้อมูลการใช้งานโทรศัพท์เคลื่อนที่ เพื่อการออกโปรโมชั่น เมื่อเข้าสู่ระบบจะพบหน้าเมนูหลัก เป็นเมนูที่สามารถใช้งานได้ทั้งไอคอนและปุ่มต่างๆ ใช้งานได้ง่าย แม้ไม่มีความรู้เรื่องการใช้โปรแกรมมาก่อน เมนูหลักของระบบแบ่งออกเป็น 5 หน้าจอ ดังนี้

5.3.1 หน้าจอเมนูหลัก

เป็นหน้าจอที่ใช้ในการแสดงผลการวิเคราะห์ข้อมูล เมนูหรือไอคอนบางปุ่มจะไม่สามารถใช้งานได้ในครั้งแรก แต่จะสามารถใช้งานได้หลังจากที่ได้โหลดข้อมูลไปแล้ว เลือกประเภทของตัวแปร เลือกตัวแปรที่จะนำมาวิเคราะห์ และกำหนดข้อจำกัดในการสร้างแบบจำลองต้นไม้เรียบร้อยแล้ว ยกเว้นในส่วนของวิว ถ้าระบบกำลังแสดงผลของสอนระบบ เมนูและไอคอนของการทดสอบ

ระบบจะไม่แสดง และเช่นเดียวกันถ้าระบบกำลังแสดงผลของการทดสอบระบบ เมนูและไอคอนของการสอนระบบก็จะไม่แสดงเช่นกัน ดังรูปที่ 5.1

การนำเข้าข้อมูลจากฐานข้อมูลจะมีเงื่อนไขดังนี้คือ ฐานข้อมูลจะต้องเป็นข้อมูลที่มาจาก Microsoft SQL Server 2000 เท่านั้น ในระบบนี้ได้มีการสร้างฐานข้อมูลไว้ให้ผู้ใช้ระบบสามารถนำข้อมูลเหล่านี้มาวิเคราะห์ได้ ซึ่งเป็นข้อมูลลูกค้าที่ใช้บริการใช้โทรศัพท์มือถือในระบบแบบใช้ก่อนจ่ายทีหลังระหว่างวันที่ 1 กรกฎาคม ถึง 30 กันยายน พ.ศ. 2547 จำนวน 1,500 เรคอร์ด

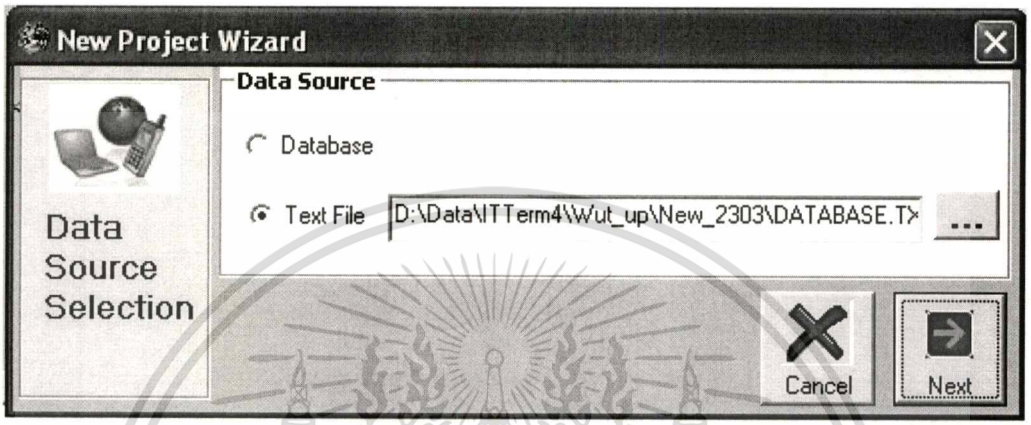
รูปที่ 5.1 เมนูหลักของระบบ

5.3.2 หน้าจอสำหรับการนำเข้าข้อมูลเข้าสู่ระบบ

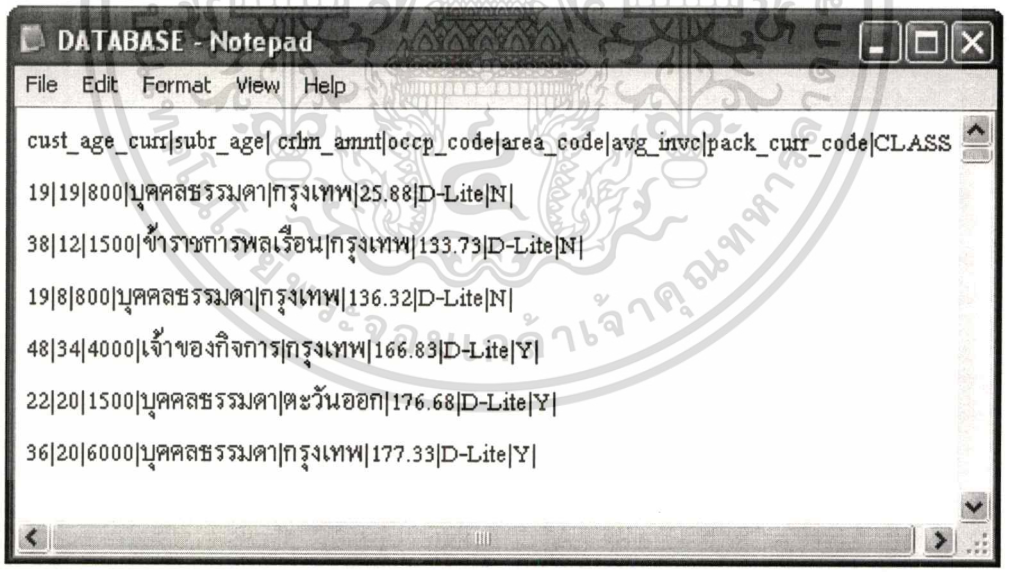
หน้าจอ Data Source Selection เป็นหน้าจอสำหรับการนำเข้าข้อมูลเข้าสู่ระบบ โดยมาได้จาก 2 แหล่ง คือ จากฐานข้อมูล หรือแฟ้มข้อความ ดังรูปที่ 5.2 ดังนี้

1. Database หมายถึง ข้อมูลที่ต้องการวิเคราะห์นั้นเก็บอยู่ในฐานข้อมูล
2. Text File หมายถึง ข้อมูลที่ต้องการวิเคราะห์เก็บอยู่ในเท็กซ์ไฟล์ เมื่อผู้ใช้ระบบเลือก Text File ระบบจะแสดงปุ่ม Browse เพื่อให้ผู้ใช้ระบบเลือกเท็กซ์ไฟล์ที่ต้องการนำมา

วิเคราะห์ การโหลดข้อมูลจากเท็กซ์ไฟล์เข้าโปรแกรมนั้น ข้อมูลจะต้องมีรูปแบบของข้อมูลดังนี้ คือ ชื่อแอตทริบิวต์ (Attribute) จะอยู่ในแถวแรก และคั่นข้อมูลแต่ละตัวด้วยเครื่องหมายไปป์ (|) ดังรูปที่ 5.3



รูปที่ 5.2 หน้าจอการนำเข้าข้อมูล

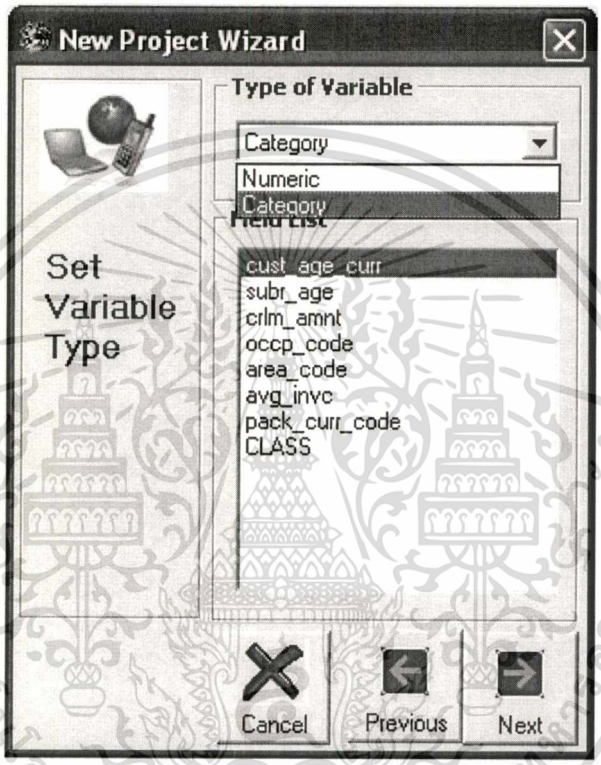


รูปที่ 5.3 ตัวอย่างแฟ้มข้อความที่ใช้ในการวิเคราะห์

5.3.3 หน้าจอกำหนดประเภทของตัวแปร

หน้าจอ Set Variable Type เป็นหน้าจอที่ใช้กำหนดประเภทของตัวแปรว่าเป็นตัวแปรแบบ Category หรือ Numeric ในหน้าจอนี้ประกอบด้วย 2 ส่วน ดังรูปที่ 5.4 ดังนี้

1. Field List แสดงตัวแปรทั้งหมดที่สามารถนำมาวิเคราะห์ได้ ซึ่งระบบจะดึงชื่อตัวแปรขึ้นมาให้ทั้งหมด
2. Type of Variable ใช้ในการกำหนดประเภทของตัวแปร ซึ่งระบบจะตั้งค่าเริ่มต้นทุกตัวแปรเป็น Category เพื่อป้องกันในกรณีที่ผู้ใช้ระบบไม่ได้ตั้งค่าตัวแปรคลาสเลเบล



รูปที่ 5.4 หน้าจอการกำหนดประเภทของตัวแปร

5.3.4 หน้าจอกำหนดตัวแปรเพื่อวิเคราะห์

หน้าจอ Tree Modeling เป็นหน้าจอในการกำหนดตัวแปรที่ใช้กำหนดคลาสเลเบล และตัวแปรที่จะนำมาวิเคราะห์แบ่งเป็น 3 ส่วน ดังรูปที่ 5.5 ดังนี้

1. Field List แสดงตัวแปรทั้งหมดที่ระบบสามารถนำมาวิเคราะห์ได้
2. Class Attribute ใช้ในการกำหนดคลาสเลเบล
3. Other Attribute(s) ใช้ในการกำหนดตัวแปร ที่จะนำมาวิเคราะห์ร่วมกับคลาสเลเบล ซึ่งแบ่งออกเป็น 2 ส่วนดังนี้

   - All Other(s) ใช้ตัวแปรทุกตัวที่เหลืออยู่ใน Field List มาวิเคราะห์ร่วมกับคลาสเลเบล

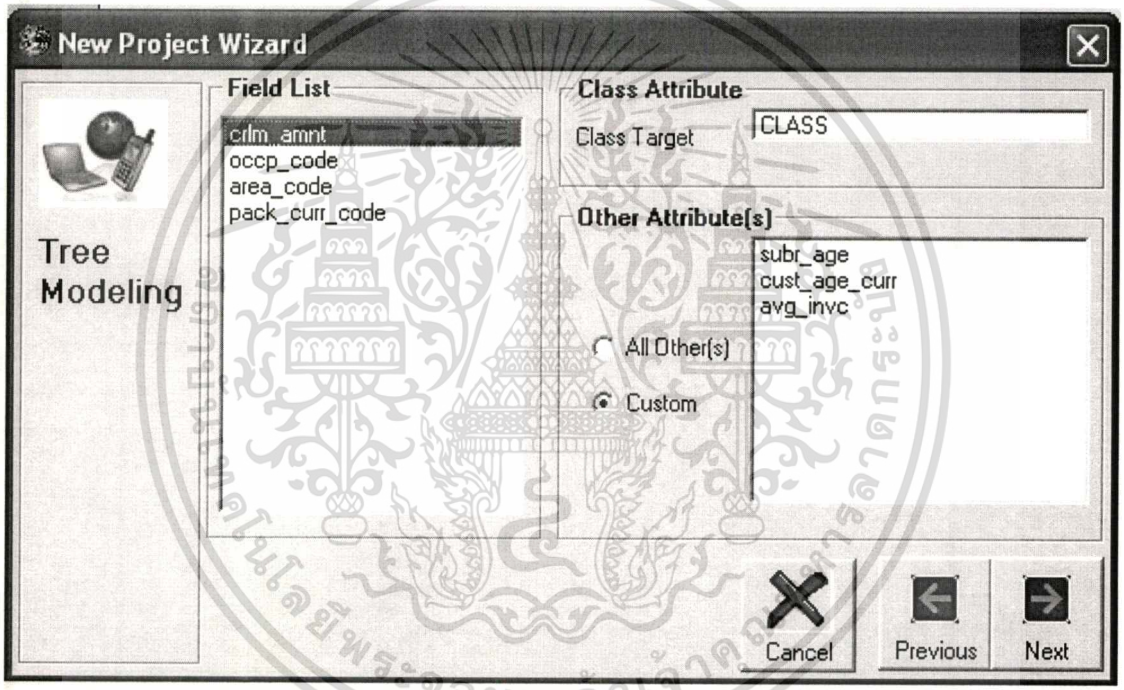
- Custom เลือกตัวแปรเพียงบางตัวมาวิเคราะห์ร่วมกับคลาสเลเบล

การทำงานในหน้าจอนี้ คือ

1. ในการเลือกตัวแปรต่างๆ ที่นำมาวิเคราะห์นั้นจะใช้วิธีลากแล้วปล่อย

2. การที่ผู้ใช้ระบบจะทำงานในหน้าจอถัดไปได้นั้น ต้องกำหนดคลาสเลเบลก่อนเสมอ ปุ่ม [Next] จึงจะสามารถทำงานได้ ซึ่งก็คือฟิลด์ CLASS หรือฟิลด์ที่ใช้แบ่งคลาสนั่นเอง

3. ในกรณีผู้ใช้ระบบเลือกที่ All Others ไม่จำเป็นที่จะต้องลากตัวแปรมาไว้ที่ช่อง Other Attributes และถ้าผู้ใช้ระบบเลือกตัวแปรเข้ามา ระบบจะเปลี่ยนค่าที่ตั้งไว้เป็น Custom ทันที



รูปที่ 5.5 หน้าจอการกำหนดตัวแปรเพื่อวิเคราะห์

5.3.5 หน้าจอหน้าจอในการตั้งค่าของแบบจำลอง

หน้าจอ Rule Setting & Partitioning เป็นหน้าจอที่ใช้ในการแบ่งข้อมูลที่จะนำมาวิเคราะห์และทดสอบ ทั้งใช้ในการการแตกแบบจำลองต้นไม้ ซึ่งแบ่งออกเป็น 2 ส่วน ดังรูปที่ 5.6 ดังนี้

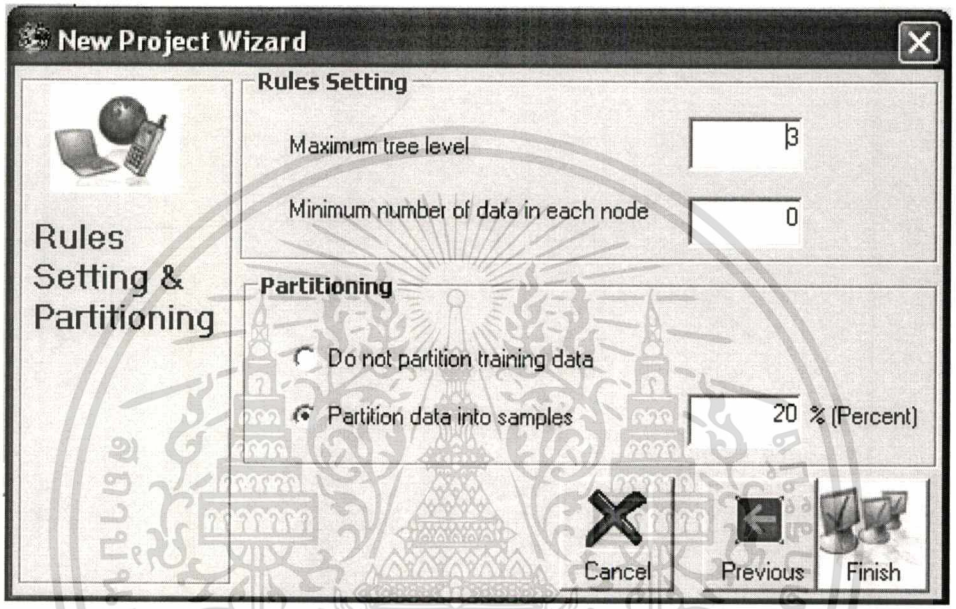
1. Partitioning ใช้แบ่งข้อมูลไว้เพื่อใช้เป็นข้อมูลในการทดสอบแบบจำลองต้นไม้ ซึ่งคิดเป็นเปอร์เซ็นต์ของจำนวนข้อมูลทั้งหมด

2. Rules Setting ใช้ในการกำหนดกฎในการหยุดแตกแบบจำลองต้นไม้

- Maximum Level ใช้ในการกำหนดระดับของแบบจำลองต้นไม้ที่สามารถแตก

ได้ซึ่งระดับรูทจะถือเป็นระดับที่ 1

- Minimum number of data in each node ใช้ในการกำหนดข้อมูลน้อยสุดของแต่ละโหนด ที่สามารถนำมาแตกออกได้ ถ้ามีจำนวนข้อมูลน้อยกว่าค่าที่กำหนดไว้ระบบก็จะหยุดแตกแบบจำลอง



รูปที่ 5.6 หน้าจอในการตั้งค่าของแบบจำลอง

5.3.6 หน้าจอแสดงผลการวิเคราะห์ข้อมูล

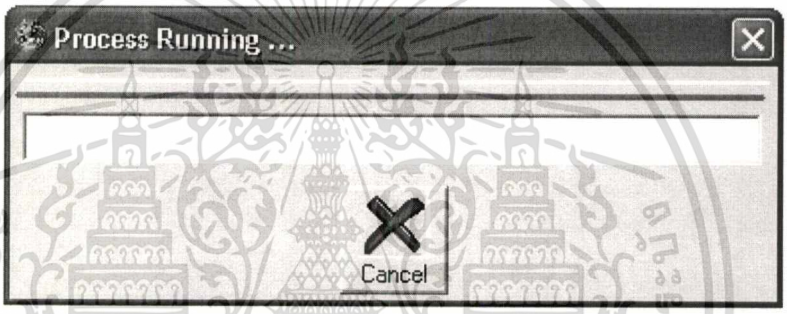
เมื่อใส่ข้อมูลต่างๆ จนครบเรียบร้อยแล้วกดปุ่ม [FINISH] โปรแกรมจะแสดงผลจากการวิเคราะห์ หรือผลของการสอนระบบออกมาในรูปแบบของต้นไม้ ซึ่งขณะที่โปรแกรมกำลังทำงานอยู่จะขึ้นหน้าจอแสดงสถานะของการประมวลผล โดยสามารถหยุดการทำงานโดยการกดปุ่มยกเลิก [Cancel] ดังรูปที่ 5.7 เมื่อโปรแกรมประมวลผลเรียบร้อยจะแสดงผลของการทำงานว่าเสร็จแล้วดังรูปที่ 5.8 ให้กดปุ่ม [Success] ระบบจะแสดงผลจากการทำงานดังรูปที่ 5.9 โดยหน้าจอจะแบ่งออกเป็น 3 ส่วนดังนี้

1. ส่วนแสดงแบบจำลองต้นไม้รูปเล็ก จะใช้ช่วยในกรณีที่แบบจำลองต้นไม้มีขนาดใหญ่เกินหน้าจอจะแสดงผลได้ ผู้ใช้สามารถที่จะคลิกไปที่หมายเลขโหนดที่ต้องการดูในแบบจำลองรูปเล็ก ระบบก็เลื่อนภาพของแบบจำลองต้นไม้ขนาดใหญ่ให้เลื่อนไปที่ยังโหนดที่ผู้ใช้ต้องการโดยอัตโนมัติ

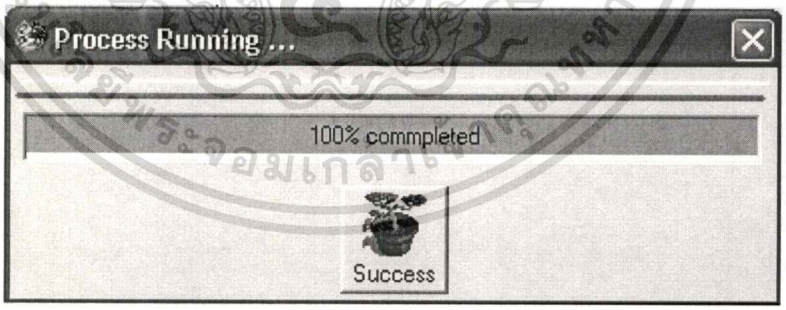
2. ส่วนแสดงแบบจำลองต้นไม้รูปเล็ก จะใช้ในกรณีที่แบบจำลองต้นไม้มีขนาดใหญ่เกินหน้าจอที่แสดงไว้ ผู้ใช้สามารถที่จะคลิกไปที่หมายเลขโหนดในแบบจำลองรูปเล็กและระบบก็เลื่อนภาพไปที่โหนดที่ผู้ใช้ต้องการโดยอัตโนมัติ

3. ส่วน STATUS REPORT จะใช้ในการแสดงรายละเอียดขั้นตอนการทำงานของโปรแกรมรวมทั้งบอกถึงข้อผิดพลาดในกรณีระบบไม่สามารถวิเคราะห์ผลได้

4. ส่วนแสดงแบบจำลองต้นไม้ ในส่วนนี้จะแสดงรายละเอียดของโหนดในแต่ละโหนดว่ามีข้อมูลอยู่เท่าไร และแสดงค่าความเชื่อมั่น ซึ่งในหน้าจอนี้จะแสดงข้อมูล 2 ส่วนคือในส่วนของ Training Tree Model และ Testing Tree Model

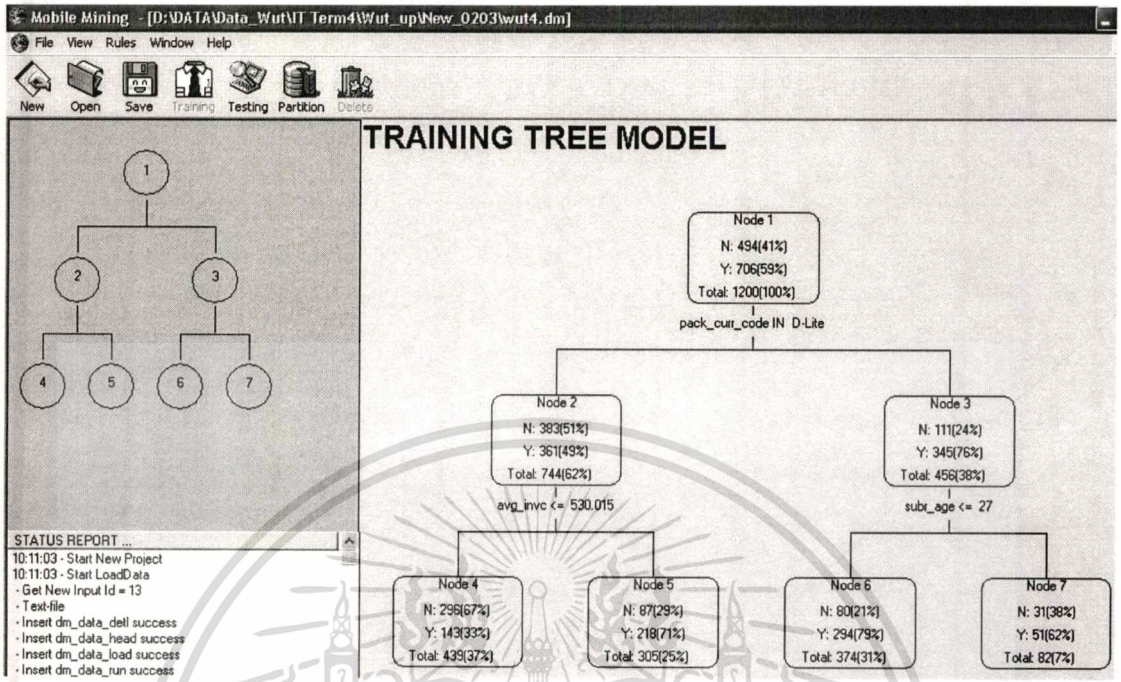


รูปที่ 5.7 หน้าจอสถานะช่วงประมวลผล



รูปที่ 5.8 หน้าจอสถานะประมวลผลเรียบร้อย

เมื่อโปรแกรมประมวลผลเรียบร้อยแล้ว สามารถเก็บแบบจำลองต้นไม้ที่ต้องการได้ เพื่อนำมาใช้วิเคราะห์ข้อมูลอีกครั้งในภายหลัง โดยการกดปุ่มบันทึก [Save] ไฟล์ที่เกิดขึ้นจะมีนามสกุล MB ส่วนกระบวนการทำงานหรือขั้นตอนการประมวลผลแบบจำลองต้นไม้ ก็จะบันทึกไปพร้อมๆ กัน โดยใช้ชื่อเดียวกันกับไฟล์แบบจำลองต้นไม้ แต่มีนามสกุล LOG ดังรูปที่ 5.10



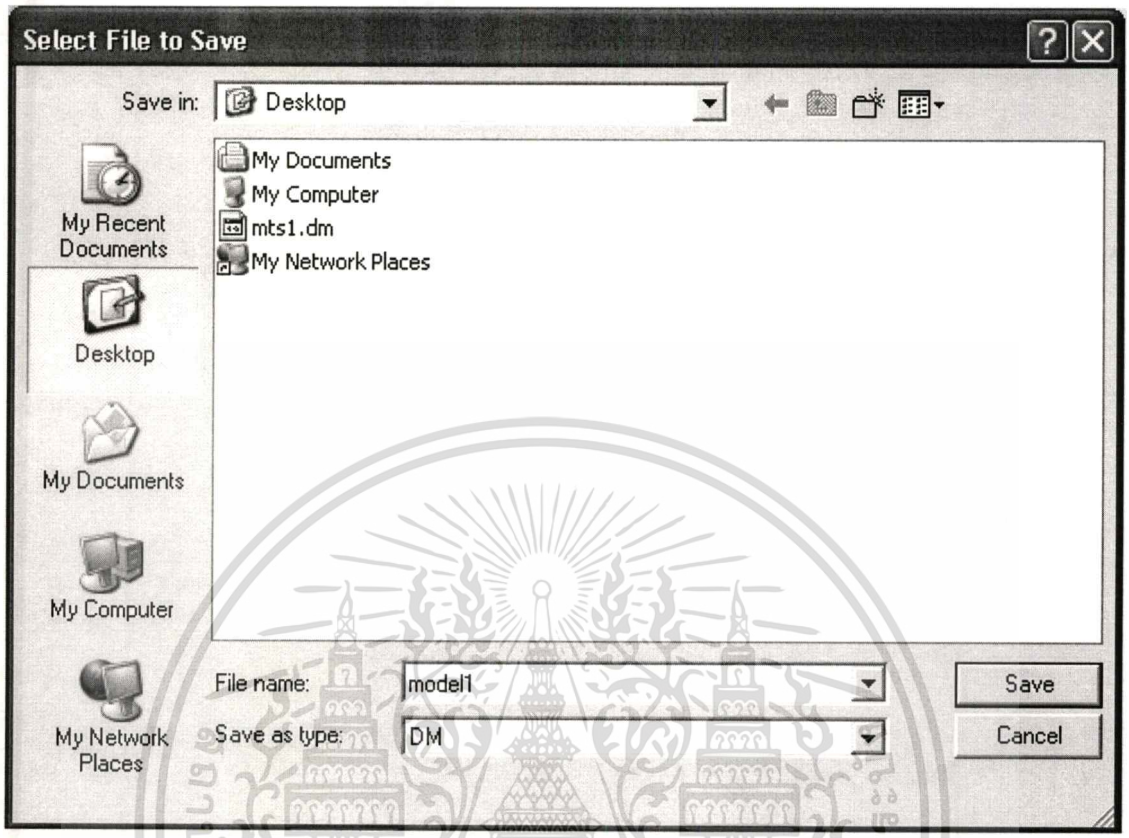
รูปที่ 5.9 หน้าจอแสดงผลการวิเคราะห์ข้อมูล

5.3.7 การแสดงข้อมูลในส่วนของ Testing Set

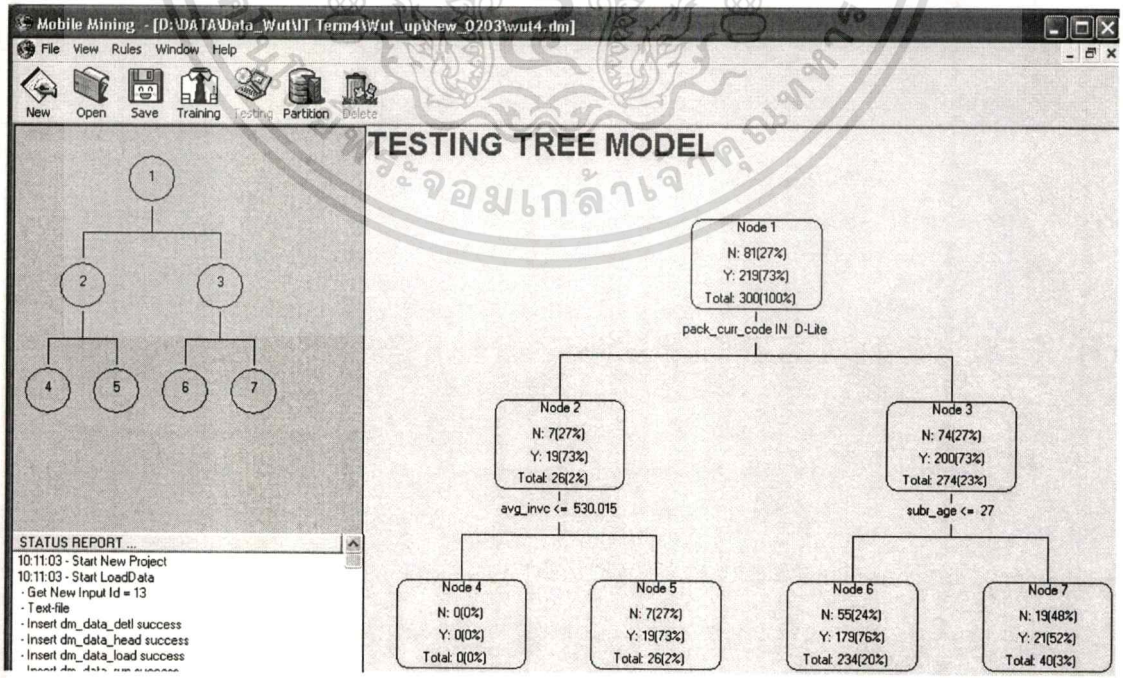
ในการดูข้อมูลในส่วนของ Testing Set นั้นให้ผู้ใช้เข้าไปที่เมนู View --> Testing Data Set หรือจะคลิกที่ตรงปุ่ม [Testing] ระบบจะนำข้อมูลในส่วนที่เป็นการทดสอบมาวิเคราะห์ให้โดยใช้เงื่อนไขแบบจำลองต้นไม้เดิมซึ่งจะได้ผลดังรูปที่ 5.11

5.3.8 การปรับแต่งแบบจำลองต้นไม้

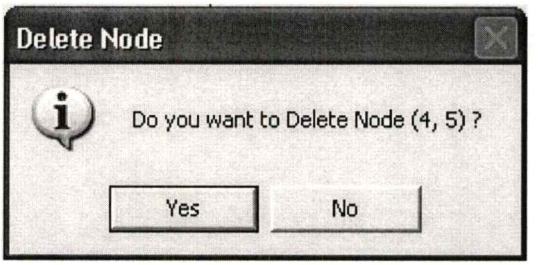
เมื่อผู้ใช้ระบบวิเคราะห์แล้วว่าโหนดบางโหนดนั้นไม่มีผลต่อการวิเคราะห์ผู้ใช้ระบบสามารถปรับแต่งแบบจำลองต้นไม้เองได้ โดยให้คลิกไปที่หมายเลขโหนดในแบบจำลองต้นไม้จะพบกล่องข้อความดังรูปที่ 5.12



รูปที่ 5.10 หน้าจอการบันทึกแบบจำลองต้นไม้

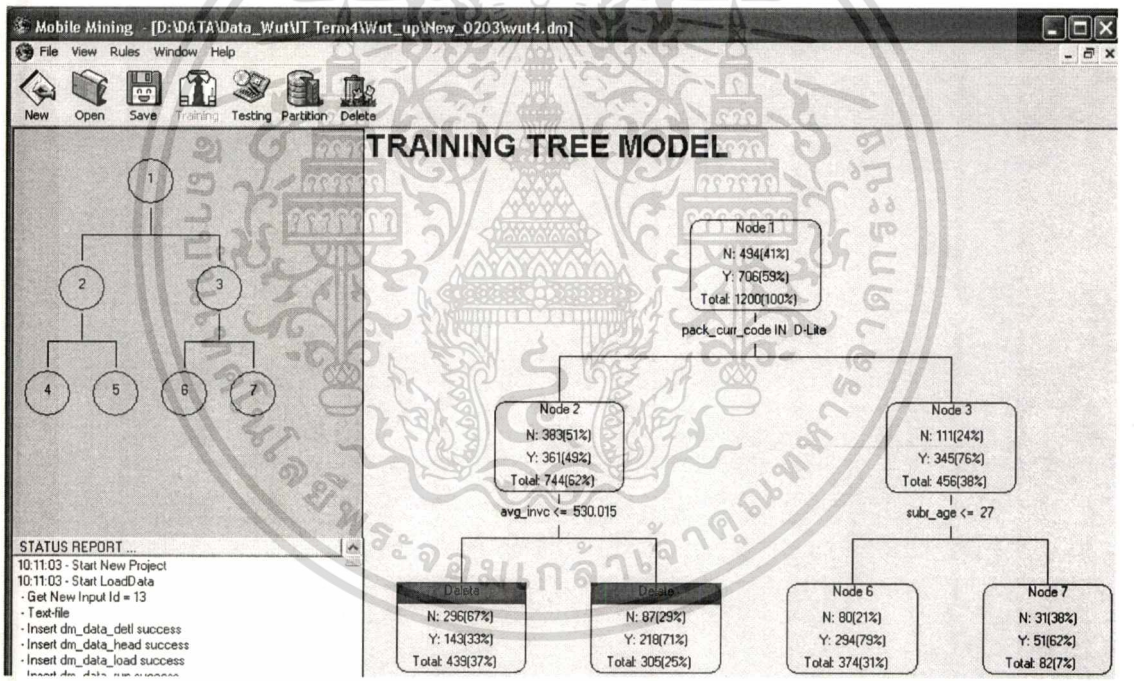


รูปที่ 5.11 หน้าจอแสดงผลของการใช้ Testing data



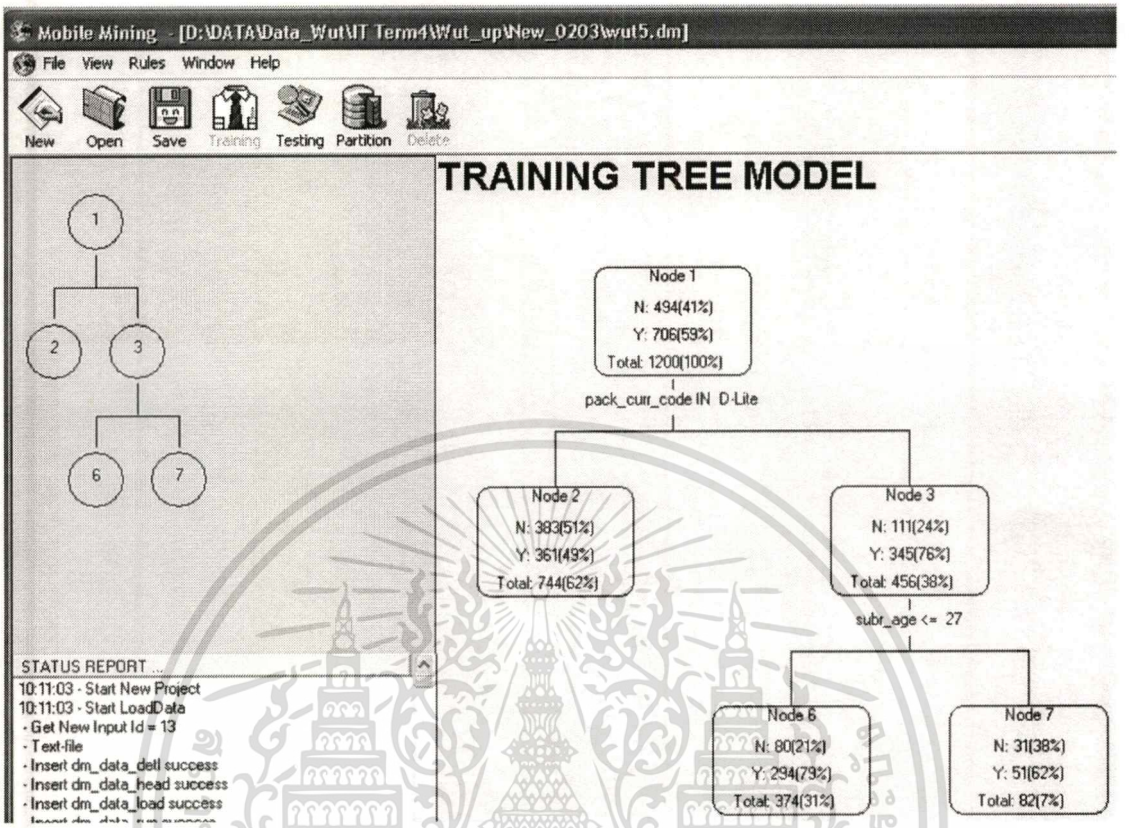
รูปที่ 5.12 กล่องข้อความการลบโหนด

ถ้าต้องการลบโหนดให้กดปุ่ม Yes ถ้าไม่ต้องการลบให้กดปุ่ม No กรณีที่กดปุ่ม Yes จะขึ้นแถบสีแดงที่โหนดที่ต้องการลบดังรูปที่ 5.13



รูปที่ 5.13 หน้าจอการลบโหนด

โหนดใดที่ผู้ใช้พิจารณาแล้วพบว่า โหนดนั้นไม่น่าจะมีผลต่อการวิเคราะห์ข้อมูลการใช้งานโทรศัพท์เคลื่อนที่ ผู้ใช้สามารถลบโหนดนั้นได้ โดยการคลิกไปยังโหนดที่ต้องการลบ ระบบจะทำแถบสีแดงแสดงคำว่า Delete ในโหนดนั้น ซึ่งถ้าผู้ใช้ยืนยันที่จะลบ ให้คลิกที่ไอคอน Delete ระบบจะลบโหนดนั้นทิ้งไปดังรูปที่ 5.14



รูปที่ 5.14 หน้าจอแสดงการลบโหนด

5.3.9 การออกจากระบบ

ในการออกจากระบบนั้นสามารถทำกดปุ่ม File - -> Exit ที่หน้าจอ

5.4 การวิเคราะห์ผลลัพธ์

เนื่องจากข้อมูลที่ได้รับมาเป็นกลุ่มลูกค้าที่ยกเลิกหรือเปลี่ยนแปลงผู้ให้บริการ เมื่อผ่านตัวแยกแยะ SLIQ ในการทำค้าต้าไมนิ่ง จากรูปที่ 5.11 พบว่า

1. ถ้าโปรโมชั่นของลูกค้าเป็น D-lite และยอดเงินเฉลี่ยย้อนหลัง 3 เดือนก่อนปิดบริการมากกว่า 530.015 บาท ลูกค้ามีแนวโน้มที่จะเปลี่ยนใจไปใช้บริการของผู้ให้บริการรายอื่น 73%
2. ถ้าโปรโมชั่นของลูกค้าเป็น D-Max, D-Medium หรือ D-Flex และระยะเวลาการใช้โทรศัพท์มือถือน้อยกว่าหรือเท่ากับ 27 เดือน ลูกค้ามีแนวโน้มที่จะเปลี่ยนใจไปใช้บริการของผู้ให้บริการเลยอื่น76%
3. เมื่อเข้าไปดูข้อมูลตั้งต้นที่ได้มาจากกลุ่มลูกค้าเหล่านี้ประกอบ พบว่าลูกค้ากลุ่มนี้ไม่

เน้นการโทรแต่ออกเน้นการรับสายเป็นหลัก ทั้งยังมีการใช้งานบริการที่เป็นแบบไม่ใช่เสียง (Non Voice) ที่น้อยมากจนถึงไม่มีเลย เช่น บริการส่งข้อความสั้น (Short Message Service) หรือ (SMS) บริการส่งข้อความเคลื่อนไหว (Multimedia Message Service) และบริการสื่อสารความเร็วสูง (GPRS) เป็นต้น

4. ควรมีบริการเสริมที่เกี่ยวกับบริการที่ไม่ใช่เสียง เข้าไปกับรายการส่งเสริมการขาย หรือแยกออกมาเป็นบริการเสริมอีกชนิดหนึ่ง ที่สามารถสมัครใช้บริการได้ต่างหากไปอีกชนิดหนึ่งไปเลยไปกับเพิ่มไป เพื่อส่งเสริมให้ผู้ใช้บริการมีปริมาณการใช้งานที่มากยิ่งขึ้น

5. สำหรับลูกค้ากลุ่มที่มีค่าใช้จ่ายสูงแตกต่างจากคนอื่นมากๆ ผู้ให้บริการควรแนะนำให้ลูกค้าเปลี่ยนโปรโมชั่นใหม่ที่เหมาะสมมากกว่า เนื่องจากธรรมชาติของโปรโมชั่น D-Lite จะเหมาะกับลูกค้าที่มีพฤติกรรมการโทรออกที่ค่อนข้างน้อย เน้นการรับสายเป็นหลัก โปรโมชั่นนี้มีค่าบริการเหมาจ่ายรายเดือน 250 บาท มีอัตราค่าบริการนาทีละ 4 บาทใน 100 แรกและคิดค่าบริการนาทีละ 3 บาทใน 150 นาทีต่อมา เมื่อโทรเกิน 250 นาทีจะคิดค่าบริการนาทีละ 2 บาท ซึ่งลูกค้าต้องมีปริมาณการโทรที่ค่อนข้างสูงกว่าจะใช้อัตราค่าบริการนาทีละ 2 บาท

5.5 ความรู้ที่ได้จากการทำไมนิ่ง

1. ดาต้าไมนิ่งเป็นวิธีการที่ให้เราสามารถหาความรู้ และทำความเข้าใจข้อมูลที่ยุ่งยากซับซ้อนต่อการวิเคราะห์ ให้สามารถเข้าใจได้ง่ายขึ้น จากแหล่งข้อมูลที่มีประมาณมหาศาล ให้กลายเป็นสารสนเทศที่มีคุณค่า สามารถนำมาใช้ประโยชน์ได้จริง

2. วิธีการหรือเทคนิคที่ใช้ในการทำไมนิ่ง เพื่อการวิเคราะห์ข้อมูลมีหลายเทคนิคด้วยกัน ขึ้นอยู่กับว่าเราจะไปวิเคราะห์กับข้อมูลประเภทไหน หรือต้องการคำตอบในแนวทางอย่างไร จึงต้องเลือกใช้เทคนิคให้ถูกต้องตรงตามวัตถุประสงค์ของทางธุรกิจ หรือเป้าเหมายทางการตลาด

3. การเลือกใช้เทคนิคของการทำไมนิ่งยังต้องมองให้ออกอีกด้วยว่า จะเน้นที่คำตอบหรือความรวดเร็วในการประมวลผลข้อมูล เช่น SLIQ จะเหมาะสมในการนำไปวิเคราะห์ข้อมูลที่มีปริมาณมาก ใช้เวลาในการประมวลผลค่อนข้างน้อยเมื่อเทียบกับอัลกอริทึมอื่นๆ และยังให้ผลการวิเคราะห์ที่มีความถูกต้องสูง

4. การทำไมนิ่งทำให้รู้ว่าข้อมูลที่อยู่รอบตัวเราล้วนแต่มีคุณค่าทั้งสิ้น หากรู้จักที่จะนำวิธีการหรือเทคนิคต่างๆ มาประยุกต์ใช้กับข้อมูล แล้วดึงหรือสกัดเอาความโดดเด่น

ของข้อมูลแต่ชุดมาใช้ได้อย่างเหมาะสมถูกต้อง ถูกเวลา ย่อมสามารถสร้างเป็นความรู้ สร้างรายได้ สร้างความได้เปรียบในเชิงธุรกิจเป็นอย่างมาก โดยเฉพาะในธุรกิจผู้ให้บริการโทรศัพท์เคลื่อนที่มีการแข่งขันกันอย่างดุเดือด การนำไมนิ่งมาใช้ในการบริหารความสัมพันธ์กับลูกค้าเป็นสิ่งที่สำคัญ และมีความจำเป็นอย่างมาก เพื่อเป็นการรักษาฐานลูกค้าเก่าให้เหนียวแน่น และสร้างผู้ใช้บริการรายใหม่

บทที่ 6
บทสรุปและข้อเสนอแนะ

ผลจากการศึกษาและพัฒนาระบบวิเคราะห์ข้อมูล กรณีศึกษาการวิเคราะห์ข้อมูลการใช้งานโทรศัพท์เคลื่อนที่ เพื่อการออกโปรโมชั่นโดยการใช้ต้นไม้การตัดสินใจนั้น ได้ข้อสรุปดังต่อไปนี้

6.1 สรุปผลโครงการ

จากตัวแยกแยะ SLIQ เพื่อการทำดาต้าไมนิ่ง กรณีศึกษาการวิเคราะห์ข้อมูลการใช้งานโทรศัพท์เคลื่อนที่ เพื่อการออกโปรโมชั่นโดยการใช้ต้นไม้การตัดสินใจพบว่าระบบสามารถทำงานได้ตามวัตถุประสงค์ คือ

1. ระบบสามารถช่วยในการวิเคราะห์ต้นไม้การตัดสินใจได้อย่างถูกต้อง ตรงตามแนวคิดของอัลกอริทึม SLIQ
2. ระบบสามารถแสดงค่าความเชื่อมั่นเพื่อเป็นส่วนช่วยในการตัดสินใจให้กับผู้ใช้ระบบได้
3. ระบบสามารถนำไปใช้ในการวิเคราะห์ข้อมูลเรื่องใดก็ได้ ไม่จำเป็นต้องใช้ในการวิเคราะห์เรื่องนี้เท่านั้น แต่ต้องอยู่ภายใต้แนวคิดต้นไม้การตัดสินใจและอัลกอริทึม SLIQ
4. ระบบสามารถแสดงผลการวิเคราะห์ให้อยู่ในรูปแบบที่ง่ายต่อความเข้าใจของผู้ใช้ระบบ โดยแสดงเป็นรูปแบบจำลองต้นไม้ (Tree Model)

6.2 ประโยชน์ที่ได้รับจากโครงการ

จากตัวแยกแยะ SLIQ เพื่อการทำดาต้าไมนิ่งกรณีศึกษาการวิเคราะห์ข้อมูลการใช้งานโทรศัพท์เคลื่อนที่ เพื่อการออกโปรโมชั่นโดยใช้ต้นไม้การตัดสินใจนั้นทำให้ผู้พัฒนาระบบได้รับประโยชน์ดังนี้

1. ทำให้ผู้พัฒนาระบบมีความรู้ความเข้าในทฤษฎีดาต้าไมนิ่งเพิ่มมากขึ้น เนื่องจากได้นำความรู้นั้นมาพัฒนาระบบ และทำให้ผู้พัฒนาระบบเห็นว่าทฤษฎีดาต้าไมนิ่งนั้นมีประโยชน์และสามารถนำไปช่วยในการตัดสินใจได้จริง

2. ทำให้ผู้พัฒนาระบบได้ศึกษาแนวคิดของอัลกอริทึม SLIQ
3. ทำให้ผู้พัฒนาระบบทราบถึงวิธีการบริหารโครงงาน ทำให้สามารถนำไปประยุกต์ใช้ในการทำงานได้เป็นอย่างดี
4. ทำให้ผู้พัฒนาระบบได้เรียนรู้การพัฒนาโปรแกรมด้วยภาษาวิชวลเบสิค มากยิ่งขึ้น
5. นอกจากนี้ระบบที่ได้พัฒนาเสร็จสิ้นแล้ว สามารถนำไปใช้ในในองค์กรเพื่อช่วยในการวิเคราะห์ข้อมูลลูกค้าได้ ซึ่งระบบนี้สามารถนำไปวิเคราะห์กับเรื่องอะไรก็ได้ ไม่จำเป็นต้องเกี่ยวกับการออกโปรโมชั่นเท่านั้น

6.3 ข้อเสนอแนะและแนวทางในการพัฒนาโครงการเพิ่มเติม

1. ในการวิเคราะห์ข้อมูลอะไรก็ตาม สิ่งที่สำคัญที่สุดก็คือข้อมูลที่ผู้ใช้ระบบนำมาใช้ ดังนั้น ข้อมูลต่าง ๆ ต้องผ่านขั้นตอนเตรียมข้อมูลมาอย่างถูกต้อง เพื่อที่จะส่งผลให้ผลการวิเคราะห์นั้นถูกต้องด้วย ซึ่งในระบบนี้ยังขาดขั้นตอนในการกรองข้อมูล
2. ข้อมูลที่ผู้พัฒนาระบบนำมาใช้ในการวิเคราะห์นั้นถือว่ามีจำนวนน้อยเกินไปเมื่อเทียบกับข้อมูลจริงที่อยู่ในองค์กร แต่เนื่องจากผู้พัฒนาระบบสามารถนำข้อมูลออกมาได้เพียงเท่านี้ จึงอาจทำให้ผลของการวิเคราะห์ข้อมูลจากฐานข้อมูลที่เตรียมไว้นั้นไม่ถูกต้องเท่าที่ควร
3. ระบบที่เกี่ยวกับการวิเคราะห์ข้อมูลนั้นควรจะสามารถมองมุมมองของรูปแบบจำลองต้นไม้ (Tree Model) ได้หลายมุมมอง และควรที่จะคำนวณค่าทางสถิติและแสดงรายงานต่างๆ เพื่อใช้ประกอบการตัดสินใจ ซึ่งยังขาดการพัฒนาในจุดนี้
4. ระบบที่ใช้ในการวิเคราะห์ดาต้าไมนิ่งนั้นมีอัลกอริทึมอยู่หลายอัลกอริทึมที่สามารถนำมาใช้ในการวิเคราะห์ แต่ในระบบนี้ใช้แค่อัลกอริทึม SLIQ เท่านั้น ซึ่งถ้าเป็นไปได้ก็ควรจะมีอัลกอริทึมที่หลากหลายมากกว่านี้

บรรณานุกรม

Berry, M.J.A., and Linoff, G.S., 2000. **Mastering Data Mining: The Art and Science of Customer Relationship Management.** New York, Wiley.

Berson, Alex, Smith , Stephen, and Thearling, Kurt. 1999. **Building Data Mining Applications for CRM.** Montreal, Canada: McGraw-Hill Companies.

Cabena, Peter, Hadjnian, Stadler, Verhees, Zanasi, and Zanasi, Alessandro. 1997. **Discovering Data: Mining From Concept to Implementation.** Upper Saddle, New Jersey: Prentice Hall.

Han, Jiawei, and Kamber, Micheline. 2001. **Data Mining: Concepts and Techniques.** San Francisco, CA: Morgan Kaufmann.

Hong, Mingsheng. 2004. **SLIQ (Supervised Learning in Quest).** [Online]. Available: http://www.cs.cornell.edu /~mshong/SLIQ.ppt.

Mehta, Manish, Agrawal, Rakesh, and Rissanen, Jorma. 2001. **SLIQ: A Fast Scalable Classifier for Data Mining.** [Online]. Available: http://www.cs.yorku.ca/~jarek/courses/6421/sliq.pdf.

Piatetsky-Shapiro, Gregory, and Frawley, William J. 2001. **Geographic Data Mining & Knowledge Discoveryowledge Discovery in Database.** New York, Taylor & Francis.

ประวัติผู้เขียน

ชื่อผู้เขียน	นายวุฒิไกร มะลิลา
ระดับมัธยมศึกษาตอนปลาย	โรงเรียนเบญจมราชูทิศ จังหวัดจันทบุรี
ระดับอุดมศึกษา	คณะเทคโนโลยีการเกษตร สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
วุฒิการศึกษาระดับปริญญาตรี	(วท.บ.) เทคโนโลยีการจัดการ
ประสบการณ์การทำงาน	บมจ. โทเทิ่ล แอ็คเซ็ส คอมมูนิเคชั่น