การเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มของอัลกอริทึม DECISION TREES และ C4.5

THE EFFICIENCY COMPARISION CLASSIFICATION ALGORITHM OF DECISION TREES AND C4.5

โดย

ศิรประภา นนทะจันทร์

SIRAPRAPA NONTACHANT

อาจารย์ที่ปรึกษา

รศ.ดร. อาริต ธรรมโน

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน
หลักสูตรวิทยาศาสตร์มหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ภาคเรียนที่ 1 ปีการศึกษา 2552

THE EFFICIENCY COMPARISION CLASSIFICATION ALGORITHM OF DECISION TREES AND C4.5

SIRAPRAPA NONTACHNANT

A REPORT SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS OF THE COURSE

SYSTEM DEVELOPMENT PROJECT

MASTER OF SCIENCE PROGRAM IN INFORMATION TECHNOLOGY

FACULTY OF INFORMATION TECHNOLOGY

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

1/ 2009

ใบรับรองโครงการพัฒนาระบบงาน (System Development Project)

เรื่อง

การเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มของอัลกอริทึม DECISION TREES และ C4.5

THE EFFICIENCY COMPARISION CLASSIFICATION ALGORITHM OF DECISION TREES AND C4.5

นางสาว ศิรประภา นนทะจันทร์

รหัสประจำตัว 48066538

ขอรับรองว่ารายงานฉบับนี้ ไม่ได้คัดลอกมาจากที่ใด

รายงานนี้ได้รับการตรวจสอบและอนุมัติให้เป็นส่วนหนึ่งของ

การศึกษาวิชาโครงการพัฒนาระบบงาน หลักสูตรวิทยาศาสตร์มหาบัณฑิต (เทคโนโลยีสารสนเทศ)

ภาคเรียนที่ 1 ปีการศึกษา 2552

..... อาจารย์ที่ปรึกษา

(รศ.ดร. อาริต ธรรมโน)

..... กรรมการสอบ

(รศ.ดร. วรพจน์ กรีสุระเดช)

..... กรรมการสอบ

(ผศ.ดร. ธนารัตน์ ชลิดาวงศ์)

หัวข้อ	การเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มของอัลกอริทึม (Decision Tree และ C4.5)
นักศึกษา	นางสาวศิรประภา นนทะจันทร์
รหัสนักศึกษา	48066538
ปริญญา	วิทยาศาสตร์มหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2552
อาจารย์ที่ปรึกษา	รศ.ดร.อาริต ธรรมโน

บทคัดย่อ

เนื่องจากในปัจจุบันข้อมูลเข้ามามีบทบาทและมีความสำคัญต่อองค์กรเป็นอย่างมาก ซึ่งข้อมูลเหล่านี้สามารถเปลี่ยนเป็นข่าวสารที่ล้วนมีประโยชน์ต่อการพิจารณาและตัดสินใจด้านธุรกิจ ซึ่งเทคนิคหนึ่งที่เข้ามาช่วยในการประมวลผลและสามารถสังเคราะห์เอาข้อมูลที่มีอยู่มาทำการวิเคราะห์ เพื่อช่วยสนับสนุนในการตัดสินใจนั้นคือ การทำดาต้าไมนิ่ง ซึ่งเทคนิคการจัดหมวดหมู่นี้ เป็นการแบ่งกลุ่มของข้อมูลที่เราสนใจว่าอยู่ในกลุ่มใด โดยมีการนำข้อมูลในอดีตที่เคยจัดกลุ่มแล้วมาสร้างเป็นโมเดลที่ใช้ในการทำนายกลุ่ม

Title	Decision Trees และ C4.5 (The efficiency comparison classification algorithm of Decision Trees and C4.5)
Student	Miss Siraprapa Nontachant
Student ID.	48066538
Degree	Master of Science
Program	Information Science
Major	Information Science
Academic Year	2009
Advisor	Assoc.Prof.Dr. Arit Thommano

ABSTRACT

Since the current day are important for business. Those data can be translate into bunch of information whish are considered very important to business for analysis and decision making. Data Mining in the technology for mining and prediction data for support decision in business. One of the most interesting technique is "Classification" technique. This technique is classifying the data to specific group by using the classified data in the past to create the model for forecasting in data group this technique use to forecasting group of classified data and group data in the future

กิตติกรรมประกาศ

ในการพัฒนาและศึกษาการทำดาต้าไมนิ่งของโครงการนี้ ได้รับความสนับสนุนอย่างดีจากหลายฝ่ายไม่ว่าจะเป็นการให้คำปรึกษาและกำลังใจผู้เขียนขอขอบคุณทุก ๆ คนที่มีส่วนร่วมในกาทำงานในครั้งนี้

นางสาวศิรประภา นนทะจันทร์

17 กันยายน 2552

สารบัญ

หน้า

สารบัญ(ต่อ)

หน้า

สารบัญภาพ

รูปที่ หน้า

สารบัญตาราง

รูปที่ **หน้า**

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันธุรกิจมีอัตราการแข่งขันที่สูงมากซึ่งส่งผลให้องค์กรต้องหาวิธีในการเพิ่มประสิทธิภาพการทำงานภายในขององค์กร ซึ่งการนำเอาเทคโนโลยีดาต้าไมนิ่งเข้ามาช่วยก็เป็นอีกทางหนึ่งที่สามารถนำมาช่วยเพิ่มประสิทธิภาพในการทำงานได้ เนื่องจากองค์กรส่วนใหญ่มักจะเผชิญกับปัญหาของข้อมูลดิบจำนวนมากแต่ข้อมูลที่ประยุกต์ใช้ได้นั้นมีน้อย ซึ่งดาต้าไมนิ่งสามารถดึงความรู้ออกมาจากข้อมูลจำนวนมากที่ถูกเก็บสะสมไว้ได้ เพื่อนำเอาข้อมูลเหล่านั้นมาทำการสังเคราะห์เพื่อให้ได้รูปแบบที่สามารถช่วยสนับสนุนการตัดสินใจ

ในโลกของธุรกิจบริษัทต่างๆ จะพยายามหาเทคนิคที่สามารถนำความสำเร็จมาสู่บริษัท เช่น ในโลกธุรกิจขนาดย่อมจะสร้างความสัมพันธ์กับลูกค้า โดยสังเกตจากความต้องการ ความชอบและความสนใจของลูกค้า และอาจมีการเรียนรู้ได้จากผลสะท้อนในอดีตว่าจะทำอย่างไรให้การบริการลูกค้ามีประสิทธิภาพดีขึ้นในอนาคต หรือบริษัทที่เป็นผู้ออกบัตรเครดิตและธนาคารต่างๆ จะมีกระบวนการที่ใช้ดาต้าไมนิ่งให้เป็นประโยชน์ ในการตัดสินใจว่าลูกค้ากลุ่มใดเป็นกลุ่มที่ดี, ทำความเข้าใจลูกค้า, ช่วยในการแยกประเภทของลูกค้าและจะทำนายกลุ่มของบุคคลที่คาดว่าจะเข้ามาเป็นลูกค้าในอนาคต เป็นต้น อย่างไรก็ตามการเรียนรู้นั้นต้องมากกว่าการเก็บสะสมข้อมูลอย่างตรงไปตรงมา ซึ่งจะทำให้การทำงานไม่เป็นประสิทธิภาพ

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

การศึกษาโครงงานนี้มีวัตถุประสงค์

1. เพื่อศึกษาและทำความเข้าใจกระบวนการทำงานของการทำดาต้าไมนิ่ง
2. เพื่อศึกษาเทคนิคการทำดาต้าไมนิ่งโดยใช้รูปแบบของการแบ่งกลุ่มของข้อมูล (Classification) โดยศึกษาเทคนิคโครงข่ายประสาทเทียม (Neural Networks) แบบ Back-propagation และวิธีการสร้างต้นไม้ตัดสินใจ (Decision Tree) โดยใช้อัลกอริทึมแบบ ID3 และ C 4.5
3. เพื่อพัฒนาโปรแกรมแบบจำลองของการจำแนกกลุ่มของข้อมูลด้วยวิธีการสร้างต้นไม้ตัดสินใจโดยเลือกใช้อัลกอริทึม ID3 และ C4.5 เพื่อเปรียบเทียบประสิทธิภาพ เพื่อเป็นแนวทางในการเลือกใช้อัลกอริทึมที่มีเหมาะสมและมีประสิทธิภาพมากที่สุด

1.3 ขอบเขตการศึกษา

การศึกษาโครงงานพัฒนาระบบนี้มีขอบเขตการศึกษาคือเป็นการพัฒนาโปรแกรมจำลองเพื่อทำการประเมินการทำงานของอัลกอริทึมการสร้างต้นไม้ตัดสินใจแบบ ID3 และ C4.5 เพื่อเปรียบเทียบประสิทธิภาพของผลการทำนายกับค่าของข้อมูลจริงที่ใช้ในการทดสอบ โดยมีขอบเขตของการศึกษาดังต่อไปนี้

1. ใช้เทคนิคการกระบวนการทำคาด้าไมนิ่ง (Data Mining)ในการพัฒนาโปรแกรม
2. เทคนิคการจัดหมวดหมู่ข้อมูล (Classification)
3. การสร้างต้นไม้ตัดสินใจ (Decision Tree) โดยใช้อัลกอริทึม ID3 และอัลกอริทึม C4.5
4. ฐานข้อมูลที่จะใช้ทดสอบการทำงานและสร้างรูปแบบใช้ฐานข้อมูล Microsoft SQL Server 2005
5. ฐานข้อมูลที่ใช้ในการทดสอบต้องมีรูปแบบของข้อมูลที่เป็นลักษณะตัวแปรที่กำหนดความเป็นไปได้ของข้อมูลอย่างชัดเจน เช่น Yes หรือ No หรือตัวแปรที่มีการจัดลำดับของข้อมูล เช่น hot, mild, cool เป็นต้น ซึ่งจะนำมาใช้กับเทคนิคของการจำแนกกลุ่มโดยใช้เทคนิคของการสร้างต้นไม้ตัดสินใจ (Decision Tree)
6. การแสดงผลของโปรแกรมจะแสดงโดยใช้รูปแบบของ tree view เพื่อใช้ในการอ่านค่าผลลัพธ์ของการทดสอบโปรแกรม
7. คาด้าเซตในการรันโปรแกรม จะแบ่งออกเป็น 70% สำหรับการสร้างโมเดลต้นไม้ตัดสินใจ และ 30% สำหรับทำการทดสอบการทำงานของโมเดลที่ได้
8. สำหรับการสร้างต้นไม้ตัดสินใจของ C4.5 จะมีการกำหนดจำนวนของระดับของต้นไม้ (โหนด) ของการสร้างต้นไม้ตัดสินใจจากผู้ใช้งาน สำหรับกรณีที่กำหนดจำนวนระดับของต้นไม้ที่มากกว่าค่าสูงสุด โปรแกรมจะสร้างต้นไม้จากค่าสูงสุดที่เป็นไปได้

1.4 ทฤษฎีหรือแนวคิดที่ใช้ในการศึกษา

โครงงานเรื่องนี้ใช้หลักการทำงานของ คาด้าไมนิ่งเป็นขั้นตอนหนึ่งที่มีความสำคัญในกระบวนการค้นหาลักษณะแฝงของข้อมูล ที่มีประโยชน์ในฐานข้อมูล (Knowledge Discovery in Database: KDD) ซึ่งโดยทั่วไปกระบวนการของ KDD นั้นประกอบด้วยขั้นตอนต่างๆ ดังนี้ การคัดเลือกข้อมูล, การรวบรวมข้อมูล, การกรองข้อมูล, การแปลงรูปแบบข้อมูล, การไมนิ่งข้อมูล, การเลือกรูปแบบ และการประเมินผลหรือวิเคราะห์

โดยจะทำการศึกษาอัลกอริทึมของ ID3 และ C4.5 เพื่อเปรียบเทียบการทำงานของทั้งสองอัลกอริทึม โดยจะทำการพิจารณาจากการกำหนดขั้นของการแตกกิ่งของ C4.5 เพื่อเปรียบเทียบกับความถูกต้องหลังจากนำโมเดลของทั้งสองอัลกอริทึมมาใช้งาน โดยจะทำการทดสอบกับข้อมูลจากการสุ่มขึ้นมา 30% จากข้อมูลทั้งหมด และใช้ข้อมูลในการสร้างโมเดล 70%

1.5 ขั้นตอนของการศึกษา

1. ศึกษากระบวนการและหลักการทำงานของดาต้าไมนิ่ง (Data Mining)
2. ศึกษาการทำดาต้าไมนิ่งโดยเลือกใช้การแบ่งกลุ่มของข้อมูลด้วยเทคนิคของการสร้างต้นไม้ตัดสินใจ (Decision Tree) โดยเลือกใช้อัลกอริทึม ID3 และ C4.5
3. ศึกษาการทำงานของอัลกอริทึม ID3 และ C4.5 เพื่อนำมาใช้ในการพัฒนาโปรแกรม ในการสร้างรูปแบบในลักษณะของ tree view
4. ศึกษาและเลือกใช้เครื่องมือต่างๆ เพื่อใช้ในการพัฒนาระบบ
   โดยทำการเลือกใช้เครื่องมือในการพัฒนาดังต่อไปนี้
   1) เลือก Application Tool ในการพัฒนาระบบ Microsoft Visual Studio.Net 2005
   2) เลือกใช้ระบบฐานข้อมูล Microsoft SQL Server 2005
   3) ระบบปฏิบัติการ Microsoft Windows XP Professional Service Pack 2
5. ออกแบบ User Interface ในการใช้งาน และสร้างฐานข้อมูลที่จะใช้ในการทดสอบการทำงานของระบบ
6. ออกแบบและวิเคราะห์ระบบที่จะพัฒนาระบบงาน
7. ทำการพัฒนาระบบเพื่อทำการสร้างแบบจำลอง
8. ทดสอบการทำงานและทำการปรับปรุงแก้ไขระบบงาน
9. สรุปผลกการดำเนินงานของโครงการ

1.6 ประโยชน์ที่คาดว่าจะได้รับ

1. เพื่อให้ทราบถึงการทำงานของกระบวนการของการทำค้าต้าไมนิ่ง
2. ใช้เป็นข้อมูลประกอบการตัดสินใจในการทำงานเกี่ยวกับการจำแนกกลุ่มของข้อมูลโดยใช้เทคนิคที่เหมาะสม
3. ทำให้ทราบถึงประสิทธิภาพของอัลกอริทึมที่เราเลือกมาพิจารณาเปรียบเทียบการทำงาน
4. สามารถใช้ระบบที่พัฒนาไปทำการจัดหมวดหมู่ของข้อมูลที่ต้องการได้

5. **เปรียบเทียบให้เห็นการทำงานที่แตกต่างกันของทั้งสองอัลกอริทึม และสามารถเปรียบเทียบความถูกต้องที่ได้จากการใช้อัลกอริทึมทั้งสอง**

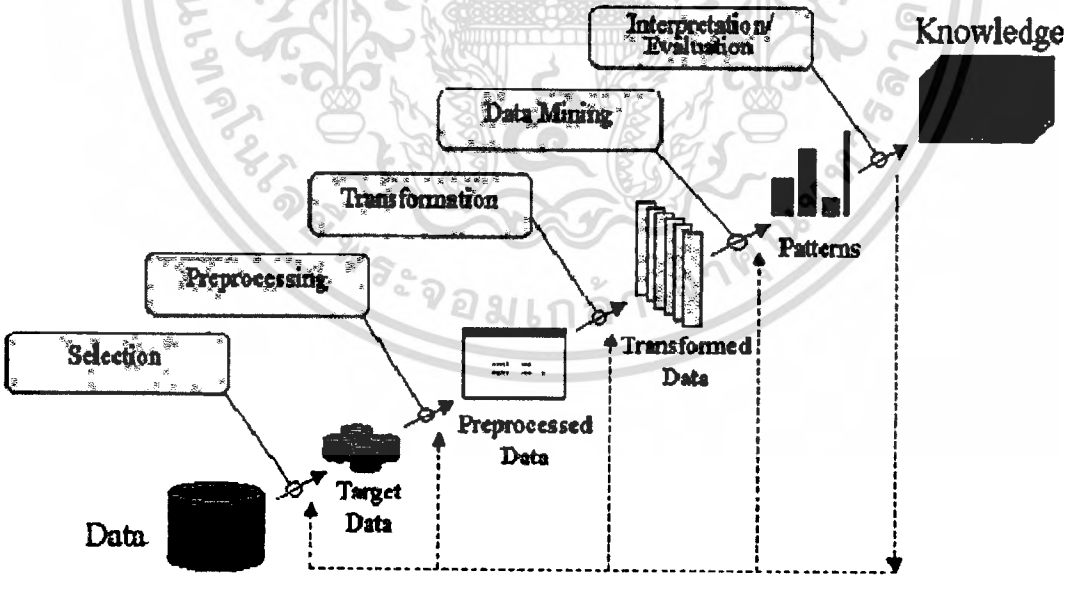
บทที่ 2
แนวคิดและทฤษฎีที่เกี่ยวข้อง

2.1 ความหมายของการทำดาต้าไมนิ่ง

ดาต้าไมนิ่งเป็นขั้นตอนหนึ่งที่มีความสำคัญในกระบวนการค้นหาลักษณะแฝงของข้อมูล ที่มีประโยชน์ในฐานข้อมูล (Knowledge Discovery in Database: KDD) ซึ่งโดยทั่วไปกระบวนการของ KDD นั้นประกอบด้วยขั้นตอนต่างๆ ดังนี้ การคัดเลือกข้อมูล, การรวบรวมข้อมูล, การกรองข้อมูล, การแปลงรูปแบบข้อมูล, การไมนิ่งข้อมูล, การเลือกรูปแบบ และการประเมินผลหรือวิเคราะห์

2.2 กระบวนการกการทำงานของต้าไมนิ่ง

ขั้นตอนการสืบค้นข้อมูลจากฐานข้อมูลประกอบด้วยกระบวนการต่างๆ แบ่งออกได้เป็นขั้นตอนต่างๆ ดังนี้



รูปที่ 2.1 ขั้นตอนการทำดาต้าไมนิ่ง

2.2.1 การกรองข้อมูล (Data Cleaning)

เป็นกระบวนการที่ทำให้เกิดความมั่นใจในคุณภาพของข้อมูล ที่จะนำมาใช้ในการวิเคราะห์ โดยการเลือกเอาเฉพาะคอลัมน์ที่สำคัญและคาดว่าจะนำไปใช้ ปรับเปลี่ยนข้อมูลให้มีค่าเหมาะสมในการตัดสินใจ และแก้ไขข้อมูลให้ถูกต้องสมบูรณ์ ระหว่างการทำขั้นตอนการกลั่นกรองข้อมูลจะมีปัญหาบ่อยๆ ที่มักพบได้ ได้แก่

- Noisy Data คือตัวแปรตัวหนึ่งหรือมากกว่ามีค่าซึ่งเกินกว่าค่าที่เราคาดไว้ ซึ่งอาจจะหมายถึงแง่ดีหรือแง่ร้ายก็ได้ ในแง่ดีก็คือมันจะแสดงอย่างชัดเจนถึงโอกาสซึ่งเรากำลังมองหาอยู่ ในแง่ร้ายคือมันอาจจะเป็นข้อมูลที่ไม่สมบูรณ์ สาเหตุที่เกิดขึ้นได้อาจจะมาจากความเลินเล่อของมนุษย์ ตัวอย่างเช่น Operator ใส่อายุให้คนเป็น 300 ปี หรือใส่ค่าของรายได้เป็นติดลบ ค่าเหล่านี้ควรจะถูกแก้ไข หรือเอาออกจากการวิเคราะห์ ควรมีขั้นตอนการเช็คข้อมูลก่อนนำมาใช้
- Missing Value คือค่าที่ไม่ได้แสดงในข้อมูลที่เราได้เลือกแล้ว หรือค่าที่ไม่สมบูรณ์ที่เราลบออกไประหว่างการทำ Noise Detection ค่าอาจจะหายไปเพราะเกิดจากความเลินเล่อของมนุษย์ เพราะว่าไม่มีข้อมูลนั้นระหว่างการทำ Input ข้อมูล การจัดการกับค่าที่หายไปนั้นสามารถจัดการได้ด้วยเทคนิคที่ต่าง ๆ กัน

2.2.2 การรวบรวมข้อมูล (Data Integration)

คือการนำเอาข้อมูลจากหลายๆ แหล่งมารวมกัน

2.2.3 การคัดเลือกข้อมูล (Data Selection)

จุดประสงค์ในการเลือกข้อมูล เพื่อที่จะทำการเลือกข้อมูลที่จะใช้ในการวิเคราะห์การทำคาต้าไมนิ่งในเบื้องต้น โดยข้อมูลที่ทำการเลือกนั้นจะขึ้นอยู่กับจุดประสงค์ของแต่ละองค์กร และโปรแกรมประยุกต์ที่เลือกใช้ ตัวแปรที่ใช้ในการพิจารณาประกอบไปด้วย

2.2.4 การแปลงรูปแบบข้อมูล (Data Transformation)

เป็นการแปลงข้อมูลที่เลือกมาให้อยู่ในรูปที่เหมาะสม สำหรับการนำไปใช้ในการวิเคราะห์อัลกอริทึม (Algorithm) และแบบจำลองที่ใช้ในการทำคาต้าไมนิ่งต่อไป

2.2.5 การทำไมนิ่ง (Data Mining)

การใช้เทคนิคต่างๆ ในกระบวนการการทำคาต้าไมนิ่ง เพื่อให้ได้รูปแบบของข้อมูลที่จะนำมาใช้ในการพิจารณาและสนับสนุนการตัดสินใจ เพื่อทำนายแนวโน้มการเกิดขึ้นของข้อมูลที่คาดว่าจะเกิดขึ้นในอนาคต

2.2.6 การเลือกรูปแบบ (Pattern Evaluation)

เป็นกระบวนการเลือกรูปแบบที่เหมาะสมจากขั้นคาต้าไมนิ่งเพื่อนำไปสู่การค้นพบความรู้

2.2.7 การประเมินผลลัพธ์ที่ได้ (Knowledge Presentation)

เป็นขั้นตอนการแปลความหมาย และการประเมินผลลัพธ์ที่ได้ว่ามีความเหมาะสม หรือตรงกับวัตถุประสงค์ที่ต้องการหรือไม่ โดยทั่วไปควรมีการแสดงผลในรูปแบบที่สามารถเข้าใจได้โดยง่าย

2.3 เทคนิคการทำดาต้าไมนิ่ง

สามารถแบ่งเทคนิคที่ใช้ในการทำไมนิ่งได้เป็น 4 เทคนิคด้วยกันคือ

2.3.1 เทคนิคการจัดหมวดหมู่ของข้อมูล (Classification)

เทคนิคที่ใช้ในการทำดาต้าไมนิ่งโดยวิธี Classification แบ่งออกเป็น 2 แบบหลักๆ คือ Decision Tree และ Neural Network ซึ่งทั้งสองวิธีนี้จะใช้วิธีการแบบ Supervised Learning คือ การสร้างแบบจำลองการจัดหมวดหมู่โดยใช้ข้อมูลจากกลุ่มที่ได้ทำการกำหนดไว้ล่วงหน้าแล้ว ที่เรียกว่า Training Set เทคนิคที่ใช้ในการจัดหมวดหมู่แบ่งออกได้เป็น

- Decision Tree เป็นเทคนิคที่ค่อนข้างแพร่หลาย เนื่องจากผู้ใช้สามารถทำความเข้าใจผลลัพธ์ได้ง่าย โดยจะลักษณะเป็นโครงสร้างเป็นแบบต้นไม้ คือจะประกอบด้วยโนดบนสุด (Root Node) ซึ่งจะแตกออกเป็นโนดลูกได้ และโนดในระดับล่างเรียกว่าโหนดใบ (Leaf Node) จากลักษณะเช่นนี้จะเห็นว่าการเดินทางจากโหนดราก (Root Node) ไปยังโนหดใบ (Leaf Node) จะเป็นการเดินทางในลักษณะเส้นทางเดียวเท่านั้น โดยเส้นทางที่ได้มานั้นจะเป็นตัวที่ใช้ใช้ในการจัดแบ่งหมวดหมู่ของข้อมูลออกเป็นกลุ่ม
- Neural Network มีพื้นฐานการจำลองมาจากการทำงานของสมองมนุษย์ โดยจะมีชั้น Input, ชั้น Hidden และชั้น Output ที่มีลักษณะเป็นการเชื่อมโยงต่อกันไป โดยการเชื่อมโยงนี้เป็นการเชื่อมโยงกันโดยอาศัยค่าความสำคัญของข้อมูล ส่วนค่าของ Output ที่ได้มาก็จะนำมาทำการเปรียบเทียบกับค่ามาตรฐานหรือค่าที่ได้การตั้งเอาไว้ ถ้าค่าที่ได้มานั้นไม่เป็นไปตามที่คาดไว้ก็จะมีการปรับเปลี่ยนค่าของ weight และทำการทำซ้ำจนกว่าจะได้ค่าที่ต้องการ

2.3.2 เทคนิคการจัดกลุ่มของข้อมูล (Clustering)

เป็นกระบวนการในการแบ่งกลุ่มข้อมูลออกเป็นกลุ่มเพื่อให้ง่ายต่อการวิเคราะห์ข้อมูล เช่น แบ่งตามอายุ รายได้ เป็นต้น ซึ่งการแบ่งกลุ่มของข้อมูลนี้ไม่สามารถกำหนดข้อมูลไว้ล่วงหน้าได้ว่าข้อมูลนี้ควรจะอยู่ในกลุ่มใดกลุ่มหนึ่ง หรือที่เรียกว่า Unsupervised Learning

2.3.3 เทคนิคกฎความสัมพันธ์ของข้อมูล (Association Rule)

เป็นวิธีในการวิเคราะห์หาความสัมพันธ์ของข้อมูลภายในกลุ่ม เพื่อใช้ลักษณะของข้อมูลหนึ่งไปหาความสัมพันธ์ของอีกข้อมูลหนึ่ง ซึ่งสามารถแบ่งออกได้เป็น

- Association Discovery ใช้วิเคราะห์การเลือกซื้อสินค้าจากรายการเดียวกัน โดยจะทำการพิจารณาถึงสินค้าที่ผู้บริโภคมีแนวโน้มว่าจะซื้อควบคู่กันไป การวิเคราะห์ในลักษณะเช่นนี้เรียกว่า Market Basket Analysis รูปแบบของกฎจะเป็นในลักษณะของ IF X then Y หรือ When X then Y โดยเหตุการณ์ทั้ง X และ Y จะเกิดขึ้นในเวลาเดียวกัน
- Sequential Pattern Discovery ใช้ระบุความเกี่ยวเนื่องกันในลักษณะที่เป็นลำดับของข้อมูลการซื้อสินค้า โดยมีจุดมุ่งหมายที่จะเข้าใจพฤติกรรมของผู้บริโภคต่อสินค้าในระยะยาว
- Similar Time Sequential Discovery ใช้ค้นหาความเกี่ยวเนื่องกันของข้อมูล 2 กลุ่ม ซึ่งจะมีการขึ้นต่อกันทางด้านของเวลา โดยจะมีรูปแบบการเคลื่อนที่ไปในทางเดียวกัน ผู้ขายมักจะใช้เพื่อดูแนวโน้มในการเตรียมสินค้าในสต็อก เช่นในกรณีที่ยอดขายของน้ำอัดลมเพิ่มขึ้น จะพบว่ายอดขายของมันฝรั่งก็สูงขึ้นตาม

2.4 กระบวนการการทำงานของ Classification

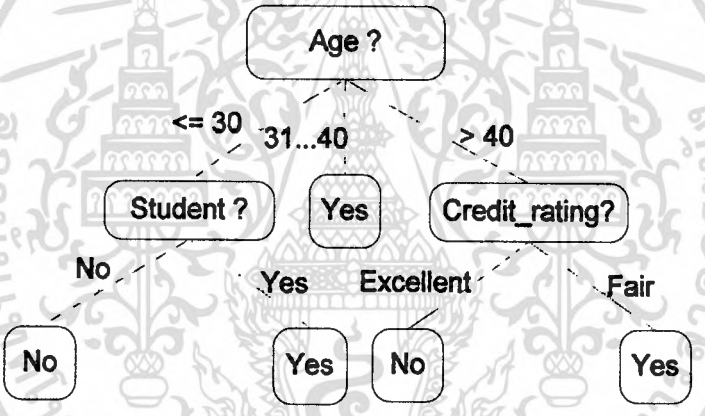
- การสร้างโมเดล (Model Construction) จะใช้ข้อมูลตัวอย่าง (training data) เพื่อใช้ในการฝึกฝน
- การใช้โมเดล (Model Evaluation) จะนำข้อมูลทดสอบ (testing data) คือข้อมูลที่เหลือจากการทดลองมาใช้ในการทดสอบโมเดล
- นำโมเดลที่ได้มาทำการจัดหมวดหมู่ของข้อมูล (Model Usage) การนำข้อมูลใหม่ที่ยังไม่เคยถูกแบ่งกลุ่มมาผ่านโมเดลเพื่อจัดหากลุ่มของข้อมูลที่เหมาะสม

บทที่ 3

การจำแนกกลุ่มของข้อมูลโดยใช้อัลกอริทึม ID3

3.1 การจัดหมวดหมู่โดยวิธี Decision Tree

เป็นแบบจำลองที่มีลักษณะคล้ายต้นไม้ ที่มีการทำงานแบบ Supervised Learning คือสามารถสร้างแบบจำลอง การจัดหมวดหมู่ได้จากกลุ่มตัวอย่างของข้อมูลที่ได้กำหนดไว้ล่วงหน้าแล้วที่เรียกว่า Training Set ได้อัตโนมัติ จะเห็นว่าจาก Root Node จนถึง Leaf Node จะมีเพียงเส้นทางเดียวเท่านั้น ซึ่งเส้นทางนี้จะอธิบายถึงกฎที่ใช้ในการจัดหมวดหมู่ของแต่ละกลุ่ม และในแต่ละ Leaf Node



รูปที่ 3.1 แสดงรูปแบบของ Decision Tree

3.1.1 หลักการทำงาน Decision Tree (อัลกอริทึม ID3)

- หา Attribute ที่สำคัญที่สุดมาทำการแบ่งข้อมูล โดย Attribute นี้จะถูกนำมาสร้างเป็น Root Node โดยจะมี Target Attribute เป็นผลลัพธ์ ซึ่งถูกกำหนดไว้ก่อน
- ถ้าข้อมูล sample อยู่ใน class เดียวกันแล้ว เราจะให้โนด นั้นเป็น leaf node โดยมีชื่อตาม class นั้น
- ถ้าข้อมูล sample ไม่ได้อยู่ใน class เดียวกัน จะทำการเลือก Attribute ที่ดีที่สุดที่สามารถทำแบ่ง ข้อมูล sample ได้ตาม class
- ทำการแบ่งข้อมูล sample ทั้งหมดที่ได้แตกออกตาม class

- วนกลับไปทำที่ขั้นตอนแรก โดยการหาแอทริบิวที่สำคัญที่สุดจากข้อมูลที่เข้ามา เพื่อหาตัวแบ่งต่อไป โดยจะหยุดเมื่อเงื่อนไขใด เงื่อนไขหนึ่งเป็นจริง
    - ข้อมูลข้อมูลตัวอย่างทั้งหมดอยู่ในกลุ่มเดียวกัน
    - ไม่มีแอทริบิวเหลืออยู่
    - ไม่มีข้อมูลตัวอย่างสำหรับการสร้างกิ่ง

3.1.2 วิธีการสร้างต้นไม้โดยใช้อัลกอริทึม ID3

วิธีการสร้างต้นไม้ตัดสินใจแบบ ID3 ใช้หลักการของ *Information Theory* โดยในการที่จะเลือกแอทริบิวเพื่อใช้ในการตัดสินใจของแต่ละโนดนั้นจะพยายามเลือกแอทริบิวที่มีค่า *Information Gain* มากที่สุด การวัดค่า *Information Gain* นั้นสามารถใช้ค่า *Entropy* ซึ่งวัดความไม่แน่นอน หรือการกระจายของข้อมูลในชุดข้อมูล โดยข้อมูลที่อยู่ในกลุ่มเดียวกัน หรือคลาสเดียวกันแล้วจะมีค่า *Entropy* มีค่าต่ำที่สุดคือใกล้เคียงกับ 0 มากที่สุด และถ้าข้อมูลไม่มีความแตกต่างกันคือทั้งหมดอยู่ในกลุ่มเดียวกันหมดแล้วค่า *Entropy* จะมีค่าเป็น 0

การคำนวณหาค่า *Information gain* โดยใช้สูตร

$$I(s_1, s_2, ...s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad (3.1)$$

S คือ set ของ sample ทั้งหมด

S_i คือ จำนวนของ Sample ของ S ใน class C_i

p_i คือ ความน่าจะเป็นที่ Sample นั้นเป็นของ class C_i

การคำนวณหาค่า *Entropy* โดยใช้สูตร

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + ... + s_{mj}}{s} I(s_{1j} + ... + s_{mj}) \qquad (3.2)$$

A คือ Attribute มีค่าทั้งหมด v ค่า $\{a_1, a_2, .., a_v\}$ซึ่ง Attribute A สามารถทำการแบ่ง S ออกเป็น v $\{s_1, s_2, ..., s_v\}$

S_j คือ Sample ใน S ที่ Attribute A มีค่า a_j

S_{ij} คือ จำนวนของ Class C_i ใน subset s_j

คำนวณหาค่า *Information Gain* ในแต่ละ Attribute โดยใช้สูตร

$$I(s_{1j}, s_{2j}, ..., s_{mj}) = -\sum_{i=1}^{m} P_{ij} \log_2(P_{ij}) \qquad (3.3)$$

$P_{ij} = \frac{s_{ij}}{|s_j|}$ คือ ความน่าจะเป็นข้อมูลตัวอย่างของ S_j ในการแบ่งกลุ่มในClass C_j หลังจากนั้นนำไปคำนวณหาค่า Gain ของ Attribute A

การคำนวณหาค่า *Gain* โดยใช้สูตร

$$Gain(A) = I(s_1, s_2 ... s_3) - E(A) \quad \textbf{(3.4)}$$

3.1.3 ข้อจำกัด Decision Tree (อัลกอริทึม ID3)

- การแบ่งกลุ่มข้อมูลของ Decision Tree กรณีเป็นข้อมูลที่มีค่าต่อเนื่อง เช่น ข้อมูลรายได้ ข้อมูลราคาต้องทำการแปลงให้อยู่ในช่วงหรือตัดเป็นกลุ่มก่อน
- เมื่ออัลกอริทึมเลือกว่าจะใช้ค่าไหนเป็นตัวแบ่งแล้ว จะไม่คำนึงถึงความสัมพันธ์ระหว่างแอทริบิว
- การจัดการกับข้อมูลที่ไม่ทราบค่า อาจมีผลกระทบกับผลลัพธ์ของ Decision Tree
- ต้นไม้ที่มีระดับชั้นมากเกินไป จะทำให้ข้อมูลที่ผ่านโหนดแตกออกเป็นชิ้นเล็กชิ้นน้อย ซึ่งข้อมูลเหล่านั้นจะไม่มีประโยชน์ในการนำมาใช้ทำการวิเคราะห์
- ปัญหาเรื่อง Over fitting / Over training เกิดจากการที่ แบบจำลองได้เรียนรู้เข้าไป ถึงรายละเอียดของข้อมูล มากเกินไปจนทำให้เกิดโหนดที่เป็นส่วนเฉพาะเจาะจงกับกลุ่มข้อมูลที่ใช้ ในการเรียนรู้ซึ่งจะต้องหาวิธีการในการตัดกิ่งนี้ออก

3.1.4 ตัวอย่างการคำนวณโดยใช้อัลกอริทึม ID3

RID	age	income	student	credit_ratin	class:buy_computer
1	<= 30	high	no	fair	no
2	<=30	high	no	excellent	no
3	31...40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	31...40	low	yes	excellent	yes
8	<=30	medium	no	fair	no
9	<=30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<=30	medium	yes	excellent	yes
12	31...40	medium	no	excellent	yes
13	31...40	high	yes	fair	yes
14	>40	medium	no	excellent	no

รูปที่ 3.2 แสดงตัวอย่างฐานข้อมูล

จากรูปแสดงข้อมูลที่ใช้เป็นตัวอย่างในการคำนวณจากตาราง AllElectronic ซึ่งเป็นข้อมูลของลูกค้าที่ได้ทำการซื้อเครื่องคอมพิวเตอร์จากบริษัท และได้ทำการแบ่ง class ออกเป็น 2 กลุ่มด้วยกัน (จาก Attribute buy_computer) คือกลุ่มของ Yes และ No (Yes: มีการซื้อคอมพิวเตอร์จากบริษัท และ No: ไม่ได้ทำการซื้อคอมพิวเตอร์) จากตารางเราจะได้ค่าของตัวอย่างข้อมูลในกลุ่มของ Yes = 9 และกลุ่มของ No = 5 จากข้อมูลทั้งหมด 14 Record สำหรับ Attribute ที่นำมาพิจารณานั้นประกอบไปด้วย age, income, student, credit_rating

จากตารางเราสามารถคำนวณค่าInformation Gain จากสมการที่ (1) ได้ดังนี้

$$I(s_1, s_2) = I(9,5) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.940$$

ต่อจากนั้นทำการคำนวณค่า *Entropy* ในแต่ละ Attribute จากมาการที่ (2) โดยจะทำการคำนวณหาค่า *Information Gain* ของ Attribute age ทำการพิจารณาจาก Class ที่เราได้ทำการแบ่งเอาไว้ คือ Yes และ No จะได้กลุ่มที่ทำการพิจารณาเพื่อคำนวณหาค่า *Information Gain* เป็น 3 กลุ่มดังนี้ โดยใช้สมการที่ (3) ในการคำนวณ

Age (<=30) พิจารณาได้ค่าของ Yes = 2 และ No = 3

$S_{11} = 2, S_{21} = 3$ $\quad I(s_{11}, s_{21}) = I(2,3) = 0.971$

Age (31...40) พิจารณาได้ค่าของ Yes = 4 และ No =0

$S_{12}=4, S_{22}= 0$ $\quad I(s_{12}, s_{22}) = I(4,0) = 0$ (แสดงว่าข้อมูลอยู่ในกลุ่มเดียวกันทั้งหมดแล้ว)

Age (>40) พิจารณาได้ค่าของ Yes = 3 และ No = 2

$S_{13}=3, S_{23}= 2$ $\quad I(s_{13}, s_{23}) = I(3,2) = 0.971$

ทำการคำนวณหาค่า *Entropy* ของ Attribute age โดยใช้สมการ (2)

$$E(age) = \frac{5}{14}I(s_{11}, s_{21}) + \frac{4}{14}I(s_{12}, s_{22}) + \frac{5}{14}I(s_{13}, s_{23})$$

$$= \frac{5}{14}(0.971) + \frac{4}{14}(0) + \frac{5}{14}(0.971) = 0.694$$

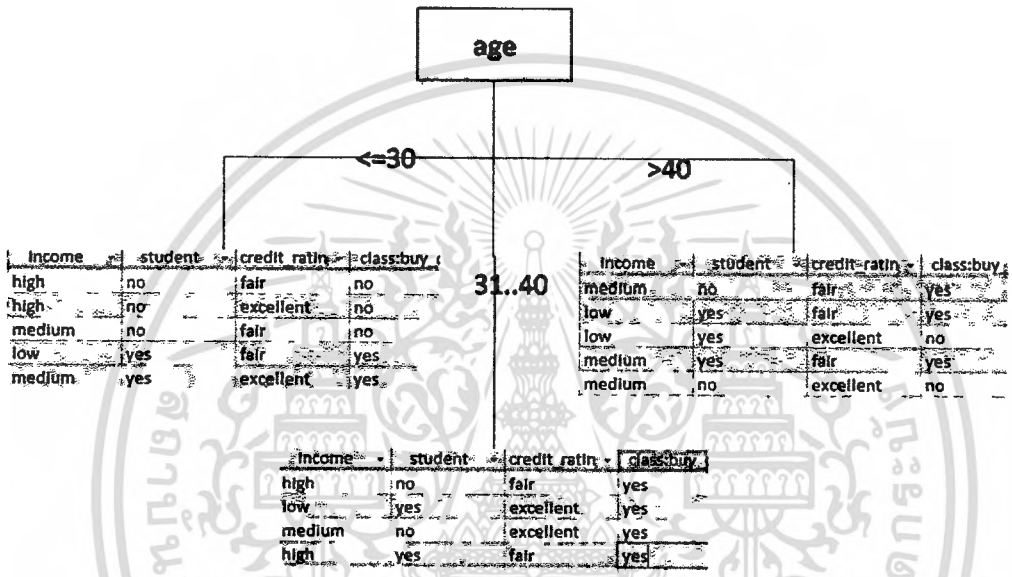
ทำการคำนวณหาค่าของ *Gain* ของ Attribute age โดยใช้สมการ (4)

$$Gain(age) = I(s_1, s_2) - E(age) = 0.246$$

ในลักษณะเดียวกันเราสามรถทำการคำนวณทุก ๆ Attribute ที่เหลืออยู่ทั้งหมดโดยใช้การคำนวณจากสมการข้างต้นในลักษณะเดียวกันกับการคำนวณหาค่าของ Attribute age ได้ค่า *Gain*ในแต่ละAttribute

$Gain(income) = 0.029, Gain(student) = 0.151, Gain(credit\_rating) = 0.048$

จากการพิจารณาค่า *Gain* ที่ได้ในแต่ละ Attribute เราจะได้ค่า *Gain* ของ Attribute age มีค่ามากที่สุด ดังนั้นเราจะทำการเลือก Attribute age เป็น Root Node ของต้นไม้ตัดสินใจที่จะใช้ในการแตกกิ่งเพื่ออธิบายถึงผลลัพธ์ของข้อมูล เราสามารแสดงผลของแตกกิ่งในรอบแรกของการคำนวณได้ดังนี้



รูปที่ 3.3 แสดงผลการแตกกิ่งในรอบแรกของการคำนวณ

จากรูปจะเห็นว่ากิ่งของ age = 31...40 มี class อยู่ในกลุ่มเดียวกันทั้งหมดคือ class Yes (*Entropy* = 0) ทำให้เราได้กิ่งของ age =31...40 เป็น leaf node และจัดอยู่ใน class Yes (โดยเราจะกำหนดให้ของ leaf Node มีชื่อตาม class นั้น) หลังจากนั้นทำการหาวนรอบทำซ้ำเพื่อทำการคำนวณจาก Attribute ที่เหลือ

จนกว่า Class ของแต่ละ Attribute จะอยู่ใน Class เดียวกันทั้งหมด และจะทำให้เราได้ผลลัพธ์ตามตารางที่ 3.1ข้างต้นที่ใช้อธิบายถึงลักษณะของข้อมูล

3.1.5 การอธิบายกฎที่ได้จากการสร้างต้นไม้ตัดสินใจ

จากการสร้างต้นไม้ตัดสินใจเราสามารถอ่านและนำเสนอข้อมูลที่ได้จากต้นไม้ โดยใช้กฎของการจำแนกข้อมูลในลักษณะของการอธิบายกฎแบบ *IF-THEN* สำหรับกฎแต่ละกฎนั้นจะได้จากการอ่านค่าเส้นทางเดินการแตกกิ่งของต้นไม้จากโนดรากไปยังแต่ละโนดใบ โดยสำหรับแต่ละค่าของแอทริบิว ที่เดินทางผ่านไปนั้นจะให้รูปแบบของเส้นทางของกฎที่จะใช้ *IF* เป็นตัวเชื่อมของประโยคกับผลลัพธ์ของกลุ่มคำ

ทำนายที่แสดงด้วยโนดใบด้วยประโยค *THEN* ที่เป็นผลอันเนื่องมาจากประโยค *IF* เหตุผลที่เลือกใช้การอธิบายโดยใช้กฎของ *IF-THEN* นี้เนื่องจากสามารถอธิบายให้เข้าใจได้ง่ายที่สุดโดยเฉพาะอย่างยิ่งกับต้นไม้ที่มีขนาดใหญ่มากๆ

จากผลของการใช้ต้นไม้ตัดสินใจ ทำให้เราได้กฎของความสัมพันธ์ในลักษณะของ *IF-THEN* โดยการพิจารณาเส้นทางของการแตกกิ่งของต้นไม้จากโนดรากไปยังใบแต่ละใบของต้นไม้ ซึ่งสามารถแสดงเป็นกฎดังนี้ (จากรูปที่ 3.1)

IF *age* = *"<=30"* AND *student* = *"no"* THEN *buys_computer* = *"no"*

IF *age* = *"<=30"* AND *student* = *"yes"* THEN *buys_computer* = *"yes"*

IF *age* = *"31...40"* THEN *buys_computer* = *"yes"*

IF *age* = *">40"* AND *credit_rating* = *"excellent"* THEN *buys_computer* = *"no"*

IF *age* = *">40"* AND *credit_rating* = *"fair"* THEN *buys_computer* = *"yes"*

บทที่ 4
การจำแนกกลุ่มของข้อโดยใช้อัลกอริทึม C4.5

4.1 วิธีการสร้างต้นไม้โดยใช้อัลกอริทึม C4.5

อัลกอริทึม C4.5 พัฒนาขึ้นมาเพื่อทำการแก้ไขปัญหาของการทำงานแบบ ID3 ที่เกิดจากการที่แบบจำลองได้เรียนรู้ถึงรายละเอียดของข้อมูลมากเกินไป (Over fitting/Overt training) เพื่อปรับปรุงการทำงานให้มีลักษณะที่ดียิ่งขึ้น โดยจะใช้วิธีการตัดกิ่งไม้ตัดสินใจนั้นๆ ออกไป

นอกจากนี้ยังพบว่าอัลกอริทึม ID3 จะมีความลำเอียงกับแอทริบิวที่มีช่วงของความเป็นไปได้ของข้อมูลมากๆ ทำให้แอทริบิวนั้นจะมีแนวโน้มในการนำมาสร้างเป็นรากหรือโนดของต้นไม้มากที่สุด เนื่องจากเมื่อแอทริบิวนั้นมีความแตกต่างของข้อมูลมากๆ จะทำให้ได้ค่า *Information gain* จากการคำนวณมีค่าเป็น 0 ซึ่งเป็นผลให้ค่า *Entropy* มีค่ามากที่สุดเสมอ อัลกอริทึมC4.5 จึงได้มีการใช้ค่า *gain ratio* ในการเลือกรากหรือโนดของต้นไม้แทนเพื่อลดความลำเอียงที่จะเกิดขึ้น การทำงานของ C4.5 มีหลักการทำงานดังนี้

4.1.1 การคำนวณหาค่า *gain ratio*

ค่า Gain ratio สามารถคำนวณได้จากสมการ

$$GainRation = \frac{Gain}{SplitInfo} \tag{4.1}$$

ค่าสารสนเทศการแบ่งแยก $= SplitInfo = -\sum_{i=1}^{n} \frac{|t_i|}{|T|} \log_2 \frac{|t_i|}{|T|}$

T = จำนวนเรคคอร์ดทั้งหมดของชุดข้อมูลทดสอบ

t_i = จำนวนเรคคอร์ดที่ถูกแบ่งคุณสมบัติตจากชุดข้อมูลย่อย

จากข้อมูลตัวอย่างข้อมูลที่ได้แสดงการสร้างต้นไม้แบบ ID3 มาแล้วนั้น เราจะใช้ค่า *Gain* จากการคำนวณจากอัลกอริทึม ID3 หลังจากนั้นจะทำการคำนวณหา *Split Info* เพิ่มเพื่อนำไปใช้ในการหาค่า *Gain ration* ซึ่งในอัลกอริทึม C4.5 จะใช้ค่านี้เป็นตัวแบ่งแทนการใช้ค่า *Gain*

ดังนั้นจะได้ค่า *Gain* ในแต่ละ Attribute ดังนี้

Gain (age) = 0.246

Gain (income) = 0.029

Gain (student) = 0.151

Gain (credit_rating) = 0.048

การคำนวณหา Split Info

Split info (age) = $-\frac{5}{14}\log_2\frac{5}{14}-\frac{4}{14}\log_2\frac{4}{14}-\frac{5}{14}\log_2\frac{5}{14}=1.577$

ดังนั้น Gain ration = 0.246/1.577 = 0.156

Split info (income) = $-\frac{4}{14}\log\ \frac{4}{14}-\frac{6}{14}\log_2\frac{6}{14}-\frac{4}{14}\log_2\frac{4}{14}=0.999$

ดังนั้น Gain ratio = 0.029/0.999 = 0.029

Split info (student) = $-\frac{7}{14}\log_2\frac{7}{14}-\frac{7}{14}\log_2\frac{7}{14}=1$

ดังนั้น Gain ratio = 0.151/1 = 0.151

Split info (credit_rating) = $-\frac{6}{14}\log_2\frac{6}{14}-\frac{8}{14}\log_2\frac{8}{14}=0.985$

ดังนั้น Gain ration = 0.048/0.985 = 0.048

จากการใช้ค่า Gain ration ในการคำนวณใหม่จะได้ Attribute age เป็นตัวที่มีค่ามากที่สุด ดังนั้นจะได้โนดรากของการสร้างต้นไม้เป็น Attribute age นั้นเอง

ทำการคำนวณซ้ำในลักษณะเดียวกันกับอัลกอริทึม ID3 เพียงแต่เราจะนำค่าของ *Gain ratio* มาเป็นค่าที่ใช้เลือกโหนดแทน เพื่อลดความลำเอียงของการใช้ *Information gain* ในการแบ่งที่จะมีแนวโน้มในการเลือกแอทริบิวที่มีคุณสมบัติของความแตกต่างของข้อมูลมากๆ

4.1.2 การตัดกิ่งไม้ตัดสินใจ

การตัดกิ่งไม้ตัดสินใจ เป็นเทคนิคที่จะช่วยในเรื่องของการที่ต้นไม้เกิดการเรียนรู้ที่จะสร้างโมเดลกับข้อมูลที่มีความเฉพาะเจาะจงมากเกินไป ดังนั้นเพื่อทำให้การทำงานของ Algorithm ให้มีประสิทธิภาพมากขึ้นจึงได้นำเทคนิคของการตัดกิ่งมาใช้ โดยเทคนิคในการตัดกิ่งมี 2 รูปแบบคือ

- การตัดกิ่งไม้ขณะเรียนรู้ (Pre-pruning) จะทำการตัดกิ่งในขณะที่มีการสร้างโมเดล โดยจะทำการหยุดการสร้างเมื่อตัววัดมีค่าต่ำกว่าเกณฑ์ที่ได้กำหนดไว้ซึ่งคำนวณจากค่าทางสถิติ เช่น Information Gain เป็นต้น
- การตัดกิ่งไม้หลังการเรียนรู้ (Post-pruning) การลดจำนวนกิ่งของต้นไม้หลังจากที่ได้โมเดลมาแล้ว โดยจะพิจารณาจากความซับซ้อนของโนดที่ได้จากการเรียนรู้

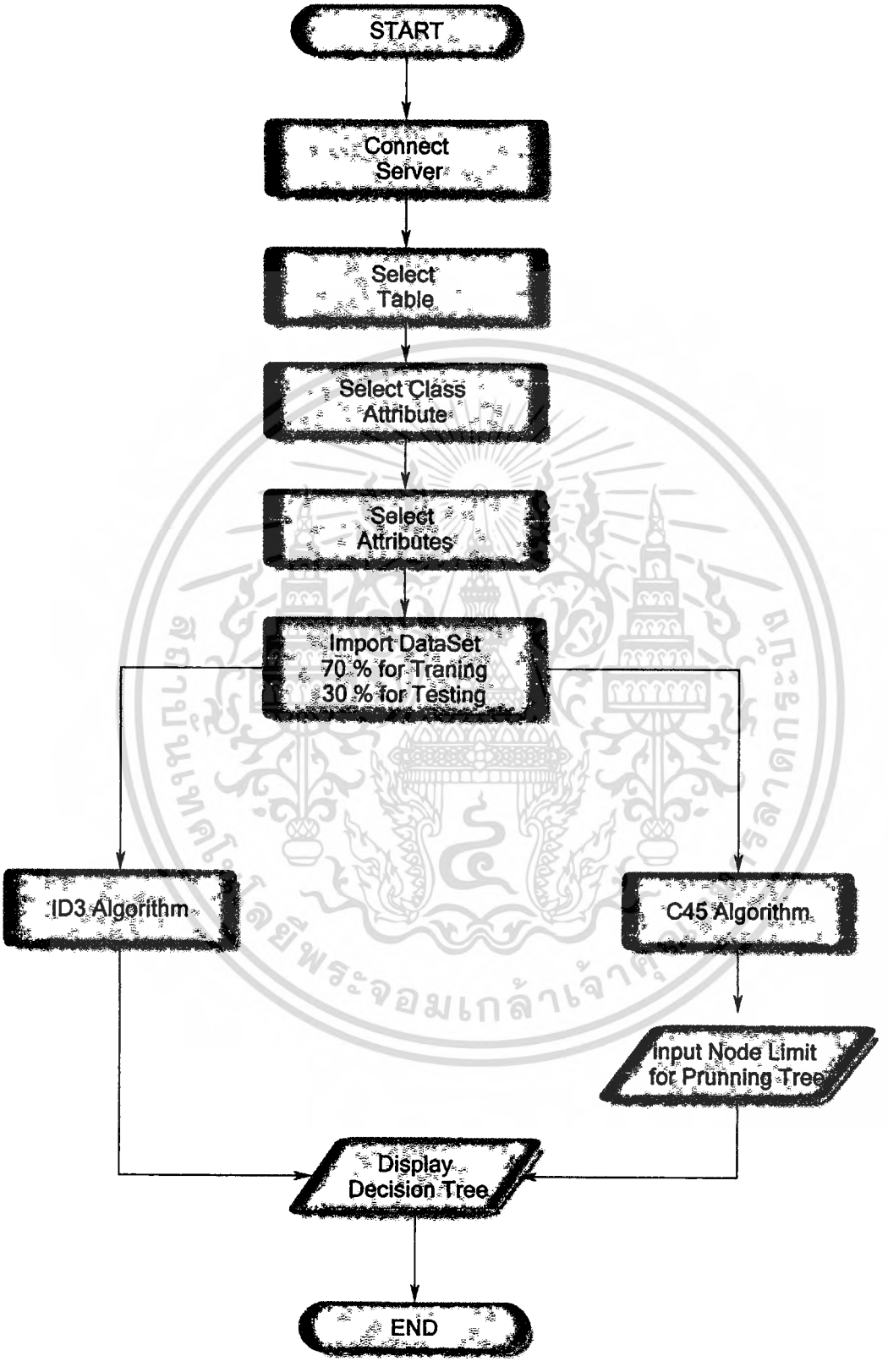
บทที่ 5
วิธีการดำเนินการศึกษา

ในการศึกษาโครงงานนี้ เป็นการนำทฤษฎีของการแบ่งกลุ่มของข้อมูล โดยใช้เทคนิคของโครงข่ายประสาทเทียม (Neural Networks) และเทคนิคการสร้างต้นไม้ (Decision Tree) และการใช้อัลกอริทึมในการทำงานแบบ Back-propagation และ C4.5 ตามลำดับเพื่อทำการประเมินประสิทธิภาพการทำงานอัลกอริทึมในแต่ละตัว โดยแบ่งขั้นตอนวิธีการในการดำเนินการศึกษาได้ดังนี้

1. โครงสร้างแบบจำลอง
2. ขั้นตอนการดำเนินงาน
3. การออกแบบโปรแกรมจำลอง
4. อัลกอริทึมในการทำงานของแบบจำลอง

5.1 องค์ประกอบในการพัฒนาโครงการ

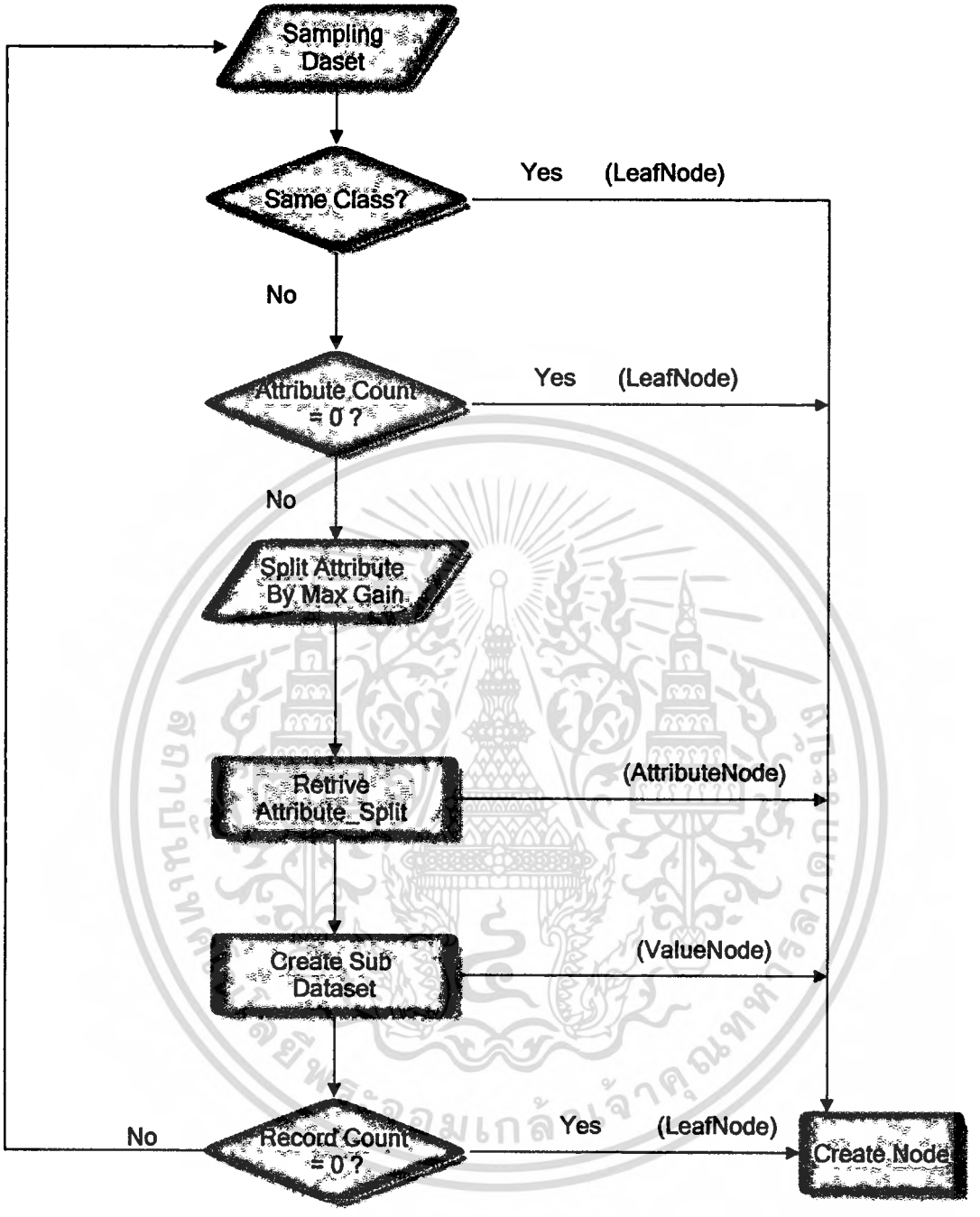
1. คอมพิวเตอร์ ACER TravelMate รุ่น 3213NWXCi
2. ระบบปฏิบัติการ Microsoft Windows XP Professional Service Pack 3
3. เลือกใช้ระบบฐานข้อมูล Microsoft SQL Server 2005
4. เครื่องมือในการพัฒนาระบบ Microsoft Visual Studio.Net 2005



รูปที่ 5.1 แสดงขั้นตอนการทำงานของโปรแกรม

ขั้นตอนการทำงาน

- เริ่มต้นทำการเชื่อมต่อกับ Sever ที่จะใช้งาน (Microsoft SQL Server 2005) เพื่อใช้งานฐานข้อมูลของระบบ
- ระบบจะทำการแสดงตารางทั้งหมดทำการเลือกตารางที่จะใช้งาน
- หลังจากเลือกตารางแล้ว จะแสดงแอททริบิวทั้งหมดของตารางที่เราได้ทำการเลือกเลือกแอททริบิวเป้าหมายที่จะทำการทำนาย
- เลือกแอททริบิวต่าง ๆ ที่จะใช้ทำนายโดยข้อมูลที่ใช้ทั้งหมดจะเป็นลักษณะ categorical
- เมื่อได้ข้อมูลที่จะใช้ทำนายครบแล้ว จะแสดงต้นไม้ตัดสินใจโดยใช้อัลกอริทึม ID3 และ C45 ในลักษระของทรีวิวเพื่อทำการแสดงผล

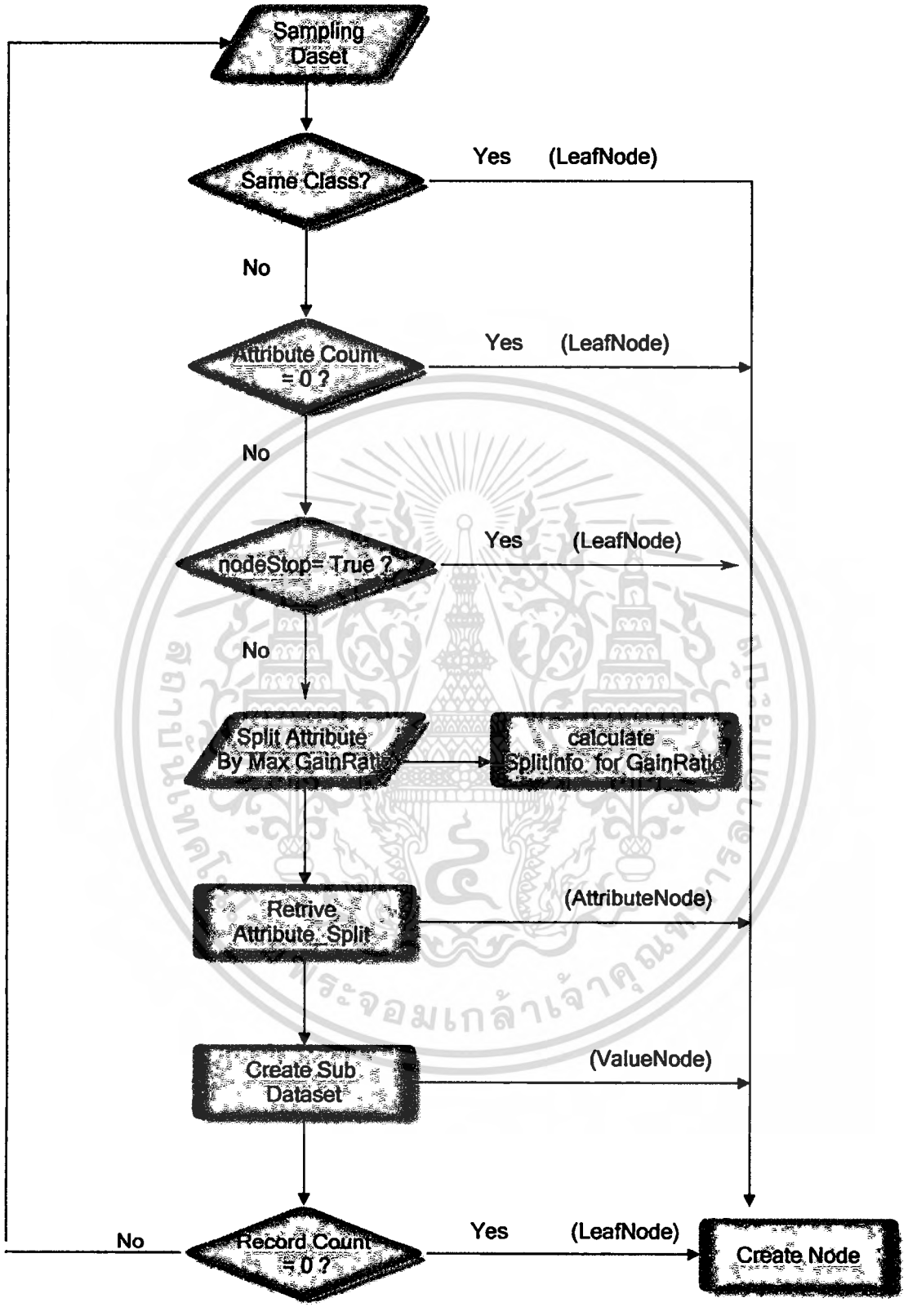


รูปที่ 5.2 แสดงหลักการของการแตกกิ่ง ID3

อธิบายหลักการแตกกิ่งโดยใช้อัลกอริทึม ID3

- ทำการนำข้อมูลที่จะใช้ในการทำไมนิ่งเข้าสู่ระบบ
- ตรวจสอบว่าข้อมูลที่นำเข้านั้นมีข้อมูลของแอทริบิวเป้าหมายอยู่ในกลุ่มเดียวกันหมดแล้วหรือไม่ ถ้าอยู่ในกลุ่มเดียวกันให้ทำการสร้างโหนด(LeafNode)

- ถ้าข้อมูลไม่ได้อยู่ในกลุ่มเดียวกัน ให้ทำการตรวจสอบว่ามีแอทริบิวที่จะใช้ในการสร้างโหนดหรือไม่ ถ้าไม่มีให้ทำการสร้างโหนด(LeafNode)
- ถ้ายังมีแอทริบิวอยู่ ให้ทำการหาแอทริบิวที่จะทำการแตกกิ่งโดยคำนวณจากค่า Gain โดยที่แอทริบิวที่มีค่า Gain มากที่สุดจะถูกนำไปสร้างโหนด(AtrributeNode)
- นำแอทริบิวที่ได้จากการสร้างโหนด(AttributeNode) มาทำการแตกกิ่งในแต่ละค่า ถ้าไม่เหลือจำนวนของเรคคอร์ดที่ได้จากการแบ่งข้อมูลแล้วให้ทำการสร้างโหนดใบ(LeafNode)
- ถ้ายังเหลือข้อมูลจากการแบ่งข้อมูล ให้ทำการวนซ้ำการทำงาน



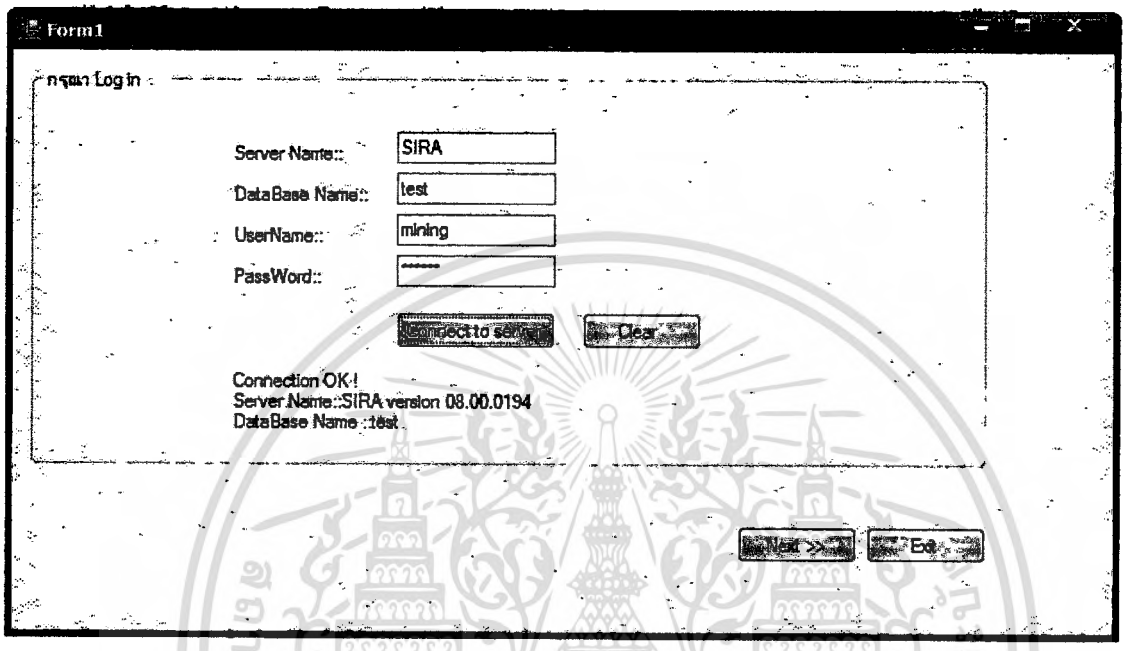
รูปที่ 5.3 แสดงหลักการของการแตกกิ่ง C45

อธิบายหลักการแตกกิ่งโดยใช้อัลกอริทึม C4.5

- ทำการนำข้อมูลที่จะใช้ในการทำไมนิ่งเข้าสู่ระบบ
- ตรวจสอบว่าข้อมูลที่นำเข้านั้นมีข้อมูลของแอทริบิวเป้าหมายอยู่ในกลุ่มเดี่ยวกันหมดแล้วหรือไม่ ถ้าอยู่ในกลุ่มเดียวกันให้ทำการสร้างโหนด(LeafNode)
- ถ้าข้อมูลไม่ได้อยู่ในกลุ่มเดียวกัน ให้ทำการตรวจสอบว่ามีแอทริบิวที่จะใช้ในการสร้างโหนดหรือไม่ ถ้าไม่มีให้ทำการสร้างโหนด(LeafNode)
- ถ้ายังมีแอทริบิวอยู่ ให้ทำการหาแอทริบิวที่จะทำการแตกกิ่งโดยคำนวณจากค่า GainRatio โดยที่แอทริบิวที่มีค่า GainRatio มากที่สุดจะถูกนำไปสร้างโหนด(AtrributeNode)
- นำแอทริบิวที่ได้จากการสร้างโหนด(AttributeNode) มาทำการแตกกิ่งในแต่ละค่า ถ้าไม่เหลือจำนวนของเรคคอร์ดที่ได้จากการแบ่งข้อมูลแล้วให้ทำการสร้างโหนด(LeafNode)
- ตรวจสอบเงื่อนไขการสร้างโหนดสำหรับทำการ Prunning tree ถ้าจำนวนโหนดตรงตามเงื่อนไขของข้อจำกัดในการสร้างโหนดให้ทำการสร้างโหนดใบ (LeafNode)
- ถ้าจำนวนโหนดของการสร้างยังไม่ตรงตามเงื่อนไขของการทำ Prunning tree แล้วให้ทำการตรวจสอบเงื่อนไขของการสร้างโหนดต่อไป
- ถ้ายังเหลือข้อมูลจากการแบ่งข้อมูล ให้ทำการวนซ้ำการทำงาน

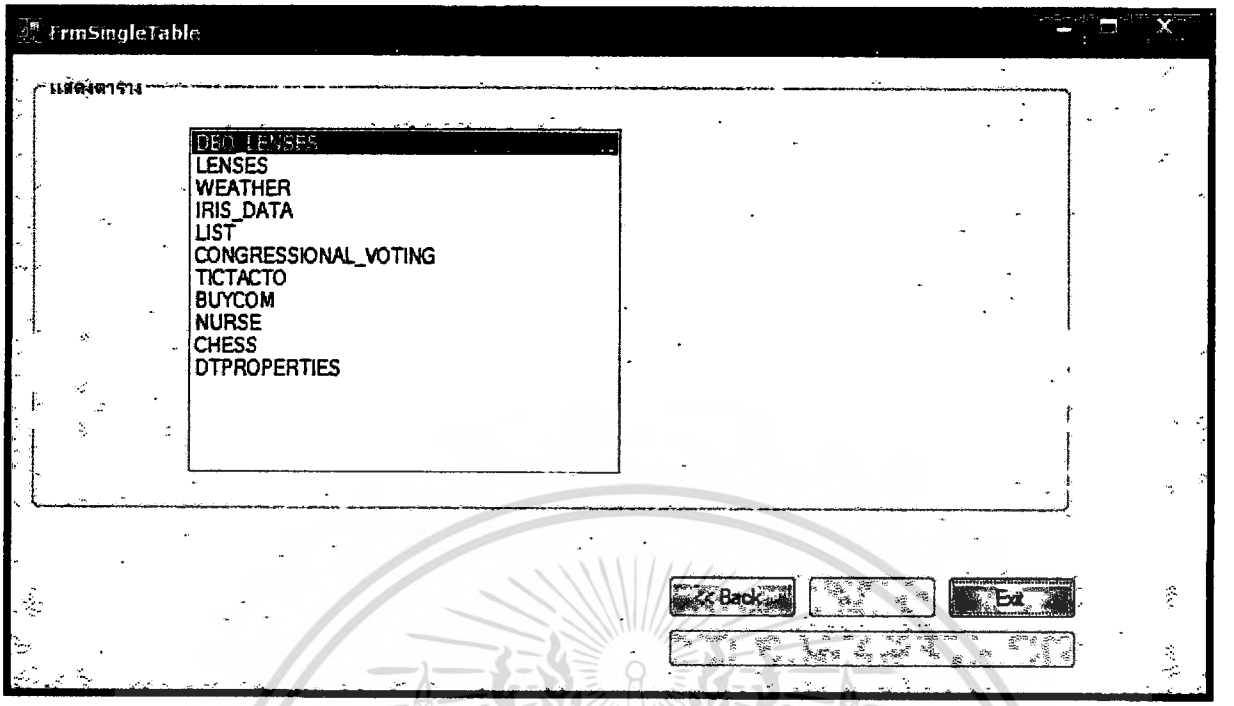
บทที่ 6

การประยุกต์ใช้โปรแกรม



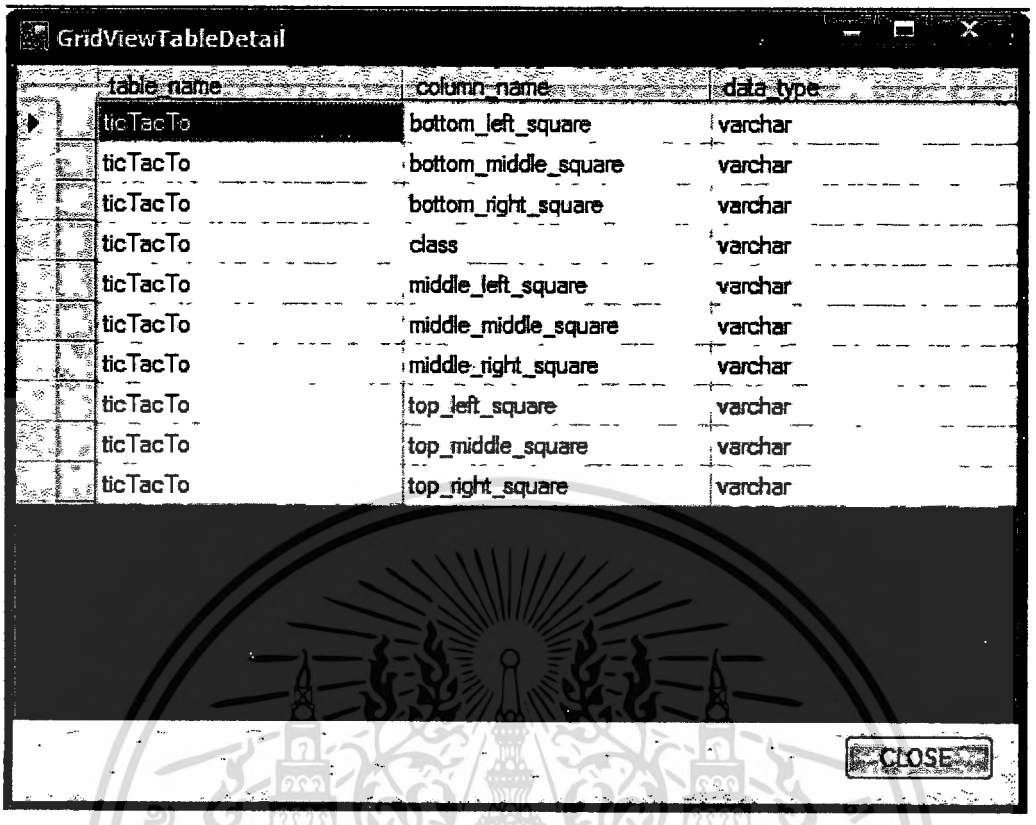
รูปที่ 6.1 หน้าจอแสดงหน้าจอการล็อกอินเข้าสู่ระบบการทำไมนิ่ง

- Server Name : ระบุ Server ที่เราจะเข้างานที่ทำการเก็บฐานข้อมูลไว้
- DataBase Name : ทำการระบุฐานข้อมูลที่ต้องการเรียกใช้
- UserName : ระบุ UserName ของผู้ใช้ที่จะทำการเข้าสู่ระบบ
- PassWord : ระบุ PassWord ที่ผู้ใช้มีเพื่อเข้าใช้งานฐานข้อมูล
- หลังจากนั้นทำการกดปุ่ม Connect to Server หลังจากที่เราได้ทำการระบุข้อมูลต่าง ๆ ที่จะใช้เข้าสู่ระบบ เมื่อข้อมูลถูกต้องจะแสดงข้อมูลต่าง ๆ ของ Server ที่เราจะทำใช้งาน
- ปุ่ม Clear ทำการล้างข้อมูลต่างๆ ที่ระบุ
- ปุ่ม Next เพื่อเรียกใช้งานหน้าต่อไปของโปรแกรม
- ปุ่ม Exit เพื่อออกจากการใช้งานโปรแกรม



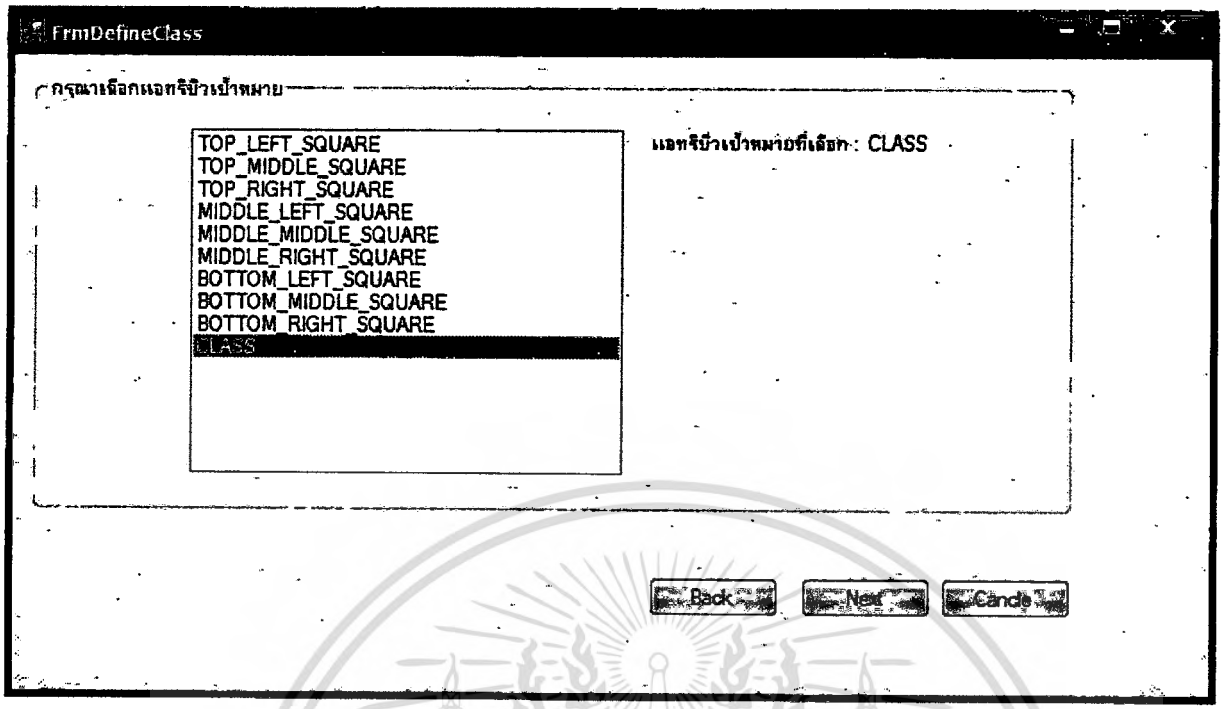
รูปที่ 6.2 หน้าจอแสดงตารางทั้งหมดของฐานข้อมูล

- แสดงชื่อตารางทั้งหมดที่มีฐานข้อมูลที่เราเรียกใช้งาน
- หลังจากทำการเลือกตารางใช้งานได้แล้ว ปุ่ม Next จะแสดงขึ้นมา
- หลังจากทำการเลือกตารางใช้งานได้แล้ว ปุ่ม ViewTableDetail จะแสดงขึ้นมา
- ปุ่ม Next เพื่อเรียกใช้งานหน้าต่อไปของโปรแกรม
- ปุ่ม Back เพื่อย้อนกลับไปยังหน้าก่อนหน้า
- ปุ่ม Exit เพื่อทำการออกจากโปรแกรม
- ปุ่ม ViewTableDetail เพื่อแสดงรายละเอียดของตารางที่เลือกประกอบไปด้วย แอทริบิว และชนิดของข้อมูลของแต่ละแอทริบิว



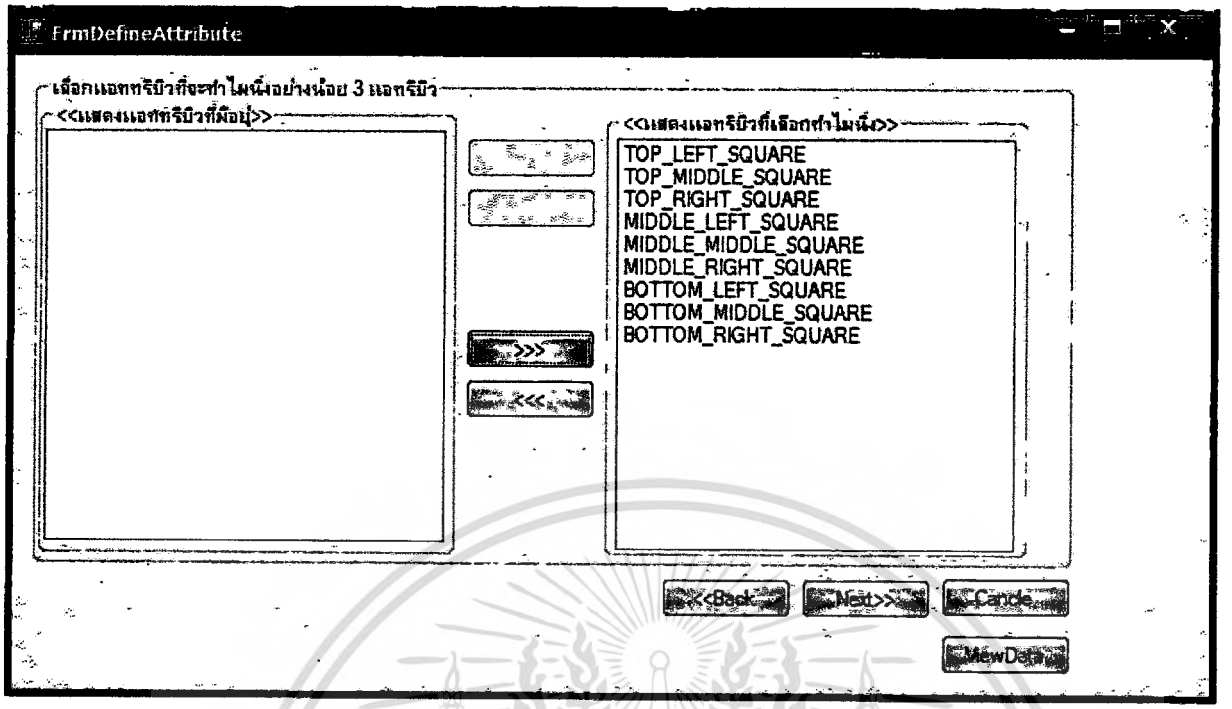
รูปที่ 6.3 หน้าจอแสดงรายละเอียดของตารางที่เลือก

- แสดงรายละเอียดของตารางที่เลือกจากการกดปุ่ม ViewTableDetail โดยจะแสดงรายละเอียดดังนี้
  - คอลัมน์Table_Name แสดงชื่อของตาราง
  - คอลัมน์ Column_Name แสดงรายละเอียดของแอทริบิวทั้งหมดของตาราง
  - คอลัมน์ Data_Type แสดชนิดของข้อมูลในแต่ละแอทริบิว
- ปุ่ม CLOSE เพื่อปิดหน้าต่างรายละเอียดของตาราง



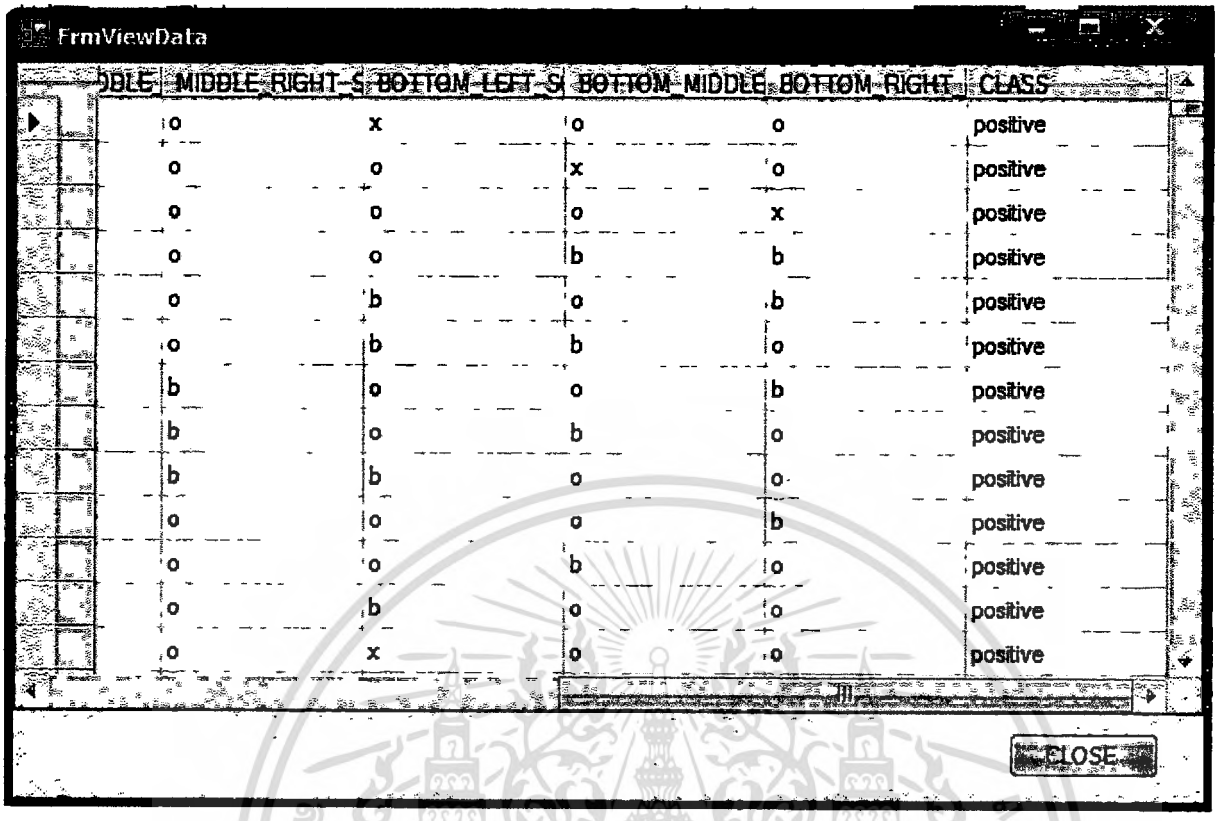
รูปที่ 6.4 หน้าจอแสดงข้อมูลของ class เป้าหมายที่ทำการเลือก

- แสดงรายการของแอทริบิวต่าง ๆ จากฐานข้อมูลที่เราเลือกใช้งาน
- ทำการคลิ๊กเลือกแอทริบิวเป้าหมายที่จะทำการทำนาย
- ปุ่ม Back เพื่อทำการย้อนกลับไปหน้าต่างก่อนของโปรแกรม
- ปุ่ม Next เพื่อไปยังหน้าต่อไปของโปรแกรม
- ปุ่ม Cancle เพื่อออกจากโปรแกรม



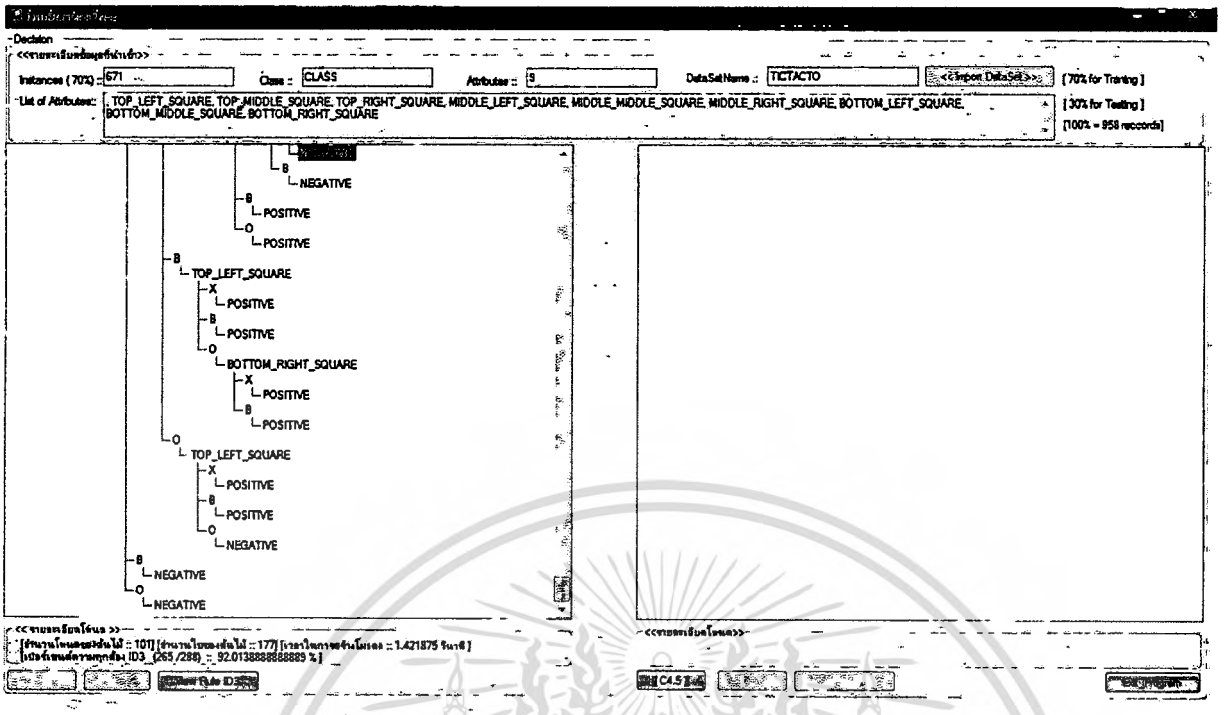
รูปที่ 6.5 หน้าจอแสดงแอททริบิวที่ทำการเลือกเพื่อทำนาย

- รายการทางซ้ายมือแสดงแอททริบิวต่าง ๆ จากตารางที่เราทำการเลือก
- รายการทางขวามือ คือแอททริบิวที่เราทำการเลือกเพื่อจะทำนาย
- ปุ่ม Back เพื่อทำการย้อนกลับไปหน้าต่างก่อนของโปรแกรม
- ปุ่ม Next เพื่อไปยังหน้าต่อไปของโปรแกรม
- ปุ่ม ViewData เพื่อแสดงรายละเอียดข้อมูลของแอททริบิวที่ได้ทำการเลือกโดยจะต้องเลือกแอททริบิวที่ทำการไมนิ่งต้องไม่น้อยกว่า 3 แอททริบิว
- ปุ่ม Cancle เพื่อออกจากโปรแกรม



รูปที่ 6.6 หน้าจอแสดงรายละเอียดของแอทริบิวที่เลือกและรายละเอียดของแอทริบิวเป้าหมาย

- แสดงรายละเอียดของข้อมูลจากแอทริบิวที่ได้ทำการเลือกและรายละเอียดของแอทริบิวเป้าหมายในรูปแบบของ Grid View
- ปุ่ม CLOSE เพื่อทำการปิดหน้าต่างการแสดงรายละเอียด



รูปที่ 6.7 หน้าจอแสดงผลลัพธ์ของการแตกกิ่งด้วยอัลกอริทึม ID3

- แสดงผลลัพธ์ของการทำนายจากการใช้อัลกอริทึม ID3
  - Instances แสดงจำนวนเรคคอร์ดทั้งหมด
  - Class แสดงแอทริบิวเป้าหมาย
  - Attributes แสดงจำนวนของแอทริบิวที่จะทำการไมนิ่ง
  - DataSetName แสดงชื่อตารางที่เลือก
  - List of Atrributes แสดงแอทริบิวทั้งหมดจากตารางที่ได้ทำการเลือก
- ปุ่ม Exit Program เพื่อออกจากโปรแกรม
- ปุ่ม Import DataSet เพื่อทำการนำข้อมูลต่าง ๆ จากที่ได้ทำการเลือกเอาไว้แล้ว

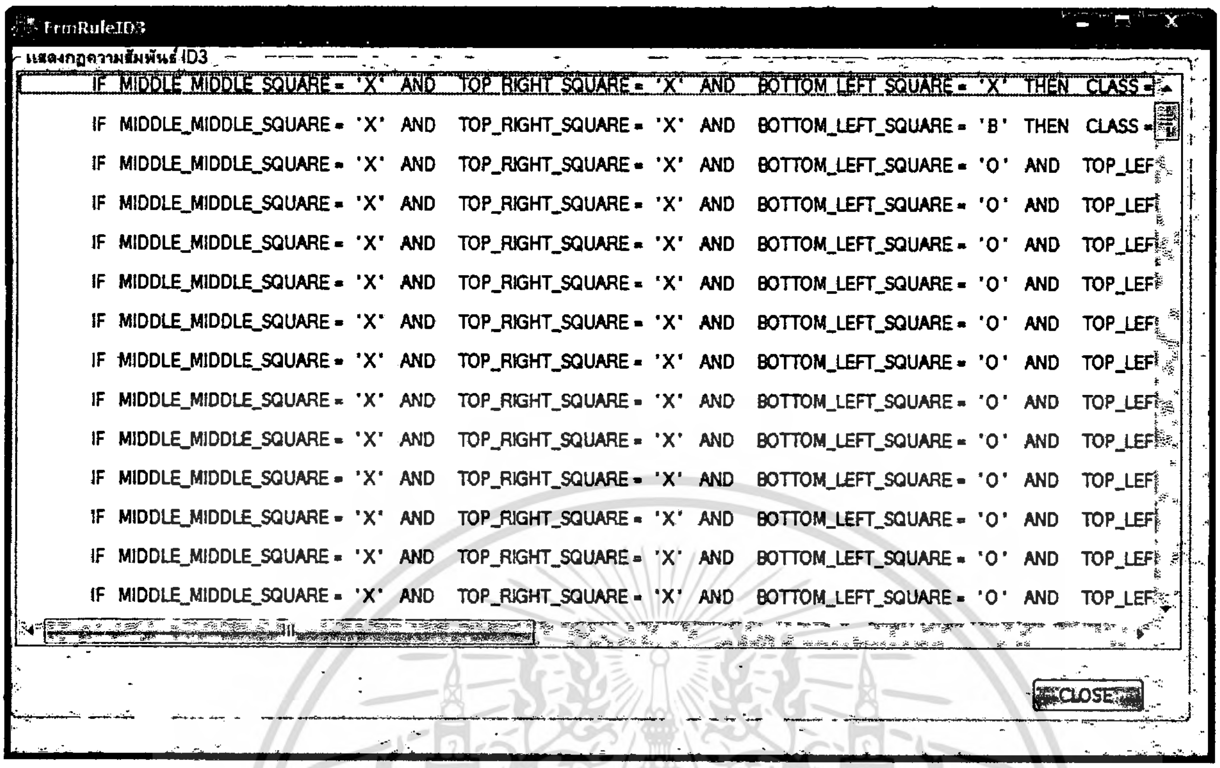
รายละเอียดของ ID3

- ปุ่ม ID3 เพื่อแสดงผลลัพธ์ของการใช้อัลกอริทึม ID3 ในรูปแบบของ TreeView
- ปุ่ม Accurancy เพื่อแสดงผลลัพธ์ของความถูกต้องจากการใช้อัลกอริทึม ID3
- ปุ่ม View Rule ID3 เพื่อแสดงกฎทั้งหมดที่ได้จากการสร้างต้นไม้จากอัลกอริทึม ID3 ในรูปแบบของ IF THEN

รายละเอียดของ C45

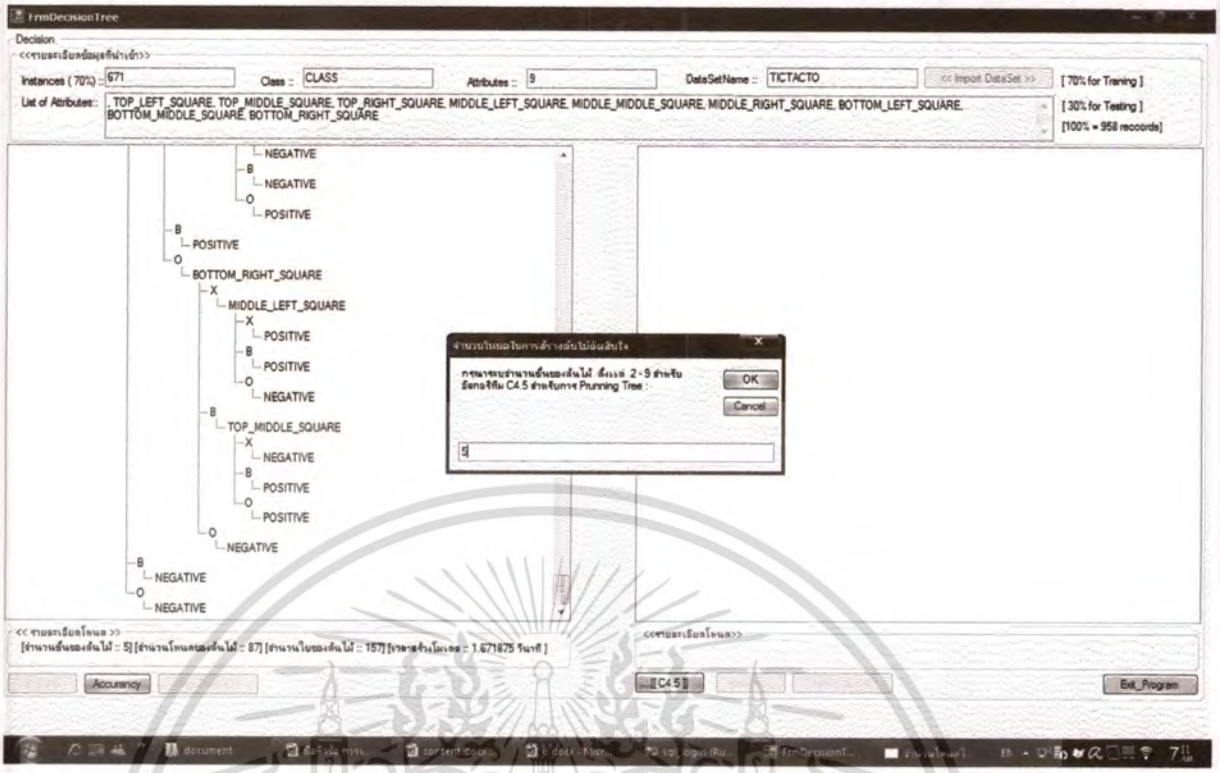
- ปุ่ม C45 เพื่อแสดงผลลัพธ์ของการใช้อัลกอริทึม C45 ในรูปแบบของ TreeView
- ปุ่ม Accurancy เพื่อแสดงผลลัพธ์ของความถูกต้องจากการใช้อัลกอริทึม C45

- ปุ่ม View Rule C45 เพื่อแสดงกฎทั้งหมดที่ได้จากการสร้างต้นไม้จากอัลกอริทึม C45 ในรูปแบบของ IF THEN



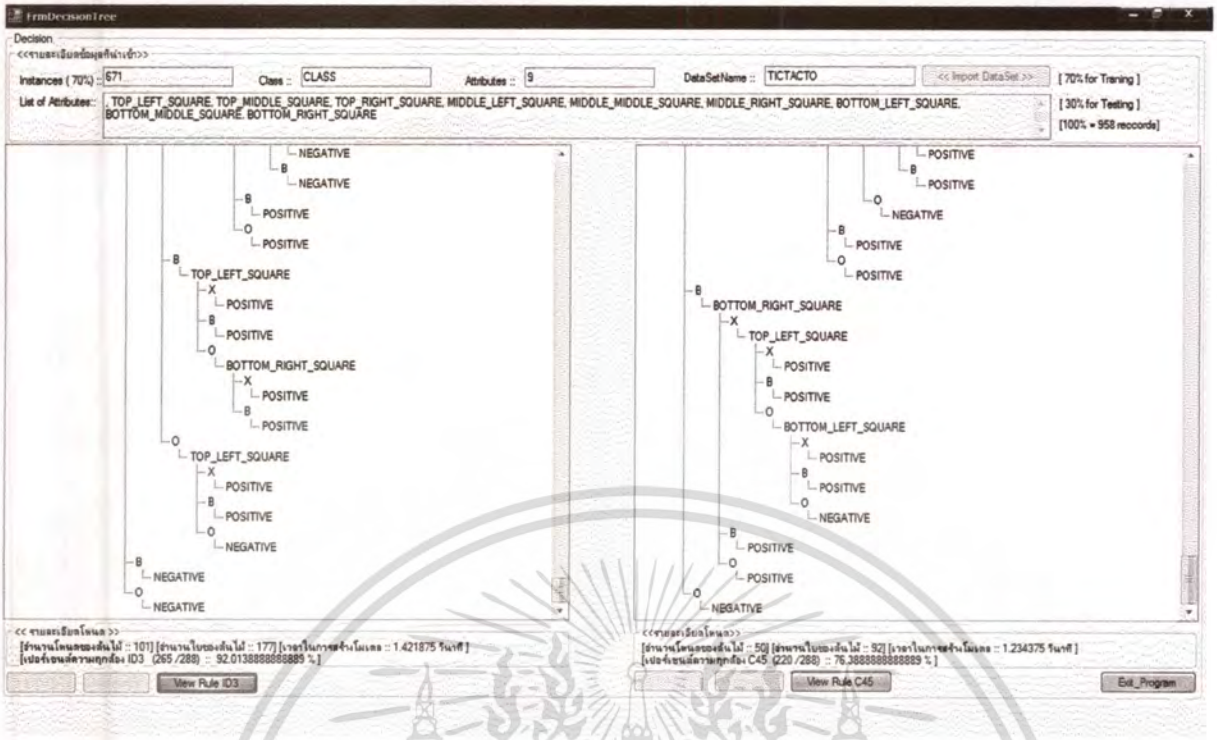
รูปที่ 6.8 หน้าจอแสดงผลลัพธ์จากการแตกกิ่งด้วยอัลกอริทึม ID3 ในรูปแบบของ IF THEN

- แสดงผลลัพธ์ของการทำนายจากการใช้อัลกอริทึม ID3 ในรูปแบบของ IF THEN จากการแตกกิ่งของต้นไม้ทั้งหมด
- ปุ่ม CLOSE เพื่อปิดหน้าต่างการแสดงผล



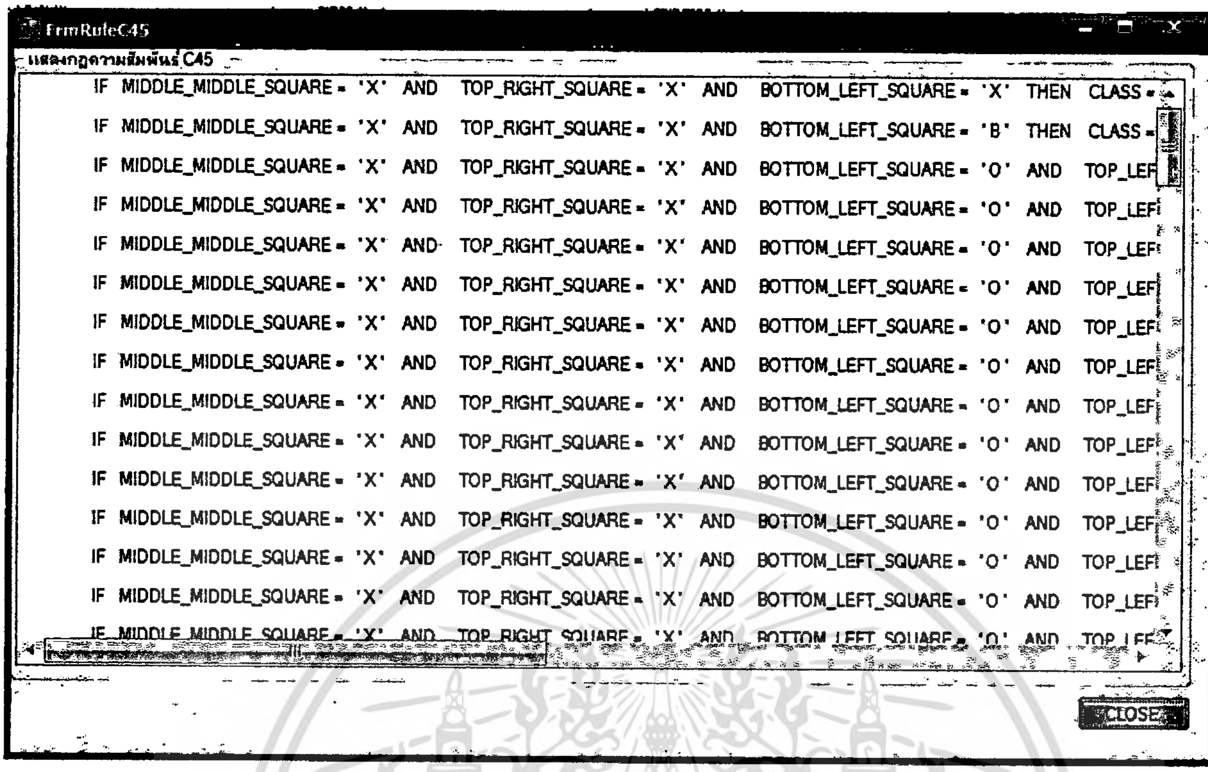
รูปที่ 6.9 หน้าจอแสดงการระบุจำนวนโหนดในการสร้างของอัลกอริทึม C4.5

- ให้ทำการระบุจำนวนขั้นของการสร้างต้นไม้ตัดสินใจ สำหรับการทำงานของอัลกอริทึม C4.5
- ปุ่ม OK เพื่อทำการยืนยันจำนวนของการสร้างชั้นของต้นไม้ที่ระบุให้กับอัลกอริทึม C4.5



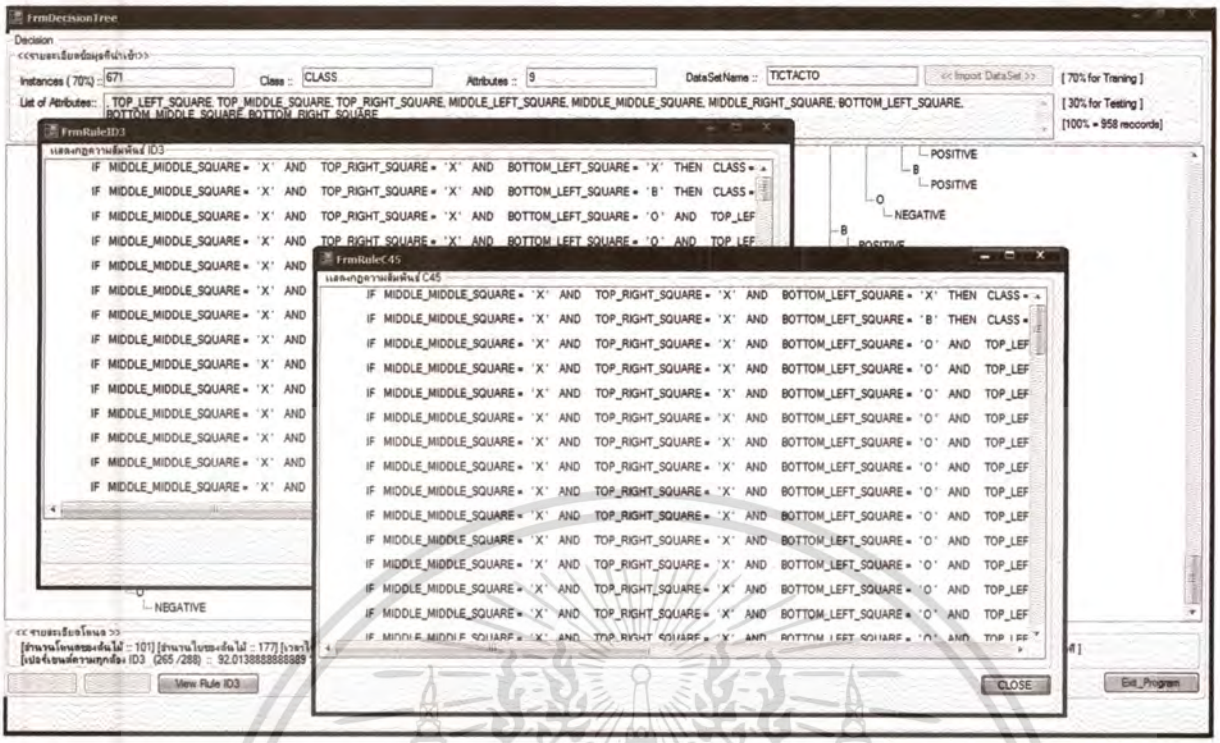
รูปที่ 6.10 หน้าจอแสดงผลลัพธ์ของกฎจากการแตกกิ่งด้วยอัลกอริทึม ID3 และ C4.5

- ปุ่ม EXIT_PROGRAM เพื่อออกจากโปรแกรมการทำงาน
- ฝั่งซ้ายแสดงการแตกกิ่งด้วยอัลกอริทึม ID3
- ฝั่งขวาแสดงการแตกกิ่งด้วยอัลกอริทึม C4.5 ที่มีการระบุจำนวนโหนดของการสร้าง



รูปที่ 6.11 หน้าจอแสดงผลลัพธ์กฎการแตกกิ่งด้วยอัลกอริทึม C45 ในรูปแบบของ IF THEN

- **แสดงผลลัพธ์ของการทำนายจากการใช้อัลกอริทึม C45 ในรูปแบบของ IF THEN จากการแตกกิ่งของต้นไม้ทั้งหมด**
- **ปุ่ม CLOSE เพื่อปิดหน้าต่างการแสดงผล**



รูปที่ 6.12 หน้าจอแสดงกฎจากการแตกกิ่งโดยใช้อัลกอริทึมของ ID3 และ C45 ในรูปแบบของ IF THEN

- หน้าจอแสดงรายละเอียดของการสร้างกฎที่ได้จากการสร้างต้นไม้ตัดสินใจจากอัลกอริทึม ID3 และ C4.5 ในรูปของ IF THEN RULE โดยจะแสดงกฎจากกิ่งทั้งหมดที่ได้มีการแตกกิ่งออกมาจนถึงโหนดใบ
- ปุ่ม CLOSE เพื่อทำการปิดการแสดงของกฎที่ได้
- ปุ่ม EXIT_PROGRAM เพื่อออกจากการทำงานของโปรแกรม

บทที่ 7
สรุปผลการศึกษาและข้อเสนอแนะ

โครงการการพัฒนาระบบของการทำไมนิ่งโดยใช้อัลกอริทึม ID3 และ C45 นี้จัดทำขึ้นเพื่อใช้เป็นเครื่องมือในการจำแนกกลุ่มของข้อมูล และช่วยเพิ่มประสิทธิภาพให้กับข้อมูลที่มีอยู่ โดยการนำข้อมูลที่มีอยู่ในอดีตมาทำการวิเคราะห์หาความสัมพันธ์และจำแนกเป็นกลุ่มออกมา โดยการอ่านค่าจากต้นไม้ที่ได้ทำการแตกกิ่งออกมา

7.1 สรุปผลการดำเนินงาน

การทำงานของโปรแกรมจะใช้ข้อมูลจากฐานข้อมูล Microsoft SQL Server 2005 และต้องเป็นข้อมูลที่จัดเก็บแบบที่มีลักษณะตัวแปรที่กำหนดความเป็นไปได้ของข้อมูลอย่างชัดเจน เช่น Yes หรือ No หรือตัวแปรที่มีการจัดลำดับของข้อมูล เช่น hot, mild, cool เป็นต้น ซึ่งจะนำมาใช้กับเทคนิคของการจำแนกกลุ่ม และตารางที่ใช้งานจะทำการเลือกได้เพียงตารางเดียว

จากการพัฒนาระบบของการจำแนกกลุ่มของข้อมูลโดยใช้อัลกอริทึม ID3 และ C45 เพื่อทำการวิเคราะห์แล้วจะพบว่า การแตกกิ่งของข้อมูลที่มีจะมีความแตกต่างกัน ในกรณีที่มีข้อมูลที่ใช้มีความแตกต่างกันมาก ๆ ในกรณีของอัลกอริทึม ID3 จะพบว่าจะมีการแตกกิ่งออกไปมากมายซึ่งเป็นการแตกกิ่งที่ลึกเกินไป จนอาจจะไม่เกิดประโยชน์ และจากการใช้อัลกอริทึม C45 เพื่อทำการเปรียบเทียบโดยใช้ข้อมูลตัวเดียวกันจะพบว่า การแตกกิ่งจะมีประสิทธิภาพมากกว่าเนื่องจากใช้จำนวนโหนดของการแตกกิ่งจะใช้จำนวนน้อยกว่าและได้ความถูกต้องมากกว่าจากการใช้โมเดลนั้นไปทำการทดสอบกับชุดข้อมูลทดสอบ ในกรณีที่ข้อมูลนั้น ๆ มีความแตกต่างกันของข้อมูลมาก ๆ ซึ่งจะทำให้การทำงานของอัลกอริทึม ID3 นั้นแตกกิ่งออกไปในทุก ๆ ความเป็นไปได้ของข้อมูล

ข้อมูลการทดสอบชุดที่ 1

DataSet Name : Car

Attributes : 6

จำนวน Instances : 1728 Records

Training Set (70%) : 1210 Records

Testing Set (30%) : 519 Records

ตารางที่ 7.1 แสดงตารางเปรียบเทียบจำนวนชั้นของดาต้าเซท Car

จำนวนชั้น	จำนวนโหนด	จำนวนใบ	เวลา	ความถูกต้อง	อัลกอริทึม
6	271	553	2.6406	87.4759 %	ID3
6	271	553	3.4063	87.4759 %	C4.5

ข้อมูลการทดสอบชุดที่ 2

DataSet Name : Congress_Voting

Attributes : 16

จำนวน Instances : 232 Records

Training Set (70%) : 163 Records

Testing Set (30%) : 70 Records

ตารางที่ 7.2 แสดงตารางเปรียบเทียบจำนวนชั้นของดาต้าเซท Congress_Voting

จำนวนชั้น	จำนวนโหนด	จำนวนใบ	เวลา	ความถูกต้อง	อัลกอริทึม
5	10	11	0.3906	97.1428 %	ID3
5	11	12	0.500	97.1428 %	C4.5

ข้อมูลการทดสอบชุดที่ 3

Dataset Names: Chesses

Attributes : 16

จำนวน Instances : 3196 Records

Training Set (70%) : 2238 Records

Testing Set (30%) : 959 Records

ตารางที่ 7.3 แสดงตารางเปรียบเทียบจำนวนชั้นของคาด้าเซท Chesses

จำนวนชั้น	จำนวนโหนด	จำนวนใบ	เวลา	ความถูกต้อง	อัลกอริทึม
11	41	43	4.1406	99.7914 %	ID3
11	40	43	8.7180	99.7914 %	C4.5

ข้อมูลการทดสอบชุดที่ 4

DataSet Name : TicTacTo

Attributes : 9

จำนวน Instances : 958 Records

Training Set (70%) : 671 Records

Testing Set (30%) : 288 Records

ตารางที่ 7.4 แสดงตารางเปรียบเทียบจำนวนชั้นของคาด้าเซท Tictacto

จำนวนชั้น	จำนวนโหนด	จำนวนใบ	เวลา	ความถูกต้อง	อัลกอริทึม
5	108	188	1.3906	86.1111 %	ID3
5	118	210	2.1718	91.6666 %	C4.5

ข้อมูลการทดสอบชุดที่ 5

DataSet Name : Weather

Attributes : 4

จำนวน Instances : 42 Records

Training Set (70%) : 30 Records

Testing Set (30%) : 13 Records

ตารางที่ 7.5 แสดงตารางเปรียบเทียบจำนวนชั้นของดาต้าเซท Weather

จำนวนชั้น	จำนวนโหนด	จำนวนใบ	เวลา	ความถูกต้อง	อัลกอริทึม
3	5	8	0.1718	100 %	ID3
3	5	8	0.1974	100 %	C4.5

- จากผลการทดลองทั้ง 5 ชุดโดยทำการเปรียบเทียบจำนวนชั้นของต้นไม้ของอัลกอริทึม ID3 และ C4.5 โดยการกำหนดจำนวนของชั้นการแตกกิ่งของอัลกอริทึม C4.5 ให้มีค่าเท่ากับจำนวนชั้นของอัลกอริทึม ID3 ที่ได้จากการทดลองข้อมูลตัวอย่างทั้งหมด 70 % ในการสร้างโมเดล และข้อมูล 30% ในการทดสอบ จากผลการทดลองทำให้ทราบว่าจำนวนโหนดที่เท่ากันนั้น อัลกอริทึม ID3 จะมีจำนวนโหนดของต้นไม้น้อยกว่าจำนวนโหนดของอัลกอริทึม C4.5 และผลความถูกต้องของข้อมูลจะได้ว่าจากข้อมูลทั้ง 5 ชุดผลการทดลองจะได้ว่ามีชุดผลการทดลองของอัลกอริทึม ID3 มีค่าความถูกต้องน้อยกว่าของอัลกอริทึม C4.5อยู่หนึ่งชุดคือชุดข้อมูลของ TICTACTO และอีก 4 ชุดข้อมูลที่เหลือจะได้ค่าความถูกต้องของข้อมูลเท่ากัน ทำให้ทราบว่าการจำกัดโหนดของการแตกกิ่งไม่มีผลต่อความถูกต้องของข้อมูลในกรณีของการทดสอบด้วยชุดข้อมูลตัวอย่างทั้ง 5 ชุดนี้

7.2 ข้อเสนอแนะ

ตัวโปรแกรมที่ทำการพัฒนายังไม่สามารถทำงานได้กับฐานข้อมูลที่เป็นตัวเลขได้ และทำงานได้กับฐานข้อมูลเพียงตารางเดียวเท่านั้น ดังนั้นในการพัฒนาระบบเพิ่มเติมควรจะเพิ่มส่วนของการทำงานที่สามารถหาความสัมพันธ์ได้ในหลายตาราง และสามารถใช้งานกับฐานข้อมูลที่เป็นตัวเลขได้

บรรณานุกรม

ศุภชัย สมพานิช . 2550 **Advance.Net Programming** กรุงเทพ : devBook

Building Classification Models : ID3 and C4.5. [Online]. Available

http://www.cis.temple.edu/~ingargio/cis587/readings/id3-c45.html

Jiawei Han and Micheline Kamber . 2001. **Data Mining Concept and Techniques.**

USA : Academic Press

ประวัติผู้เขียน

ชื่อ : นางสาวศิรประภา นนทะจันทร์

วันเดือนปีเกิด : 24 กรกฎาคม 2524

สถานที่เกิด จังหวัด เพชรบูรณ์

ประวัติการศึกษา :

มัธยมต้น : โรงเรียนขอนแก่นวิทยายน

มัธยมปลาย : โรงเรียนขอนแก่นวิทยายน

ปริญญาตรี : มหาวิทยาลัยขอนแก่น