การพัฒนาระบบการวิเคราะห์หาสาเหตุการเสียของกระบวนการผลิต โดยใช้ Classification Tree

System Development of Diagnostic system of Hard Disk manufacturing by using Classification Tree

โดย

นางสาววิไล แม้นถาวรสิริ

รหัส 43067034

อาจารย์ที่ปรึกษา

ผศ.ดร.วรพจน์ กรีสุระเดช

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน

หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ

ภาคเรียนที่ 2 ปีการศึกษา 2544

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ชื่อหัวข้อ	การพัฒนาระบบการวิเคราะห์หาสาเหตุการเสียของกระบวนการผลิต โดยใช้ Classification Tree
นักศึกษา	นางสาววิไล แม้นถาวรสิริ
อาจารย์ที่ปรึกษา	ผศ.ดร.วรพจน์ กรีสุระเดช
ระดับการศึกษา	วิทยาศาสตรมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2544

บทคัดย่อ

Data Mining ได้ถูกนำมาใช้งานอย่างกว้างขวางในปัจจุบัน เนื่องจากเทคนิคและอัลกอริธึมต่างๆ ที่ได้ปรับปรุง เพื่อนำมาใช้ค้นหาข้อมูลที่เป็นประโยชน์ ทำให้สามารถนำมาใช้กับข้อมูลต่างๆ ใน database หรือจากแหล่งอื่นๆ ได้อย่างมีประสิทธิภาพ

โครงการนี้ได้ทำการศึกษา และพัฒนาระบบการวิเคราะห์ สาเหตุการเสียของกระบวนการผลิตที่เกิดขึ้น ในกระบวนการผลิตฮาร์ดดิสก์ โดยนำเอาข้อมูลที่ได้จากกระบวนการผลิต มาทำการ Mining โดยใช้ Classification Tree เพื่อหารูปแบบความสัมพันธ์ที่เป็นประโยชน์ เพื่อที่จะนำข้อมูลที่ได้ไปใช้ในการแก้ไขปัญหาที่เกิดขึ้นได้อย่างรวดเร็วและมีประสิทธิภาพ ซึ่งจะเป็นการช่วยปรับปรุงกระบวนการผลิตให้มีประสิทธิภาพยิ่งขึ้น

Title	System development for diagnostic system Hard Disk manufacturing by using Classification Tree
Student	Ms. Vilai Manthavornsiri
Advisor	Dr. Worapoj Kreesuradej
Level of Study	Master of Science in Information Technology
Major	Information Science
Academic Year	2001

ABSTRACT

Data Mining's increased popularity in present is due partly to technological improvements that permit faster, more effective analyses the data and can link to data from database or other source.

In this project have studied Data mining in Classification tree techniques for developing the diagnostic system for Hard Disk manufacturing. This system will use data from Hard disk process to mining by using Classification tree and discover the useful information. This information will help us to find the root cause of defective and problem in process so we can improve the hard disk process to be more efficiency.

กิตติกรรมประกาศ

ในการทำโครงการพัฒนาระบบการวิเคราะห์หาสาเหตุการเสียของกระบวนการผลิต โดยใช้ Classification Tree ผู้เขียนขอขอบพระคุณท่าน ผศ.ดร.วรพจน์ กรีสุระเดช อาจารย์ที่ปรึกษา ที่ได้กรุณาให้คำปรึกษาและแนะนำแนวทางในการแก้ไขปัญหาต่าง ทำให้โครงงานนี้ สำเร็จลุล่วงมาได้ ขอขอบคุณเพื่อนๆ และพี่ๆ ทุกคน ที่ให้ความคำแนะนำ และคอยช่วยเหลือในการทำโครงงานนี้

วิไล แม้นถาวรสิริ

14 กุมภาพันธ์ 2545

สารบัญ

บทที่1
บทนำ

1.1 บทนำ

วงการอุตสาหกรรมต่างๆ ในปัจจุบัน มีการแข่งขันกันสูงมากเพื่อสามารถผลิตสินค้าให้มีคุณภาพ และได้ทันต่อความต้องการของลูกค้า เพื่อให้เป็นผู้นำในตลาด ทำให้มีการนำเสนอกลยุทธ์ต่างๆมาใช้ช่วยในการพัฒนาและปรับปรุงประสิทธิภาพของกระบวนการผลิต

Data mining เองก็เป็นอีกเทคนิคหนึ่งซึ่งไม่เพียงแต่สามารถนำมาใช้ช่วยวิเคราะห์ข้อมูลในเชิงธุรกิจของ ธนาคาร สถาบันการเงิน ความสัมพันธ์ของลูกค้าและบริการเท่านั้น แต่ก็ยังสามารถนำมาประยุกต์ใช้กับอุตสาหกรรมการผลิต โดยสามารถนำมาใช้ช่วยในการค้นหาข้อมูลจากกระบวนการผลิต เพื่อใช้ช่วยในการตัดสินใจเพื่อกระทำการควบคุมกระบวนการผลิต หรือปรับปรุงเพื่อเพิ่มประสิทธิภาพของกระบวนการผลิต รวมถึงเพิ่มคุณภาพของผลิตภัณฑ์เอง โดยการใช้ data mining ยังใช้เวลาอันสั้นที่จะค้นหาข้อมูลจะช่วยในการตัดสินใจที่ก่อให้เกิดประโยชน์ต่อกระบวนการผลิตในอุตสาหกรรม ทำให้ผู้ผลิตสามารถที่จะแข่งขันในธุรกิจปัจจุบันที่เวลาก็เป็นอีกหนึ่งปัจจัยที่สำคัญ

1.2 วัตถุประสงค์ในการพัฒนาระบบงาน

1. ศึกษา Data Mining โดยการใช้ Classification Tree เพื่อนำมาใช้กับข้อมูลการผลิต
2. นำความรู้ทาง Data Mining และ อัลกอลิธึมที่ศึกษา มาพัฒนาระบบการวิเคราะห์หาสาเหตุของเสียของกระบวนการผลิตในอุฒสาหกรรม ฮาร์ดิสก์ โดยใช้ Classification Tree
3. สามารถนำผล หรือข้อมูลที่ได้จากการทำ Data Mining มาใช้ช่วยในการพัฒนาและปรับปรุงกระบวนการผลิตให้ดียิ่งขึ้น

1.3 ขอบเขตของระบบงาน

ในโครงการพัฒนาระบบงานนี้ จะทำการพัฒนาระบบที่สามารถนำข้อมูลจากกระบวนการผลิตมาใช้ในการทำ Data Mining โดยจะทำการพัฒนาระบบเป็นรูปแบบApplication online บน Web สามารถทำการเชื่อมต่อกับ Database เพื่อใช้ข้อมูลจาก Database หรือสามารถนำข้อมูลที่มีอยู่ในรูปแบบของไฟล์มาทำการ Mining

1.4 ขั้นตอนการดำเนินงาน

1. กำหนดวัตถุประสงค์และเป้าหมายการดำเนินการ โดยมีเป้าหมายจะทำการวิเคราะห์สาเหตุของเสียของกระบวนการผลิต
2. ทำการหาแหล่งของข้อมูล โดยจะใช้ข้อมูลที่ได้จากกระบวนการผลิตฮาร์ดดิสก์
3. ทำการศึกษาขั้นตอนการทำ Data Mining โดยใช้ Classification Tree ทำการศึกษาอัลกอลิธึมที่เหมาะสมที่จะนำมาใช้
4. ทำการเตรียมข้อมูลที่จะใช้ในรูปแบบที่เหมาะสมกับอัลกอลิธึม
5. ทำการออกแบบระบบ กำหนดขอบเขตการทำงานของระบบที่จะทำการพัฒนา
6. พัฒนาระบบ โดยนำอัลกอลิธึมของ Classification Tree ที่ทำการศึกษามาประยุกต์เพื่อให้เหมาะสมกับข้อมูลกระบวนการผลิต
7. นำข้อมูลที่เตรียมไว้มาทำการทดสอบกับระบบที่ได้พัฒนาไว้
8. นำผลลัพธ์ที่ได้จากระบบมาทำการวิเคราะห์ความถูกต้องและประสิทธิภาพของระบบ
9. ทำการสรุปผล และแนวทางการปรับปรุงระบบให้ดีขึ้น

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. ได้เรียนรู้การทำงานของ Data Mining โดยใช้ Classification Tree
2. ได้ระบบที่ใช้ในการวิเคราะห์สาเหตุของเสียในกระบวนการผลิตโดยใช้ Data Mining
3. สามารถนำข้อมูลที่ได้ไปปรับปรุงกระบวนการผลิตได้อย่างทันท่วงที
4. เพิ่มทักษะในการเขียนและพัฒนาโปรแกรม

บทที่ 2

การวิเคราะห์สาเหตุการเสียในกระบวนการผลิต

2.1 ลักษณะกระบวนการผลิตในอุตสาหกรรมฮาร์ดดิสก์

ในอุตสาหกรรมการผลิตทั่วไป ต่างก็มีจุดมุ่งหมายขององค์กรที่จะควบคุมกระบวนการผลิต ให้มีความสามารถในการผลิต ได้ผลผลิตตามความต้องการ และมีประสิทธิภาพที่ดี ซึ่งก็ขึ้นอยู่การควบคุมกระบวนการผลิต และ การหาแนวทางในการปรับปรุงกระบวนการให้ดียิ่งขึ้น เพื่อให้เกิดผลกำไรสูงสุด โดยยังคงรักษาคุณของผลิตภัณฑ์นั้นไว้ได้

โครงการนี้จะเป็นการนำข้อมูลในกระบวนการผลิตในอุตสาหกรรมฮาร์ดดิสก์ โดยข้อมูลที่นำมาเป็นกระบวนการผลิตขั้นตอนของการทดสอบ Electrical Test (ET) ของ HGSA (Head gimbal/stack assembly) และข้อมูลการผลิตในขั้นตอนกระบวนการผลิต Slider เนื่องจากกระบวนการผลิตฮาร์ดดิสก์ ก็มีอีกหลายกระบวน สามารถแบ่งออกเป็นกระบวนการใหญ่ๆ ได้ 4 ขั้นตอน ดังรูปที่ 2.1 คือ

1. WAFER
2. SLIDER
3. HGSA
4. DRIVE

รูปที่ 2.1 แสดงกระบวนการผลิต ฮาร์ดดิสก์

กระบวนการทำ HGSA นั้น ก็เป็นกระบวนการผลิตที่สำคัญก่อนที่จะนำ HGSA ที่ได้ไปทำการประกอบเพื่อเป็น ฮาร์ดดิสก์ต่อไป ซึ่งในปัจจุบันนี้ ในกระบวนการผลิตก็จะมีการประกอบวัตถุดิบต่างๆเข้าด้วยกัน และมีการขั้นตอนการทดสอบคุณสมบัติของ HGSA ว่าได้ตามคุณสมบัติของผลิตภัณฑ์ที่กำหนดไว้หรือไม่ โดยในขั้นตอนการทดสอบ ET นี้จะมีค่าพารามิเตอร์ต่างๆ ที่กำหนดว่าจะทำการทดสอบ และมีค่าที่เป็นขอบเขตการยอมรับของแต่ละผลิตภัณฑ์ (Specification) ซึ่งค่าพารามิตเตอร์เหล่านี้อาจจะเกี่ยวข้องกับค่าที่กระบวนการผลิตอื่นๆก่อนหน้านี้ และ HGSA ที่มีค่า

พารามิตเตอร์ต่างๆ ผ่านตามขอบเขตการยอมรับของ HGSA ในทุกข้อที่กำหนดไว้แล้วเท่านั้นจึงจะถูกส่ง ต่อไปเพื่อประกอบเป็นฮาร์ดดิสก์ต่อไป ซึ่งการตัดสินถึงคุณภาพของ

ค่าพารามิตเตอร์ต่างๆ เหล่านี้จะถูกเก็บลงใน ดาต้าเบส (Database) เป็นช่วงระยะเวลาหนึ่ง ซึ่งจะมีขนาดจำนวนข้อมูลขึ้นอยู่กับกำลังการผลิตของแต่ละผลิตภัณฑ์ ซึ่งมักจะเป็นข้อมูลจำนวนมาก

Parameters	SPEC		
	Unit	Min	Max
BLPOP	dB	-1	4
COLD_RD_RES	Ohm		158
HFA_AVG	uVp-p	845	
LF_AMP_SYM	-	0.75	1.33
LF_RAN	%		10
MAR_P (W+E)	u"		46.7
MAR_P2			-4.05
MAXDW_3070	-		3
MAXTP_SLR	-		2
MR_RES	Ohm	38	64
OTC_AVG	uin	5.84	
OTC_EFL	log		-7
OTC_OFF		-23.4	23.4
OVW_AVG	dB		-25
PW_UINCH	ns		7.9
RD_W_ASYM	%	-15	11
RD_WDT	uin	18	29.7
SGAW	log		-4.05
TP_NLUMP	%		5
TP_PLUMP	%		5
Value17 (Super BR)			-0.1
Value9 (Delta OVW)	dB		12.5
WR_RES	Ohm	4	17
X_TRK_SYM	-	-0.5	0.5

รูปที่ 2.2 ตัวอย่างค่าพารามิตเตอร์ต่างๆ ที่ใช้ในการทดสอบ ET ในขั้นตอน HGSA

รูปที่ 2.2 เป็นตัวอย่างค่าพารามิเตอร์ต่างๆ ที่ใช้ทดสอบในกระบวนการ ET ของ HGSA รวมทั้งค่าขอบเขตการยอมรับได้ของแต่ละพารามิตเตอร์ ซึ่งนอกจากข้อมูลเหล่านี้แล้วยังมีปัจจัยอื่นๆที่ต้องใช้ในการวิเคราะห์ถึงสาเหตุของปัญหาที่เกิดขึ้นในกระบวนการผลิตอีก เช่น การควบคุมกระบวนการผลิต ความผิดพลาดที่อาจเกิดจากคนหรือเครื่องมือวัด จากค่าของแฟคเตอร์ที่เราใช้ในการปรับแต่งเพื่อให้สามารถทำการอ่านค่าได้ถูกต้องในมาตราฐานเดียวกัน หรือเกิดจากแผ่นmedia ที่ใช้ในการทดสอบการอ่านค่า ซึ่งปัจจัยเหล่านี้เราอาจจะอนุมานได้ว่าเป็นพารามิเตอร์ส่วนหนึ่งของผลิตภัณฑ์ เพียงแต่ไม่มีการกำหนดค่าขอบเขตการยอมรับได้ที่แน่นอน

นอกจากการคำนึงถึงคุณภาพของผลิตภัณฑ์แล้ว สิ่งสำคัญอีกสิ่งหนึ่งที่ต้องคำนึงถึงอยู่เสมอในกระบวนการผลิตก็คือ ผลผลิต (Yield) ที่ได้จากกระบวนการผลิต มีสูตรการหาผลผลิตดังสมการข้างล่างนี้ ซึ่งเป็นส่วนที่จะบอกว่ากระบวนการผลิตนั้นมีประสิทธิภาพมากน้อยเพียงไร โดยปกติแล้ว กระบวนการผลิตทั่วๆไปต่างก็มีจุดมุ่งหมายในการที่จะผลิตเพื่อให้เกิดของเสียน้อยที่สุด เพื่อก็ให้เกิดผลกำไรสูงสุด จากสมการผลผลิตนี้ ผลิตภัณฑ์ที่ยอมรับได้ก็คือผลิตภัณฑ์ที่ย่านเกณฑ์ตามขอบเขตการยอมรับได้ ซึ่งแตกต่างกันตามแต่ละประเภทผลิตภัณฑ์

$$\text{ผลผลิต (Yield)} = \frac{\text{จำนวนผลิตภัณฑ์ที่ยอมรับได้ (Output)}}{\text{จำนวนผลิตภัณฑ์ทั้งหมดในกระบวนการผลิต (Input)}}$$

รูปที่ 2.3 ตัวอย่างสมาการคำนวณหาค่าผลผลิตที่ได้

การคำนวณหาค่าผลผลิตนี้ก็ทำโดยการดึงข้อมูลของกระบวนการผลิตที่เก็บอยู่ใน Database มาทำการคำนวณ ตามสมการข้างต้นนี้ โดยคำนึงถึงค่าขอบเขตการยอมรับได้ ดังนั้น เราจึงสามารถหา ผลผลิตแยกแต่ละพารามิตเตอร์ได้เพื่อช่วยสะดวกในการวิเคราะห์ถึงประสิทธิภาพของกระบวนการผลิต ซึ่งผลผลิตของผลิตภัณฑ์ต้องเกิดจากการคำนึงถึงทุกพารามิตเตอร์ที่ทำการทดสอบ ว่าต้องอยู่ในขอบเขตการยอมรับได้ของผลิตภัณฑ์

การวิเคราะห์ถึงประสิทธิภาพ (Performance analysis) ของกระบวนการผลิต เป็นกระบวนการที่กระทำไปโดยที่มีจุดมุ่งหมายที่สำคัญเพื่อให้ลูกค้าพอใจ (Customer satisfied) เป็นกระบวนการที่ มองถึงหัวใจของปัญหาที่เกิดหรือ คือการวิเคราะห์ถึงสาเหตุของปัญหา และโอกาสที่มีอยู่ที่จะสามารถนำไปสู่หนทางการแก้ไขหรือปรับปรุงกระบวนการผลิตให้มีประสิทธิภาพมากยิ่งขึ้น

วิธีการวิเคราะห์ปัญหาที่นิยมแบบเดิมเช่น การใช้หลักการทางสถิติ การใช้การหลักการทดสอบสมมุติฐาน หรือ การออกแบบการทดลอง ในบางครั้งไม่สามารถที่จะวิเคราะห์ข้อมูลได้อย่างมีประสิทธิภาพได้ อีกทั้งยังต้องใช้โดยกลุ่มคนที่มีความรู้ทางสถิติ หรือมีความรู้และประสบการณ์เกี่ยวข้องกับระบบที่จะทำการวิเคราะห์ข้อมูล ทำให้ Data mining เป็นอีกกลยุทธ์หนึ่งที่สามารถนำมาใช้ช่วยในการวิเคราะห์หาสาเหตุการเสียของกระบวนการผลิตทั่วๆไป

2.2 วิธีการหาสาเหตุของปัญหาในกระบวนการผลิต

ลักษณะการแก้ไขปัญหาทั่วไปที่ใช้อยู่ในปัจจุบัน เป็นการใช้หลักการวิเคราะห์ข้อมูลจากข้อมูลที่มีอยู่ ซึ่งเป็นส่วนหนึ่งที่สำคัญของกระบวนการผลิต ที่จะทำให้กระบวนการผลิตดำเนินงานได้อย่างราบรื่น โดยมีวิธีการดำเนินการทั่วๆ ไป ดังนี้

- วิเคราะห์ปัญหาและสาเหตุของปัญหาที่เกิดขึ้น
- หาความสัมพันธ์จากข้อมูลที่มีอยู่โดยประสบการณ์ที่มีอยู่ในการตั้งสมมุติฐาน และใช้หลักการทางสถิติ และใช้เครื่องมือทางสถิติต่างๆ เช่น MinitTab หรือ Jmp ในการพิสูจน์สมมุติฐานนั้นๆ
- นำความสัมพันธ์ที่หาได้มา มาทำการพยากรณ์สิ่งที่จะเกิดในอนาคต เพื่อหาแนวทางที่จะรองรับ
- หาวิธีการปรับปรุงกระบวนการผลิตให้ดีขึ้น

ซึ่งในปัจจุบันกระบวนการเหล่านี้ต้องกระทำโดยใช้คนที่มีพื้นฐานความรู้เกี่ยวกับขั้นตอนการผลิตทั้งหมด มีความรู้เกี่ยวกับเทคโนโลยีในอุตสาหกรรมฮาร์ดดิสก์ และต้องมีการเรียนรู้เพิ่มเติมตลอดเวลา มีความรู้ทางสถิติพอสมควร และต้องมีประสบการณ์ในเรื่องเกี่ยวกับการวิเคราะห์ข้อมูล เพื่อที่จะสามารถใช้ เครื่องมือ(Tools) ทางสถิติต่างๆ มาช่วยในการวิเคราะห์ข้อมูลได้ เพราะการที่จะทำการวิเคราะห์ข้อมูลต้องทราบ ถึงความเกี่ยวข้องกันของค่าพารามิเตอร์ต่างๆ และโมเดลทางสถิติที่สามารถนำมาใช้กับข้อมูลที่มีอยู่ได้

จากคุณสมบัติของผู้ที่จะสามารถทำการวิเคราะห์ปัญหาที่เกิดขึ้นในกระบวนการผลิตนี้ ทำให้ขั้นตอนการแก้ไขปัญหาเห็นไปได้อย่างค่อนข้างช้า เนื่องจากขาดบุคลากรที่มีคุณสมบัติดังกล่าว และการที่จะฝึกฝนให้คนมีประสบการณ์ขนาดนั้นก็เป็นเรื่องที่ต้องใช้เวลาอย่างมาก

นอกจากนี้อุตสาหกรรมฮาร์ดดิสก์ ก็ยังเป็นอุตสาหกรรมที่มีการพัฒนาอย่างรวดเร็ว มีการวิจัยและพัฒนาเพื่อนำเทคโนโลยีใหม่ๆ มาใช้ เนื่องจากมีคู่แข่งทางการตลาด หลายราย และเพื่อตอบสนองความต้องการของผู้บริโภค จึงทำให้ต้องการเครื่องมือที่จะสามารถช่วยในการวิเคราะห์ข้อมูลได้อย่างมีประสิทธิภาพ และรวดเร็ว

บทที่3

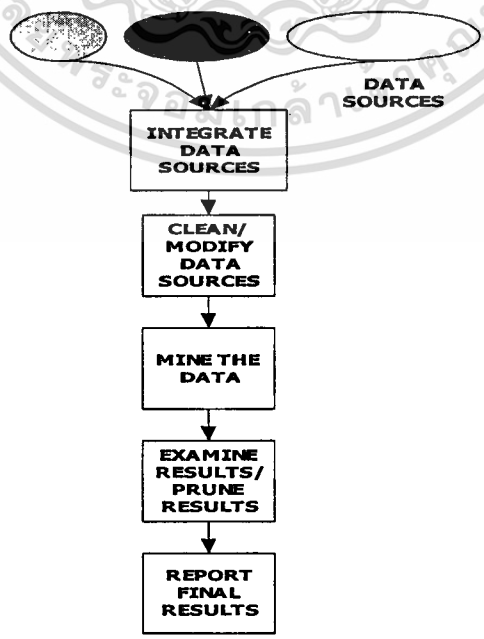
Data Mining

ดาต้าไมน์นิ่ง เป็นกระบวนการที่เป็นการนำเอาข้อมูลที่ซ่อนอยู่อยู่ภายใต้ข้อมูลซึ่งข้อมูลเหล่านี้อาจมาจากฐานข้อมูลขนาดใหญ่ ข้อมูลที่ได้เป็นข้อมูลที่ไม่ทราบมาก่อน และข้อมูลที่ได้จากการทำ Data mining เหล่านี้ก็เป็นข้อมูลที่มีประโยชน์ สามารถนำข้อมูลนี้ไปใช้ช่วยเป็นแนวทางในการตัดสินใจใดๆที่ก่อให้เกิดผลประโยชน์ในทางธุรกิจ ซึ่งถือว่าเป็นจุดประสงค์หลักของการทำ Data Mining

ข้อมูลที่ได้จากการทำ Mining ไม่ได้เกิดจากสมมุติฐาน หรือ จากการคาดคะเนจากประสบการณ์ แต่เป็นข้อมูลหรือความสัมพันธ์ที่เกิดขึ้นจริงที่ซ่อนอยู่ภายใต้ข้อมูลที่เรามีอยู่ ดังนั้นการทำ Data Mining มักจะไม่ใช้การตั้งสมมุติฐาน แต่จะเป็นการดูผลลัพธ์ที่ได้จากการทำ Mining เลย ซึ่งจะแตกต่างจากวิธีการวิเคราะห์ข้อมูลทางสถิติแบบอื่นๆ

3.1 การทำ Data Mining

การทำ Data Mining ประกอบด้วยกระบวนการต่างๆ ซึ่งสามารถแบ่งออกได้เป็นขั้นตอนใหญ่ๆ ได้ 5 ขั้นตอน ดังรูปที่ 3.1 ดังนี้



รูปที่ 3.1 ขั้นตอนการทำ Data Mining [Bhavani. 1998]

ขั้นตอนการทำ Data Mining ประกอบด้วย 5 ขั้นตอนใหญ่ ดังนี้

1. ขั้นตอนกำหนดวัตถุประสงค์และแหล่งข้อมูลของการทำ Data Mining

เป็นขั้นตอนส่วนสำคัญที่จะกำหนดถึง ความต้องการหรือปัญหา ที่ต้องการทราบ ซึ่งมักเป็นความต้องการที่มุ่งเพื่อนำคำตอบที่ได้ ให้เกิดประโยชน์ แต่จะไม่ใช่เกิดจากการตั้งสมมุติฐาน และเป็นการกำหนดถึงแหล่งที่มาของข้อมูลที่จะทำการ Mining ซึ่งก็เป็นแหล่งข้อมูลที่คาดว่าจะได้คำตอบจากสิ่งที่ต้องการทราบ

2. ขั้นตอนการจัดเตรียมข้อมูลที่จะทำการ Mining

มักขั้นตอนที่ต้องใช้เวลามากที่สุด เนื่องจากต้องมีการพิจารณาข้อมูลในแทบจะทุกเรื่องเช่น ความเกี่ยวข้องของข้อมูลกับวัตถุประสงค์ของการทำ ชนิดของข้อมูล ประเภทของข้อมูล จำนวนของข้อมูล การตรวจสอบข้อมูลว่าเป็นข้อมูลที่เหมาะสมหรือไม่ ทั้งนี้ต้องคำนึงถึงอายุของข้อมูลด้วย โดยอาจต้องมีการกำจัดข้อมูลที่ไม่จำเป็นหรือไม่ถูกต้องออกไป (Noisy data) รวมทั้งเป็นการเตรียมข้อมูลให้พร้อมที่จะทำการ Mining โดยการปรับเปลี่ยนรูปแบบของข้อมูล (Data Transform) เพื่อให้เหมาะสมกับอัลกอลิธึมที่จะเลือกใช้ ซึ่งเป็นการจัดการเพื่อให้การMiningทำไปได้อย่างมีประสิทธิภาพ

3. ขั้นตอนการทำ Data Mining

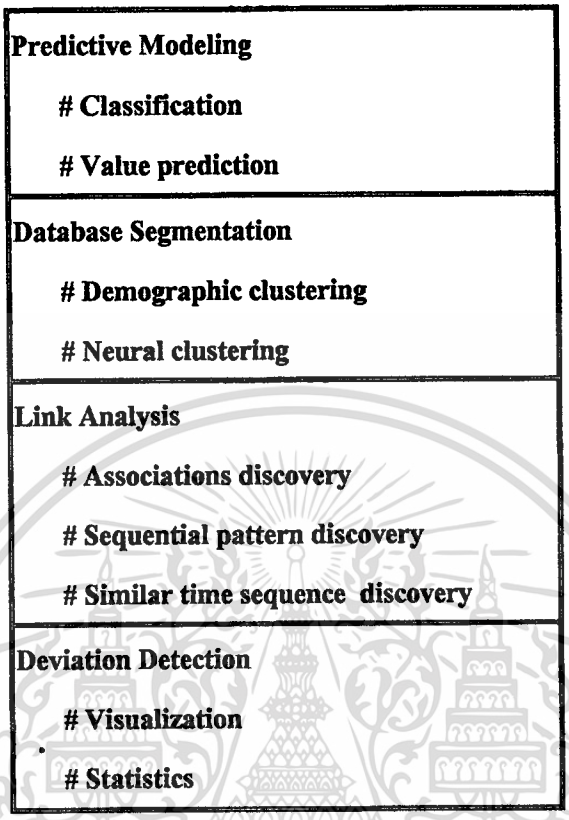
ขั้นตอนการทำ Data mining ถือว่าเป็นหัวใจหลักของการทำ Data Mining เพราะการเลือกเอาวิธีการและกระบวนการอัลกอลิธึม (Algorithm) ในการทำ Mining ที่เหมาะสม ก็จะเป็นการทำให้การ Mining ได้ผลอย่างรวดเร็วและถูกต้องตามจุดประสงค์ที่ต้องการ ซึ่งเทคนิคต่างๆของการทำ Data mining จะอธิบายรายละเอียดให้หัวข้อถัดไป

4. ขั้นตอนการประเมินผลที่ได้จากการทำ Data Mining

เป็นเสมือนขั้นตอนการอธิบายและประเมินถึงผลที่ได้จากการทำ Mining ว่าสามารถนำมาใช้ให้บรรลุถึงจุดประสงค์ที่ต้องการหรือไม่ รวมทั้งเป็นการประเมินถึงความถูกต้องของผลที่ได้จากการทำ ซึ่งก็นับว่าเป็นสิ่งสำคัญอย่างหนึ่งเช่นกัน เพราะบางครั้งผลที่ได้จากการทำ mining อาจมีข้อผิดพลาด ซึ่งอาจเกิดได้จากหลายสาเหตุที่เราอาจไม่คาดคิด จึงต้องมีการตรวจสอบผลที่ได้

5. ขั้นตอนการนำเสนอความรู้ที่ได้

การนำเสนอความรู้ได้เป็นขั้นตอนสุดท้ายของกระบวนการทั้งหมด เป็นการนำเสนอถึงผลที่ได้จากการทำ Data mining และนำเสนอถึงวิธีการที่จะนำผลที่ได้นี้ไปใช้ให้เกิดประโยชน์



รูปที่ 3.2 กระบวนการและเทคนิคต่างๆของData Mining [Peter. 1997]

3.2 เทคนิคของการทำ Data Mining

Data Mining มีเทคนิคและอัลกอลิธึมที่สามารถนำมาใช้งานอยู่หลายประเภท ขึ้นอยู่กับรูปแบบ Application ที่ต้องการนำมาใช้งาน แต่สามารถ แบ่งออกเป็น รูปแบบ ต่างๆ ได้ดังรูปที่ 3.2

- Predictive Modeling

เป็นการคาดคะเน ทำนายถึงความเป็นไปได้ โดยใช้การสังเกตจากรูปแบบของข้อมูลที่มีอยู่ โดยมีลักษณะเป็นการเรียนรู้จากกลุ่มข้อมูลที่ได้กำหนดไว้ แล้วจึงนำไปวิเคราะห์กลุ่มข้อมูลที่ต้องการ

ซึ่งรูปแบบนี้ เป็นลักษณะแบบ Supervised leaning จึงมีรูปแบบการพัฒนาในสองช่วงคือ

- ช่วงการเรียนรู้ (Training Phase) เป็นการสร้างโมเดลโดยการใช้ข้อมูลในอดีต และมีจำนวนข้อมูลจำนวนมาก
- ช่วงการทดสอบ (Testing Phase) เป็นการตรวจสอบความน่าเชื่อถือและประสิทธิภาพของโมเดลที่สร้างขึ้น จึงเป็นข้อมูลที่มีจำนวนไม่มากนัก

Predictive Modeling ยังสามารถแบ่งย่อยได้อีก เป็น 2 เทคนิคคือ

1. Classification ซึ่งเป็นการแบ่งกลุ่มของข้อมูลตามชนิดของกลุ่มข้อมูลที่ควรจะเป็น สามารถแบ่งกลุ่มข้อมูลได้อย่างชัดเจน ซึ่งมีอัลกอลิธึมที่นิยมคือ Tree Induction และ Neural Induction

2. Value prediction เป็นการทำนายถึง ค่าความต่อเนื่องของข้อมูล เป็นการทำนายค่าที่เป็นตัวเลข โดยมีเทคนิคที่นิยมใช้คือ Linear regression และ Nonlinear regression

นอกจากนี้แล้วยังมีเทคนิค RBF (Radial basis function) ซึ่งเป็นเทคนิคใหม่ของ Value prediction โดยเป็นเทคนิคที่ให้ความยืดหยุ่นมากกว่าแบบ linear regression และ Nonlinear regression โดยที่เทคนิคนี้จะไม่ได้เป็นแค่ single nonlinear function อย่างเดียว แต่จะมีการ weight ผลรวมของเซ็ทของ nonlinear ด้วย ทำให้เรียกว่าเป็น Radial basis function

- Database Segmentation

Segmentation หรือ Clustering เป็นการทำการแบ่งกลุ่มย่อยข้อมูลเพื่อทำการแยกออกให้ทราบว่าข้อมูลชุดนี้มีทั้งหมดกี่กลุ่ม ซึ่งการแบ่งกลุ่มข้อมูลนี้ไม่สามารถกำหนดได้ว่าข้อมูลนี้ควรจะอยู่กลุ่มใด แต่เป็นการกำหนดกลุ่มของข้อมูลจากธรรมชาติของข้อมูลเอง ไม่ได้ใช้ความรู้สึกหรือประสบการณ์ในการตัดสินใจ แบ่งกลุ่มข้อมูลแต่จะจัดการ โดยอัลกอลิธึมที่เหมาะสมของแต่ละกลุ่มข้อมูล จึงเป็นเรียกว่าเป็นรูปแบบของ unsupervised learning ซึ่งก็แบ่งได้ย่อยได้ตามเทคนิคที่ใช้ คือ Demographic clustering และ Neural clustering

- Link Analysis

เป็นการศึกษาความสัมพันธ์ของข้อมูลหรือกลุ่มของข้อมูลว่ามีความสัมพันธ์กันในรูปแบบลักษณะใด โดยเรียกความสัมพันธ์นี้ว่าเป็น 'Association' เป็นโมเดลที่นิยมมากในการวิเคราะห์เพื่อหาความสัมพันธ์ระหว่าง ลูกค้ากับ สินค้าหรือบริการ

สามารถแบ่งย่อยได้อีก 3 ลักษณะตามการวิเคราะห์ข้อมูลคือ

1.Association discovery เป็นการวิเคราะห์ข้อมูลที่เกิดขึ้นพร้อมกันภายในกลุ่มข้อมูลเดียวกัน เป็นเทคนิคที่นิยมมากชนิดหนึ่ง มักในการวิเคราะห์ถึงพฤติกรรมการซื้อของผู้บริโภค จึงเป็นเทคนิคที่มีอีกชื่อหนึ่งว่า Market basket analysis

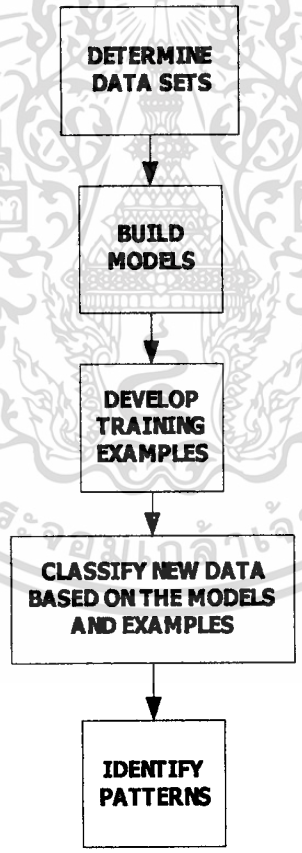
2. Sequential pattern discovery เป็นการศึกษาความสัมพันธ์ระหว่างข้อมูล โดยเทียบข้อมูลกับเวลา ซึ่งเป็นการศึกษาพฤติกรรมในระยะยาว (long term behavior)

3. Similar time sequence discovery เป็นการศึกษาพฤติกรรมของข้อมูลที่เกิดขึ้นทั้งหมดหรือเกิดขึ้นในช่วงเวลาเดียวกัน เพื่อหาความสัมพันธ์ระหว่างกลุ่มของข้อมูลเหล่านี้

- Deviation Detection

เป็นโมเดลที่จะใช้เทคนิคทางสถิติ (Statistics) และ การทำให้เห็นภาพ (Visualization) ซึ่งเป็นรูปแบบการสรุปข้อมูลให้ออกมาในรูปแบบการแสดงผลแบบกราฟฟิก (graphic) เช่นการใช้ graph เช่น Histograms Scatter plots หรือ กราฟ วงกลม เพื่อให้เข้าใจง่าย ก็ทำให้เป็นที่นิยมอย่างมาก ในการใช้แสดงผล นอกจากนี้ visualization ยังสามารถนำไปใช้ร่วมกับเทคนิคอื่นๆ โดยใช้ในการแสดงผลที่ได้ในรูปแบบของกราฟฟิก เพื่อให้เข้าใจได้ง่ายขึ้น

Statistics ก็ใช้เพื่อการวัดถึงความชัดเจนหรือความน่าเชื่อถือของข้อมูล ทำให้มักจะถูกนำไปใช้ในกระบวนการเตรียมข้อมูล หรือทำการ วิเคราะห์ผลที่ได้จากการทำ Data mining ว่ามีความน่าเชื่อถือมากน้อยเพียงไร



รูปที่ 3.3 กระบวนการเอาเทคนิคของ Data mining มาใช้ [Bhavani. 1998]

นอกจากนี้แต่ละเทคนิคของดาต้าไมน์นิ่งยังสามารถเลือกที่จะนำเอาอัลกอลิธึมต่างๆมาใช้ได้ และรูปที่ 3.3 ก็เป็นรูปแสดงกระบวนการนำเอาเทคนิคของ Data mining มาใช้ ซึ่งการเลือกใช้อัลกอ

ลิธึมต่างๆ ก็ยังขึ้นอยู่กับปัจจัยอื่นๆ อีก เช่น จุดมุ่งหมายของการทำ ลักษณะของข้อมูล ชนิดของข้อมูล และจำนวนข้อมูลที่มีอยู่ ในการทำ Data mining บางครั้งก็อาจต้องมีการเปลี่ยนเทคนิคการทำ ถ้าเทคนิคนั้นไม่เหมาะสม

ขั้นตอนที่สำคัญของกระบวนการนำมาใช้อยู่ที่การกำหนดถึงกลุ่มของข้อมูลที่จะนำมา และการสร้าง model ที่เหมาะสมกับข้อมูล ซึ่งจะทำให้ผลของการทำ Data mining ได้ผลอย่างถูกต้องและรวดเร็ว ในการทำ Data mining อาจจะได้ข้อมูลหรือรูปแบบของข้อมูลมากมาย แต่ไม่ทุกรูปแบบของข้อมูลที่เราสนใจ เราก็จะเลือกสนใจแต่รูปแบบที่ตรงกับจุดมุ่งหมายหรือคำตอบที่เราต้องการ

3.3 ความหมายของ Classification

Classification เป็นเทคนิคหนึ่งของ Data mining ใน Predictive modeling ซึ่งเป็นเทคนิคที่เรียกอีกอย่างว่าเป็น Supervised learning โดยมีกระบวนการที่ประกอบด้วย 2 ขั้นตอนใหญ่ๆ คือ

1. การเรียนรู้ (Learning)

เป็นการนำเอากลุ่มของข้อมูล (Training Data set) ที่จะนำมาทำการศึกษาโดยใช้ อัลกอลิธึมของ classification เพื่อทำการเรียนรู้เพื่อทำการสร้าง โมเดล (Model) ที่จะสามารถอธิบายถึงลักษณะของข้อมูล ที่ซ่อนอยู่อยู่ภายใต้ข้อมูล ซึ่งโมเดลนี้จะมีลักษณะที่กลุ่มของข้อมูลถูกทำการแจกแจงออกเป็น class ต่างๆ ด้วย Classification rule และ class แต่ละ class นี้ก็จะมีลักษณะเฉพาะกลุ่มที่สามารถจะสรุปออกมาได้ ซึ่งจะเป็นการ mapping จาก attributed ไปเป็นกลุ่มที่สามารถระบุได้

2. การทดสอบข้อมูล (Testing data)

Test data จะถูกนำมาใช้เพื่อวัดถึงความถูกต้องของ Classification rule ที่ถูกสร้างมาจากขั้นตอนการ learning โดยความถูกต้องนี้จะเป็นตัวพิจารณาว่า Classification rule ที่ถูกสร้างนี้ สามารถนำมาให้กับกลุ่มข้อมูลใหม่ๆ ได้หรือไม่

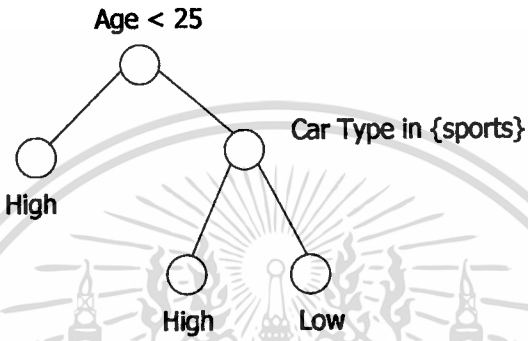
Classification ยังสามารถใช้เทคนิคแบบต่างได้อีก เช่น

- Decision Tree มีลักษณะการแบ่งกลุ่มของข้อมูลเป็นโครงสร้างแบบต้นไม้

- Neural Network เป็นแบบจำลองที่สร้างขึ้นเพื่อเลียนแบบการทำงานของ เครือข่ายประสาทสิ่งมีชีวิต (Biological Neural) โดยเป็นชุดของการเชื่อมต่อของอินพุทและเอาต์พุท โดยแต่ละเส้นทางการเชื่อมต่อ มีการweight ค่าค่าหนึ่ง ซึ่งจะถูกปรับให้มีค่าที่เหมาะสมที่สุดในแต่ละเส้นทาง ในขั้นตอนการเรียนรู้

Classification Tree

เป็นโครงสร้างที่แสดงผลอยู่ในรูปแบบของแผนภูมิต้นไม้ โดยแต่ละกิ่งของต้นไม้ก็เกิดจากการทดสอบเพื่อการจัดประเภท (Classification) ที่ปลายสุดของแต่ละโหนด (Leaves node) ของต้นไม้ ก็เป็นกลุ่มของข้อมูลที่ถูกจัดกลุ่มตามประเภทของข้อมูลที่มีอยู่ รูปที่ 3.4 คือตัวอย่างของ Classification Tree



รูปที่ 3.4 ตัวอย่างClassification Tree

สำหรับโครงงานนี้เลือกใช้ Data Mining โดยใช้ Classification tree เนื่องจากเหตุผลหลายประการ ดังนี้

1. เป็นเทคนิคที่ใช้เวลาในการทำงานน้อย เมื่อเปรียบเทียบกับ Classification แบบอื่นๆ
2. ผลที่ได้จากการทำ Classification tree นี้ง่ายต่อการเข้าใจของมนุษย์
3. ความถูกต้องแม่นยำในการทำนาย (Predictive) ของข้อมูลมีความถูกต้องเท่าเทียม หรือมากกว่า classification แบบอื่นๆ
4. เหมาะกับดาต้าเบสขนาดใหญ่ (Very large database) และ สามารถใช้กับ highly dimensional ได้

ขั้นตอนการสร้าง Classification Tree ยังสามารถแบ่งย่อยได้เป็นสองส่วนคือ

1.Tree building เป็นขั้นตอนการสร้าง tree แบบจากบนลงล่าง Top- Down โดยเริ่มต้นสร้างจากตำแหน่งที่รากของต้น (root node) แล้วใช้ algorithm ในการค้นหาพารามิตเตอร์ที่เหมาะสมที่สุดที่จะทำการแตก node ออกเป็นต้นไม้

2.Tree pruning เป็นขั้นตอนที่จะตรวจสอบ Tree ที่สร้างขึ้นมาแรก ไม่ให้มีขนาดใหญ่เกินไป ซึ่งอาจเกิดปัญหาในการเข้าถึงข้อมูลที่แตกย่อยมากเกินไป หรือปัญหาเรื่อง Overfitting ได้

Classification Tree มี algorithm ต่างๆที่สามารถเลือกใช้ในการสร้าง Tree โดยทั่วไปจะแตกต่างกันที่หลักการในการสร้าง เลือกพารามิตเตอร์ที่จะทำการแตก Node เพื่อที่จะสร้าง Tree หรือหลักการในการ pruning

ซึ่งมีตัวอย่างของอัลกอลิธึมต่างๆของ Classification Tree ดังนี้

- ID3 : Induction Decision Tree เป็นอัลกอลิธึมที่พัฒนาโดย Quilan
- C4.5 : Decision tree induction algorithm เป็นตัวที่พัฒนามาจาก ID3 อีกทีโดย Quilan
- CART : Classification and regression trees
- SLIQ : A Fast Scalable Classifier for data mining
- SPRINT: A Scalable Parallel Classifier for data mining
- RainForest : A Framework of Fast Decision Tree Construction of Large Dataset
- CMP : A Fast Decision Tree Classifier Using Multivariate Predictions

สำหรับโครงงานพัฒนาระบบนี้เลือกที่จะใช้อัลกอลิธึม SLIQ ซึ่งจะกล่าวถึงรายละเอียดของ SLIQ ในบทต่อไป

บทที่ 4

ทฤษฎีของ Classification Tree โดยใช้ SLIQ

สำหรับโครงงานนี้เลือกที่จะใช้อัลกอลิธึม SLIQ ซึ่งเป็นอัลกอลิธึมหนึ่งใน Decision Tree classifier ที่สามารถรองรับได้ทั้งข้อมูลAttribute ที่เป็นตัวเลข Numeric หรือเป็น Categorical ซึ่งแบ่งข้อมูลเป็นกลุ่มๆ อย่างชัดเจน และเป็นอัลกอลิธึมที่มีการปรับปรุงเรื่องความเร็วในการประมวลผลให้สามารถใช้งานได้กับข้อมูลจำนวนมาก ทำให้เหมาะกับระบบที่มีการขยายตัวสูง

4.1 SLIQ (A Fast Scalable Classifier for Data Mining)

SLIQ เป็นอัลกอลิธึมที่มีข้อดีหลายอย่างดังนี้

1. สามารถรองรับทั้งข้อมูลแบบ เป็น Numeric หรือเป็น Categorical
2. ใช้เทคนิค Pre-sorting ในการจัดการกับข้อมูลแบบตัวเลข ในช่วงของการสร้าง Tree ทำให้ไม่ต้องมีการจัดเรียงข้อมูลทุกๆ ครั้งในประมวลผล ช่วยลดเวลาในการประมวลผล
3. โครงสร้างการจัดการข้อมูล มีการนำเทคนิคของการแบ่งข้อมูลออกเป็น Class list และ Attribute list แล้วใช้ Index เข้ามาช่วยในการอ้างอิงถึงกัน ทำให้ช่วยลดขนาดของข้อมูลที่ใช้ในการประมวลผล
4. เป็นอัลกอลิธึมที่สามารถใช้งานกับข้อมูลจำนวนมากได้ โดยยังคงมีความถูกต้องในการทำงานค่อนข้างสูง ทำให้สามารถใช้กับระบบที่มีการขยายตัวสูง

ขั้นตอนการสร้าง Classification Tree ของ SLIQ ก็สามารถแบ่งย่อยได้เป็นสองส่วนคือ Tree Building และ Tree Pruning

Tree Building ก็ประกอบไปด้วยขั้นตอนย่อยๆ 2 ส่วนคือ

1. การหาค่าที่ใช้ในการ Split ของแต่ละ Attribute และ การเลือกค่าที่ดีที่สุดในการ Spilt
2. การแบ่งกลุ่มของข้อมูลออกจากกันโดยเลือกจากค่าที่ดีที่สุดในการ Split

ซึ่งการทำงานของการสร้างTree นี้ก็จะเป็นวนลูปทำซ้ำไปเรื่อยๆ จนกระทั่งสามารถทำการแบ่งกลุ่มข้อมูลจนได้กลุ่มข้อมูลเดียวกันหมดในแต่ละกลุ่มย่อยๆ ซึ่งรูปที่ 4.1 คือตัวอย่างอัลกอลิธึมที่แสดงการทำงานในการสร้าง Tree

```
MakeTree(Training 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);
```

รูปที่ 4.1 ตัวอย่างอัลกอลิธึมในการสร้าง Tree [Mehta et.at. 1996]

สำหรับอัลกอลิธึม SLIQ นี้จะทำการหาค่าที่ใช้ในการ Split ของแต่ละ Attribute โดยใช้ Gini Index เป็นตัววัดความสามารถในการ Split ข้อมูล และ SLIQ ยังมีการใช้เทคนิค Pre-sorting เพื่อช่วยให้ประมวลผลได้เร็วขึ้น ดังนั้นเวลาในการประมวลผลก็จะขึ้นอยู่กับเทคนิคที่ใช้ในการ Sort ข้อมูล

นอกจากนี้ เพื่อเป็นการขจัดข้อจำกัดในการประมวลผลที่เกิดจากขีดความสามารถในการเก็บข้อมูลของหน่วยความจำ และเพื่อให้เทคนิคของ Pre-sorting ใช้งานได้อย่างมีประสิทธิภาพ ทำให้ SLIQ มีการสร้างลิสต์ (list) ของแต่ละ Attribute ของข้อมูลชุดเรียนรู้ (Training Set) แยกออกจากกัน และยังมีลิสต์อีกหนึ่งลิสต์ที่ชื่อว่า Class list เป็นตัวเก็บค่า Class ของแต่ละข้อมูล ซึ่งในแต่ละลิสต์จะประกอบด้วย 2 ส่วนหลักๆ คือ Attribute Value และ Index โดยใช้ Index เป็นตัวอ้างอิงถึงกับ Attributes ค่าอื่นๆ และค่าของ Class ที่ได้ทำการแยกจัดเก็บไว้ที่ Class list ซึ่ง Class List จะเป็นที่เก็บตำแหน่งของ node ต่างๆ บนTreeด้วย นอกจากนี้ Attribute list นี้สามารถนำไปเก็บไว้ในดิสก์ได้ถ้าข้อมูลมีจำนวนมากและข้อมูลส่วนนั้นยังไม่ต้องทำการประมวลผล เพื่อเป็นการเพิ่มความเร็วในการประมวลผล

รูปที่ 4.2 เป็นตัวอย่างของข้อมูลที่จะนำมาทำ Classification Tree ซึ่งประกอบด้วย 2 Attribute คือ อายุ (Age) และประเภทของรถ (Car Type) โดยมี Class เป็นระดับของความเสี่ยงที่จะเกิดอุบัติเหตุของลูกค้าเป้าหมาย ซึ่งจากข้อมูลนี้เราสามารถนำข้อมูลมาสร้างเป็น Attribute list ได้ 2 list คือ List ของ Age และ List ของ Car Type โดยใช้ Class list Index เป็นตัวอ้างอิงถึง Class list และสำหรับข้อมูลที่เป็น Numeric หลังจากที่ทำการสร้างเป็น Attribute list แล้วเราก็สามารถทำการ

Sort ข้อมูลของแต่ละ list ได้เลย โดยใช้เทคนิค Pre-sorting คือทำการจัดเรียงข้อมูลแค่เพียงครั้งเดียว ก่อนทำการประมวลผล

Age	Car Type	Class
23	Family	High
17	Sports	High
43	Sports	High
68	Family	Low
32	Truck	Low
20	Family	High

รูปที่ 4.2 ตัวอย่างข้อมูล

รูปที่ 4.3 เป็นตัวอย่างของข้อมูล หลังจาก การสร้าง Attribute list และ Class list และทำการ Pre-sorting แล้วของข้อมูลจากรูปที่ 4.2 ซึ่งจะเห็นว่ามีการอ้างอิงข้อมูลถึงกัน โดยใช้ Class List Index

Age	Class List Index
17	2
20	6
23	1
32	5
43	3
68	4

Car Type	Class List Index
Family	1
Sports	2
Sports	3
Family	4
Truck	5
Family	6

	Class	Leaf
1	High	N1
2	High	N1
3	High	N1
4	Low	N1
5	Low	N1
6	High	N1

รูปที่ 4.3 ตัวอย่างการสร้าง Attribute list และ Class List

4.2 การแตกกิ่งของ Tree

ในสำหรับเทคนิคในการเลือกค่าของ Attibute ที่จะมาทำการแตกกิ่งนั้นมีได้หลายเทคนิค ขึ้นอยู่กับ อัลกอลิธึม ที่เลือกใช้งาน

Numeric Attribute

การแตกกิ่งของค่าตัวเลขโดยปกติที่นิยมใช้มี 3 วิธี

1.ทำการแบ่งออกเป็น Range เช่น High Medium และ Low หรือ [0,15K) , [15K,60K) , [60K,100K) , [100,….)

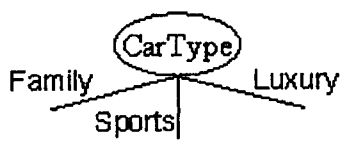
2.ใช้ค่า Thresholds ค่าใดค่าหนึ่งในการแตกกิ่ง เช่น A <= a และ A > a ซึ่งจะทำให้การแตกกิ่งแบ่งออกเป็น 2 กลุ่มเป็น Binary decision เสมอ

3.ใช้การหาค่าความสัมพันธ์เชิงเส้นของค่าข้อมูลต่าง ๆ และเลือกใช้การแบ่งข้อมูลออกเป็น Range หรือการแบ่งเป็นBinary เข้ามาใช้ร่วมกับความสัมพันธ์ที่ได้

สำหรับอัลกอลิธึม SLIQ นี้จะใช้ค่า Thresholds ในการแตกกิ่ง โดยจะเลือกค่าที่อยู่ในรูปของ A <= v เมื่อ v เป็นจำนวนจริง นอกจากนี้ยังใช้เทคนิคของ Pre-sorting ในการหาค่า Split ของตัวเลข ดังนั้น ขั้นตอนแรกของการทำงานก็คือการจัดเรียงข้อมูลของข้อมูลชุด Training จากน้อยไปหามาก เพื่อให้ได้ข้อมูลที่อยู่ในรูปของ V1 , V2 ,V3 ,….. Vn แล้วจึงทำการหาค่าที่จะทำการแตกกิ่งโดยเลือกเอาค่ากลางระหว่างข้อมูลชุด Training ที่มีอยู่เป็น $(V_i+V_{i+1})/2$ ดังนั้นถ้ามีข้อมูลทั้งหมด n ค่า ก็จะทำการคำนวณหาค่าที่เป็นไปได้ในการแตกกิ่งทั้งหมด N-1 ค่า

Categorical Attributes

ถ้า S(A) เป็น Set ของค่าที่เป็นไปได้ของ Categorical ดังนั้น Split test ที่เป็นไปได้จะอยู่ในรูปของ $A \in S'$ เมื่อ $S' \subset S$ ดังนั้นจำนวนของซับเซ็ทที่เป็นไปได้(Possible Subset) ของ Attribute ที่เป็นไปได้ซึ่งมีค่าทั้งหมด n ค่าคือ 2^n



รูปที่ 4.4 ตัวอย่างการแตกค่าของ Categorical

```
EvaluateSplits()
    For each attribute A do
        Traverse attribute 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 4.5 ตัวอย่างอัลกอลิธึมการแตกกิ่ง [Mehta et.at. 1996]

วิธีการเลือกค่าที่ดีที่สุดในการแบ่งกลุ่มข้อมูล

โดยทั่วไปที่นิยม มี 2 เทคนิคคือ Information Gain ซึ่งเป็นวิธีที่ใช้ในอัลกอลิธึม ID3และ C4.5 และ Gini Index ซึ่งเป็นวิธีที่ ใช้ในอัลกอลิธึม SLIQ นี้

Gini Index

Gini Index เป็นวิธีที่ใช้หาค่าแตกกิ่งที่ใช้ในอัลกอลิธึม CART , SLIQ และ SPRINT โดยค่า Gini Index เป็นการวัดค่าความมีสิ่งเจือปนในการแตกกิ่ง และมีสมการการคำนวณดังนี้

$$GINI(t) = 1 - \sum_{j} [p(j \mid t)]^2$$

โดยที่ Pj | t คือค่าความถี่ที่เกี่ยวข้องกับ Class j ใน node t

ซึ่งค่าที่ได้จากการคำนวณนี้ ค่าที่มากที่สุดจะอยู่ในรูปของ (1- 1/ Nc) เมื่อแต่ละข้อมูลกระจายเท่ากันในทุกๆ Class ซึ่งเป็นการบอกว่าการแตกกิ่งแบบนั้นเป็นข้อมูลที่ไม่น่าสนใจ

ค่าที่น้อยที่สุดก็คือ 0 ซึ่งจะเกิดเมื่อแต่ละกลุ่มข้อมูลถูกแยกออกจากกันอย่างชัดเจนในแต่ละ Class ซึ่งเป็นค่าที่เราสนใจมากที่สุด

ในตารางที่ 4.1 แสดงตัวอย่างการคำนวณค่า Gini Index ของข้อมูล 4 รูปแบบซึ่งมาจากข้อมูลทั้งหมด 6 ข้อมูล แต่ละแบบประกอบด้วย Class ที่สนใจ 2 Class คือ C1 และ C2

Class	Group 1	Group 2	Group 4	Group 4
C1	0	1	2	3
C2	6	5	4	3
Gini Index	$1-(0/6)^2-(6/6)^2$	$1-(1/6)^2-(5/6)^2$	$1-(2/6)^2-(4/6)^2$	$1-(0/6)^2-(6/6)^2$
	0	0.278	0.44	0.5

ตารางที่ 4.1 ตัวอย่างการคำนวณ Gini Index

ตัวอย่างกลุ่มแรก ประกอบด้วย Class C2 อย่างเดียวจำนวน 6 ข้อมูล สามารถสรุปการแบ่งข้อมูลออกเป็น C1 = 0 และ C2 = 6 ดังนั้นค่า Gini Index = $1-(0/6)^2-(6/6)^2 = 0$ ซึ่งเป็นค่าน้อยที่สุด เพราะเป็นรูปแบบการแบ่งกลุ่มข้อมูลของ C1 และ C2 ออกจากกันอย่างชัดเจน

ดังนั้นเมื่อ โหนด P ทำการแตกกิ่งออกเป็น K กลุ่ม ค่าคุณภาพของการแตกกิ่งก็สามารถทำการคำนวณได้ดังสมการข้างล่างนี้

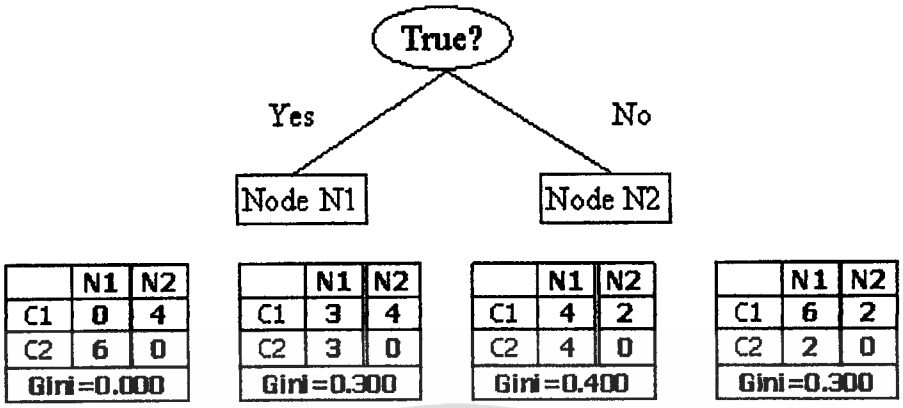
$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

โดยที่ค่า n_i คือ จำนวนของข้อมูลที่ i

n คือ จำนวนของข้อมูลที่ P

รูปที่ 4.6 ก็เป็นตัวอย่างการแตกกิ่งและการคำนวณ Gini Index ในแต่ละกรณี ซึ่งการแตกกิ่งนี้ได้แตกออกเป็น 2 ส่วนคือ N1 และ N2 และแต่ละส่วนประกอบด้วย 2 Class คือ C1 และ C2

สำหรับเกณฑ์การเลือกที่ค่า Gini Index ที่ดีที่สุดก็คือการเลือกรูปแบบที่ให้ค่า Gini Index ที่น้อยที่สุดในการแตกกิ่ง

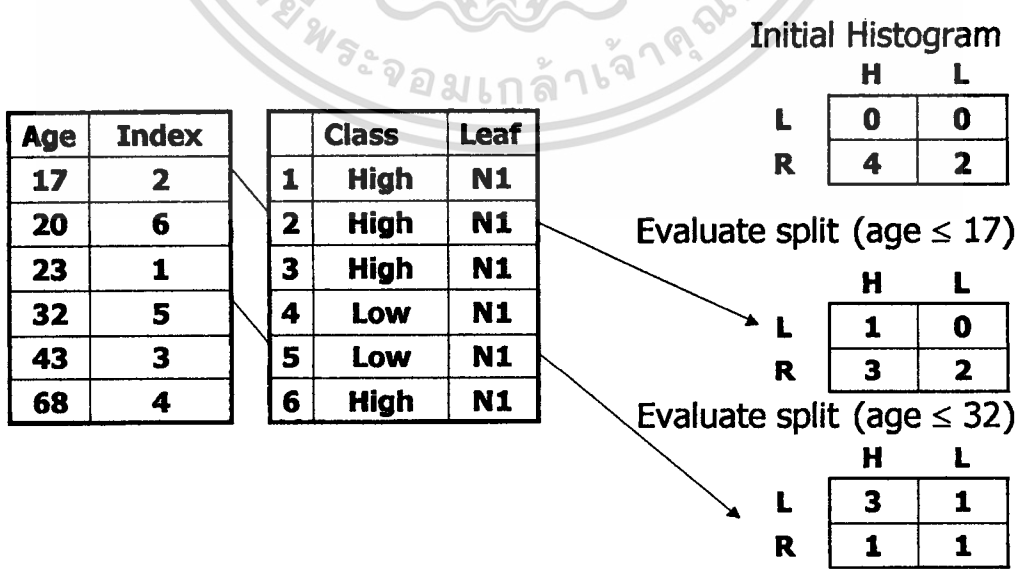


รูปที่ 4.6 ตัวอย่างแสดงการคำนวณ Gini Index ของข้อมูลรูปแบบต่างๆกัน

Class Histogram

SLIQ ยังมีการใช้ Class Histogram ในการเก็บค่าความถี่ของข้อมูลในแต่ละ Class สำหรับแต่ละค่าของ Attribute ซึ่งสำหรับค่าที่เป็นตัวเลข Class Histogram ก็จะประกอบไปด้วย ค่า Class และค่าความถี่ของข้อมูลใน Class นั้น แต่ถ้าเป็น Categorical Attribute ก็จะมีการเก็บค่า Attribute ด้วย

รูปที่ 4.7 ก็เป็นตัวอย่างการใช้ Class Histogram เก็บค่าเมื่อมีการคำนวณหาค่า Gini Index ในการแตกกิ่งของ SLIQ ของแต่ละค่าที่เป็นไปได้

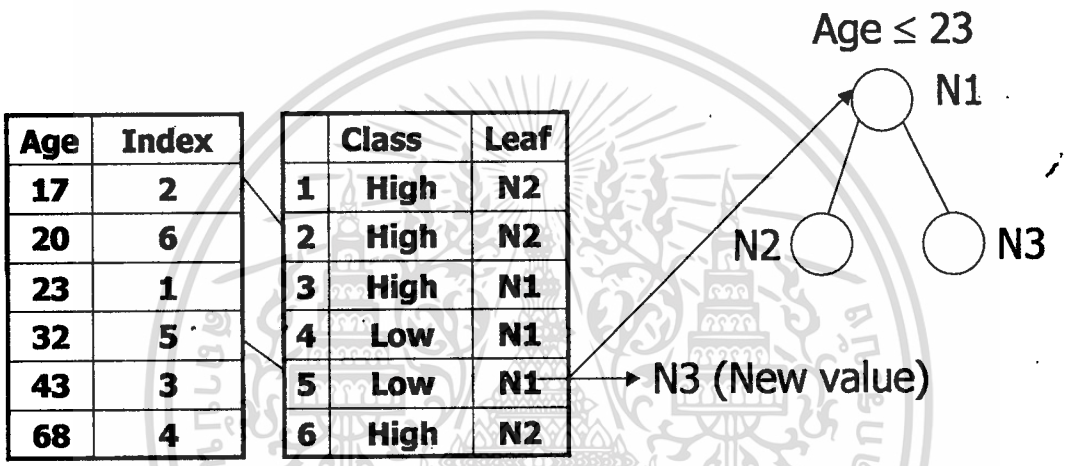


รูปที่ 4.7 ตัวอย่างการใช้ Class Histogram

การ Update Class list

เมื่อทำการหาค่าที่ดีที่สุดในการแตกกิ่งโดยใช้ค่า Gini Index แล้ว ก็จะมีการ Update ค่าใน Class list เพื่อให้ข้อมูลที่อยู่ใน Class list สามารถบอกได้ทำการแบ่งข้อมูลนี้ไปอยู่ในตำแน่งใดแล้วของ Tree

รูปที่ 4.8 เป็นตัวอย่างการ update ค่าใน Class list เมื่อมีการแตกกิ่งที่ค่า Age <=23 ค่าต่างๆ ที่ N1 ก็จะต้องทำการ update ใหม่ให้เป็นค่าที่ถูกต้องตามการแตกกิ่งนี้



Age	Index
17	2
20	6
23	1
32	5
43	3
68	4

	Class	Leaf
1	High	N2
2	High	N2
3	High	N1
4	Low	N1
5	Low	N1
6	High	N2

รูปที่ 4.8 แสดงการ Update class list

UpdateLabels()

For each attribute A used in a split **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 the class c

รูปที่ 4.9 ตัวอย่างอัลกอลิธึม การ Update Class list [Mehta et.at. 1996]

ซึ่งหลักการทำงานของการ Pruning แบบ Cross Validation ก็คือ จะทำการเช็คค่า Accuracy ของโหนดที่ประกอบด้วย Sub Tree ว่าเมื่อทำการเปรียบเทียบค่า Accuracy ระหว่างการมี Sub Tree กับการไม่มี Sub Tree แบบไหนให้ค่า Accuracy ดีกว่ากัน

ดังนั้นถ้าทำการพบว่าค่า Accuracy ของการไม่มี Sub Tree ดีกว่าการมี Sub Tree ก็จะทำการ Pruning ตัด Sub Tree นั้นทิ้งไป และก็จะทำการทำซ้ำไปเรื่อยโดยเริ่มจากด้านล่างของสู่ด้านบนของ Tree จนกระทั่งไม่มี Sub Treeที่จะก่อให้เกิดการ Pruning

4.4 การวัด Accuracy

ความแม่นยำ (Accuracy) เป็นสิ่งสำคัญที่ใช้เป็นตัววัดค่าความแม่นยำในการแบ่งกลุ่มข้อมูลในอนาคต หรือ กลุ่มข้อมูลที่ไม่ได้ผ่านการเรียนรู้มาก่อน ซึ่งจะเป็นข้อมูลที่ใช้ในการ Pruning และบอกถึงค่าความแม่นยำของโมเดลที่ได้

การวัดความแม่นยำ กรณีที่ถ้าบอกว่ามีความแม่นยำถึง 90 % ของค่าที่ตอบที่เป็น Positive sample อาจจะไม่ถูกต้องนัก ถ้าข้อมูลที่ใช่มีเพียง 3-4 % ของ Training data ดังนั้นวิธีที่น่าจะคำนวณความแม่นยำให้การแบ่งกลุ่มข้อมูลควรพิจารณาในทุกๆค่าที่เป็นไปได้ของข้อมูล

Sensitivity เป็นการวัดความสามารถในการจดจำรูปแบบของ Positive sample ในขณะที่ Specificity เป็นความสามารถในการจดจำรูปแบบของ Negative sample ซึ่งมีสูตรดังสมการข้างล่างนี้ โดยมี Precision เป็นการวัดเปอร์เซ็นของจำนวนข้อมูลที่คาดว่าจะเป็น Positive sample แล้วค่าจริงๆ ก็เป็น Positive sample

$$\text{Sensitivity} = t\_pos \;/\; Pos$$

$$\text{Specificity} = t\_neg \;/\; Neg$$

$$\text{Precision} = t\_pos \;/\; (\, t\_pos + f\_pos \,)$$

โดยที่ t_pos คือ จำนวนของค่าที่เป็นจริงของข้อมูล Positive

t_neg คือ จำนวนของค่าที่เป็นจริงของข้อมูล Negative

f_pos คือ จำนวนของค่าที่เป็นเท็จของข้อมูล Positive

Pos คือ จำนวนของข้อมูล Positive

Neg คือ จำนวนของข้อมูล Negative

ซึ่งจากการคำนวณหาค่า Sensitivity และ Specificity ก็จะทำให้เราสามารถทำการหาค่า Accuracy ได้โดย สมการการหาค่า Accuracy จะอยู่ในรูปของฟังก์ชั่นกับค่า Sensitivity และค่า Specificity ดังสมการข้างล่างนี้

$$\text{Accuracy} = \text{Sensitivity} * \text{Pos} / (\text{Pos} + \text{Neg}) + \text{Specificity} * \text{Neg} (\text{Pos} + \text{Neg})$$

โดยที่ Pos คือ จำนวนของข้อมูล Positive

Neg คือ จำนวนของข้อมูล Negative

นอกจากนี้เรายังสามารถทำการเปรียบเทียบโมเดลที่สร้างว่าโมเดลใดมีความถูกต้องในการสร้างโมเดล จากข้อมูลชุดเดียวกัน โดยการเปรียบเทียบจากค่า Accuracy ที่ได้ โดยโมเดลที่มีค่า Accuracy สูงก็แสดงว่ามีความถูกต้องในการ Predict ค่าของข้อมูลสูง

บทที่ 5
ระบบการวิเคราะห์หาสาเหตุการเสียของกระบวนการผลิต

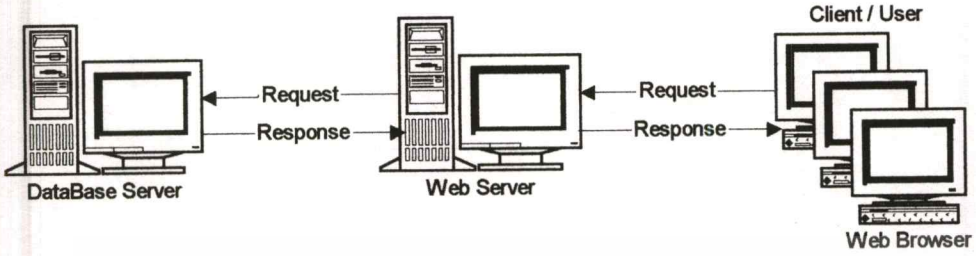
สำหรับในบทนี้จะเป็นรายละเอียดทั้งหมดของระบบการวิเคราะห์หาสาเหตุการเสียของกระบวนการผลิต ลักษณะการทำงานของระบบที่จะสามารถทำงานเพื่อให้ได้ผลออกมา และนำผลนั้นไปช่วยในการวิเคราะห์หาสาเหตุของเสียที่เกิดขึ้นในกระบวนการผลิตฮาร์ดิสก์ สำหรับข้อมูลที่จะใช้ภายในระบบนี้ก็เน้นที่ข้อมูลที่ได้จากการทดสอบชิ้นงานในกระบวนการผลิต ซึ่งในบทนี้จะกล่าวถึงลักษณะโครงสร้างการทำงานของระบบและรายละเอียดการทำงานของระบบที่ได้ออกแบบไว้ ภาพประกอบแสดงหน้าจอของระบบ การทดสอบระบบ และผลจากการทดสอบระบบนี้

5.1 เครื่องมือที่ใช้ในการพัฒนาระบบ

ในการพัฒนาระบบการวิเคราะห์หาสาเหตุการเสียของกระบวนการผลิตนี้ ได้ใช้ Microsoft Visual Basic สำหรับการพัฒนาระบบทั้งหมด เนื่องจาก Visual Basic เป็นเครื่องมือที่พัฒนาบนระบบปฏิบัติการ Windows และเป็นการพัฒนาโปรแกรมแบบ Visual Programming ทำให้ผู้พัฒนาสามารถเลือกออกแบบการทำงานของระบบได้อย่างสะดวกและรวดเร็ว อีกทั้งยังสามารถทำการเขียนโปรแกรมเพื่อการทำงานที่ติดต่อกับระบบข้อมูล Database จึงเหมาะกับระบบนี้ที่ต้องมีการติดต่อเอาข้อมูลจาก Database

5.2 แนวทางการออกแบบระบบ

เนื่องจากปัจจุบันแอปพลิเคชั่น (Application) ที่ใช้กันอยู่ทั่วไปหรือระบบที่ช่วยในการวิเคราะห์ข้อมูลต่างในกระบวนการผลิตในองค์กร ได้ทำออกมาในรูปแบบของ Application บน Web เนื่องจากมีผู้ใช้ในองค์กรเป็นจำนวนมาก และมีผู้ใช้งานจากหลากหลายสถานที่ ซึ่งก็มีข้อดีคือเมื่อมีการเปลี่ยนแปลง หรือปรับปรุงรูปแบบการทำงานต่างๆ ของระบบ ก็จะสามารถทำได้ โดยผู้ใช้บริการระบบไม่ต้องทำการปรับเปลี่ยนอะไรที่เครื่องของผู้ใช้ระบบเอง และเมื่อต้องการติดต่อหรือใช้งานข้อมูลจาก Database ผู้ใช้ไม่จำเป็นต้องทราบถึงลักษณะการทำงานที่ติดต่อกับ Database เลย โดยระบบนี้จะใช้ Web Server เป็นตัวกลางในการดึงข้อมูลจาก Database มาทำการประมวลผลที่เครื่องผู้ใช้ และ ผู้ใช้บริการมีเพียง Web Browser และต่อเข้ากับระบบเน็ทเวอร์ (Network) ของบริษัท ก็สามารถทำการเข้าใช้บริการของระบบได้ ซึ่งทำให้ง่ายและสะดวกต่อการใช้งาน



รูปที่ 5.1 เป็นลักษณะโครงสร้างของระบบที่ออกแบบไว้

5.3 โครงสร้างการทำงานของระบบ

Input ข้อมูล

สามารถทำการรับข้อมูลจาก Database ที่มีการกำหนดชื่อ Table และ Attribute ของแต่ละข้อมูลไว้แล้ว และสามารถทำการรับข้อมูลจาก Text file

ข้อมูลจาก Text file

โดยที่ Text file นี้ต้องมีรูปแบบ (format) ของข้อมูลคือ ต้องมี Header เป็นชื่อของ Attribute และใช้เครื่องหมาย ” , ” คั่นระหว่างข้อมูล และสามารถทำการรับข้อมูลที่มีลักษณะดังนี้ คือ

- Categorical แบบ Nominal ซึ่งลำดับของข้อมูลจะไม่มีผลกับค่าของข้อมูล
- Quantitative แบบ Continuous คือเป็นค่าที่เป็นเลขจำนวนจริง และลำดับของข้อมูลถือว่าสำคัญ

ข้อมูลจาก Database

มีหลักการการเลือกตัวแปรหรือพารามิตเตอร์ต่างๆของข้อมูลดังนี้

- เป็นข้อมูลที่เป็นตัวแปรที่น่าจะเป็นประโยชน์ในการแก้ไขปัญหาที่เกิดขึ้นในกระบวนการผลิต เช่น เครื่องมือที่ใช้
- เป็นค่าพารามิตเตอร์ที่ใช้ในการวัดคุณภาพของผลิตภัณฑ์ ว่าคุณภาพตามต้องการหรือไม่
- เป็นค่าพารามิตเตอร์ที่จากหลักการทาง ทฤษฎีคาดว่าน่าจะมีผลต่อค่าพารามิตเตอร์ที่เกี่ยวข้องกับกระบวนการผลิต และคุณภาพของผลิตภัณฑ์

การเตรียมข้อมูล (Data Processing)

ข้อมูลจาก Database ที่ใช้ภายในระบบนี้ ได้ผ่านขั้นตอนของการ Cleaning และ Transform ข้อมูลไว้แล้วซึ่งมีรายละเอียดดังนี้

Data integration

- ข้อมูลที่จะทำการ Mining มาจากหลายแหล่งของข้อมูล ซึ่งแบ่งกลุ่มได้ตามขั้นตอนการผลิต โดยเลือกที่จะนำข้อมูลทั้งหมด มารวมกันสร้างเป็น coherent data ใหม่ โดยมีการเลือกข้อมูลบางพารามิตเตอร์เท่านั้น เนื่องจากแต่จะแหล่งของข้อมูลก็ประกอบไปด้วยข้อมูลมากมายหลายพารามิตเตอร์ โดยการเลือกข้อมูล ก็อาศัยหลักการพิจารณาการเลือกตัวแปรดังหัวข้อที่นำเสนอก่อนหน้านี้

Data cleaning

- Missing Values ข้อมูลที่มีค่าบางค่าที่หายไป (missing value) เลือกที่จะทำการตัดข้อมูลชุดนั้นออกไปเลย เนื่องจากมีข้อมูลอยู่เป็นจำนวนมาก เปอร์เซ็นต์ของค่า missing value จะน้อยมากเมื่อเทียบกับจำนวนข้อมูลทั้งหมด
- Noisy data มีเนื่องจากข้อมูลที่นำมานี้ ได้ทำการสรุปรวมข้อมูลให้ออกมาอยู่ในรูปของ Chunk แล้ว จึงเป็นการช่วยกำจัด Noisy data ออกไปจากข้อมูลด้วยอีกทางหนึ่ง เพื่อให้ข้อมูลมีความถูกต้องมากยิ่งขึ้น

Data Transformation

- เนื่องจากข้อมูลส่วนใหญ่เป็นข้อมูลที่เป็นตัวเลขเพราะเป็นค่าที่เกิดจากการทดสอบคุณสมบัติทางไฟฟ้า และส่วนใหญ่ก็จะเป็นเลข ทศนิยมขนาด 4-5 ตำแหน่ง ซึ่งเป็นข้อมูลที่ค่อนข้างละเอียด ถ้านำมาทำการคำนวณต้องมีการคำนวณเป็นจำนวนมาก ดังนั้นจะเลือกใช้เทคนิคการทำ Smoothing โดยเลือกทำการปัดเลขทศนิยม ให้เหลือเพียง 1-2 ตำแหน่ง โดยจะพิจารณาค่าของข้อมูลทั้งหมดที่มีอยู่

Data Reduction

- เนื่องจากข้อมูลการผลิตที่จะนำมาใช้นั้น มีข้อมูลเป็นจำนวนมาก เนื่องจากการผลิตในแต่ละวันมีจำนวนมาก จึงได้ทำการลดจำนวนของข้อมูลให้อยู่ในรูปของ Chunk ของ Wafer ก่อนที่จะนำข้อมูลมาใช้

การทำงานของระบบ

สำหรับ Classification Tree ของระบบนี้สามารถแบ่ง ลักษณะการทำงานออกเป็น 2 รูปแบบคือ การทำ Classification และ การทำ Prediction ซึ่งลักษณะของข้อมูลที่จะมาใช้งานก็จะต่างกันไป

Classification

การทำ Classification จะใช้ข้อมูลในอดีตที่เราทราบค่าของ Class ที่เราสนใจแล้ว ซึ่งในระบบนี้ Class ที่สนใจก็คือลักษณะการ Pass และ Fail ของชิ้นงาน และนำข้อมูลที่ทราบค่าแล้วมาทำการสร้างเป็นโมเดลของ Classification Tree โดยใช้วิธีการของอัลกอลิธึม SLIQ

การรับข้อมูลสำหรับการทำ Classification Tree ก็สามารถเลือกได้ ทั้งแบบใช้ข้อมูลจาก Database ซึ่งจะมีการให้เลือก ผลิตภัณฑ์ที่สนใจ วันที่ และ ขั้นตอนกระบวนการผลิตที่สนใจ หรือจะทำการนำข้อมูลที่มีอยู่ซึ่งเก็บอยู่ในรูปแบบของ Text file มาใช้ในการสร้างโมเดลก็ได้ โดย Tree ที่ไม่ได้ใช้ข้อมูลจาก Database

ผู้ใช้สามารถทำการกำหนดค่าพารามิตเตอร์ต่างๆ ในการสร้าง Treeซึ่งได้แก่

- จำนวน Level สูงสุดของ Tree ที่จะทำการสร้าง
- จำนวนของข้อมูลที่น้อยที่สุดที่ต้องมีในแต่ละโหนด
- ทำการระบุ Class Attribute ของโมเดลที่สนใจ คือข้อมูลที่เป็น Pass และ Fail
- ทำการระบุข้อมูลของ Class 1 และ 2
- จำนวน% ข้อมูลสำหรับการ Training data

สำหรับข้อมูลจาก Database จะมีการกำหนดค่าพารามิตเตอร์ที่สนใจของแต่ละ Product ไว้แล้วเพื่อให้การทำงานเป็นไปอย่างรวดเร็ว สะดวกและถูกต้องยิ่งขึ้น ดังนั้นผู้ใช้จึงเพียงทำการเลือกว่าจะใช้ข้อมูลจาก Database ซึ่งก็จะมีอีกหน้าจอหนึ่งมาให้ผู้ใช้เลือก ประกอบด้วย ชื่อผลิตภัณฑ์ (Product) วันที่ (Date) และ Operation ของฮาร์ดิสก์ที่สนใจ ระบบก็จะแสดงค่า Attribute ต่างๆ ที่เตรียมไว้ให้ และเมื่อทำการตอบตกลง ระบบก็จะทำการนำข้อมูลตามข้อกำหนดที่เลือกไว้มาแสดงผล ที่หน้าจอเดิมของ Classification

เมื่อผู้ใช้ระบบสั่งให้ระบบทำการสร้างโมเดลแล้ว ผลที่ได้จากการทำ Classification ก็จะแสดงผลใน 2 ส่วนคือ ในส่วนของ Classification Tree และ Classification Rule โดยระบบจะสามารถทำการ Save file โมเดลที่สร้างขึ้นไว้ได้ โดยจะเป็น File ที่อยู่ในรูปแบบของ *.tree

สำหรับรายละเอียดของหน้าจอจะกล่าวในหัวข้อถัดไป

Prediction

เนื่องจาก Classification tree ก็เป็นหนึ่งในอัลกอลิธึมในเทคนิคของ Predictive Model ดังนั้นเราจึงสามารถนำโมเดลที่สร้างจากการทำ Classification มาใช้ในการ Predict ข้อมูลที่สนใจได้ โดยสำหรับในระบบนี้ การ Prediction ก็จะเป็นการนำเอาที่ข้อมูลที่ยังไม่ทราบค่าการ Pass และ Fail มาทำการ Prediction การ Pass และ Fail โดยจำเป็นต้องใช้ข้อมูลของโมเดลที่สร้างมาจากขั้นตอนของ Classification ซึ่งลักษณะการ Prediction ของระบบนี้จะเป็นลักษณะการ Prediction ของกลุ่มข้อมูลจำนวนหนึ่ง โดยจะมีการแสดงผลของแต่ละข้อมูลที่ได้ทำการ Prediction ไว้ต่อท้ายจากคอลัมน์สุดท้ายของชุดข้อมูลที่ใช้ในการ Predict และจะนำผลที่ได้จากการ Prediction ทั้งชุดข้อมูลนี้มาทำการสรุปผลโดยแสดงในรูปแบบของผลจากการคำนวณทางคณิตศาสตร์เพื่อให้ได้เป็นเปอร์เซ็นต์การทำนายของแต่ละ Class ที่ได้ของชุดข้อมูล

การทำงานของ Prediction ก็ต้องประกอบด้วยโมเดลของ Tree ที่จะใช้ในการ Predict และชุดของข้อมูลที่จะทำการ Predict โดย ข้อมูลทั้งสองชนิดนี้ก็สามารถเลือกได้ว่าจะนำข้อมูลจาก Database มาใช้หรือนำไฟล์ข้อมูลที่มีอยู่มาใช้ แต่ข้อมูลที่จะมาทำการ Prediction ต้องมีโครงสร้างเหมือนข้อมูลที่ใช้ในการสร้างโมเดลที่ใช้ในการ Prediction เพื่อความถูกต้องในการ Prediction

ลักษณะการเลือกข้อมูลจาก Database ที่จะใช้ในการ Prediction ก็จะคล้ายกับส่วนการเลือกข้อมูลที่จะใช้ในการทำ Classification แต่เป็นข้อมูลที่ยังไม่ได้ทำการทดสอบในขั้นตอน HGA จริงจึงจะไม่ Operation HGA ให้ทำการเลือก

Out put ที่ได้จากการทำงาน

สำหรับ Out put ที่ได้จากระบบก็แบ่งออกตามลักษณะการทำงานของระบบคือ ในส่วนของ Classification ก็จะได้ผลออกมาเป็นโมเดลของ Classification Tree ซึ่งสามารถทำการจัดเก็บโมเดลที่ได้ในรูปของ Binary file ซึ่งมีขนาดเล็กได้

ในส่วนของ Prediction ก็จะได้ออกมาในรูปแบบของ Prediction Class ของแต่ละข้อมูลและผลสรุปรวมของชุดข้อมูล

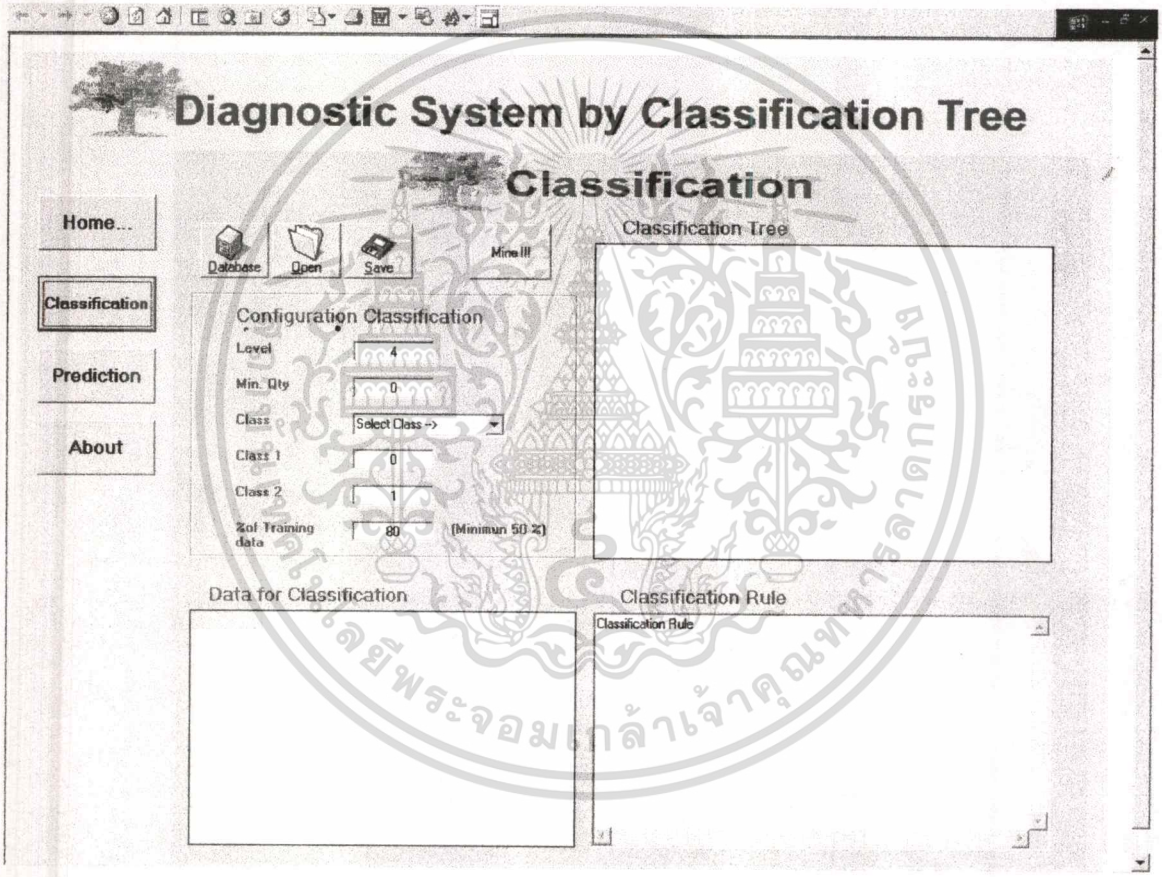
5.4 รายละเอียดของหน้าจอการทำงาน

หน้าจอการทำงานของระบบการวิเคราะห์หาสาเหตุการเสียของกระบวนการผลิต ประกอบไปด้วยทั้งสิ้น 7 หน้าจอดังนี้

1. หน้าจอแรก ของระบบการวิเคราะห์หาสาเหตุการเสียของกระบวนการผลิต ซึ่งจากหน้าจอนี้ ก็สามารถเลือกได้ว่าจะเข้ามาใช้งานในส่วนของ Classification หรือ Prediction ตัวอย่างหน้าจอแสดงในรูปที่ 5.2

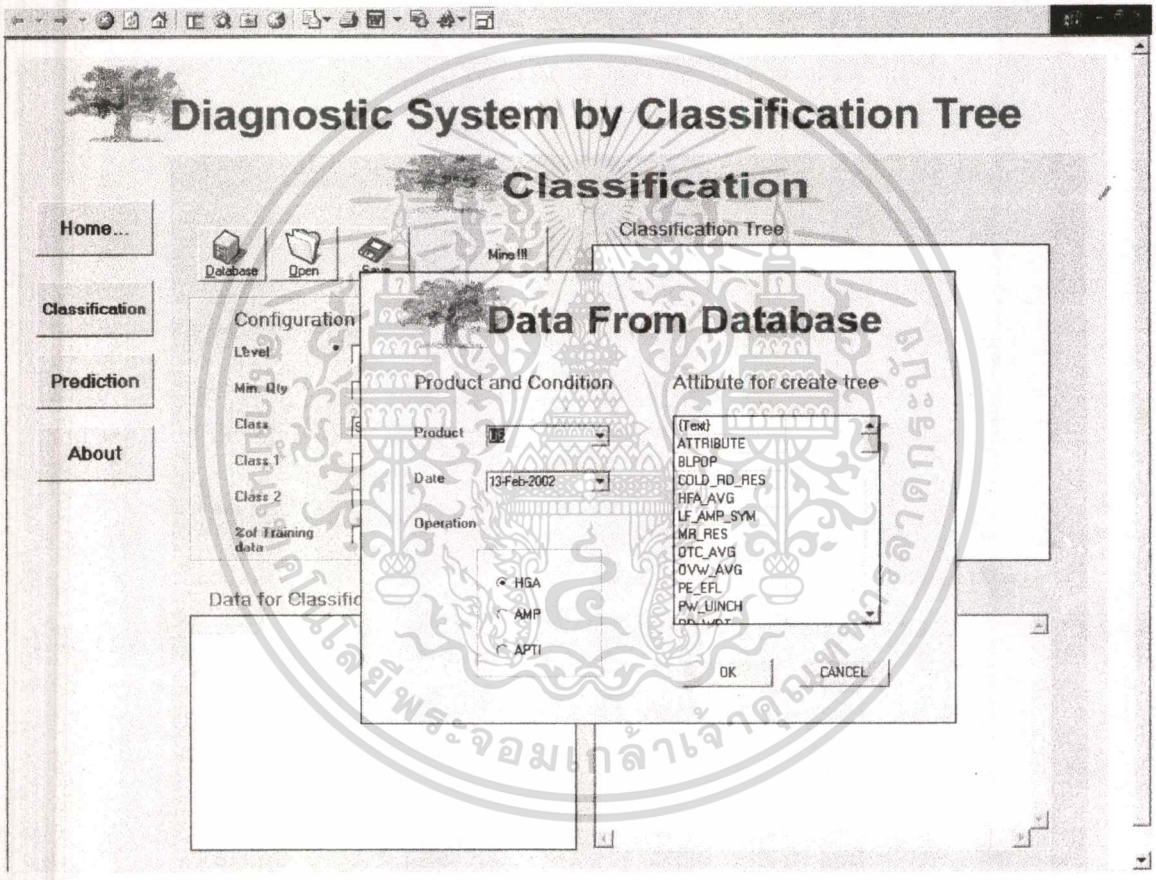
รูปที่ 5.2 หน้าจอแรกของระบบ

2. หน้าจอที่สองเป็นหน้าจอของส่วนการทำงานของ Classification ซึ่งในหน้าจอนี้ ก็จะสามารถทำการ Setup ค่า Configuration ต่างๆ ที่จะใช้ในการทำ Classification และสามารถทำการโหลดไฟล์ข้อมูลมาแสดงไว้ที่ส่วนของ Data for Classification ซึ่งจะเป็นข้อมูลที่จะนำมาใช้ในการทำ Classification หรือจะทำการเลือกข้อมูลจาก Database และผลโมเดลที่ได้จะแสดงในส่วนของ Classification Tree และ Classification Rule ตัวอย่างหน้าจอแสดงในรูปที่ 5.3



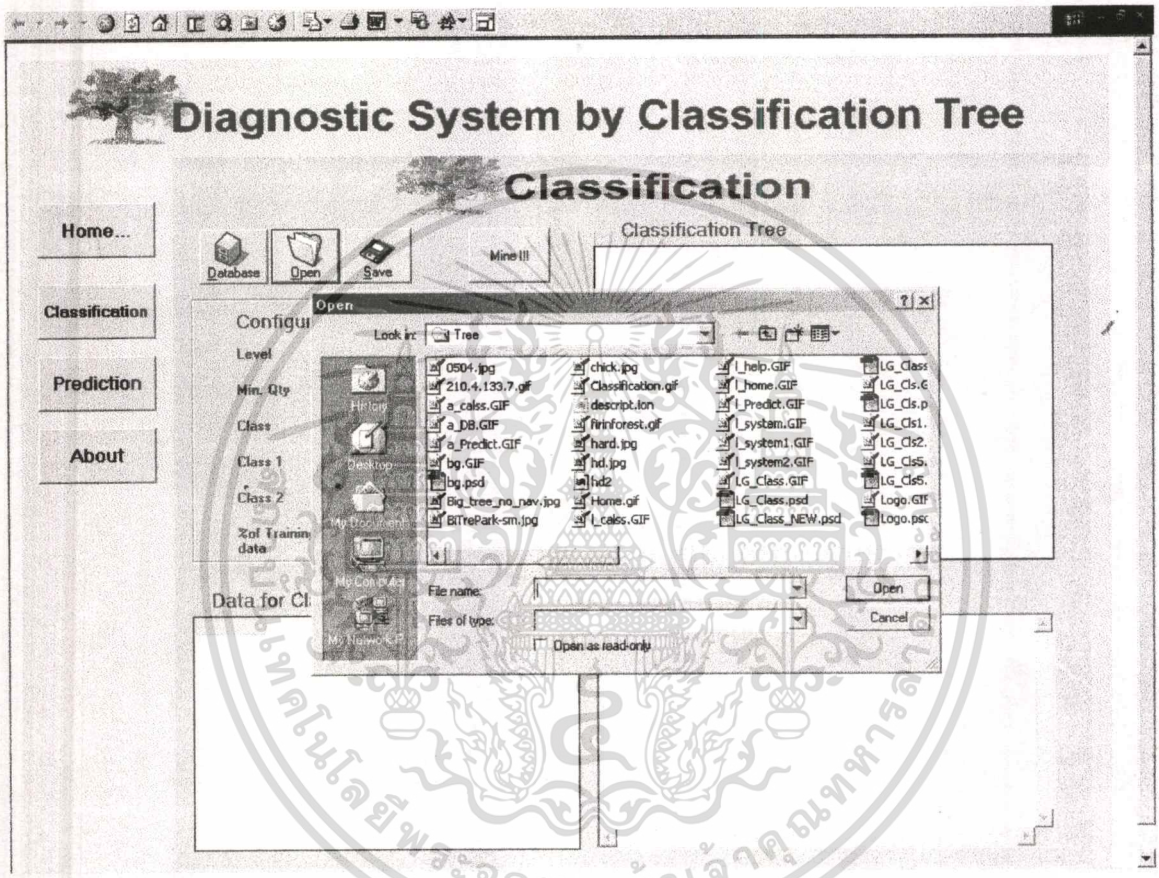
รูปที่ 5.3 หน้าจอแสดงส่วนการทำงานของ Classification

3. หน้าจอการเลือกใช้ข้อมูลจาก Database จากขั้นตอนการทำ Classification ซึ่งจะมีหน้าจอแสดงข้อมูลจาก Database ที่มีอยู่ขึ้นมาให้ทำการเลือก ซึ่งเมื่อทำการเลือกแล้วจะมี Attribute ที่ได้เตรียมไว้ของแต่ละผลิตภัณฑ์แสดงให้ดูในส่วนของ Attribute for Create Tree และเมื่อทำการตกลง ระบบก็จะไปทำการดึงข้อมูลจาก Database มาแสดงผลในส่วน Data for Classification ตัวอย่างหน้าจอแสดงในรูปที่ 5.4



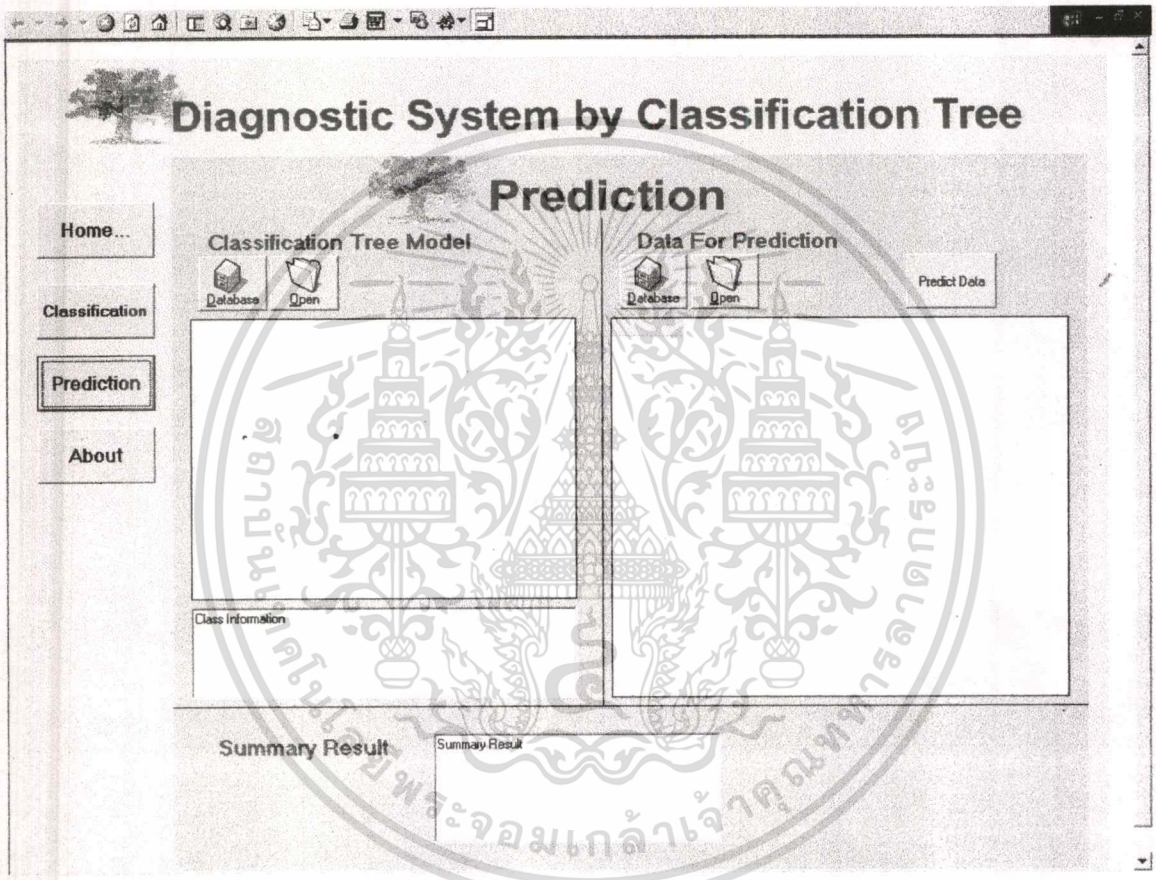
รูปที่ 5.4 หน้าจอแสดงส่วนการเลือกข้อมูลจาก Database เพื่อใช้ในการทำงานของ Classification

4. หน้าจอในส่วนของการนำข้อมูลที่ใช้ใน Classification จากไฟล์ข้อมูลที่มีอยู่ เมื่อเลือกทำการเปิดไฟล์ จะมีหน้าจอการทำงานขึ้นหน้าให้ทำการเลือกไฟล์ข้อมูล ซึ่งเป็นลักษณะหน้าจอการเปิดไฟล์ของระบบปฏิบัติการ Windows โดยจะทำการรับไฟล์ข้อมูลที่มีรูปแบบตามที่กำหนดไว้แล้ว ตัวอย่างหน้าจอแสดงในรูปที่ 5.5



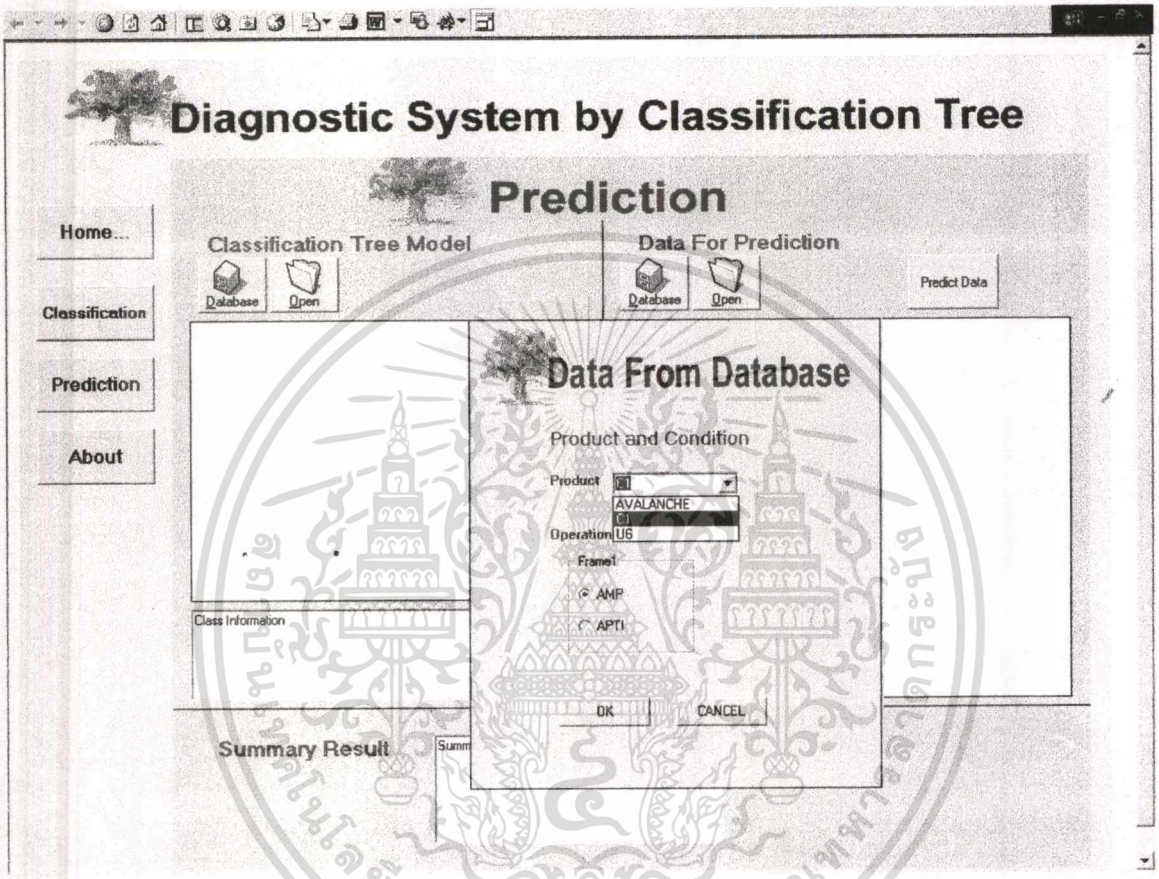
รูปที่ 5.5 หน้าจอแสดงส่วนการเลือกข้อมูลจากการเปิด File เพื่อใช้ในการทำงานของ Classification

5. หน้าจอในส่วนของ Prediction จะเป็นลักษณะการ Predict กลุ่มของข้อมูล การทำงานจะประกอบด้วยข้อมูลที่จะใช้ 2 ส่วนคือ Tree Model และ Data for Prediction และจะแสดงผลลัพธ์ออกมาในรูปแบบของ Prediction ที่ได้จากชุดข้อมูล และมีผลสรุปการ Prediction ที่ได้จากชุดข้อมูล ตัวอย่างหน้าจอแสดงในรูปที่ 5.6



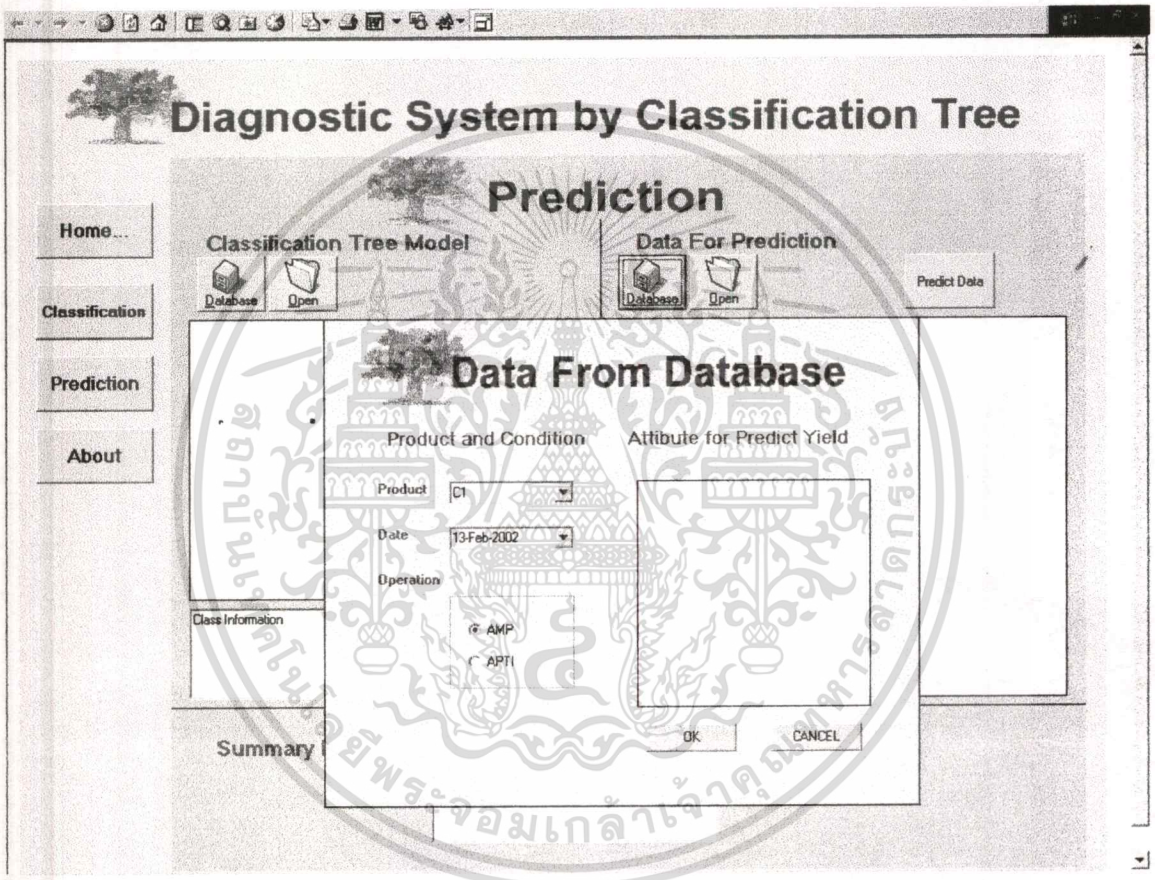
รูปที่ 5.6 หน้าจอแสดงส่วนการทำงานของ Prediction

6. หน้าจอแสดงการเลือก Tree Model จาก Database ซึ่งต้องทำการเลือกให้ตรงกับข้อมูลที่จะนำมาใช้ในการ Prediction ตัวอย่างหน้าจอแสดงในรูปที่ 5.7



รูปที่ 5.7 หน้าจอหน้าจอแสดงส่วนการเลือกโมดลจาก Database

7. หน้าจอสุดท้ายเป็นหน้าจอ สำหรับการเลือกข้อมูลในการทำ Prediction จาก Database ซึ่งเราก็สามารถทำการเลือกข้อมูลจากขั้นตอนการผลิตในส่วนของ Slider ซึ่งเป็นขั้นตอนก่อนหน้าของ HGA จาก Database โดยลักษณะการเลือกจะเหมือนกับ การเลือกข้อมูลที่ใช้ในการทำ Classification ตัวอย่างหน้าจอแสดงในรูปที่ 5.8



รูปที่ 5.8 หน้าจอหน้าจอแสดงส่วนการเลือกข้อมูลจาก Database เพื่อใช้ในการทำงานของ Prediction

5.5 การทดสอบทำงานของระบบ

ข้อมูลชุดทดสอบ

เพื่อเป็นการตรวจสอบการทำงานของระบบจึงมีการทดสอบการทำงานของระบบ โดยใช้ข้อมูลของผลิตภัณฑ์ Ultra6 ใน Operation AMP ของ Slider ประกอบไปด้วยข้อมูล Attribute จำนวน 12 ซึ่งมีตัวอย่างของชุดข้อมูลและรายละเอียดดังรูปที่ 5.9

ATTRIBUTE	VALUE		Description
ET_STATUS	1	0	ค่าที่สนใจ
DATE_TIME	4-Feb-02	3-Feb-02	วันที่ทำการทดสอบ
MAC	MACH_518	MACH_592	หมายเลขเครื่อง
SNOISE	0	10	ค่าผลของทดสอบ SNOISE
F1	950	1100	ค่าผลของทดสอบ F1
VNOISE	5	4	ค่าผลของทดสอบ VNOISE
GNOISE	1	6	ค่าผลของทดสอบ GNOISE
PROP_SYMMETRY	2	1	ค่าผลของทดสอบ PROP_SYMMETRY
COLD_RESISTANCE	62	54	ค่าผลของทดสอบ COLD RESISTANCE
RESISTANCE	65	57	ค่าผลของทดสอบ RESISTANCE
HYSTERESIS	1	1	ค่าผลของทดสอบ HYSTERESIS
BNOISE	0	0	ค่าผลของทดสอบ BNOISE
SNR	43	45	ค่าผลของทดสอบ SNR
WAFER	GAC	PXC	ชื่อ WAFER ของข้อมูล

รูปที่ 5.9 ตัวอย่างข้อมูลที่ใช้ทดสอบระบบในส่วน Classification

การทดสอบ

จากชุดข้อมูลที่ทำการทดสอบได้ทำการแบ่งออกเป็น 2 ส่วนคือ ส่วนหนึ่งใช้ข้อมูลจำนวน 80 % โดยทำการเลือกแบบสุ่ม มาใช้ในการทดสอบการทำ Classification และข้อมูลที่เหลืออีก 20 % นำไปใช้ในการทำ Prediction

ซึ่งข้อมูลที่นำไปใช้ในการทำ Classification ก็ทำการกำหนดค่าให้ทำการสร้างโมเดลจากข้อมูล 80% ของข้อมูลที่มีอยู่ และอีก 20% ใช้ในการ Test Model

และเมื่อได้โมเดลแล้วก็จะทำการนำโมเดลที่ได้ไปทำการ Prediction ข้อมูลที่เหลืออีก 20% เพื่อตรวจดูความถูกต้องในการ Prediction

ข้อมูลชุดที่ 1

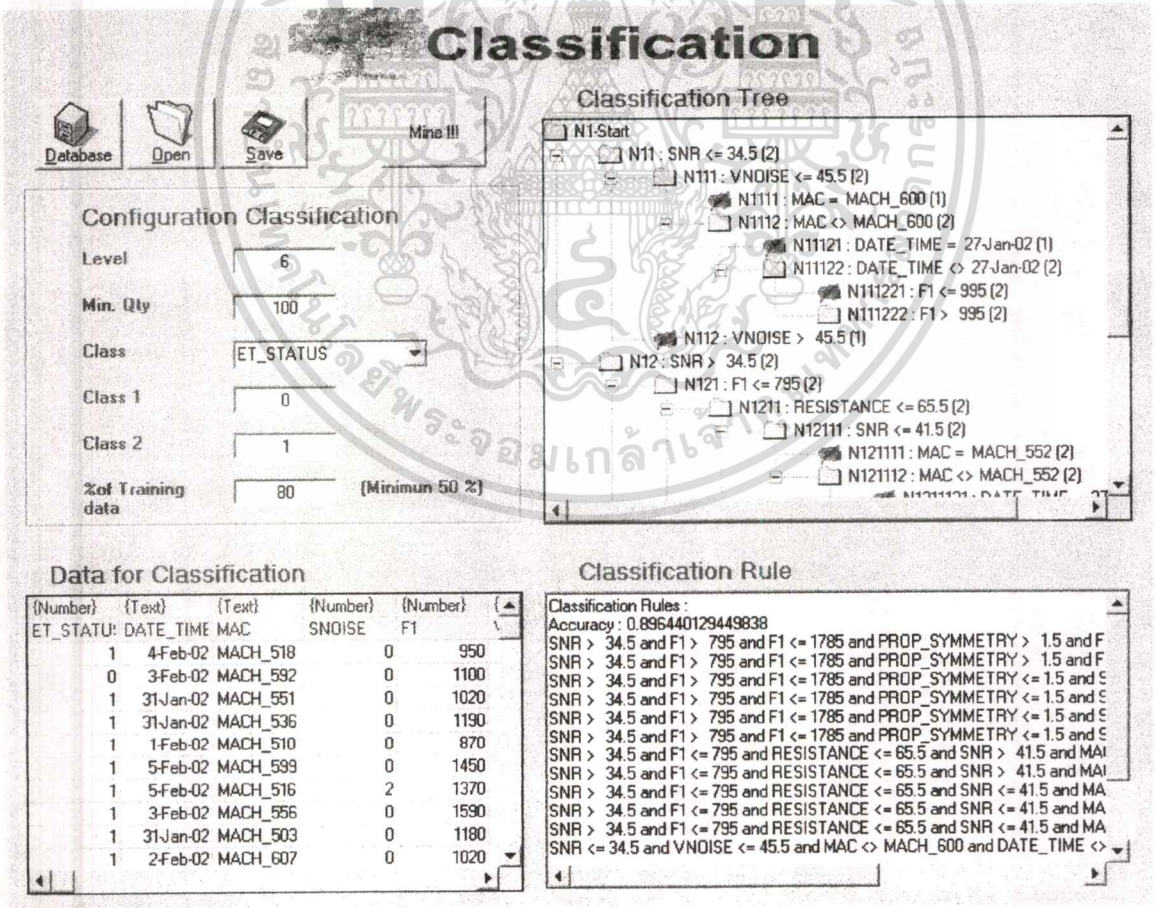
ข้อมูลทั้งหมด 5796 HGA ในช่วงวันที่ 8 กุมภาพันธ์ แบ่งออกเป็นข้อมูลที่ Pass จำนวน 5086 ข้อมูล ข้อมูลที่ Fail จำนวน 710 ข้อมูล ใช้ข้อมูลทดสอบClassification จำนวน 4635 ข้อมูล

และกำหนดให้ Level ของที่ไม่เกิน 6 Level และให้จำนวนข้อมูลอย่างน้อย 100 ข้อมูลถึงจะทำการ Split ข้อมูลได้

ผลการทดสอบข้อมูลชุดที่1

จากการทดสอบระบบได้ผลออกมาดังนี้ ในส่วน Classification สามารถทำการสร้าง Classification Tree ที่มีค่า Accuracy = 89% ซึ่งประกอบไปด้วย Rule ทั้งหมด 16 rule

สำหรับในส่วนของการ Prediction พบว่า % ความถูกต้องในการ Prediction อยู่ที่ โดยสามารถทำการ Prediction ค่า Pass ได้ถูกต้องถึง 1087 ข้อมูลที่เป็น Pass ทั้งหมด 1012 ข้อมูล ซึ่งเมื่อทำการสรุปแล้ว Yield ของกลุ่มข้อมูลที่ Predict ได้คือ 93.63% ในขณะที่ข้อมูลจริง Yield ที่ได้คือ 88% ซึ่งผิดพลาดไปประมาณ 5% ซึ่งก็เนื่องจาก โมเดลนี้มีค่า Accuracy อยู่ที่ 89% ดังนั้นการที่จะทำการ Prediction ได้ถูกต้องก็ต้องการโมเดลที่มีค่า Accuracy ที่มากกว่านี้ ตัวอย่างผลการทดสอบระบบแสดงในรูปที่ 5.10- 5.11



รูปที่ 5.10 หน้าจอแสดงผลที่ได้จากการทดสอบระบบในส่วน Classification

Classification Rules :

Accuracy : 89%

SNR > 34.5 and F1 > 795 and F1 <= 1785 and PROP_SYMMETRY > 1.5 and F1 > 935 and MAC <> MACH_561 --> 1

SNR > 34.5 and F1 > 795 and F1 <= 1785 and PROP_SYMMETRY > 1.5 and F1 > 935 and MAC = MACH_561 --> 0

SNR > 34.5 and F1 > 795 and F1 <= 1785 and PROP_SYMMETRY <= 1.5 and SNR > 41.5 and RESISTANCE > 76.5 --> 1

SNR > 34.5 and F1 > 795 and F1 <= 1785 and PROP_SYMMETRY <= 1.5 and SNR > 41.5 and RESISTANCE <= 76.5 --> 1

SNR > 34.5 and F1 > 795 and F1 <= 1785 and PROP_SYMMETRY <= 1.5 and SNR <= 41.5 and MAC <> MACH_625 --> 1

SNR > 34.5 and F1 > 795 and F1 <= 1785 and PROP_SYMMETRY <= 1.5 and SNR <= 41.5 and MAC = MACH_625 --> 1

SNR > 34.5 and F1 <= 795 and RESISTANCE <= 65.5 and SNR > 41.5 and MAC <> MACH_613 and MAC <> MACH_626 --> 1

SNR > 34.5 and F1 <= 795 and RESISTANCE <= 65.5 and SNR > 41.5 and MAC <> MACH_613 and MAC = MACH_626 --> 0

SNR > 34.5 and F1 <= 795 and RESISTANCE <= 65.5 and SNR <= 41.5 and MAC <> MACH_552 and DATE_TIME <> 27-Jan-02 --> 1

SNR > 34.5 and F1 <= 795 and RESISTANCE <= 65.5 and SNR <= 41.5 and MAC <> MACH_552 and DATE_TIME = 27-Jan-02 --> 0

SNR > 34.5 and F1 <= 795 and RESISTANCE <= 65.5 and SNR <= 41.5 and MAC = MACH_552 --> 1

SNR <= 34.5 and VNOISE <= 45.5 and MAC <> MACH_600 and DATE_TIME <> 27-Jan-02 and F1 <= 995 --> 1

SNR <= 34.5 and VNOISE <= 45.5 and MAC <> MACH_600 and DATE_TIME = 27-Jan-02 --> 0

SNR > 34.5 and F1 <= 795 and RESISTANCE > 65.5 --> 1

SNR <= 34.5 and VNOISE <= 45.5 and MAC = MACH_600 --> 0

SNR <= 34.5 and VNOISE > 45.5 --> 0

รูปที่ 5.11 ผล Classification Rule ที่ได้จากการทำ Classification

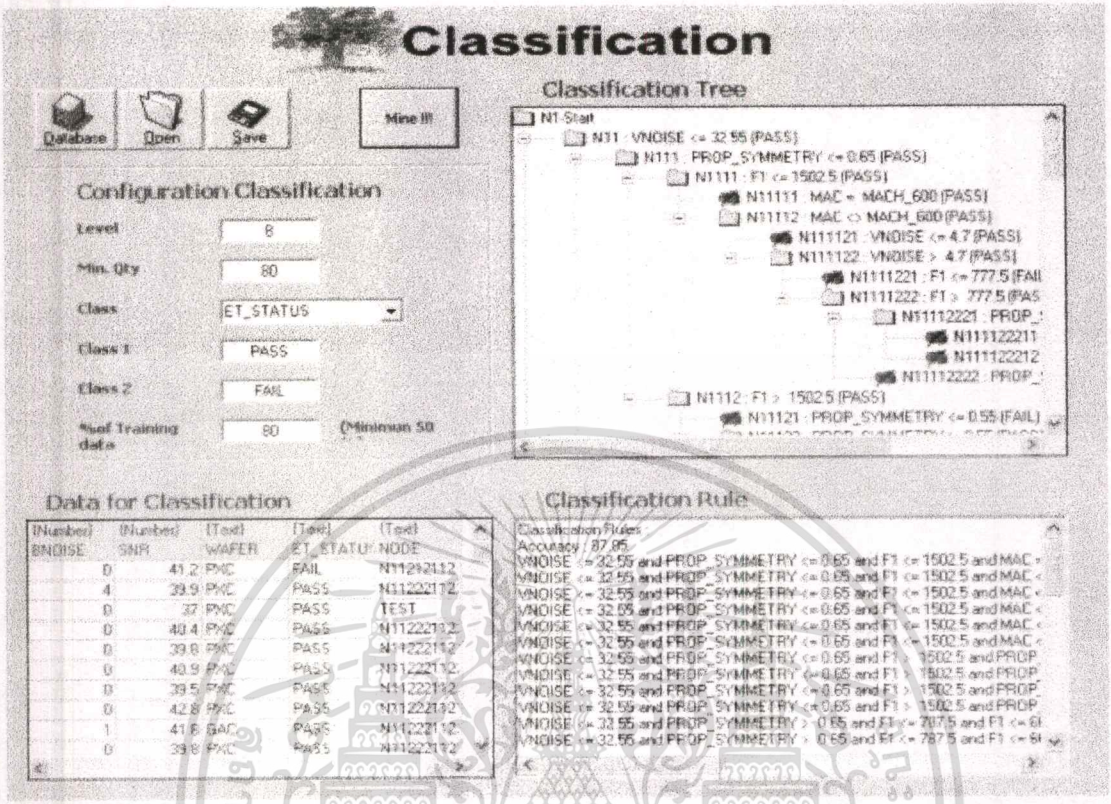
ข้อมูลชุดที่ 2

ทำการทดสอบข้อมูลจำนวนมาก โดยมีข้อมูลทั้งหมด 26,685 HGA โดยใช้ข้อมูลทดสอบ Classification จำนวน 20,000 ข้อมูล และกำหนดให้ Level ของที่ไม่เกิน 8 Level และให้จำนวนข้อมูลอย่างน้อย 80 ข้อมูลถึงจะทำการ Split ข้อมูลได้

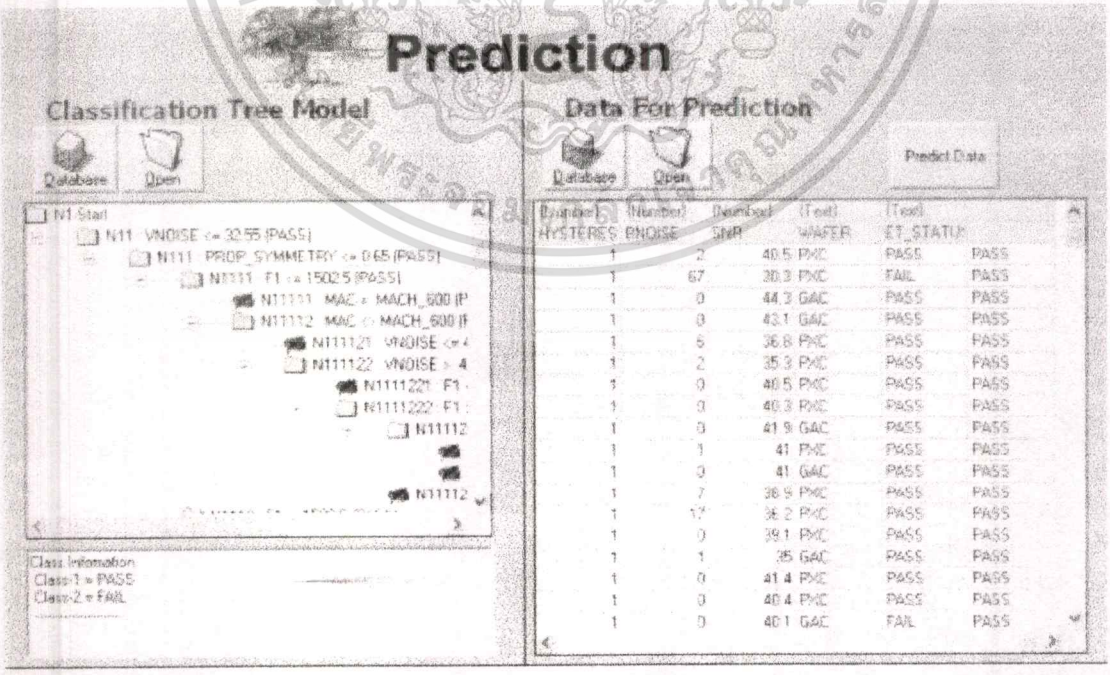
ผลการทดสอบข้อมูลชุดที่2

ในส่วน Classification สามารถทำการสร้าง Classification Tree ที่มีค่า Accuracy = 87.85% ซึ่งประกอบไปด้วย Rule ทั้งหมด 44 rule

สำหรับในส่วนของการ Prediction พบว่า % ความถูกต้องในการ Prediction อยู่ในช่วงของความผิดพลาดของโมเดลที่ได้ ซึ่งก็ได้ผลในลักษณะเดียวกับข้อมูลชุดที่ 1 ตัวอย่างผลการทดสอบระบบการ Prediction แสดงในรูปที่ 5.12-5.13



รูปที่ 5.12 แสดงผลที่ได้จากการทดสอบระบบในส่วน Classification

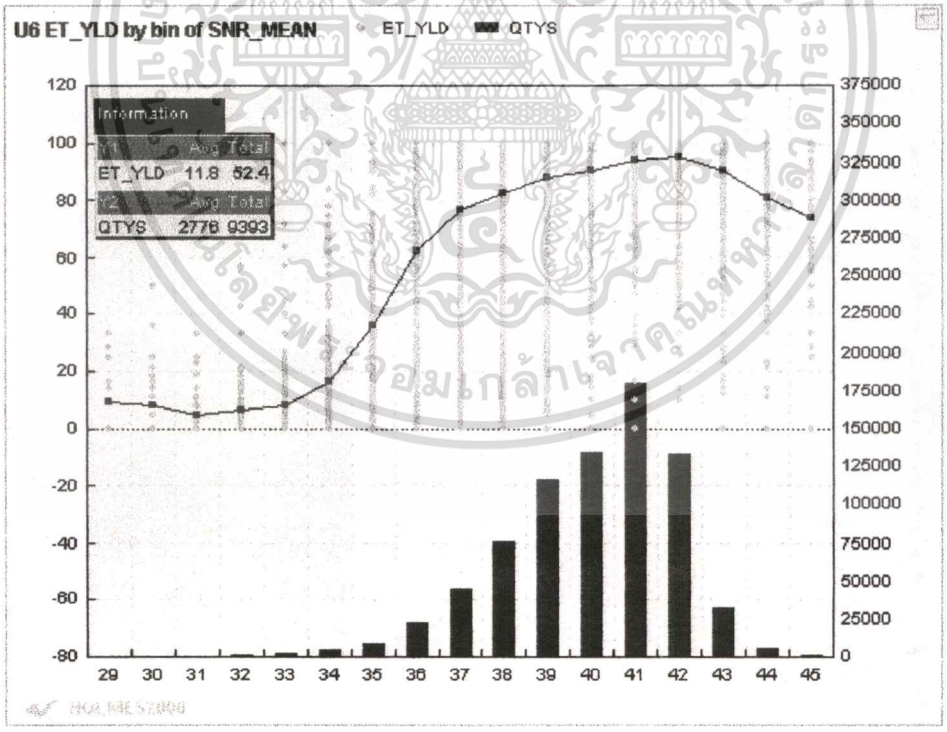


รูปที่ 5.13 ผล Prediction ที่ได้จากการใช้โมเดลจากการทำ Classification

การตรวจสอบทำงานของระบบ

จากผลของการทำ Mining ที่ได้ผลลัพธ์มาแล้วนั้น ก็จะทำการประเมินถึงผลที่ได้จากการทำ Mining อีกครั้งว่าถูกต้องหรือไม่ จากข้อมูลชุดแรกพบว่าค่าที่เลือกเป็นค่าแรกในการแตกกิ่งคือค่า SNR mean ซึ่งเป็นค่าของอัตราส่วนของ Signal/Noise ซึ่งถึงแม้ว่าเป็นพารามิตเตอร์ที่ไม่มี Spec แต่ก็ที่มักมีปัญหาถ้าพบว่าค่าน้อยมากๆสูง ดังตัวอย่างการวิเคราะห์ข้อมูลหาความสัมพันธ์ระหว่าง Yield กับค่า SNR ในรูปที่ 5.14 แสดงว่าค่า SNR ที่น้อยมากๆ จะส่งผลต่อค่า Yield ทำให้พอจะสรุปได้ว่าผลของการทำ Mining มีความน่าเชื่อถือในระดับหนึ่ง ซึ่งก็เป็นแนวทางอย่างหนึ่งในการทำการปรับปรุงประสิทธิภาพของกระบวนการผลิตว่าควรจะทำการเริ่มศึกษาที่ค่า SNR นี้ และก็ถือว่าสามารถนำมาใช้ให้บรรลุถึงจุดประสงค์ที่ต้องการได้ในระดับหนึ่ง

สำหรับการ Prediction ที่ได้ก็สามารถทำการ Predict ค่าได้ผลถูกต้องในระดับที่น่าพึงพอใจ แต่ก็ยังคงต้องการโมเดลที่จะใช้ที่มีค่า Accuracy ที่ดีที่สุด



รูปที่ 5.14 ตัวอย่างการวิเคราะห์ข้อมูลความสัมพันธ์ระหว่าง SNR กับ Yield

5.6 สรุปผลการทำงานของระบบ

จากการทดสอบการทำงานและตรวจสอบผลการทำงานของระบบพบว่า ระบบนี้สามารถทำงานได้เป็นที่น่าพอใจ สามารถทำการสร้างโมเดลของข้อมูลที่มีค่า Accuracy ค่อนข้างสูง และสามารถนำรองรับการทำงานของข้อมูลจำนวนมากๆได้ โดยเวลาที่ใช้ในการประมวลผลค่อนข้างเร็ว แต่ก็ยังมีปัญหาบางอย่างที่เกิดขึ้นจากการงาน ระบบ เช่น กรณีที่ข้อมูลจำนวนมากๆ และข้อมูลส่วนใหญ่นั้น อยู่ใน Class ใด Class หนึ่งและทำการเลือกกำหนดให้ Level ในการสร้าง Tree น้อยเกินไป ก็จะทำให้ ไม่สามารถทำการสร้างจนกระทั่งเห็นลักษณะของข้อมูล Class ที่มีจำนวนน้อยนั้น แต่การที่เลือกให้ทำการสร้าง Tree ที่มีระดับความลึกมากเกินไป ผลที่ได้ก็อาจไม่สามารถนำไปใช้งานเพื่อแก้ไขปัญหาได้ อีกทั้งระบบก็ต้องการหน่วยความจำเพิ่มขึ้นในการทำงาน ซึ่งก็จะส่งผลต่อเวลาในการสร้างโมเดลด้วย

และการทำการ Prediction ผลของการ Prediction ก็ขึ้นอยู่กับโมเดลที่ใช้ว่าน่าเชื่อถือเพียงไร และข้อมูลที่นำมาทำการ Prediction ก็ต้องเป็นข้อมูลที่มีโครงสร้างเหมือนข้อมูลที่ใช้ในการสร้างโมเดล

บทที่ 6
สรุปผล

ในบทนี้จะทำการประเมินถึงระบบที่ได้ทำการพัฒนา ในทั้งจุดดี และจุดด้อย ปัญหาต่างๆ ที่เกิดขึ้นในการพัฒนาระบบ รวมถึงข้อเสนอแนะเกี่ยวกับระบบ ที่จะสามารถช่วยเพิ่มประสิทธิภาพของระบบให้ดียิ่งขึ้น

6.1 สรุปผลการศึกษา

ระบบการวิเคราะห์หาสาเหตุการเสียของกระบวนการผลิตโดยใช้ Classification Tree ที่ได้ทำการพัฒนาขึ้นมา ระบบนี้ เป็นระบบที่สามารถนำมาใช้ช่วย ในการวิเคราะห์หาสาเหตุการเสียของกระบวนการผลิตได้จริง โดยใช้เวลาในการหาโมเดลของข้อมูลน้อยมากเมื่อเทียบกับรูปแบบการวิเคราะห์หาสาตุการเสียของกระบวนการผลิต ที่ใช้กับอยู่ในปัจจุบัน เนื่องจากอัลกอลิธึมที่เลือกใช้เป็นอัลกอลิธึมที่เหมาะกับข้อมูลจำนวนมาก และ ยังมีกลไกลการทำงาน ที่จะสามารถเลือกค่า Attribute แบบตัวเลขได้อย่างรวดเร็ว จึงทำให้ระบบซึ่งข้อมูลส่วนใหญ่เป็นข้อมูลตัวเลขที่เกิดจากการทดสอบการทำงานของชิ้นงาน สามารถทำงานได้อย่างรวดเร็ว และสามารถรองรับการทำงานของข้อมูลจำนวนมากได้

สำหรับการทำงานของระบบก็สามารถแบ่งออกเป็น 2 ส่วนคือ

1. ส่วนของ Classification จะทำการสร้างโมเดล Classification Tree จากข้อมูลที่ทราบค่า Pass หรือ Fail อยู่แล้ว ซึ่งการทำงานในส่วนนี้จะช่วยในการวิเคราะห์ถึงสาเหตุการเสียที่เกิดขึ้นในกระบวนการผลิต รวมถึงสามารถช่วยในการวิเคราะห์ถึงข้อมูลของกระบวนการก่อนหน้านี้ว่าส่งผลต่อการเกิดของเสียในกระบวนการผลิต HGA อย่างไร
2. ส่วนของ Prediction เป็นกระบวนการที่ต้องใช้ข้อมูล 2 ส่วน คือข้อมูลของโมเดล Classification Tree ที่สร้างมาจากส่วนของการทำ Classification และข้อมูลของกระบวนการ Slider ที่จะนำมา Predict ว่าชิ้นงานนั้นจะ Pass หรือ Fail อย่างไร ซึ่งจากข้อมูลที่ทำการ Prediction ได้ในส่วนนี้ สามารถช่วยในการพยากรณ์ Yield ได้ และยังสามารถนำมาใช้ในการปรับ Target ต่างๆ ของพารามิตเตอร์ในส่วนของ Slider เพื่อให้ได้ Yield สูงสุดในกระบวนการ HGA

แต่อย่างไรก็ดี ระบบการวิเคราะห์หาสาเหตุการเสียของกระบวนการผลิตนี้ ก็เป็นเพียงเครื่องมือที่ใช้ช่วยในการวิเคราะห์ข้อมูล ในอีกแนวทางหนึ่งที่ใช้หลักการทำงานของ Data Mining การที่จะใช้งานอย่างถูกต้อง มีประสิทธิภาพ และประสบความสำเสร็จในการใช้ช่วยในการวิเคราะห์ข้อมูล และนำข้อมูลที่ได้ไปใช้ให้เกิดประโยชน์ได้นั้น ก็ต้องใช้งานโดยผู้ใช้ที่มีความรู้ทั้งในเรื่องของกระบวนการผลิต กลไกลการทำงานของ Data mining รวมถึงในเรื่องของหลักการทำงานของระบบนี้ เนื่องจากการทำ Data Mining ยังต้องประกอบด้วยกระบวนต่างอื่นๆอีกหลายขั้นตอน ดังนั้นจึงต้องมีการให้ความรู้เบื้องต้นเกี่ยวกับ Data Mining และหลักการทำงานของระบบ ให้กับผู้ใช้งาน

6.2 ปัญหาที่เกิดขึ้น

ปัญหาที่เกิดขึ้นส่วนใหญ่ เกิดจากความผิดพลาดในการเขียนโปรแกรมที่ต้องให้ถูกต้องตามวิธีการของอัลกอลิธึมที่เลือกใช้ ซึ่งทำให้ต้องใช้เวลา และประสบการณ์ในการแก้ไขปัญหาต่างๆที่เกิดขึ้น นอกจากนี้ยังมีปัญหาที่เกิดจากรูปแบบการทำงานของระบบ ซึ่งสามารถสรุปได้ดังนี้

1. ผู้ใช้ระบบโดยทำการโหลดนำข้อมูลที่มีอยู่มาใช้ในการวิเคราะห์หาสาเหตุการเสียที่เกิดขึ้นในกระบวนการผลิต ควรมีความรู้ความเข้าใจทั้งในเรื่องของกระบวนการผลิต ข้อมูลที่นำมาใช้ และกระบวนการทำงานของ Data mining มิฉะนั้นผลที่ได้จากการทำ Data mining อาจไม่ถูกต้องเนื่องจากขาดกระบวนการใดกระบวนการหนึ่งใน หลักการทำงานของ Data Mining
2. ข้อมูลที่จะใช้ในการ Prediction ต้องมีโครงสร้างของข้อมูลเหมือนกับ ข้อมูลที่ใช้ในการสร้างโมเดล Classification ที่จะใช้ในการ Prediction ถ้าข้อมูลที่ได้ไม่สมบูรณ์ตามโครงสร้างก็อาจทำให้การ Prediction ผิดพลาด หรือไม่ก็จะไม่สามารถทำการ Predict ข้อมูลได้
3. ระบบค่อนข้างไม่ยืดหยุ่นทำให้ผลการทำงานที่ต้องใช้ข้อมูลจาก Database ยังไม่สามารถทำการเพิ่มหรือลดจำนวนข้อมูลหรือเพิ่มเงื่อนไขของข้อมูลได้
4. ค่าโมเดลที่ได้จากการสร้าง Classification แม้ว่าจะค่อนข้างสูง แต่ถ้านำไปใช้ในการ Prediction กลุ่มข้อมูลจำนวนมากๆ ก็ยังเกิดความผิดพลาดค่อนข้างสูง เมื่อเทียบกับค่า Variation ของ Yield ที่ยอมรับได้ ของระบบ ซึ่งอาจต้องมีการปรับปรุงหรือเลือกใช้ อัลกอลิธึมในการ Pruning ที่สามารถให้ค่า Accuracy มากกว่านี้

6.3 ข้อเสนอแนะ

ระบบที่ออกแบบไว้ลักษณะการทำงานที่ได้ยังไม่ค่อยสมบูรณ์มากนัก ระบบจึงยังมีข้อจำกัดอยู่หลายประการที่ควรจะต้องปรับปรุงแก้ไข เพื่อให้โปรแกรมมีความยืดหยุ่นอย่างเหมาะสม และสามารถนำไปประยุกต์ใช้ให้เหมาะกับระบบงานในองค์กร ได้อย่างมีประสิทธิภาพมากยิ่งขึ้น สิ่งที่ควรแก้ไขสำหรับผู้ที่สนใจจะพัฒนาต่อไปมีดังนี้

1. ควรมีการเพิ่มฟังก์ชันการใช้งาน ให้สามารถทำการใช้งานได้อย่างยืดหยุ่นมากขึ้น ซึ่งจะช่วยทำให้สามารถนำระบบไปประยุกต์ใช้งานได้มากขึ้น
2. อาจจะมีการปรับปรุงโครงสร้างการทำงานของ โปรแกรมจากการ ที่ทำการประมวลผลที่ผู้ใช้เป็นการประมวลผลที่ Server เมื่อต้องการใช้ข้อมูลจาก Database เพราะจะทำให้ระบบสามารถทำงานได้รวดเร็วขึ้น
3. อาจจะมีการนำไปใช้งานร่วมกับระบบการทำงานที่มีอยู่ เช่น ระบบทางสถิติของ SPC ที่มีการตรวจเช็คว่าการทำงานอยู่ในขอบเขตที่ยอมรับได้หรือไม่ ถ้ามีเหตุการณ์ผิดปกติเกิดขึ้น อาจจะมีการให้นำข้อมูลในช่วงเวลานั้นมาทำการ Classification เพื่อใช้เป็นแนวทางของข้อมูลที่จะทำกระบวนการอื่นๆ อีกต่อไป

บรรณานุกรม

Andrew Kusiak. 2000. **" Decomposition in Data mining : An Industrial case Study".** 345-352. In IEEE Transactions on electronics packaging manufacturing. Vol23

Ashok, N. Srivastava,Ph.D. 1999. **"Data Mining for Semiconductor Yield Forecasting and Enhancement Manufacturing Productivity"**. Future Fab International Magazine.

Bhavani, Thuraisingham,PH.D. 1998. **Data mining Technologies, Techniques, Tools, and Trends.** Boca Raton : CRC Press LLC.

J.C. Shafer, R. Agrawal, and M. Mehta. 1996. **"SPRINT:A Scalable Parallel Classifier for Data Mining".** In Proc. of the 22th VLDB Conference.

Jiawei, Han and Micheline, Kamber. 2000. **Data Mining Concepts and Techniques.** Morgan Kaufman

M. Mehta, R. Agrwal ,and J. Rissanen. 1996. **"SLIQ : A fast Scalable Classifier for Data Mining".** 18-32. In Proc. of the fifth Int'l Conference on Extending Database Technology.

Peter, Cabena. et al. 1997. **Discovering data mining from concept to implementation.** New Jersey : Prentice Hall PTR.

ประวัติผู้เขียน

ชื่อ นามสกุล	นางสาววิไล แม้นถาวรสิริ
วัน เดือน ปีเกิด	15 ตุลาคม พ.ศ. 2520
สถานที่เกิด	จังหวัดสมุทรปราการ
วุฒิการศึกษา	ปริญญาตรีวิศวกรรมศาสตร์บัณฑิต สาขาอิเล็กทรอนิกส์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
การทำงาน	วิศวกร บริษัทซีเกท เทคโนโลยี ประเทศไทยจำกัด