



ใบรับรองวิทยานิพนธ์
บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์
วิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมไฟฟ้า)
ปริญญา

วิศวกรรมไฟฟ้า

วิศวกรรมไฟฟ้า

สาขา

ภาควิชา

เรื่อง การทำนายสาเหตุของเหตุการณ์กระแสไฟฟ้าขัดข้องโดยใช้เทคนิคการทำเหมืองข้อมูล
ในระบบจำหน่ายของการไฟฟ้าส่วนภูมิภาค เขต 1 ภาคกลาง

Prediction of Power Interruption Cause Using Data Mining Techniques
for PEA Distribution Systems in Central 1 Area

นามผู้วิจัย นายจรัสศักดิ์ ศรีอ่อน

ได้พิจารณาเห็นชอบโดย

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(ผู้ช่วยศาสตราจารย์ดุลย์พิเชษฐ์ ฤกษ์ปรีดาพงศ์, Ph.D.)

อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม

(อาจารย์นันทยา คล้ายเรือง, วศ.ค.)

หัวหน้าภาควิชา

(รองศาสตราจารย์มงคล รักษาพัชรวงษ์, Ph.D.)

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว

(รองศาสตราจารย์กัญญา ชีระกุล, D.Agr.)

คณบดีบัณฑิตวิทยาลัย

วันที่ _____ เดือน _____ พ.ศ. _____

วิทยานิพนธ์

เรื่อง

การทำนายสาเหตุของเหตุการณ์กระแสไฟฟ้าขัดข้อง โดยใช้เทคนิคการทำเหมืองข้อมูล
ในระบบจำหน่ายของการไฟฟ้าส่วนภูมิภาค เขต 1 ภาคกลาง

Prediction of Power Interruption Cause Using Data Mining Techniques
for PEA Distribution Systems in Central 1 Area

โดย

นายจรัสศักดิ์ ศรีอ่อน

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์
เพื่อความสมบูรณ์แห่งปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมไฟฟ้า)

พ.ศ. 2552

ขจรศักดิ์ ศรีอ่อน 2552: การทำนายสาเหตุของเหตุการณ์กระแสไฟฟ้าขัดข้อง
โดยใช้เทคนิคการทำเหมืองข้อมูลในระบบจำหน่ายของการไฟฟ้าส่วนภูมิภาค เขต 1
ภาคกลาง ปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมไฟฟ้า) สาขาวิชา
วิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก:
ผู้ช่วยศาสตราจารย์ดุลย์พิเชษฐ์ ฤกษ์ปรีดาพงศ์, Ph.D. 65 หน้า

วิทยานิพนธ์นี้นำเสนอการวิเคราะห์หาสาเหตุของเหตุการณ์กระแสไฟฟ้าขัดข้อง ในส่วน
เหตุการณ์ที่ไม่ทราบสาเหตุซึ่งเกิดในระบบจำหน่ายแรงสูง 22 เควี ของการไฟฟ้าส่วนภูมิภาค
เขต 1 ภาคกลาง โดยใช้เทคนิคการทำเหมืองข้อมูล (Data Mining) โดยใช้ 3 เทคนิคที่ต่างกัน
อันประกอบด้วยต้นไม้ตัดสินใจ (Decision Tree) เทคนิคแบบนาอิวเบย์เซียน (Naïve Bayesian)
และกฎของริปเปอร์ (RIPPER) ข้อมูลกระแสไฟฟ้าขัดข้องที่ทราบสาเหตุถูกนำมาวิเคราะห์
หาแบบจำลอง (Model) โดยวิธีการจำแนกประเภท (Classification Method) โดยใช้เทคนิคข้างต้น
จากนั้นนำข้อมูลมาแบ่งเป็น 2 ชุด คือ ชุดข้อมูลเรียนรู้ (Training set) ใช้ในกระบวนการค้นหา
แบบจำลอง และชุดข้อมูลทดสอบ (Test set) ใช้ตรวจสอบความถูกต้องของแบบจำลอง ซึ่งผลที่ได้
จะเห็นได้ว่าการใช้เทคนิคต้นไม้ตัดสินใจมีค่าความถูกต้องสูงที่สุด มีค่าความคลาดเคลื่อน (Root
Mean Squared Error: RMSE) ต่ำ เมื่อทดสอบกับชุดข้อมูลทดสอบได้ค่าความถูกต้องเฉลี่ย 82.32
เปอร์เซ็นต์ ซึ่งเป็นที่น่าพอใจ สามารถนำข้อมูลที่ได้ไปใช้ในงานวิจัย ตลอดจนงานวิเคราะห์
วางแผนได้อย่างมีประสิทธิภาพ

Khajonsak Sri-on 2009: Prediction of Power Interruption Cause Using Data Mining Techniques for PEA Distribution Systems in Central 1 Area. Master of Engineering (Electrical Engineering), Major Field: Electrical Engineering, Department of Electrical Engineering. Thesis Advisor: Assistant Professor Dulpichet Rerkpreedapong, Ph.D. 65 pages.

This thesis presents the data mining analysis of unknown interruption cause on 22 kV Power Distribution Systems in Central 1 Area of Provincial Electricity Authority (PEA), Thailand. Prediction of the unknown cause interruptions by data mining is useful for improving efficiency of PEA activities such as the power system planning and power system analysis. Three data mining techniques; Decision Tree, Naïve Bayesian, and JRip, were applied to build the best model of classification to classify the cause interruptions. The data of interruptions from PEA software data base is used to analyze through the process of classification method based on data mining. The data was separated into training set and test set. The result showed that decision tree technique built the best model with the highest percentage of correctly instances rate and the lowest RMSE. When this model was used to predict the test data with the unknown cause interruptions, it was found that the average accuracy percentage was 82.32. The result was appreciated and this model can be applied to improve the power system research, planning and analyzing.

Student's signature

Thesis Advisor's signature

____ / ____ / ____

กิตติกรรมประกาศ

ขอขอบพระคุณผู้ช่วยศาสตราจารย์ ดร. ดุลย์พิเชษฐ์ ฤกษ์ปรีดาพงศ์ ประธานกรรมการ
ที่ปรึกษา อ.พิสุทธิ รพีศักดิ์ ที่กรุณาให้คำปรึกษาแนะนำ ช่วยเหลือและแก้ไขปัญหาวิตยานิพนธ์
ฉบับนี้ และขอขอบพระคุณรองศาสตราจารย์ ดร. ชำนาญ ห่อเกียรติ ที่กรุณาให้โอกาสให้ข้าพเจ้า
ได้พบสิ่งที่ดีในชีวิตข้าพเจ้า

ขอขอบพระคุณ รศ.ดร. ณัฐภา หอมทรัพย์ ประธานการสอบ ดร. นาดยา คล้ายเรือง
กรรมการร่วมและ ดร. ประมุข อุณหเลขกะ ผู้ทรงคุณวุฒิภายนอก ที่กรุณาช่วยแนะนำและแก้ไข
วิทยานิพนธ์เล่มนี้ให้มีความสมบูรณ์มากยิ่งขึ้น

ขอขอบคุณการไฟฟ้าส่วนภูมิภาค ที่ให้โอกาสและสนับสนุนในการเล่าเรียนรวมทั้งข้อมูล
ที่ใช้ในการทำวิทยานิพนธ์เล่มนี้จนสำเร็จ

ขอขอบพระคุณพ่อ แม่ ครอบครัวและคุณมยุรา วัฒนานนท์ ที่เป็นกำลังใจและให้
การสนับสนุนมาโดยตลอด และขอบคุณเพื่อนๆ และน้องๆ ในโครงการวิจัยและพัฒนา
ความชำนาญด้านไฟฟ้ากำลังที่สนับสนุน และให้คำแนะนำในการเขียนวิทยานิพนธ์ฉบับนี้จนสำเร็จ

ท้ายที่สุดนี้ ข้าพเจ้าขอความรู้ความสามารถและประสบการณ์ที่มีค่าในการศึกษาครั้งนี้
ไปใช้ให้เกิดประโยชน์สูงสุดต่อสังคมต่อไป

ขจรศักดิ์ ศรีอ่อน

พฤษภาคม 2552

สารบัญ

หน้า

สารบัญ	(1)
สารบัญตาราง	(2)
สารบัญภาพ	(5)
คำนำ	1
วัตถุประสงค์	2
การตรวจเอกสาร	3
อุปกรณ์และวิธีการ	22
อุปกรณ์	22
วิธีการ	22
ผลและวิจารณ์	44
ผล	44
วิจารณ์	58
สรุปและข้อเสนอแนะ	61
สรุป	61
ข้อเสนอแนะ	62
เอกสารและสิ่งอ้างอิง	63
ประวัติการศึกษา และการทำงาน	65

สารบัญตาราง

ตารางที่		หน้า
1	จำนวนข้อมูลทั้งหมดที่นำมาใช้วิเคราะห์	23
2	จำนวนข้อมูลสาเหตุ 11 สาเหตุ ที่นำมาใช้วิเคราะห์	28
3	จำนวนข้อมูลสาเหตุ 6 สาเหตุ ที่มีจำนวนข้อมูลไม่เกิน 1 เปอร์เซ็นต์ ของข้อมูลทั้งหมด	29
4	จำนวนข้อมูลสาเหตุ 6 สาเหตุ ที่รวมชุดข้อมูลที่มีจำนวนน้อยเข้าด้วยกัน เพื่อใช้วิเคราะห์	30
5	จำนวนข้อมูลสาเหตุ Attribute Time-outage ที่ทำการจัดรูปแบบให้เป็น ลักษณะข้อมูล Nominal	32
6	จำนวนข้อมูลสาเหตุ Attribute Total-Time-outage ที่ทำการจัดรูปแบบ ให้เป็นลักษณะข้อมูล Nominal	34
7	จำนวนการแบ่งชุดข้อมูล	36
8	ชุดข้อมูลเรียนรู้	42
9	ชุดข้อมูลทดสอบจำนวน 5 ชุดๆ ละ 500 ข้อมูล	42
10	ชุดข้อมูลทดสอบที่ไม่ทราบสาเหตุจำนวน 100 ข้อมูล	43
11	ชุดข้อมูลทดสอบที่ไม่ทราบสาเหตุทั้งหมด 1,745 ข้อมูล สุ่มเลือกมา 5 ข้อมูล	43
12	แสดงค่าความถูกต้องและค่าเฉลี่ยความผิดพลาด จากสร้างแบบจำลอง ในการเรียนรู้ จากชุดข้อมูลเรียนรู้ ที่ 1-3	45
13	ค่า Confusion Matrix ของการหาแบบจำลองจากข้อมูลชุดที่ 1 โดยอัลกอริทึมกฎของริบเปอร์	46
14	ค่า Confusion Matrix ของการหาแบบจำลองจากข้อมูลชุดที่ 2 โดยอัลกอริทึมกฎของริบเปอร์	46
15	ค่า Confusion Matrix ของการหาแบบจำลองจากข้อมูลชุดที่ 3 โดยอัลกอริทึมกฎของริบเปอร์	47
16	ค่า Confusion Matrix ของการหาแบบจำลองจากข้อมูลชุดที่ 1 โดยอัลกอริทึมนาอ์ฟเบย์เซียน	47

สารบัญตาราง (ต่อ)

ตารางที่		หน้า
17	ค่า Confusion Matrix ของการหาแบบจำลองจากข้อมูลชุดที่ 2 โดยอัลกอริทึมนาอิวเบย์เซียน	48
18	ค่า Confusion Matrix ของการหาแบบจำลองจากข้อมูลชุดที่ 3 โดยอัลกอริทึมนาอิวเบย์เซียน	48
19	ค่า Confusion Matrix ของการหาแบบจำลองจากข้อมูลชุดที่ 1 โดยอัลกอริทึมต้นไม้ตัดสินใจ	49
20	ค่า Confusion Matrix ของการหาแบบจำลองจากข้อมูลชุดที่ 2 โดยอัลกอริทึมต้นไม้ตัดสินใจ	49
21	ค่า Confusion Matrix ของการหาแบบจำลองจากข้อมูลชุดที่ 3 โดยอัลกอริทึมต้นไม้ตัดสินใจ	50
22	ผลการทดสอบชุดข้อมูลทดสอบ 5 ชุดๆละ 500 ข้อมูล	52
23	ค่า True Positive rate (TP rate) จากผลทดสอบชุดข้อมูลทดสอบ 5 ชุดๆ ละ 500 ข้อมูล	52
24	ค่า Confusion Matrix จากผลทดสอบชุดข้อมูลทดสอบ 5 ชุดๆ ละ 500 ข้อมูล ของชุดที่ 1	52
25	ค่า Confusion Matrix จากผลทดสอบชุดข้อมูลทดสอบ 5 ชุดๆ ละ 500 ข้อมูล ของชุดที่ 2	53
26	ค่า Confusion Matrix จากผลทดสอบชุดข้อมูลทดสอบ 5 ชุดๆ ละ 500 ข้อมูล ของชุดที่ 3	53
27	ค่า Confusion Matrix จากผลทดสอบชุดข้อมูลทดสอบ 5 ชุดๆ ละ 500 ข้อมูล ของชุดที่ 4	54
28	ค่า Confusion Matrix จากผลทดสอบชุดข้อมูลทดสอบ 5 ชุดๆ ละ 500 ข้อมูล ของชุดที่ 5	54
29	ผลรวมในการทำนายหาสาเหตุ อื่นๆ จาก Confusion Matrix ของชุดข้อมูล ทดสอบ 5 ชุดๆ ละ 500 ข้อมูล	55

สารบัญตาราง (ต่อ)

ตารางที่		หน้า
30	ผลรวมในการทำนายหาสาเหตุ ภัยธรรมชาติ จาก Confusion Matrix ของชุดข้อมูลทดสอบ 5 ชุดๆ ละ 500 ข้อมูล	55
31	ผลทดสอบข้อมูลที่ไม่ทราบสาเหตุจำนวน 100 ข้อมูล	56
32	ค่า Confusion Matrix ของผลทดสอบข้อมูลที่ไม่ทราบสาเหตุจำนวน 100 ข้อมูล	57
33	ผลการทดสอบชุดข้อมูล 5 ข้อมูล	58

สารบัญภาพ

ภาพที่		หน้า
1	อัลกอริทึมพื้นฐานของกฎของริปเปอร์	7
2	อัลกอริทึมสำหรับการจำแนกประเภทข้อความโดยใช้การเรียนรู้แนอ์ฟเบย์	10
3	อัลกอริทึมพื้นฐานในการหาต้นไม้ตัดสินใจ	13
4	การแบ่งชุดข้อมูลเพื่อทำนนวนค่าความผิดพลาดโดยวิธีครอสวาลิเดชัน	16
5	หน้าจอหลักของโปรแกรม WEKA Explore	18
6	ลักษณะ Confusion Matrix แสดงผลการวิเคราะห์และทดสอบ	20
7	ลำดับขั้นตอนการนำข้อมูลกระแสไฟฟ้าขัดข้องมาวิเคราะห์โดยวิธีเหมืองข้อมูล	22
8	ข้อมูลกระแสไฟฟ้าขัดข้องปี พ.ศ. 2547 – 2551	24
9	ไฟล์สถิติเหตุการณ์ไฟฟ้าขัดข้องปี พ.ศ. 2547 – 2551 ในรูป Main database file (file.mdb)	24
10	ลักษณะฐานข้อมูลไฟฟ้าขัดข้องจากโปรแกรม จฟ. 3	25
11	จำนวนข้อมูล ที่ทำให้เกิดเหตุการณ์ทั้งหมด 11 สาเหตุ	29
12	จำนวนข้อมูลสาเหตุ ที่ทำให้เกิดเหตุการณ์ทั้งหมด 6 สาเหตุ	30
13	การเลือกค่าสั่ง Discretize กับข้อมูลที่เป็นชนิดตัวเลขในโปรแกรม WEKA	31
14	กราฟแสดงจัดกลุ่มข้อมูลสาเหตุ Time out	33
15	กราฟแสดงจัดกลุ่มข้อมูลสาเหตุ Total time out	35
16	ลักษณะข้อมูลทั้งหมดที่เตรียมพร้อมสำหรับวิเคราะห์	37
17	ลักษณะชุดข้อมูลทดสอบโดยนำข้อมูล ปี 2551 ที่ทราบสาเหตุ 500 ข้อมูล	38
18	ลักษณะชุดข้อมูลทดสอบโดยนำข้อมูล ปี 2551 ที่ทราบสาเหตุ 100 ข้อมูล	39
19	ลักษณะชุดข้อมูลทดสอบโดยนำข้อมูล ปี 2551 ที่ทราบสาเหตุ 5 ข้อมูล	40
20	กระบวนการของการจำแนกข้อมูล (Classification)	41
21	ลักษณะผลที่ได้ในโปรแกรม จากการหาแบบจำลอง	44
22	ลักษณะการแสดงผลในโปรแกรม WEKA ที่ได้จากการนำแบบจำลองทดสอบกับชุดข้อมูลทดสอบ	51

คำอธิบายสัญลักษณ์และคำย่อ

SAIFI	=	System Average Interruption Frequency Index
SAIDI	=	System Average Interruption Duration Index
MAIFI	=	Momentary Average Interruption Frequency Index
RIPPER	=	Repeated Incremental Pruning to Produce Error Reduction
IREP	=	Incremental Reduced Error Pruning
MDL	=	Minimal Description Length
DL	=	Description Length
GUI	=	Graphical User Interface
TP rate	=	True Positive rate
FP rate	=	False Positive rate
Tto	=	Total Time-outage
To	=	Time-outage
RMSE	=	Root Mean Square Error

**การทำนายสาเหตุของเหตุการณ์กระแสไฟฟ้าขัดข้องโดยใช้เทคนิคการทำเหมืองข้อมูล
ในระบบจำหน่ายของการไฟฟ้าส่วนภูมิภาค เขต 1 ภาคกลาง**

**Prediction of Power Interruption Cause Using Data Mining Techniques
for PEA Distribution Systems in Central 1 Area**

คำนำ

ระบบจำหน่ายส่วนใหญ่ของการไฟฟ้าส่วนภูมิภาค เป็นระบบเหนือนดิน พาดไปตามแนวถนนทั่วประเทศ จึงมีปัจจัยเสี่ยงที่ส่งผลให้เกิดปัญหาการกระแสไฟฟ้าขัดข้องได้บ่อยครั้งกว่าระบบใต้ดิน โดยเหตุการณ์กระแสไฟฟ้าขัดข้องมักมีสาเหตุและระยะเวลาแตกต่างกันไปในแต่ละพื้นที่ การไฟฟ้าส่วนภูมิภาคได้ดำเนินการจัดเก็บข้อมูลกระแสไฟฟ้าขัดข้อง ด้วยโปรแกรม จฟ.3 ตั้งแต่ปี พ.ศ. 2537 เพื่อใช้ในการคำนวณดัชนีความเชื่อถือได้ คือค่าดัชนี SAIFI SAIDI และMAIFI เพื่อบอกถึงคุณภาพของการจ่ายไฟฟ้าให้กับผู้ใช้ไฟฟ้าและใช้ในการวางแผนระบบไฟฟ้า นอกจากนี้ยังมีการนำข้อมูลดังกล่าวไปใช้ในโครงการวิจัยเพื่อปรับปรุงด้านต่างๆของการไฟฟ้าส่วนภูมิภาค

ข้อมูลเหตุการณ์กระแสไฟฟ้าขัดข้องของการไฟฟ้าส่วนภูมิภาค ปีละประมาณ 30% ไม่สามารถระบุสาเหตุที่ชัดเจนของการเกิดเหตุการณ์กระแสไฟฟ้าขัดข้องได้ จึงใช้การสันนิษฐานสาเหตุจากประสบการณ์ของผู้บันทึกในการเกิดเหตุการณ์แต่ละครั้ง ซึ่งอาจทำให้ข้อมูลมีความคลาดเคลื่อนจากความเป็นจริงและเกิดผลเสียเมื่อนำข้อมูลดังกล่าวไปใช้ในการวางแผนปรับปรุงความเชื่อถือได้ในภายหลัง การศึกษาและใช้การทำเหมืองข้อมูล (Data Mining) ในการวิเคราะห์ข้อมูลเหตุการณ์กระแสไฟฟ้าขัดข้อง เพื่อระบุสาเหตุที่ไม่ทราบ ทำให้ได้ข้อมูลกระแสไฟฟ้าขัดข้องมีความน่าเชื่อถือมากขึ้นและสามารถนำไปใช้ได้อย่างถูกต้องในการวิเคราะห์และวางแผนระบบไฟฟ้าของการไฟฟ้าส่วนภูมิภาคต่อไปในอนาคต

วัตถุประสงค์

1. ศึกษาฐานข้อมูลเหตุการณ์ไฟฟ้าขัดข้องของ กฟภ. และสาเหตุต่างๆ ที่ทำให้เกิดเหตุการณ์กระแสไฟฟ้าขัดข้อง
2. ศึกษาวิธีการทำเหมืองข้อมูล ในส่วนของเทคนิคการจำแนกประเภท
3. หาแบบจำลองที่มีประสิทธิภาพในการทำนายสาเหตุจาก 3 เทคนิคที่แตกต่างกัน
4. หาสาเหตุที่ทำให้เกิดกระแสไฟฟ้าขัดข้องในกรณีที่ไม่ทราบสาเหตุ โดยใช้แบบจำลองที่เหมาะสมที่สุด

การตรวจเอกสาร

1. การทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูลนั้น มีหลายนิยามที่ให้คำจำกัดความของเหมืองข้อมูลไว้ โดยสามารถสรุปหลักสำคัญได้ว่า เหมืองข้อมูล เป็นกระบวนการของการค้นกรองสารสนเทศ (Information) ที่ซ่อนอยู่ในฐานข้อมูลใหญ่ เพื่อทำนายแนวโน้ม และพฤติกรรม โดยอาศัยข้อมูลในอดีต โดยค้นหาแบบรูปความสัมพันธ์ และความรู้ใหม่ๆ จากข้อมูล Witten et al. (2005) มีขั้นตอนดังนี้

1.1 การทำความเข้าใจปัญหา คือการวางแผนในการเลือกข้อมูลที่มีความเหมาะสมกับอัลกอริทึมที่ใช้งาน จำนวนข้อมูลที่ต้องการ และเป้าหมายเพื่อให้ได้ผลลัพธ์ที่ต้องการ

1.2 การทำความเข้าใจข้อมูล คือ การรวบรวม ตรวจสอบความถูกต้องและกำหนดคุณสมบัติที่ต้องการ

1.3 การเตรียมข้อมูล เป็นการคัดเลือกข้อมูลเพื่อทำการแปลงให้อยู่ในรูปแบบที่เหมาะสมสำหรับการนำไปใช้วิเคราะห์

1.4 การสร้างแบบจำลอง แบ่งได้เป็น 2 ประเภทใหญ่ๆ คือ

- การสร้างแบบจำลองในการทำนาย (Predictive Data Mining) เป็นการคาดคะเนลักษณะหรือประมาณค่าที่ชัดเจนของข้อมูลที่จะเกิดขึ้น โดยใช้พื้นฐานจากข้อมูลที่ผ่านมาในอดีต

- การสร้างแบบจำลองในการบรรยาย (Descriptive Data Mining) เป็นการหาแบบจำลองเพื่ออธิบายลักษณะบางอย่างของข้อมูล ที่มีอยู่ ซึ่งโดยส่วนมากจะเป็นลักษณะการแบ่งกลุ่มให้กับข้อมูล

เนื่องจากการทำเหมืองข้อมูลเป็นเทคนิคในการค้นความรู้จากข้อมูลขนาดใหญ่ การทำเหมืองข้อมูลจึงเป็นการรวมเอาศาสตร์ต่างๆ หลายแขนงมารวมไว้ด้วยกันโดยไม่จำกัดวิธีการที่จะใช้ ตัวอย่างศาสตร์ที่ใช้ เช่น เทคโนโลยีฐานข้อมูล (Database technology) วิทยาศาสตร์สารสนเทศ

(Information science) สถิติ (Statistics) และระบบการเรียนรู้ (Machine learning) เป็นต้น ซึ่งศาสตร์ต่างๆ เหล่านี้จะทำให้เกิดกระบวนการค้นความรู้ในแบบต่างๆ โดยรูปแบบการค้นความรู้หลักมีดังนี้

- การวิเคราะห์ความสัมพันธ์ (Association Analysis)
- การจำแนกประเภทและการทำนาย (Classification and Prediction)
- การจัดกลุ่มข้อมูล (Clustering Analysis)

1.4.1 การวิเคราะห์ความสัมพันธ์ (Association Analysis)

การวิเคราะห์ความสัมพันธ์เป็นรูปแบบการค้นความรู้โดยการหาสิ่งที่เรียกว่า “กฎของความสัมพันธ์ (Association rules)” ซึ่งแสดงความสัมพันธ์ของค่าที่มีความสัมพันธ์และมีเงื่อนไขที่ตรงกับข้อกำหนดและลักษณะของข้อมูลที่มีการเรียนรู้ในรูปของตะกร้าจ่ายตลาด (Market Basket) หรือการซื้อขาย (Transaction) ในรูปทั่วไปกฎความสัมพันธ์จะเขียนในรูป $X \Rightarrow Y$ (ถ้า X แล้ว Y ด้วย) โดยที่ X คือ เงื่อนไขเริ่มต้นที่บอกว่าอาจจะเกิด Y ด้วยกัน, Y คือ ผลที่จะเกิดด้วยกันหากมีการเกิด X ในการทำการวิเคราะห์ความสัมพันธ์กฎที่เกิดขึ้นนี้จำเป็นที่จะต้องกำหนดค่า สนับสนุน (Support) และค่าความมั่นใจ (Confidence) ซึ่งเป็นตัวกำหนดว่ากฎที่เกิดขึ้นนั้นมีความสัมพันธ์กันในระดับใด และยังเป็นการช่วยยับยั้งการเกิดกฎที่ไม่จำเป็นหรือกฎที่มีความเกี่ยวข้องกันน้อยมาก ตัวอย่างเทคนิคของการวิเคราะห์ความสัมพันธ์ได้แก่ การวิเคราะห์แบบตะกร้าสินค้า (Market Basket Analysis) การคำนวณแบบแอฟพริออรี (The Apriori Algorithm) และกฎความสัมพันธ์แบบหลายระดับ (Multilevel Association Rules) เป็นต้น

1.4.2 การจำแนกประเภทและการทำนาย (Classification and Prediction)

การจำแนกประเภท เป็นกระบวนการที่ใช้ในการหารูปแบบของชุดข้อมูลที่มีความใกล้เคียงกันหรือเหมือนกันมากที่สุด เพื่อใช้ในการทำนายชุดข้อมูลว่าอยู่ในประเภทใดของชุดข้อมูลที่ได้ทำการแบ่งไว้แล้ว ซึ่งชุดข้อมูลที่แบ่งไว้เกิดจากการเรียนรู้จากชุดข้อมูลที่มีอยู่แล้ว (Training data) แบบจำลองที่เกิดจากการเรียนรู้ สามารถแสดงได้หลายรูปแบบ เช่น กฎการแบ่ง (Classification rules, IF-THEN) การคำนวณแบบต้นไม้ตัดสินใจ (Decision Tree) การใช้สูตรทางคณิตศาสตร์ (Mathematical formulae) หรือโครงข่ายประสาทเทียม เป็นต้น ในส่วนของการทำต้นไม้ตัดสินใจ จะแสดงออกมาในลักษณะของแผนภูมิโครงสร้างต้นไม้ ซึ่งก้านของต้นไม้แสดงถึง

ความรู้ที่ได้ และใบไม้จะแสดงถึงประเภทชุดข้อมูลที่ถูกแบ่งออกมา แผนภูมิต้นไม้สามารถแปลงเป็นกฎการแบ่งได้ง่ายเพราะลักษณะของแผนภูมิสามารถเข้าใจได้ง่าย ในส่วนของโครงข่ายประสาทเทียมนั้น ถูกแสดงในลักษณะของการเชื่อมต่อระหว่างหน่วยที่เกิดขึ้น การทำการจำแนกประเภทนั้นมักใช้ประโยชน์ร่วมกับการทำนายโดยเฉพาะข้อมูลที่เป็นตัวเลข เราจึงอาจมองได้ว่าการทำนายเป็นการบอกถึงค่าตัวเลขและการบ่งบอกประเภทของข้อมูลนั้นในลักษณะของการดูแนวโน้ม (Trends) ที่จะเกิดขึ้น ตัวอย่างเทคนิคของการจำแนกประเภทและการทำนาย ได้แก่ การคำนวณแบบพันธุกรรม (Genetic Algorithm) การคำนวณแบบต้นไม้ตัดสินใจ และโครงข่ายประสาทเทียม (Neural Network) เป็นต้น

1.4.3 การวิเคราะห์เพื่อจัดกลุ่ม (Clustering Analysis)

การวิเคราะห์เพื่อจัดกลุ่มแตกต่างกับการทำการจำแนกประเภทและการทำนาย ซึ่งวิเคราะห์กลุ่มข้อมูลที่มีความคล้าย ซึ่งเป็นการจัดกลุ่มที่จำแนกประเภทโดยไม่มีการระบุชื่อกลุ่ม ในช่วงของการสอนแบบจำลอง โดยทั่วไปแล้ววิธีแบบนี้ใช้กับการจัดการแบ่งข้อมูลที่ไม่รู้ว่าจะจัดประเภทไว้ด้วยกันอย่างไรดี และการทำการวิเคราะห์นี้สามารถทำการบ่งบอกถึงชื่อของกลุ่มที่แบ่งขึ้นได้ด้วย ในการทำการวิเคราะห์เพื่อจัดกลุ่มนั้นอาศัยพื้นฐานของความเหมือนกันมากที่สุดและความเหมือนกันน้อยที่สุดของกลุ่ม นั่นคือข้อมูลที่ถูกจัดไว้ในกลุ่มเดียวกันจะมีความคล้ายกันสูงมาก แต่จะแตกต่างกันกับข้อมูลที่ถูกจัดไว้ในกลุ่มละกลุ่ม และตัวอย่างของการวิเคราะห์เพื่อจัดกลุ่ม ได้แก่ การหาค่าเฉลี่ย k (k-mean Algorithm) การรวมและการแบ่งกลุ่มโดยจัดลำดับชั้น (Agglomerative and Divisive Hierarchical Clustering) และการลำดับตำแหน่งเพื่อแสดงโครงสร้างการจัดกลุ่ม (Ordering Points to Identify the Clustering Structure) เป็นต้น

มีรูปแบบการค้นความรู้อีกมาก แต่จะกล่าวถึงเพียงแค่ 3 แบบ เพราะเป็นรูปแบบที่มีการใช้กันอย่างกว้างขวาง และใน 3 รูปแบบนี้ยังมีเทคนิคย่อยของแต่ละรูปแบบอีกมากมาย

1.5 การวิเคราะห์และประเมินผลลัพธ์ เป็นขั้นตอนการแปล ความหมาย และการประเมินผลลัพธ์ที่ได้ว่ามีความเหมาะสมหรือตรงกับวัตถุประสงค์ที่ต้องการหรือไม่ โดยทั่วไปควรมีการแสดงผลในรูปแบบ ที่สามารถเข้าใจได้โดยง่าย

แบบจำลองหรือความสัมพันธ์ที่หามาได้ ต้องถูกนำมาทดสอบอัตราความผิดพลาด และวิเคราะห์ความซับซ้อนของรูปแบบแบบจำลองถ้าอัตราความผิดพลาดยังสูงเกินไป อาจจะต้องปรับปรุงแบบจำลองให้ถูกต้องยิ่งขึ้น ในทำนองเดียวกันถ้าแบบจำลองที่หามาได้มีรูปแบบที่ซับซ้อนเกินไปจนยากต่อการทำความเข้าใจ อาจจะต้องหาแบบจำลองใหม่ที่มีความถูกต้องเท่าเดิม แต่มีรูปแบบที่ซับซ้อนน้อยลง

ดังนั้นขั้นตอนของการค้นหาแบบจำลองเป็นหัวใจสำคัญของกระบวนการเหมืองข้อมูล แต่ทั้งนี้แบบจำลองหรือความรู้ที่ค้นหามาได้ จะมีคุณภาพและเป็นประโยชน์เพียงใดนั้น ปัจจัยสำคัญที่สุดคือข้อมูล คุณภาพและจำนวนของข้อมูลจะมีผลโดยตรงต่อผลลัพธ์ที่จะได้จากการทำเหมืองข้อมูล คุณภาพของข้อมูลสะท้อนมาจากลักษณะต่างๆ (Attribute or Feature) ที่ประกอบกันขึ้นเป็นข้อมูลหนึ่งรายการ ถ้าลักษณะที่รวบรวมมาเป็นลักษณะหลักที่สามารถตัดสินใจพฤติกรรมของข้อมูล ผลลัพธ์ของการทำเหมืองข้อมูลมีความถูกต้อง จำนวนของข้อมูลเป็นอีกปัจจัยสำคัญที่ช่วยยืนยันว่าผลลัพธ์ที่ได้นั้นมีความเที่ยงตรงสูงเพียงใด

2. อัลกอริทึมสังเคราะห์ความรู้ (Learning algorithm)

อัลกอริทึมสังเคราะห์ความรู้ในรูปแบบการจำแนกประเภทและการทำนายนำมาใช้ 3 อย่างสำหรับหาแบบจำลองเพื่อใช้ทำนายสาเหตุของเหตุการณ์กระแสไฟฟ้าขัดข้องที่ไม่ทราบสาเหตุได้แก่

2.1 กฎของริปเปอร์ (RIPPER)

กฎของริปเปอร์ หรือ RIPPER (Repeated Incremental Pruning to Produce Error Reduction) เป็นการจำแนกประเภทโดยใช้หลักการเรียนรู้ด้วยกฎ หลักการนี้ได้พัฒนาขึ้นโดย William W. Cohen ใน ค.ศ. 1995 เป็นการพัฒนามาจากวิธี IREP (Incremental Reduced Error Pruning) โดยเพิ่มการหาค่าที่เหมาะสมที่สุด (Optimization) กฎของริปเปอร์ เป็นวิธีการเรียนรู้ชุดของกฎจากข้อมูลซึ่งกฎจะอยู่ในลักษณะ “if-then-else” จากข้อมูลกระบวนการแบ่งคลาสชนิดสองคลาสสำหรับวิธีกฎของริปเปอร์พื้นฐานเป็นไปตามวิธี IREP ซึ่งทำได้ตามอัลกอริทึมของ IREP ตามภาพที่ 1

```

procedure IREP(Pos,Neg)
begin
Ruleset:=∅
while Pos ≠ ∅ do
  /* grow and prune a new rule* /
  split (Pos,Neg) into (GrowPos,GrowNeg)
  and (PrunePos,PruneNeg)
  Rule := GrowRule(GrowPos,GrowNeg)
  Rule := PruneRule(Rule,PrunePos,PruneNeg)
  if the error rate of Rule on
  (PrunePos,PruneNeg) exceeds 50% then
    return Ruleset
  else
    add Rule to Ruleset
    remove examples covered by Rule
    from (Pos,Neg)
  endif
endwhile
return Ruleset
end

```

ภาพที่ 1 อัลกอริทึมพื้นฐานของกฎของริบเปอร์

สำหรับการแบ่งข้อมูลแบบสองคลาส (2-class Data) สามารถแบ่งเป็น ข้อมูลกลุ่มบวก (Positive) และ กลุ่มลบ (Negative) เริ่มต้นโดยกำหนดเซตของกฎ (Rule-set) ให้ว่าง แบ่งข้อมูลเรียนรู้ (Training Set) เป็นชุดข้อมูลเติบโต (Growing Set) และชุดข้อมูลการตัดกฎ (Pruning Set) โดยมีขนาด $2/3$ และ $1/3$ ของชุดข้อมูลเรียนรู้ ตามลำดับ สำหรับการแบ่งคลาส (Class) ที่มีมากกว่าสองกลุ่ม เช่น $C_1, C_2, \dots, C_k$ เรียงตามลำดับขนาดกลุ่มเล็กไปใหญ่ IREP ใช้สำหรับการแยกกลุ่มต่างๆ ออกจากกัน ข้อมูลที่ถูกครอบคลุมโดยกฎที่ได้จากการเรียนรู้ (Learned Rules) จะเป็นตัวแยกข้อมูลนั้นออกจากชุดข้อมูลเรียนรู้ การแยกกลุ่มเริ่มจากกลุ่มที่เล็กที่สุด เมื่อกลุ่ม C_1 ถูกแยกออกมาแล้วก็ใช้ IREP ทำการซ้ำสำหรับกลุ่ม C_2 จนถึง C_k ทำการกำหนดเงื่อนไขเพื่อสร้างกฎจากชุดข้อมูลเติบโตสำหรับการแยกกลุ่มปกติกับผิดปกติ แต่ละกฎที่สร้างโดยวิธีกฎของริบเปอร์ ประกอบด้วยการรวมกันของเงื่อนไข เงื่อนไขจะอยู่ในรูป $A_n = v, A_c \leq t$ หรือ $A_c \geq t$ โดยที่ A_n เป็น แอททริบิวต์แบบนาม (Nominal Attribute) และ v เป็นค่ากำหนดกฎ (Legal Value) สำหรับ A_c เป็นแอททริบิวต์แบบตัวเลข (Continuous numerical attribute) และ t เป็นค่าใดๆ สำหรับ A_c กฎโดยทั่วไปจะครอบคลุมทั้งข้อมูลที่เป็นกลุ่มบวกและลบ เมื่อเงื่อนไขถูกเพิ่มขึ้นมาเรื่อยๆ

การครอบคลุมจะเฉพาะเจาะจงข้อมูลที่เป็นกลุ่มปกติหรือกลุ่มผิดปกติมากขึ้น เมื่อเงื่อนไขถูกเพิ่มขึ้นมาเรื่อยๆ การครอบคลุมจะเฉพาะเจาะจง ข้อมูลที่เป็นกลุ่มบวกหรือกลุ่มลบมากขึ้น แต่ครั้งของการเพิ่มเงื่อนไขที่ให้ Foil's Information Gain สูงสุด ตามสมการที่ 1 กฎก็จะสมบูรณ์ขึ้น

$$Gain(R', R) = s \cdot (\log_2 \frac{N'+}{N'} - \log_2 \frac{N+}{N}) \quad (1)$$

เมื่อ

R กฎเดิม

R' กฎใหม่หลังจากมีการเพิ่มเงื่อนไข

N จำนวนของ instance ที่ถูกรอบคลุมโดยกฎ R

N' จำนวนของ instance ที่ถูกรอบคลุมโดยกฎ R'

$N+$ จำนวนข้อมูลกลุ่มปกติที่ถูกรอบคลุมโดยกฎ R

$N'+$ จำนวนข้อมูลกลุ่มปกติที่ถูกรอบคลุมโดยกฎ R'

s จำนวนข้อมูลปกติที่ถูกรอบคลุมทั้งกฎ R และ R' หลังจากการเพิ่ม

เมื่อมีการสร้างกฎขึ้นมาจะมีการตัดกฎทันที เพื่อเป็นการป้องกันข้อมูลมากเกินไป (Overfitting) กฎจะถูกประเมินโดยชุดข้อมูลการตัดกฎ ตามเงื่อนไข v^* ซึ่งค่า v^* คือ เมตริกซ์ที่ใช้พิจารณาการตัดกฎ แสดงในสมการที่ 2

$$v^*(Rule, prunePos, pruneNeg) = \frac{p - n}{p + n} \quad (2)$$

โดย p เป็นจำนวนข้อมูลกลุ่มบวกที่ถูกรอบคลุมโดยกฎ และ n เป็นจำนวนข้อมูลกลุ่มลบที่ถูกรอบคลุมโดยกฎ การตัดกฎจะพิจารณาตัดเงื่อนไขเพื่อเพิ่มค่าของ v^* และจะหยุดตัดเมื่อได้ค่า v^* สูงสุด

กระบวนการจะวนซ้ำตั้งแต่การสร้างกฎและตามด้วยการตัดกฎ ในกฎของริบเปอร์ การวนซ้ำจะหยุดเมื่อพบจุดที่เหมาะสมที่สุดโดยพิจารณาจาก Minimal Description Length (MDL) เมื่อ Description Length (DL) ยาวเกิน DL ของกลุ่มกฎที่เล็กที่สุดการวนซ้ำจะหยุด ต่างจากวิธี

IREP ที่หยุดกระบวนการเมื่อค่าความคลาดเคลื่อน (Error) เกิน 50 เปอร์เซ็นต์ สำหรับการแบ่งคลาสที่มีมากกว่าสองกลุ่ม ก็จะทำการแยกออกมาทีละกลุ่มจนกว่าจะครบกลุ่มที่ต้องการ

2.2 กฎของเบย์ (Bayes's Rule)

เทคนิคแบบนาอิวเบย์เซียน (Naïve Bayesian) ทำนายผลโดยหลักการของการจำแนกประเภทโดยใช้ทฤษฎีของเบย์ (Bayes Theorem) การเรียนรู้เบย์อย่างง่าย (Naïve-Bayes) เป็นวิธีจำแนกประเภทข้อมูลที่มีประสิทธิภาพวิธีหนึ่ง โดยที่ใช้งานได้ดี เหมาะกับกรณีของเซตตัวอย่างมีจำนวนมากและคุณสมบัติ (Attribute) ของตัวอย่างไม่ขึ้นต่อกันมีการจำแนกประเภทเบย์อย่างง่ายไปประยุกต์ใช้งานในด้านการจำแนกประเภทข้อความ (Text Classification) การวินิจฉัย (Diagnosis) และพบว่าใช้งานได้ดีไม่ต่างจากการจำแนกประเภทวิธีการอื่นทำให้ผู้วิจัยเลือกวิธีการนี้มาใช้ในงานวิจัย เนื่องจากเป็นวิธีการจำแนกข้อมูลได้อย่างมีประสิทธิภาพและมีอัลกอริทึมในการทำงานที่ไม่ซับซ้อนเหมือนวิธีการอื่นๆ โดยความน่าจะเป็นของกลุ่มข้อมูลที่จะเป็นกลุ่ม C_i สำหรับข้อมูลที่มีคุณสมบัติ n ตัว $X = \{ A_1, A_2, \dots, A_n \}$ หรือ $P(C_i | A_1, \dots, A_n)$ จากกฎของเบย์ (Bayes' Rule) จะได้สมการที่ 3

$$P(C_i | A_1, \dots, A_n) = \frac{\prod_{j=1}^n P(A_j | C_i) P(C_i)}{P(A_1, \dots, A_n)} \quad (3)$$

โดยหา $P(C_i | A_1, \dots, A_n)$ จากสมการสำหรับทุกๆกลุ่ม i , คุณสมบัติตัวที่ j แล้วนำค่าที่ได้มาเปรียบเทียบกับกลุ่มที่มีค่าความน่าจะเป็นมากที่สุดคือคำตอบ เทคนิคแบบนาอิวเบย์เซียน ยังไม่รองรับข้อมูลที่เป็นข้อมูลต่อเนื่อง (Continuous Data) ด้วย ดังนั้นตัวแปรอิสระหรือตัวแปรตามที่มีค่าเป็นค่าต่อเนื่องจะต้องถูกแบ่งเป็นช่วง ถ้าแบ่งไม่เหมาะสมก็จะมีผลต่อคุณภาพของแบบจำลองที่สร้างขึ้น แต่ถ้าไม่คำนึงถึงข้อจำกัดนี้แล้ว จะสามารถให้ผลลัพธ์ที่ดีและรวดเร็วได้ ความง่ายและความเร็วทำให้เทคนิคนี้เป็นเครื่องมือที่ดีในการสร้างแบบจำลองและหารูปแบบความสัมพันธ์ที่ไม่ซับซ้อน

• **Naïve_Bayes_Learn(examples)**
FOR EACH *target value v* **DO**
 $\overline{P}(v_j) \leftarrow \text{estimate } P(v_j)$

FOR EACH *attribute value a of each attribute* **DO**
 $\overline{P}(a_i | v_j) \leftarrow \text{estimate } P(a_i | v_j)$

• **classify_New_Example(x)**

$$V_{NB} = \underset{v_j \in V}{\operatorname{argmax}} \overline{P}(v_j) \times \prod_{i=1}^n \overline{P}(a_i | v_j)$$

ภาพที่ 2 อัลกอริทึมสำหรับการจำแนกประเภทข้อความโดยใช้การเรียนรู้แนอ์เบย์

2.3 ต้นไม้ตัดสินใจ (Decision Tree)

ต้นไม้ตัดสินใจนับว่าเป็นวิธีการเรียนรู้ที่ใช้มากที่สุดในกระบวนการเรียนรู้ของเครื่อง ต้นไม้ตัดสินใจสามารถเรียนรู้โดยการแยกแยะข้อมูลในกลุ่มตัวอย่างออกเป็นกลุ่มย่อยต่างๆ โดยที่ผลลัพธ์ของข้อมูลในกลุ่มย่อยแต่ละกลุ่มเป็นอย่างเดียวกัน การแยกแยะข้อมูลออกเป็นกลุ่มย่อยนี้ทำได้โดยใช้คุณสมบัติ (Attributes) ของข้อมูลเป็นตัวแยกแยะ ต้นไม้ตัดสินใจที่ได้จากการเรียนรู้ทำให้ทราบว่า คุณสมบัติไหนของข้อมูลที่เป็นตัวกำหนดผลลัพธ์ และคุณสมบัติแต่ละตัวของข้อมูลมีความสำคัญมากน้อยต่างกันอย่างไร ซึ่งเป็นประโยชน์ช่วยให้ผู้ใช้สามารถวิเคราะห์ข้อมูลและตัดสินใจได้ถูกต้องยิ่งขึ้นการสร้างต้นไม้ตัดสินใจจะทำจากบนลงล่าง (Top-Down) โดยเริ่มต้นด้วยการเลือกคุณสมบัติให้เป็นโหนด คุณสมบัติที่ถูกเลือกมาเป็นโหนดจะเป็นคุณสมบัติที่เมื่อแยกตัวอย่างเป็นกลุ่มย่อยแล้วทำให้สมาชิกส่วนใหญ่ในแต่ละกลุ่มย่อยมีผลลัพธ์เป็นอย่างเดียวกันมากที่สุดในการวัดความสามารถในการแยกตัวอย่างค่าเกน (Gain) ของแต่ละโหนดนั้น ทำได้โดยอาศัยทฤษฎีสารสนเทศ (Information Theory) ที่นำค่าเอนโทรปี (Entropy) มาใช้เป็นตัววัดความไร้ระเบียบของข้อมูล

โดยหลักการพื้นฐานของการสร้างต้นไม้ตัดสินใจ เป็นการสร้างในลักษณะจากบนลงล่าง (Top-Down) คือเริ่มจากการสร้างรากของต้นไม้ก่อนแล้วจึงแตกกิ่งไปจนถึงใบ โดยแสดงขั้นตอนการสร้างต้นไม้ตัดสินใจได้ดังนี้ Han and Kamber (2001)

- 1) ต้นไม้เริ่มต้นโดยมีโหนดเพียงโหนดเดียวแสดงถึงชุดข้อมูลฝึก (Training Set)
- 2) ถ้าข้อมูลทั้งหมดอยู่ในกลุ่มเดียวกันแล้ว ให้โหนดนั้นเป็นใบและตั้งชื่อแยกตามกลุ่มของข้อมูลนั้น
- 3) ถ้าในโหนดมีข้อมูลหลายกลุ่มปะปนอยู่ จะต้องวัดค่าเกน (Gain) ของแต่ละแอททริบิวต์เพื่อที่จะใช้เป็นเกณฑ์ในการคัดเลือกแอททริบิวต์ ที่มีความสามารถในการแบ่งแยกข้อมูลออกเป็นกลุ่มต่างๆ ได้ดีที่สุด โดยแอททริบิวต์ที่มีค่าเกนมากที่สุดจะถูกเลือกให้เป็นตัวทดสอบหรือแอททริบิวต์ที่ใช้ในการตัดสินใจ โดยแสดงในรูปของโหนดบนต้นไม้
- 4) กิ่งของต้นไม้ ถูกสร้างขึ้นจากค่าต่างๆ ที่เป็นไปได้ของโหนดทดสอบ และข้อมูลจะถูกแบ่งออกตามกิ่งต่างๆ ที่สร้างขึ้น
- 5) ทำการวนซ้ำเพื่อหาแอททริบิวต์ที่มีค่าเกนมากที่สุด สำหรับข้อมูลที่ถูกแบ่งแยกออกมาในแต่ละกิ่งเพื่อนำแอททริบิวต์นี้มาสร้างเป็นโหนดตัดสินใจต่อไป โดยที่แอททริบิวต์ที่ถูกเลือกมาเป็นโหนดแล้วจะไม่ถูกเลือกใช้อีก สำหรับโหนดในระดับต่อ ๆ ไป
- 6) ทำการวนซ้ำเพื่อแบ่งข้อมูลและแตกกิ่งของต้นไม้ไปเรื่อยๆ โดยการวนซ้ำจะสิ้นสุดก็ต่อเมื่อเงื่อนไขข้อใดข้อหนึ่งต่อไปนี้จริง

ต้นไม้ตัดสินใจ (Decision Tree) ต้นไม้ตัดสินใจเป็นโครงสร้างที่ใช้แสดงกฎที่ได้จากเทคนิคการจำแนกประเภทข้อมูล โดยต้นไม้ตัดสินใจจะมีลักษณะคล้ายโครงสร้างต้นไม้ ที่แต่ละโหนดแสดงคุณลักษณะ (Attribute) ในการสร้างต้นไม้ตัดสินใจ ปัญหาสำคัญที่ต้องพิจารณาคือ ควรจะตัดสินใจเลือกแอททริบิวต์ใดมาทำหน้าที่เป็นโหนดราก ในแต่ละขั้นตอนของการสร้างต้นไม้และต้นไม้ย่อย (Subtree) ของต้นไม้ตัดสินใจ เกณฑ์ที่ใช้ช่วยประกอบการเลือกแอททริบิวต์คือการคำนวณค่ามาตรฐานเกน (Gain Criterion) ซึ่งเป็นค่าที่บ่งบอกว่าแอททริบิวต์นั้นสามารถจำแนกกลุ่มของข้อมูลได้ดีเพียงใด โดยทดลองเลือกแต่ละแอททริบิวต์ที่เป็นไปได้จากชุดข้อมูลมาทำหน้าที่เป็นโหนดราก ถ้าแอททริบิวต์ใดให้ค่าเกนสูงที่สุด แสดงว่าแอททริบิวต์นั้นสามารถจำแนกกลุ่มของข้อมูลได้ดีที่สุด การใช้ค่า Information Gain จะช่วยลดจำนวนครั้งของการทดสอบในการแยกแยะข้อมูลอีกทั้งยังรับประกันว่าต้นไม้ตัดสินใจที่ได้ไม่มีความซับซ้อนมากเกินไป

ซึ่งค่า Information Gain นั้นสามารถคำนวณได้จากสมการดังสมการที่ 4

$$I(s_1, s_2, \dots, s_n) = - \sum_{i=1}^n \frac{s_i}{s} \log_2 \frac{s_i}{s} \quad (4)$$

เมื่อ

S เป็นเซตของข้อมูลซึ่งประกอบด้วยข้อมูล s เรคคอร์ด

n เป็นจำนวนกลุ่มทั้งหมดที่ต่างกันของข้อมูลชุดนั้น

C_i แทนกลุ่มในลำดับ ที่ i โดย ที่ i มีค่าระหว่าง 1 ถึง n

s_i แทนจำนวนข้อมูลที่เป็นสมาชิกของ S และอยู่ในกลุ่ม C_i

s_{ij} แทนจำนวนข้อมูลที่เป็นสมาชิกของ S ในกลุ่ม C_i จากการแบ่งข้อมูลด้วยค่าที่เป็นไปได้ ของแอททริบิวต์ A

j ค่าระหว่าง 1 ถึง v

ค่าเอ็นโทรปีของแอททริบิวต์ A ซึ่งมีค่าของแอททริบิวต์เป็น $(a_1, a_2, a_3, \dots, a_v)$

หาได้ในสมการที่ 5

$$E(A) = \sum_{j=1}^v \frac{s_{1j} + \dots + s_{nj}}{s} I(s_{1j}, s_{2j}, \dots, s_{nj}) \quad (5)$$

ดังนั้นจะสามารถพิจารณาค่ามาตรฐานเกณฑ์ดังสมการที่ 6

$$Gain(A) = I(s_1, s_2, \dots, s_n) - E(A) \quad (6)$$

อัลกอริทึม C4.5 เป็นอัลกอริทึมพัฒนาโดย Quinlan (1993) ที่ได้พัฒนาต่อมาจากอัลกอริทึม ID3 โดยได้เพิ่มการใช้ค่ามาตรฐานอัตราส่วนเกณฑ์ (Gain Ratio Criterion) ในการตัดสินใจเลือกแอท-ทริบิวต์ ที่จะใช้เป็นโหนดรากด้วย เนื่องจากค่ามาตรฐานเกณฑ์จะมีความลำเอียงกับข้อมูลที่ประกอบด้วยแอททริบิวต์ที่มีค่าที่เป็นไปได้จำนวนมากๆ โดยการปรับค่าเกณฑ์ให้ถูกต้อง โดยใช้ค่าสารสนเทศการแบ่งแยก (Split Information) โดยหาได้ในสมการที่ 7 ของแต่ละแอททริบิวต์เพื่อใช้คำนวณค่ามาตรฐานอัตราส่วนเกณฑ์

$$\text{ค่าสารสนเทศการแบ่งแยก} = - \sum_{i=1}^v \frac{|t_i|}{|T|} \log_2 \left[\frac{|t_i|}{|T|} \right] \quad (7)$$

โดยให้ T แทนชุดของข้อมูลเรียนรู้ $\{T_1, T_2, \dots, T_v\}$ เป็นชุดของตัวอย่างย่อยในแต่ละกิ่งจำนวน v ชุด ค่าสารสนเทศการแบ่งแยกนี้จะแสดงถึงระดับการกระจายของข้อมูล เมื่อนำไปหารค่าเกณฑ์จะได้ค่ามาตรฐานส่วนเกิน แสดงในสมการที่ 8 ซึ่งช่วยแก้ปัญหาเรื่องการลำเอียงของค่ามาตรฐานเกินได้จะได้ว่า

$$\text{ค่ามาตรฐานอัตราส่วนเกิน} = \frac{\text{ค่ามาตรฐานเกิน}}{\text{ค่าสารสนเทศการแบ่งแยก}} \quad (8)$$

- (1) create a node N ;
- (2) if *samples* are all of the same class C then
- (3) return N as a leaf node labeled with the class C ;
- (4) if *attribute-list* is empty then
- (5) return N as a leaf node labeled with the most common class in *samples*; // majority voting
- (6) select *test-attribute*, the attribute among *attribute-list* with the highest information gain;
- (7) label node N with *test-attribute*;
- (8) for each known value a_i of *test-attribute*
- (9) grow a branch from node N for the condition *test-attribute* = a_i ;
- (10) let s_i be the set of samples in *samples* for which *test-attribute* = a_i ;
- (11) if s_i is empty then
- (12) attach a leaf labeled with the most common class in *samples*;
- (13) else attach the node returned by Generate decision tree(s_i , *attribute-list*, *test-attribute*);

ภาพที่ 3 อัลกอริทึมพื้นฐานในการหาต้นไม้ตัดสินใจ

อัลกอริทึมจะคำนวณค่าเอ็นโทรปีสำหรับแต่ละแอทริบิวต์ แอทริบิวต์ที่มีค่าเอ็นโทรปีสูงสุดจะถูกเลือกให้เป็นแอทริบิวต์ทดสอบในเซต s จากนั้นจะสร้างโหนด และกิ่งที่แสดงค่าในแอทริบิวต์ และแบ่งข้อมูลต่อไปตามลำดับ

การตัดกิ่งต้นไม้ (Pruning) ในการสร้างต้นไม้ตัดสินใจ ปัญหาการเกิดข้อมูลมากเกินไป คือการเข้ากันเกินไปของชุดข้อมูลเรียนรู้ที่มีค่าความถูกต้องสูง แต่เมื่อนำไปใช้กับชุดข้อมูลทดสอบแล้วได้ค่าความถูกต้องออกมาน้อย ดังนั้นจึงต้องมีการตัดเล็มเพื่อให้ได้ผลที่ดี การตัดเล็มต้นไม้อาจแบ่งได้สองประเภทคือ การตัดกิ่งขณะเรียนรู้ (Pre-pruning) และการตัดกิ่งหลังการเรียนรู้ (Post-pruning) จากต้นไม้ถูกสร้างเสร็จแล้ว

3. การตัดกิ่งต้นไม้ (Pruning)

ในขณะที่กำลังสร้างต้นไม้ตัดสินใจในแต่ละกิ่งอาจเกิดการสร้างอย่างผิดพลาด อันเนื่องมาจากข้อมูลเรียนรู้ที่มีข้อมูลรบกวน (Noise) ซึ่งเกิดจากการบันทึกข้อมูลผิดพลาดหรือความผิดพลาดที่เกิดจากระบบเอง หรือในชุดข้อมูลอาจจะมีข้อมูลที่ผิดปกติจากข้อมูลส่วนใหญ่ (Outlier) ปะปนมาด้วย การตัดกิ่งต้นไม้ตัดสินใจเป็นเทคนิคที่ใช้สำหรับแก้ปัญหาเหล่านี้ และช่วยลดการเกิดปัญหาการเจาะจงโมเดลกับข้อมูลมากเกินไปได้ โดยปัญหานี้ทำให้ได้โครงสร้างต้นไม้ที่สามารถจำแนกข้อมูลได้ดีกับชุดข้อมูลที่ใช้สร้างต้นไม้ตัดสินใจเท่านั้น แต่เมื่อนำไปใช้กับข้อมูลใหม่ ประสิทธิภาพในการจำแนกกลุ่มข้อมูลจะลดลง การตัดกิ่งต้นไม้ตัดสินใจจะใช้ค่าทางสถิติในการตัดกิ่งที่มีความน่าเชื่อถือน้อยที่สุดออกไป เพื่อให้ต้นไม้ใหม่ที่ได้สามารถทำงานได้รวดเร็วขึ้นและยังเป็นการปรับปรุงขีดความสามารถของต้นไม้ในการทำนายข้อมูลใหม่ๆ ได้แม่นยำมากยิ่งขึ้นอีกด้วย โดยการตัดกิ่งต้นไม้ตัดสินใจที่เป็นที่นิยมมีอยู่ 2 ประเภท ดังต่อไปนี้

3.1 การตัดกิ่งขณะเรียนรู้ (Pre-pruning)

เป็นการตัดกิ่งต้นไม้หรือหยุดการแตกกิ่งในขั้นตอนการสร้างต้นไม้ตัดสินใจโดยการทำให้โหนดที่ถูกตัดนั้นเปลี่ยนเป็นใบ และให้ใบนั้นแสดงกลุ่มที่มีจำนวนข้อมูลสนับสนุนหรือมีความน่าจะเป็นที่ข้อมูลจะอยู่ในกลุ่มนั้นมากที่สุด (Breslow and Aha, 1997)

ในขณะที่ทำการสร้างต้นไม้ตัดสินใจนั้น ต้องมีการคำนวณ หรือวัดค่าทางสถิติที่สำคัญต่างๆ เพื่อใช้ประเมินว่าควรที่จะสร้างหรือแตกกิ่งของต้นไม้อย่างไร ถ้าค่าที่วัดได้ไม่ถึงจุดที่กำหนดไว้ก็จะถือว่าโหนดนั้นไม่สมควรที่จะทำการแตกกิ่งต่อไป ซึ่งเป็นการยากที่จะกำหนดว่าค่าที่ใช้เป็นเกณฑ์เหล่านั้นควรจะมีค่าเป็นเท่าไร ถ้ากำหนดค่านั้นสูงเกินไปก็จะทำให้ได้ต้นไม้ที่มีความซับซ้อน แต่ถ้ากำหนดค่าต่ำเกินไปก็จะทำให้ได้ต้นไม้ที่มีขนาดเล็กจนไม่สามารถนำไปใช้งานได้

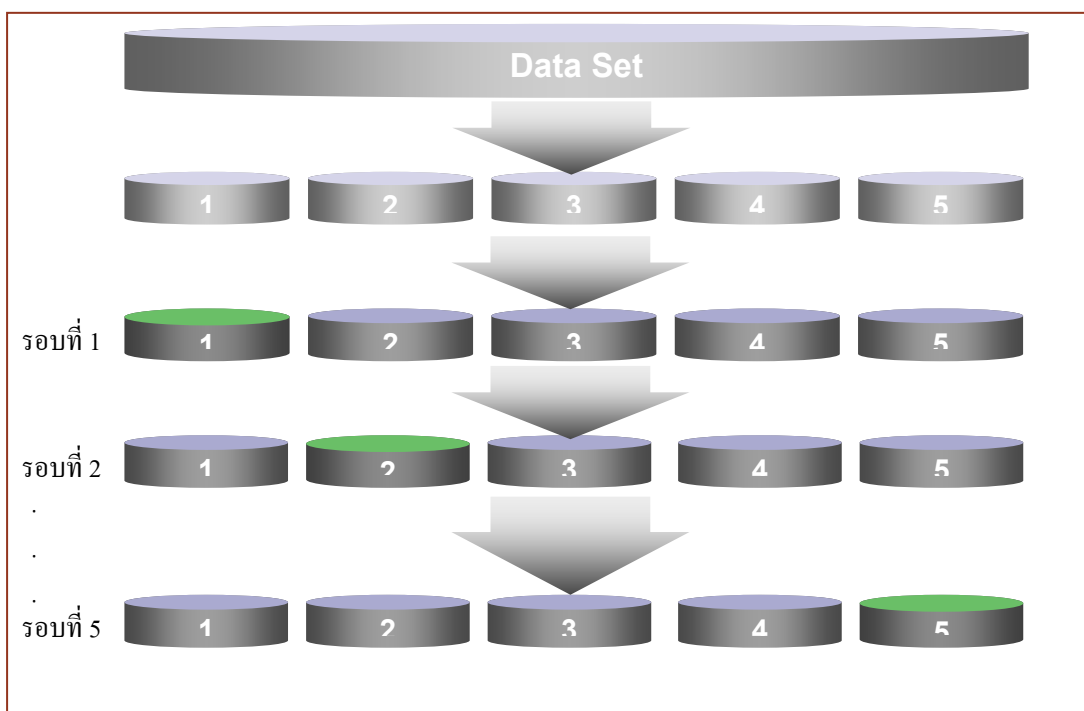
3.2 การตัดกิ่งหลังการเรียนรู้ (Post-pruning)

เป็นการตัดกิ่งของต้นไม้ตัดสินใจที่ถูกสร้างขึ้นสมบูรณ์แล้ว โดยใช้การวัดค่าความซับซ้อนของแต่ละโหนด หลังจากทำการตัดกิ่งของต้นไม้แล้วโหนดที่อยู่ล่างสุดที่ไม่ได้ถูกตัดจะถูกเปลี่ยนไปเป็นใบและจะแสดงกลุ่มที่มีจำนวนข้อมูลสนับสนุนมากที่สุด Breslow and Aha (1997) สำหรับทุกๆ โหนดที่ไม่ใช่ใบของต้นไม้ จะมีการคำนวณค่าอัตราความผิดพลาดที่คาดหวังไว้ ซึ่งเป็นค่าที่แสดงถึงความผิดพลาดที่จะเกิดขึ้น ถ้าโหนดของต้นไม้ย่อนั้นถูกตัดออกไปโดยที่ค่าความผิดพลาดของโหนดที่ไม่ถูกตัดจะถูกคำนวณโดยใช้ค่าผลรวมความผิดพลาดของแต่ละกิ่ง และให้ค่าน้ำหนักตามสัดส่วนของกิ่งนั้นๆ ถ้าการตัดโหนดนั้นนำไปสู่การเกิดความผิดพลาดที่สูงขึ้น โหนดของต้นไม้ย่อนั้นก็จะต้องยังคงไว้ แต่ถ้าการตัดนั้นทำให้ได้ค่าความผิดพลาดเป็นที่ยอมรับได้ โหนดนั้นก็จะถูกตัดออกไป หลังจากทำการตัดกิ่งต้นไม้ตัดสินใจแล้วจะต้องทำการวัดค่า ความแม่นยำ (Accuracy) ของต้นไม้ที่ทำการตัดกิ่งแล้วด้วย โดยที่ต้นไม้ที่ให้ค่าความผิดพลาดน้อยที่สุดจะถูกเลือก นอกจากการตัดกิ่งต้นไม้ตัดสินใจโดยอาศัยการวัดค่าความผิดพลาดที่จะเกิดขึ้นแล้ว ยังมีเทคนิคอื่น ๆ เช่น การใช้ค่าการเข้ารหัส (Encode) ในการพิจารณาตัดกิ่งของต้นไม้โดยใช้หลักการ ของ Minimum Description Length (MDL) ด้วย Quinlan and Rivest (1989) เป็นต้น

4. การตรวจสอบ (Cross Validation)

การตรวจสอบ เป็นวิธีการในการคาดเดาค่าความผิดพลาดของแบบจำลอง หรือการสุ่มตัวอย่าง (Resampling) โดยเริ่มจากแบ่งชุดข้อมูลออกเป็นส่วนๆ และนำบางส่วนจากชุดนั้นมาตรวจสอบ ผลลัพธ์จากการทำตรวจสอบมักถูกใช้เป็นตัวเลือกในการกำหนดแบบจำลอง

ในกรณีการทำ K -fold Cross Validation เราจะแบ่งข้อมูลออกเป็น K ชุดเท่าๆ กัน และทำการคำนวณค่าความผิดพลาด K รอบ โดยแต่ละรอบการคำนวณข้อมูลชุดหนึ่งจากข้อมูล K ชุด จะถูกเลือกออกมาเพื่อเป็นข้อมูลทดสอบ และข้อมูลอีก $K-1$ ชุดจะถูกใช้เป็นชุดสำหรับการเรียนรู้ ดังภาพที่ 4



ภาพที่ 4 การแบ่งชุดข้อมูลเพื่อคำนวณค่าความผิดพลาดโดยวิธีครอสวาไลเดชั่น

5. การทำดิสครีไทซ์ (Discretization)

คือการแบ่งค่าต่อเนื่องออกเป็นช่วง เพื่อเป็นการลดจำนวนลำดับชั้นไม่ให้มีมากจนเกินไป ซึ่งโดยทั่วไปวิธีการทำดิสครีไทซ์ มีดังนี้

- วิธีแบ่งเก็บในถัง (Binning Discretization)
- วิธีใช้กราฟแสดงค่าทางสถิติ (Histogram based Discretization)
- วิธีใช้เอ็นโทรปี (Entropy based Discretization)

5.1 วิธีแบ่งเก็บในถัง แบ่งออกเป็น 2 วิธีดังนี้

5.1.1 วิธีแบ่งด้วยความกว้างที่เท่ากัน (Equal-width partitioning)

การทำดิสครีไทซ์ วิธีนี้เป็นวิธีการที่ง่ายที่สุดและพื้นฐานที่สุด โดยเป็นการแบ่งช่วงทั้งหมดของข้อมูลออกเป็นช่วง และแต่ละช่วงนั้นกินพื้นที่ขนาดเท่าๆ กัน (ใช้ระยะห่างเป็น

เกณฑ์) ออกเป็นจำนวน n ช่วง แล้วเวลาเลือกข้อมูลมาว่าจะใส่ถึงไหนให้ดูว่า ข้อมูลนั้นอยู่ตรงกับช่วงของถึงใดก็นำไปใส่ถึงนั้น

5.1.2 วิธีแบ่งด้วยความสูงที่เท่ากัน (Equal-height Partitioning)

การทำดิสครีไทซ์ วิธีนี้ เรียกอีกอย่างหนึ่งว่าเป็นการแบ่งด้วยความถี่ที่เท่ากัน ซึ่งคล้ายกับวิธีแรก แต่จะช่วยแก้ปัญหาการกระจุกตัวของข้อมูลได้ โดยเป็นการแบ่งช่วงทั้งหมดของข้อมูลออกเป็นช่วง และแต่ละช่วงนั้นมีจำนวนสมาชิกเท่าๆ กัน (ใช้ความถี่เป็นเกณฑ์) ออกเป็นจำนวน n ช่วง ซึ่งทั้ง n ช่วงไม่จำเป็นต้องมีความกว้างเท่ากันก็ได้

5.2 วิธีดิสครีไทซ์ โดยใช้กราฟแสดงค่าทางสถิติ

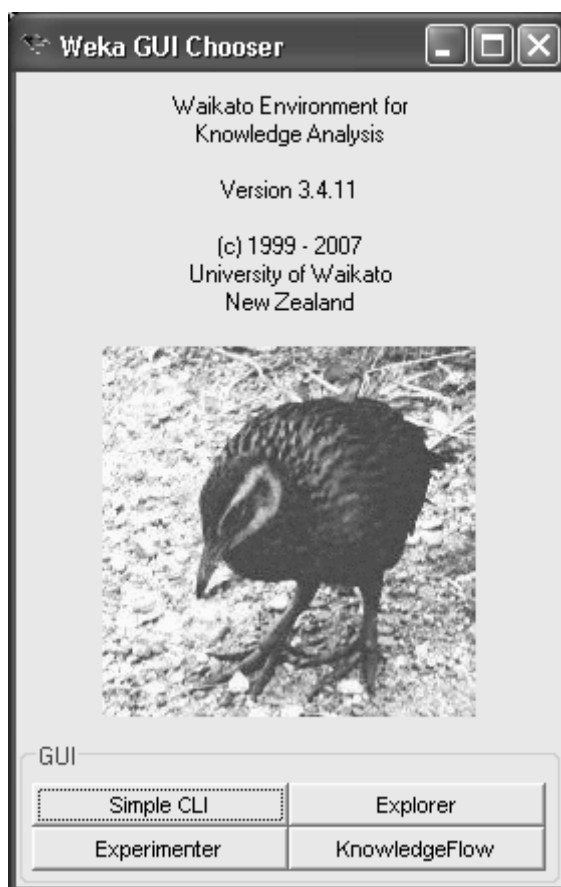
Wu *et al.* (1975) เป็นเทคนิคการลดข้อมูลที่ได้รับการนิยมนอย่างมาก โดยวิธีการทำก็จะทำการแบ่งข้อมูลแบบกลุ่ม ข้อมูลจะถูกแบ่งเป็นช่วงตามความต้องการ โดยอาศัยกราฟแสดงค่าทางสถิติ (Histogram) ช่วยในการแบ่งช่วงข้อมูล การกระจายของความน่าจะเป็นที่แอทริบิวต์ ที่เป็นคลาสใดๆ จะถูกคำนวณ จากนั้นจะได้กราฟแสดงการกระจายของความน่าจะเป็น และใช้วิธีหาค่าความผิดพลาดกำลังสองที่ต่ำที่สุด (Minimum Square Error) ช่วยในการหาสัมประสิทธิ์ ซึ่งใช้สำหรับการหาจุดแบ่งคลาส

5.3 วิธีดิสครีไทซ์ โดยใช้เอ็นโทรปี

Fayyad and Irani (1993) วิธีนี้ใช้กับอัลกอริทึมบางประเภท ได้แก่ ID3 หรือ C4.5 ซึ่งใช้หลักการหาค่าเอ็นโทรปีที่น้อยที่สุดเป็นจุดแบ่งสำหรับการดิสครีไทซ์หรือการแบ่งค่าต่อเนื่องของข้อมูลที่เป็นค่าต่อเนื่อง วิธีนี้เป็นการหาจุดแบ่งสองช่วงในแต่ละแอทริบิวต์ รวมไปถึงการแบ่งเป็นหลายช่วงในแต่ละแอทริบิวต์ด้วยเช่นกัน ส่วนเกณฑ์ที่ใช้สำหรับการหยุดการแบ่งช่วงนั้น ใช้หลักการ Minimal Description Length (MDL)

6. ซอฟต์แวร์ Weka

Weka ย่อมาจาก Waikato Environment for Knowledge Analysis ซึ่งเป็นซอฟต์แวร์สำเร็จภาพประกอบประเภทซอฟต์แวร์ให้เปล่า (Free Ware) ที่สามารถดาวน์โหลดได้จากเว็บไซต์ ซึ่งอยู่ภายใต้การควบคุมของ GPL License ซึ่งโปรแกรม Weka ได้ถูกพัฒนามาจากภาษาจาวาทั้งหมด ซึ่งเขียนมาโดยเน้นกับงานทางด้านการเรียนรู้ด้วยเครื่อง (Machine Learning) และ การทำเหมืองข้อมูลโปรแกรมประกอบไปด้วยโมดูลย่อยๆ สำหรับใช้ในการจัดการข้อมูล และเป็นโปรแกรมที่สามารถใช้ Graphic User Interface (GUI) และ ใช้คำสั่งในการให้ซอฟต์แวร์ประมวลผลดังภาพที่ 5



ภาพที่ 5 หน้าจอหลักของโปรแกรม Weka Explore

โปรแกรมหลักของซอฟต์แวร์ Weka

- Simple CLI (Command Line Interface) เป็นโปรแกรมรับคำสั่งการทำงานผ่านการพิมพ์
- Explorer เป็นโปรแกรมที่ออกแบบในลักษณะ GUI
- Experimenter เป็นโปรแกรมที่ออกแบบการทดลองและการทดสอบผล
- KnowledgeFlow เป็นโปรแกรมออกแบบผังการไหลของความรู้
- ArffViewer เป็นโปรแกรมที่ใช้สำหรับแก้ไขเพิ่มประเภท Arff
- Log เป็นโปรแกรมที่ใช้อ่านข้อความบันทึกเก็บระหว่างการทำงาน

เมนูหลักของ Explorer

- Preprocess การเตรียมข้อมูล
- Classify รวมโมดูลการทำเหมืองข้อมูลแบบจัดจำแนกประเภท
- Cluster รวมโมดูลการทำเหมืองข้อมูลแบบการเกาะกลุ่ม
- Associate รวมโมดูลการทำเหมืองข้อมูลแบบกฎเชื่อมโยง
- Select attributes รวมโมดูลสำหรับการวิเคราะห์ความเกี่ยวพันของลักษณะประจำ
- Visualize นำเสนอข้อมูลด้วยภาพนามธรรมสองมิติ

ประเภทของแฟ้มข้อมูลที่ได้รับได้

- แฟ้มข้อมูลที่ต้องอยู่ในรูปแบบ ASCII อาจเป็น arff, csv, C45
- ในกรณีแฟ้มข้อมูลอยู่ในเครือข่ายผู้ใช้สามารถเรียกใช้โดยอาศัย URL
- หรืออาจใช้ข้อมูลที่อยู่ในฐานข้อมูลที่เชื่อมโยงผ่าน JDBC

ในการศึกษาอัลกอริทึม C4.5 ของระบบ Weka ซึ่งพัฒนาด้วยภาษา Java แบ่งออกเป็น 2 ส่วนด้วยกัน คือ การแยกเอาเฉพาะคลาสที่เกี่ยวข้องกับอัลกอริทึม C4.5 เท่านั้นออกมาจากทั้งระบบของ Weka เพื่อศึกษาโครงสร้างของอัลกอริทึม และอีกส่วนก็คือการศึกษารายละเอียดภายในซอร์สโค้ดของอัลกอริทึม และเพื่อให้ง่ายต่อการศึกษารายละเอียดจึงเลือกใช้ Eclipse เวอร์ชัน 2.1.3 ซึ่งเป็นแพลตฟอร์มที่พัฒนาขึ้นแบบ Open Source ใช้สำหรับออกแบบและพัฒนา Java

Application และ Web-based Application ที่มีประสิทธิภาพ นำมาเป็นเครื่องมือช่วยในการศึกษา โครงสร้างและการทำงานของอัลกอริทึมที่ใช้ในการสร้างต้นไม้ตัดสินใจ

7. ตัววัดความน่าเชื่อถือ

เมื่อสร้างแบบจำลองเสร็จแล้ว จะต้องวัดความน่าเชื่อถือของแบบจำลองที่เราสร้าง โดยพิจารณาจากค่าทางสถิติที่ได้จากการวิเคราะห์และทดสอบดังนี้

7.1 Confusion Matrix เป็นส่วนที่ทำให้เห็นรายละเอียดของผลการทำนายดังภาพที่ 6

โดยที่

- a, d คือจำนวนข้อมูลที่มีในการทำนายถูก

- b, c คือจำนวนข้อมูลที่มีในการทำนายผิด

		<i>PREDICTED CLASS</i>	
		<i>Class=Yes</i>	<i>Class=No</i>
<i>ACTUAL CLASS</i>	<i>Class=Yes</i>	<i>a</i>	<i>b</i>
	<i>Class=No</i>	<i>c</i>	<i>d</i>

ภาพที่ 6 ลักษณะ Confusion Matrix แสดงผลการวิเคราะห์และทดสอบ

7.2 Correctly Classified Instances เป็นส่วนที่บอกว่ามีการทำนายข้อมูลถูกต้องค่าความแม่นยำในการทำนายคำนวณได้จากสมการที่ 9

$$\text{Correctly Classified Instances} = \frac{(a+d)}{(a+b+c+d)} \quad (9)$$

7.3 True Positive rate (TP rate) เป็นการดูว่าข้อมูลที่เป็นผลจากการทำนายด้วยแบบจำลอง Decision Tree ที่สร้างขึ้นมาแล้วให้คำตอบตรงกับความเป็นจริง คือ จำนวนข้อมูลที่ตอบถูกในคลาสนั้น/จำนวนข้อมูลทั้งหมดในคลาสนั้นสมการที่ 10

$$\text{true positive (TP) rate} = \frac{a}{(a+d)} \quad (10)$$

7.4 False Positive rate (FP rate) ก็จะคล้ายๆ กับ TP rate แต่การคิดจะไม่เหมือนกัน False Positive คือ ข้อมูลที่ไม่ได้อยู่ในคลาสนั้นแต่คำตอบจากโมเดลบอกว่าอยู่ในคลาสนั้นสมการที่ 11

$$\text{false positive (FP) rate} = \frac{b}{(a+d)} \quad (11)$$

7.5 Root Mean Squared Error (RMSE) เป็นดัชนีที่ใช้วัดความคลาดเคลื่อนระหว่างค่าจริงและค่าที่พยากรณ์ได้ โดยคำนวณได้จากสมการที่ 12

$$RMSE = \sqrt{\frac{1}{n} \left\{ \sum_{i=1}^n (O - O_{cal})^2 \right\}} \quad (12)$$

เมื่อ

n คือจำนวนข้อมูลทั้งหมด

O คือค่าของข้อมูลจริง

O_{cal} คือค่าที่ได้จากโมเดล

อุปกรณ์และวิธีการ

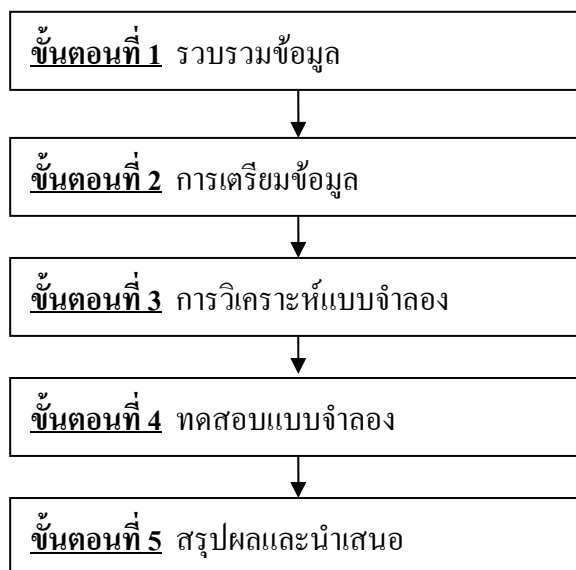
อุปกรณ์

วัสดุอุปกรณ์ที่จำเป็นสำหรับทำวิทยานิพนธ์นี้ ประกอบด้วย

1. เครื่องคอมพิวเตอร์ส่วนบุคคล
2. ระบบปฏิบัติการ Microsoft Windows XP
3. โปรแกรม Microsoft Office 2003 (Word, Excel, PowerPoint และ Access)
4. โปรแกรม Weka version 3.4.11
5. เครื่องพิมพ์ และอื่นๆ

วิธีการ

การวิเคราะห์หาสาเหตุที่ทำให้เกิดเหตุการณ์กระแสไฟฟ้าขัดข้องที่ระบุไม่ทราบสาเหตุในระบบจำหน่ายแรงสูง 22 เควี ของการไฟฟ้าส่วนภูมิภาค เขต 1 ภาคกลาง ในวิทยานิพนธ์นี้สรุปไว้ในภาพที่ 7 โดยนำข้อมูลซึ่งได้จากโปรแกรม จฟ.3 มาทำการเตรียมความพร้อมสำหรับนำไปใช้หาแบบจำลองที่เหมาะสมที่สุดโดยใช้อัลกอริทึมที่แตกต่างกัน 3 อัลกอริทึมได้แก่ ต้นไม้ตัดสินใจ กฎ-ของเบย์ และกฎของริบเปอร์ นำผลมาเปรียบเทียบประสิทธิภาพจากค่าความผิดพลาดที่เกิดขึ้นโดยพิจารณาจากค่าที่น้อยที่สุด เพื่อความแม่นยำในการทำนายผล โดยทดสอบกับชุดข้อมูลทดสอบที่เตรียมไว้ เมื่อได้แบบจำลองดังกล่าวแล้วสามารถนำไปทำนายสาเหตุที่ทำให้เกิดเหตุการณ์กระแสไฟฟ้าขัดข้องที่ระบุไม่ทราบสาเหตุในระบบจำหน่ายแรงสูง ต่อไป



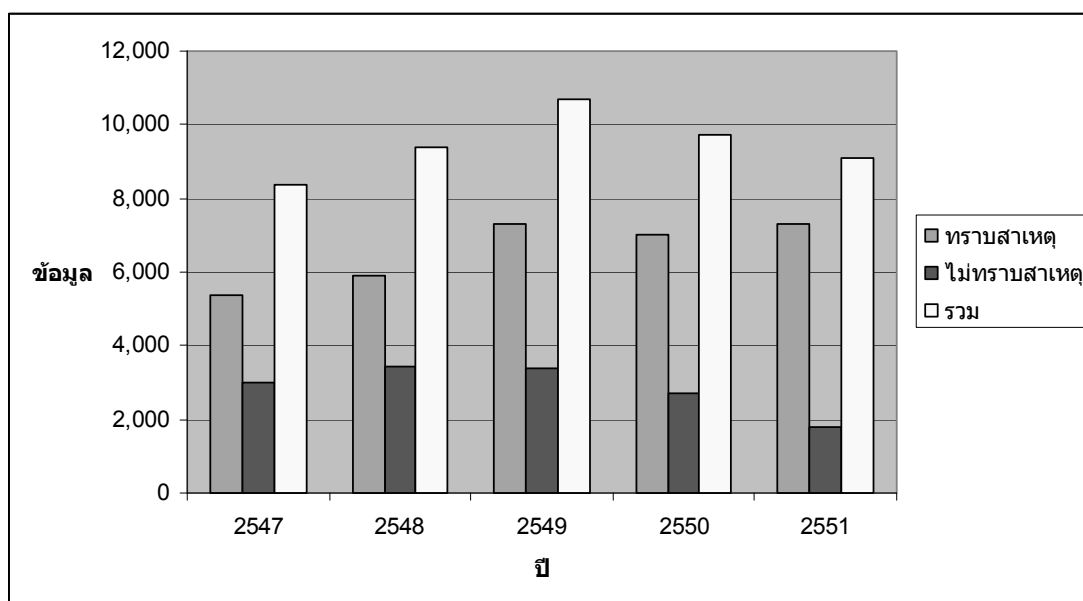
ภาพที่ 7 ลำดับขั้นตอนการนำข้อมูลกระแสไฟฟ้าเข้าเครื่องวิเคราะห์โดยวิธีเหมืองข้อมูล

1. ข้อมูลกระแสไฟฟ้าเข้าเครื่อง

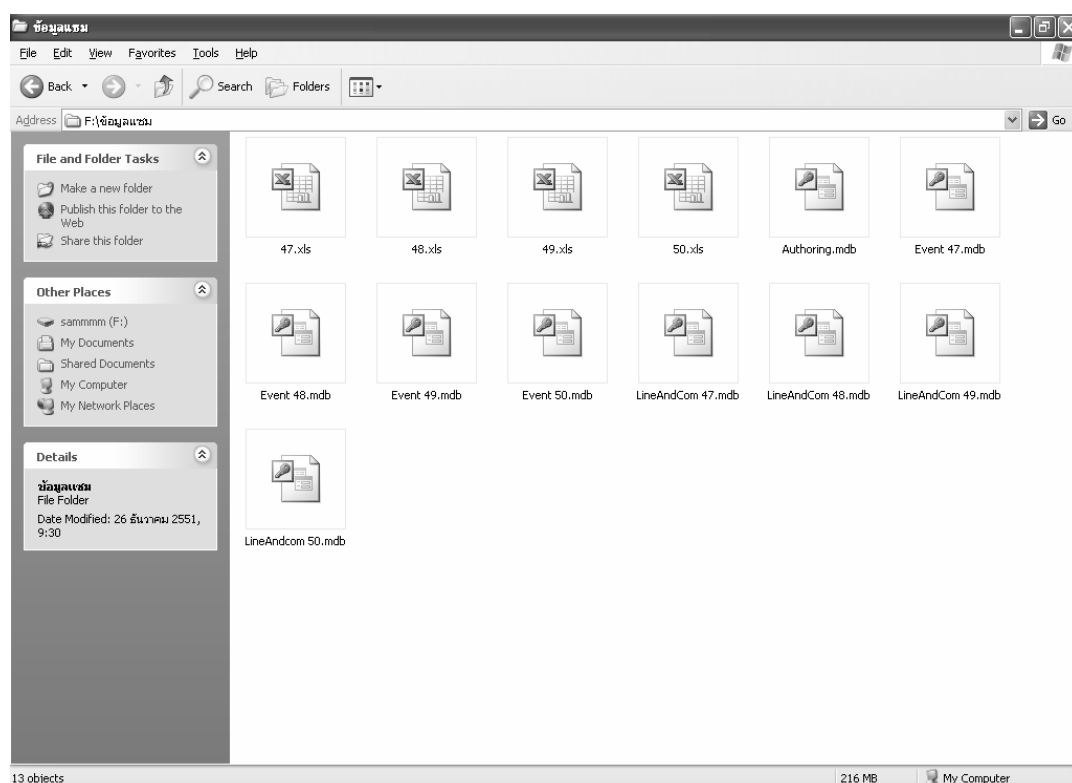
จำนวนครั้งเหตุการณ์ไฟฟ้าเข้าเครื่องสามารถหาได้จากข้อมูลเหตุการณ์ไฟฟ้าเข้าเครื่องในระบบจำหน่ายที่บันทึกไว้ในฐานข้อมูลโปรแกรมจ่ายไฟ (จฟ.3) เป็นโปรแกรมที่พนักงานการไฟฟ้าส่วนภูมิภาค เขต1 ภาคกลางบันทึกไว้ตั้งแต่เดือนมกราคม พ.ศ. 2547 ถึงเดือนธันวาคม พ.ศ. 2551 จำนวน 47,259 ข้อมูล มีรายละเอียดตามตารางที่ 1 โดยได้รับข้อมูลมาในรูปแบบการจัดเก็บ Main Database File (file.mdb) ซึ่งจะคัดเลือกเฉพาะคอลัมน์ที่ต้องนำมาใช้ในวิทยานิพนธ์นี้ดังแสดงในภาพที่ 1

ตารางที่ 1 จำนวนข้อมูลทั้งหมดที่นำมาใช้วิเคราะห์

ชนิดข้อมูล	จำนวนข้อมูลปี พ.ศ. 2547-2551					รวม
	2547	2548	2549	2550	2551	
ทราบสาเหตุ	5,375	5,926	7,308	7,017	7,328	32,954
ไม่ทราบสาเหตุ	2,982	3,450	3,374	2,714	1,785	14,305
รวม	8,357	9,376	10,682	9,731	9,113	47,259



ภาพที่ 8 ข้อมูลกระแสไฟฟ้าขัดข้องปี พ.ศ. 2547 – 2551



ภาพที่ 9 ไฟล์สถิติเหตุการณ์ไฟฟ้าขัดข้องปี พ.ศ. 2547 – 2551 ในรูป Main Database File (file.mdb)

Microsoft Access - [แยกไลน์เฟนไลน์แยก : แบบผสมตามแบบใช้เลือกข้อมูล]

เพิ่ม แก้ไข มุมมอง แทรก รูปแบบ ฐานข้อมูล เครื่องมือ หน้าต่าง วิสัยทัศน์

พิมพ์ค่าตามข้อความช่วยเหลือ

	SubID	FeederID	ComID2	Dateoutage	Timeoutage	Caption	TotalTime	Known	typename	Weather
▶	ARA	ARA03	ARA03VB-01	02-ม.ย.-50	16:30	T/L(Function)	18	☐	อื่นๆ	WC01
	ARA	ARA04	ARA04VB-01	08-ม.ค.-50	12:05	T/L(Function)	33	☒	สัตว์	WC01
	ARA	ARA04	ARA04VB-01	17-ส.ค.-50	8:33	T/L(Function)	90	☒	ยานพาหนะ	WC01
	ARA	ARA09	ARA09VB-01	10-พ.ค.-50	10:50	T/L(Function)	60	☒	สัตว์	WC01
	ATA	ATA01	ATA01VB-01	19-ม.ย.-50	21:28	T/L(Function)	42	☒	ต้นไม้	WC06
	ATA	ATA02	ATA02VB-01	23-ก.ค.-50	15:24	T/L(Function)	14	☐	อุปกรณ์	WC05
	ATA	ATA03	ATA03VB-01	06-เม.ย.-50	15:42	T/L(Function)	46	☐	อื่นๆ	WC03
	ATA	ATA03	ATA03VB-01	27-เม.ย.-50	21:32	T/L(Function)	34	☒	อุปกรณ์	WC06
	ATA	ATA07	ATA07VB-01	07-มี.ค.-50	23:16	T/L(Function)	22	☐	อื่นๆ	WC01
	ATA	ATA07	ATA07VB-01	26-มี.ค.-50	13:43	T/L(Function)	28	☐	อื่นๆ	WC01
	ATA	ATA07	ATA07VB-01	17-ก.ค.-50	15:13	T/L(Function)	12	☒	อื่นๆ	WC03
	ATA	ATA07	ATA07VB-01	23-ก.ค.-50	15:20	T/L(Function)	30	☐	ต้นไม้	WC05
	ATA	ATA09	ATA09VB-01	15-ม.ย.-50	3:27	T/L(Instantenc	3	☐	อื่นๆ	WC01
	ATA	ATA09	ATA09VB-01	15-ม.ย.-50	4:06	T/L(Instantenc	11	☐	อื่นๆ	WC01
	ATA	ATA09	ATA09VB-01	28-ส.ค.-50	0:19	T/L(Function)	26	☒	สภาพสิ่งแวดล้อม	WC06
	ATA	ATA09	ATA09VB-01	28-ส.ค.-50	1:13	T/L(Instantenc	1	☒	สภาพสิ่งแวดล้อม	WC06
	ATA	ATA10	ATA10VB-01	01-พ.ค.-50	9:08	T/L(Function)	39	☒	อุปกรณ์	WC04
	ATA	ATA10	ATA10VB-01	24-ม.ย.-50	16:40	T/L(Function)	6	☒	ต้นไม้	WC03
	ATA	ATA10	ATA10VB-01	25-ม.ย.-50	9:15	T/L(Function)	26	☒	อุปกรณ์	WC02
	ATA	ATA10	ATA10VB-01	11-เม.ย.-50	14:46	T/L(Function)	54	☐	อื่นๆ	WC03
	ATB	ATB01	ATB01VB-01	19-ม.ย.-50	21:43	T/L(Function)	69	☒	ต้นไม้	WC06
	ATB	ATB01	ATB01VB-01	20-ส.ค.-50	15:12	T/L(Function)	20	☒	อุปกรณ์	WC01
	ATB	ATB01	ATB01VB-01	20-ส.ค.-50	15:13	T/L(Function)	20	☒	สัตว์	WC01
	ATB	ATB01	ATB01VB-01	23-ก.ค.-50	15:45	T/L(Function)	19	☒	อื่นๆ	WC01
	ATB	ATB02	ATB02VB-01	20-เม.ย.-50	13:06	T/L(Function)	32	☒	อื่นๆ	WC01
	ATB	ATB02	ATB02VB-01	27-เม.ย.-50	18:50	T/L(Function)	220	☒	ภัยธรรมชาติ	WC05
	ATB	ATB02	ATB02VB-01	27-เม.ย.-50	18:58	T/L(Function)	24	☒	อุปกรณ์	WC06
	ATB	ATB02	ATB02VB-01	24-ม.ย.-50	17:32	T/L(Function)	18	☐	อื่นๆ	WC03
	ATB	ATB02	ATB02VB-01	25-ม.ย.-50	3:18	T/L(Function)	11	☒	อุปกรณ์	WC02
	ATB	ATB02	ATB02VB-01	25-ม.ย.-50	3:58	T/L(Instantenc	5	☒	อุปกรณ์	WC02
	ATB	ATB02	ATB02VB-01	25-ม.ย.-50	4:30	T/L(Instantenc	4	☒	อุปกรณ์	WC02
	ATB	ATB03	ATB03VB-01	13-พ.ค.-50	22:09	T/L(Function)	16	☒	อุปกรณ์	WC02
	ATB	ATB06	ATB06VB-01	18-ม.ค.-50	15:56	T/L(Instantenc	4	☒	อุปกรณ์	WC01
	ATB	ATB07	ATB07VB-01	05-...-50	12:00	T/L(Function)	20	☒	ส.พ.ช.	WC04

จะเขียน: 1 จาก 918

มุมมองแบบข้อมูล

NUM

ภาพที่ 10 ลักษณะฐานข้อมูลไฟฟ้าขัดข้องจากโปรแกรม จฟ.3

2. เตรียมข้อมูล

2.1 การเลือกข้อมูล

การเตรียมข้อมูลสำหรับวิทยานิพนธ์นี้ ใช้ข้อมูลกระแสไฟฟ้าขัดข้องจากโปรแกรม จฟ.3 ซึ่งได้คัดเลือกเฉพาะคุณลักษณะหรือแอททริบิวต์ที่มีลักษณะของข้อมูลที่เหมาะสมกับลักษณะการวิเคราะห์ มีความถูกต้องในระดับที่เหมาะสม มีความสัมพันธ์กับผลลัพธ์ที่ต้องการ โดยได้คัดเลือกมาทั้งหมด 11 แอททริบิวต์ ดังนี้

- SubID หมายถึงสถานีไฟฟ้าที่จ่ายไฟให้อุปกรณ์ป้องกันที่ทำงานเปิดวงจร แต่ละสถานีไฟฟ้าจะใช้สัญลักษณ์เป็นรหัสตัวอักษรภาษาอังกฤษ 3 ตัว เช่น SRD DOP เป็นต้น

- Date outage หมายถึง วันที่เกิดเหตุการณ์ไฟฟ้าขัดข้อง ในรอบปี

- Time outage หมายถึง เวลาที่เกิดเหตุการณ์ไฟฟ้าขัดข้อง ในรอบ 24 ชั่วโมง

- Total Time หมายถึง ระยะเวลาที่เกิดเหตุการณ์ไฟฟ้าขัดข้อง ตั้งแต่เริ่มเกิดเหตุการณ์กระแสไฟฟ้า ขัดข้องจนจ่ายไฟได้ปกติ

- Caption หมายถึง ลักษณะการทำงานตามฟังก์ชันของอุปกรณ์ป้องกันที่ทำงานเปิดวงจร ได้แก่

1) T/L (Function) อุปกรณ์ป้องกันอัตโนมัติทำงานครบตามฟังก์ชันที่ตั้งไว้แล้วจึงเปิดวงจรออก

2) T/L (Instantaneous) อุปกรณ์ป้องกันอัตโนมัติทำงานเปิดวงจรออกทันทีเนื่องจากกระแสลัดวงจรรุนแรง

3) T/R1 อุปกรณ์ป้องกันอัตโนมัติทำการเปิดวงจรออก 1 ครั้ง แล้วปิดวงจรกลับจ่ายได้ตามปกติ

4) T/R2 อุปกรณ์ป้องกันอัตโนมัติทำการเปิดวงจรออก 2 ครั้ง แล้วปิดวงจรกลับจ่ายได้ตามปกติ

5) T/R3 อุปกรณ์ป้องกันอัตโนมัติทำการเปิดวงจรออก 3 ครั้ง แล้วปิดวงจรกลับจ่ายได้ตามปกติ

6) Not Work อุปกรณ์ป้องกันอัตโนมัติ ฝืนไม่ทำงานเพื่อกำจัดฟอลต์ออกจากระบบ เนื่องจากอุปกรณ์ป้องกันด้าน High Side ทำงานเปิดวงจรทำให้ไฟดับ

7) D/F ฟิวส์ขาด

- Weather หมายถึง สภาพอากาศในขณะที่เกิดเหตุการณ์กระแสไฟฟ้าขัดข้องหรือสภาพ อากาศเมื่อเข้าถึงจุดเกิดเหตุมีสภาพอากาศ 6 ลักษณะ คือ อากาศปกติ อากาศชื้น หมอก ลมแรง ฝนตกฟ้าคะนอง

- Pha_A หมายถึง สถานะรีเลย์ของอุปกรณ์ป้องกันที่ทำงานตามฟังก์ชันการทำงานของเฟส A โดยมีลักษณะการทำงานแบบ Time (ทำงานตามฟังก์ชันที่กำหนดระยะเวลาไว้) หรือ Instantaneous (ทำงานแบบทันทีทันใด)

- Pha_B หมายถึง สถานะรีเลย์ของอุปกรณ์ป้องกันที่ทำงานตามฟังก์ชันการทำงานของเฟส B โดยมีลักษณะการทำงานแบบ Time หรือ Instantaneous

- Pha_C หมายถึง สถานะรีเลย์ของอุปกรณ์ป้องกันที่ทำงานตามฟังก์ชันการทำงานของเฟส C โดยมีลักษณะการทำงานแบบ Time หรือ Instantaneous

- Pha_G หมายถึง สถานะรีเลย์ของอุปกรณ์ป้องกันที่ทำงานตามฟังก์ชันการทำงานของกราวด์ โดยมีลักษณะการทำงานแบบ Time หรือ Instantaneous

- Type name หมายถึง สาเหตุของเหตุการณ์ไฟฟ้าขัดข้อง การจัดการข้อมูลเพื่อใช้วิเคราะห์หามีทั้งหมด 11 สาเหตุ คือ สัตว์ สภาพสิ่งแวดล้อม อุปกรณ์ ต้นไม้ ภัยธรรมชาติ วัสดุแปลกปลอม ยานพาหนะ บุคคลภายนอก พนักงาน กฟภ. จ่ายเกินพิกัด และอื่นๆ ซึ่งการระบุสาเหตุต่างๆจะแบ่งออกเป็น 2 ลักษณะ ได้แก่

1) ทราบสาเหตุ ในกรณีที่สามารถทราบจุดเกิดเหตุและระบุสาเหตุที่ทำให้เกิดเหตุการณ์กระแสไฟฟ้าขัดข้องได้

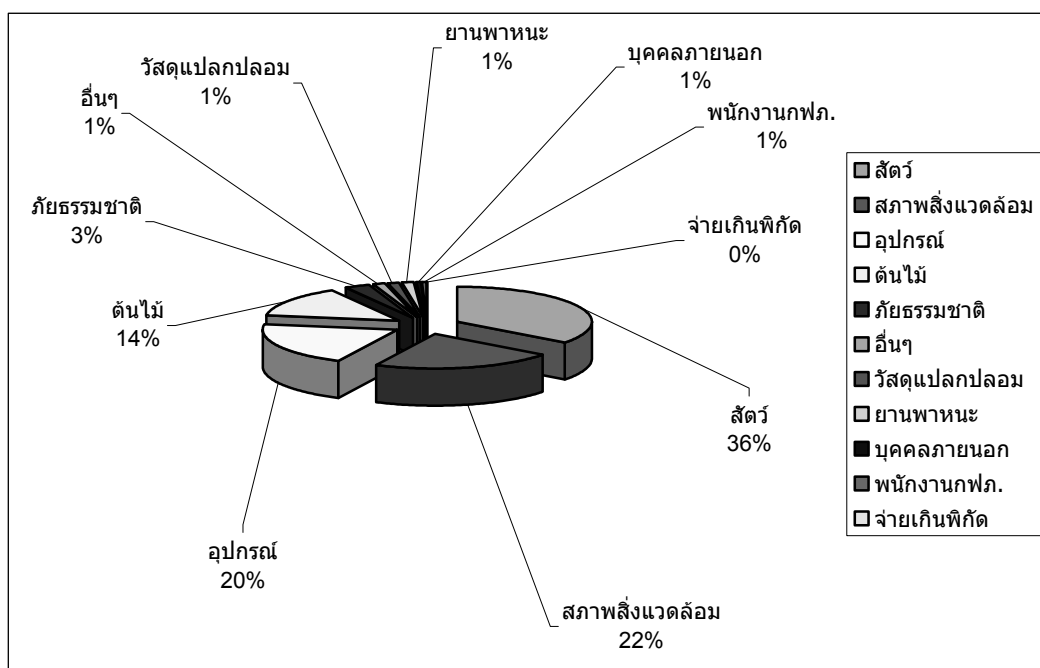
2) สันนิษฐาน ให้ใช้ดุลพินิจประกอบกับสภาพหน้างานในการสันนิษฐานหาสาเหตุ ในกรณีที่ไม่สามารถหาจุดเกิดเหตุและระบุสาเหตุที่ทำให้เกิดเหตุการณ์กระแสไฟฟ้าขัดข้อง

2.2 รวมข้อมูล

โดยพิจารณาจากชุดข้อมูลของปี พ.ศ. 2547 – 2550 เป็นชุดข้อมูลที่ระบุกรณีทราบสาเหตุทั้งหมด 25,626 ข้อมูล ใช้เป็นชุดข้อมูลเรียนรู้ เพื่อเตรียมสำหรับใช้วิเคราะห์หาแบบจำลองการวิเคราะห์สาเหตุโดยเลือกใช้แอททริบิวต์ Type name เป็นแอททริบิวต์เป้าหมาย ซึ่งสาเหตุของกระแสไฟฟ้าขัดข้องของแอททริบิวต์ Type name มีจำนวนข้อมูลแบ่งตามสาเหตุ ดังแสดงในตารางที่ 2 และคิดเป็นลักษณะเปอร์เซ็นต์ตามภาพที่ 11

ตารางที่ 2 จำนวนข้อมูลสาเหตุ 11 สาเหตุ ที่นำมาใช้วิเคราะห์

สาเหตุ	จำนวนข้อมูล
สัตว์	9070
สภาพสิ่งแวดล้อม	5583
อุปกรณ์	5196
ต้นไม้	3667
ภัยธรรมชาติ	729
อื่นๆ	366
วัสดุแปลกปลอม	363
ยานพาหนะ	356
บุคคลภายนอก	147
พนักงาน กฟภ.	140
จ่ายเกินพิกัด	9



ภาพที่ 11 จำนวนข้อมูลแยกตามสาเหตุ 11 สาเหตุโดยเฉลี่ย

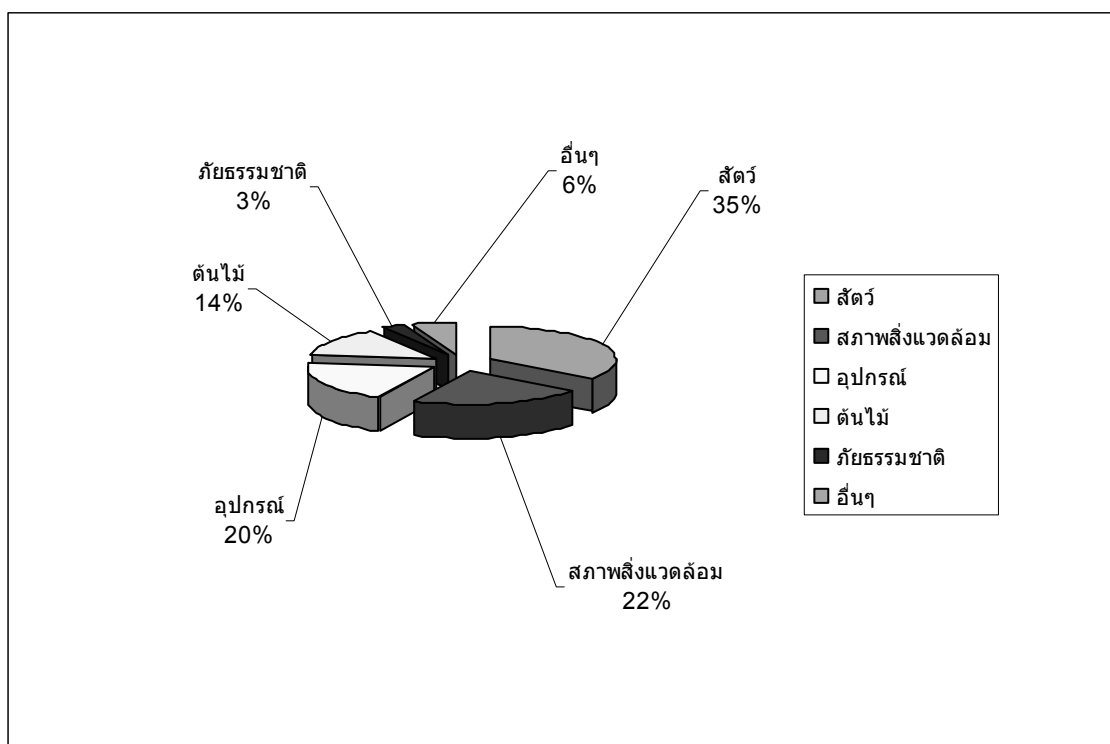
จากชุดข้อมูลในตารางที่ 3 พิจารณาได้ว่ามีบางสาเหตุมีจำนวนน้อย ได้แก่ สาเหตุอื่นๆ วัสดุแปลกปลอม ยานพาหนะ บุคคลภายนอก พนักงาน กฟภ. และจายเกินพิกัด จึงพิจารณาสาเหตุเหล่านี้รวมเป็นสาเหตุอื่นๆ ดังแสดงในตารางที่ 3 ซึ่งจะเหลือเพียง 6 สาเหตุที่นำมาใช้ โดยพิจารณา รวมจากชุดข้อมูลที่มีจำนวนน้อยทั้งหมด กำหนดให้เป็นเหตุการณ์ อื่นๆ ได้ 1,320 ข้อมูล ดังนั้น จำนวนข้อมูลทั้งหมดที่นำมาใช้วิเคราะห์ แบ่งเป็นตามตารางที่ 4

ตารางที่ 3 จำนวนข้อมูลสาเหตุ 6 สาเหตุ ที่มีจำนวนข้อมูลไม่เกิน 1 เปอร์เซนต์ของข้อมูลทั้งหมด

สาเหตุ	จำนวนข้อมูล
อื่นๆ	366
วัสดุแปลกปลอม	363
ยานพาหนะ	356
บุคคลภายนอก	147
พนักงาน กฟภ.	140
จายเกินพิกัด	9

ตารางที่ 4 จำนวนข้อมูลสาเหตุ 6 สาเหตุ ที่รวมชุดข้อมูลที่มีจำนวนน้อยเข้าด้วยกันเพื่อวิเคราะห์

สาเหตุ	จำนวนข้อมูล
สัตว์	9,070
สภาพสิ่งแวดล้อม	5,583
อุปกรณ์	5,196
ต้นไม้	3,667
ภัยธรรมชาติ	729
อื่นๆ	1,381

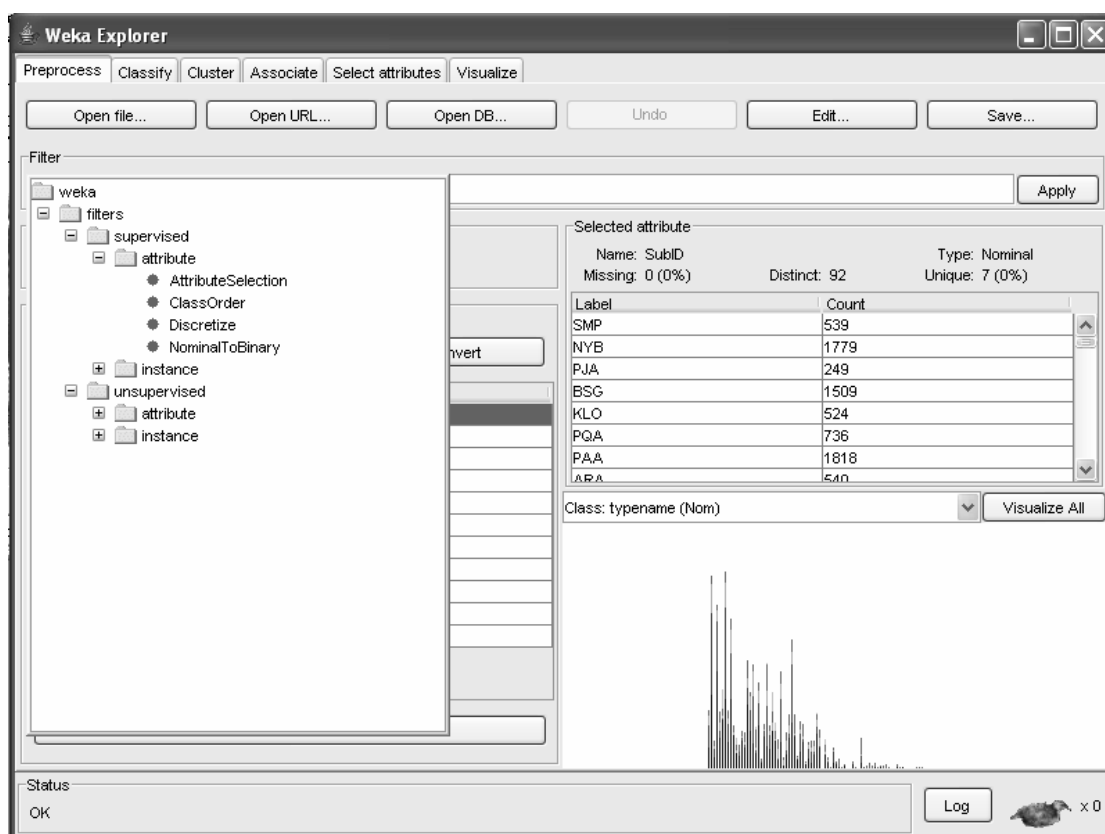


ภาพที่ 12 จำนวนข้อมูลสาเหตุ ที่ทำให้เกิดเหตุการณ์ทั้งหมด 6 สาเหตุ

2.3 แบ่งช่วงข้อมูล

โดยการพิจารณาชุดข้อมูลเรียนรู้ที่มีคุณสมบัติเป็นชนิดตัวเลข (Numeric) นำมาเปลี่ยนให้เป็นข้อมูลที่เป็นชนิดรูปแบบนาม (Nominal) ซึ่งแอททริบิวต์ที่มีลักษณะของข้อมูลเป็นชนิดตัวเลขอยู่ 2 แอททริบิวต์ คือ Time outage และ Total Time จึงต้องทำการจัดข้อมูลเพื่อให้เหมาะสมกับการใช้ ในการวิเคราะห์สำหรับอัลกอริทึม

การแปลงข้อมูลทำได้โดยการใช้วิธี Discretize โดยการเปิดข้อมูลที่ต้องการแปลงให้อยู่ในรูปแบบนามในโปรแกรม Weka ใส่ฟิลเตอร์ให้กับข้อมูลโดยทำที่หน้าแรก (Preprocess) โดยเลือกใช้ฟิลเตอร์ Discretize และทำการเปลี่ยนค่าพารามิเตอร์ของ Discretize ให้ได้ค่าที่เหมาะสม

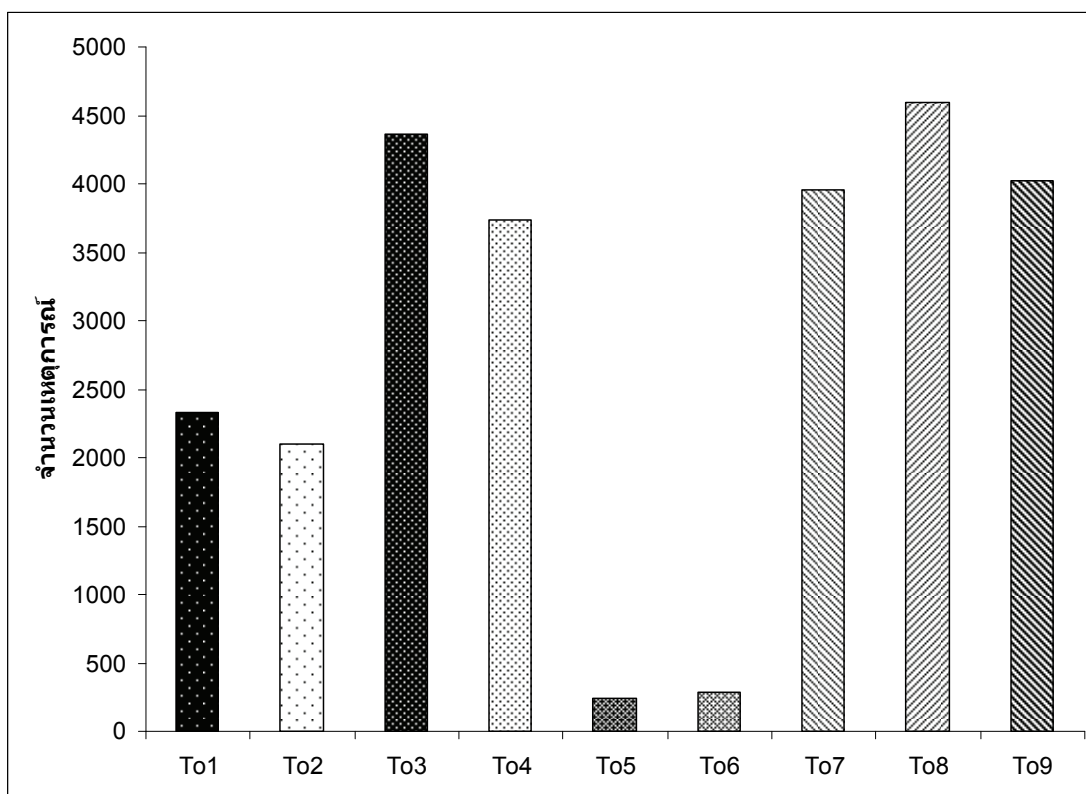


ภาพที่ 13 การเลือกคำสั่ง Discretize กับข้อมูลที่เป็นชนิดตัวเลขในโปรแกรม Weka

จากชุดข้อมูลเรียนรู้พิจารณาได้ว่าจากข้อมูลเดิม **Time-outage** : ตั้งแต่เวลา 00.00 น. ถึง เวลา 23.59 น. สามารถแบ่งช่วงได้ตามตารางที่ 5

ตารางที่ 5 จำนวนข้อมูลสาเหตุ Attribute Time-outage ที่ทำการจัดรูปแบบให้เป็นลักษณะข้อมูล Nominal

ชื่อช่วง	ตั้งแต่เวลา (น.)	ถึงเวลา (น.)	จำนวนข้อมูล
To1	00.00	05.59	2,331
To2	06.00	07.29	2,100
To3	07.30	11.59	4,360
To4	12.00	15.29	3,740
To5	15.30	15.59	245
To6	16.00	16.19	287
To7	16.20	17.27	3,956
To8	17.28	19.33	4,588
To9	19.34	23.59	4,019

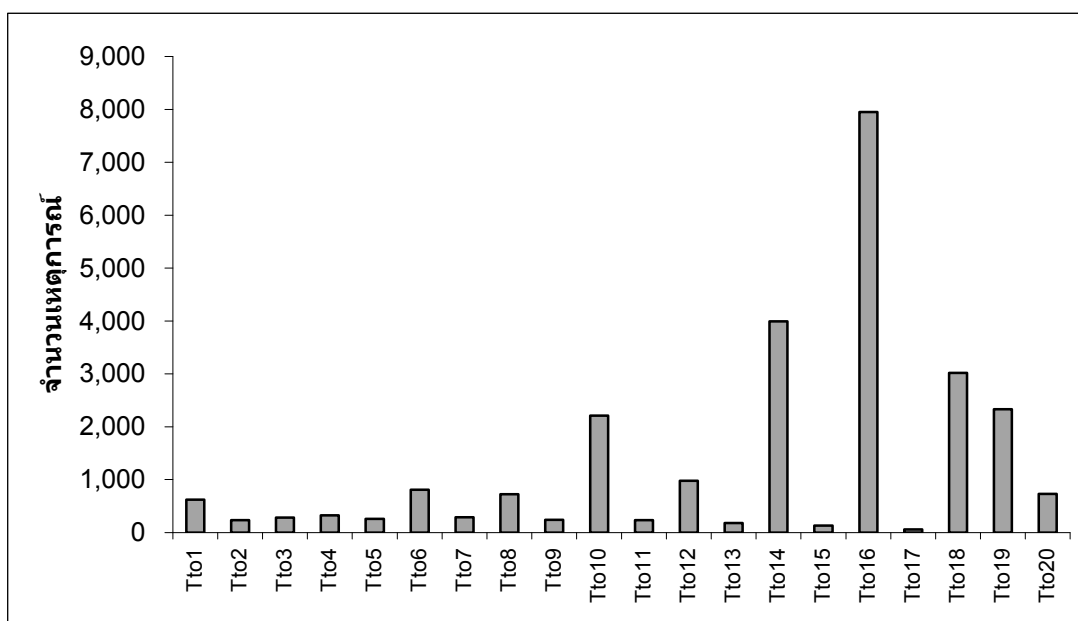


ภาพที่ 14 กราฟแสดงจัดกลุ่มข้อมูลสาเหตุ Time outage

จากชุดข้อมูลเรียนรู้พิจารณาได้ว่าจากข้อมูลเดิม Total-Time ตั้งแต่ 0 นาทีเป็นต้นไป สามารถจัดแบ่งช่วงได้ตามตารางที่ 6

ตารางที่ 6 จำนวนข้อมูลสาเหตุ Attribute Total-Time ที่ทำการจัดรูปแบบให้เป็นลักษณะข้อมูล
Nominal

ชื่อช่วงเวลา	ขอบล่าง (นาทีที่)	ขอบบน (นาทีที่)	จำนวนข้อมูล
Tto1	0.00	9.4	621
Tto2	9.5	10.4	238
Tto3	10.5	14.4	283
Tto4	14.5	15.4	329
Tto5	15.5	19.4	261
Tto6	19.5	20.4	811
Tto7	20.5	24.4	291
Tto8	24.5	25.4	727
Tto9	25.5	29.4	239
Tto10	29.5	30.4	2,211
Tto11	30.5	34.4	237
Tto12	34.5	35.4	981
Tto13	35.5	39.4	181
Tto14	39.5	45.4	3,992
Tto15	45.5	49.4	134
Tto16	49.5	65.4	7,949
Tto17	65.5	69.4	61
Tto18	69.5	80.4	3021
Tto19	80.5	130.4	2,330
Tto20	130.5	inf	729



ภาพที่ 15 กราฟแสดงจัดกลุ่มข้อมูลสาเหตุ Total time

- ชุดข้อมูล **Date outage** ลักษณะของข้อมูลถูกระบุเป็นระยะเวลาภายในปีโดยแสดงตั้งแต่ วันที่ 1 มกราคม ถึง วันที่ 31 ธันวาคม ซึ่งเป็นข้อมูลชนิดตัวเลข จะกำหนดให้แบ่งช่วงได้อยู่ในช่วงของเดือนทั้ง 12 เดือน

จากกระบวนการจัดการข้อมูลที่ผ่านมาจะใช้วิธีเช่นเดียวกันกับชุดข้อมูลปี 2551 ตั้งแต่เดือน มกราคม ถึงเดือนธันวาคม จำนวน 9,114 ข้อมูล โดยกระบวนการดังกล่าวจะช่วยให้จัดข้อมูลให้มีความเหมาะสมในการวิเคราะห์เช่นกัน

2.4 เลือกตัวอย่างข้อมูล

สำหรับแบบจำลองและทดสอบในการวิเคราะห์ จะทำการแบ่งข้อมูลจากชุดข้อมูลสามารถแบ่งเพื่อใช้ในการวิเคราะห์ดังนี้

2.4.1 สร้างแบบจำลอง ใช้ข้อมูลตั้งแต่ ปี 2547-2550 มีจำนวน 38,146 ข้อมูล แยกข้อมูลที่ทราบสาเหตุทั้งหมด 25,626 ข้อมูล โดยสุ่มข้อมูลมา 3 ชุดๆละ 16,656 ข้อมูลเหตุการณ์คิดเป็น 65 เปอร์เซนต์

2.4.2 ชุดข้อมูลทดสอบ ใช้ข้อมูลปี 2551 มีจำนวน 9,113 ข้อมูล แบ่งมีข้อมูลที่ทราบสาเหตุทั้งหมด 7,328 ข้อมูล และข้อมูลที่ไม่ทราบสาเหตุจำนวน 1,785 ข้อมูล ใช้เป็นชุดข้อมูลทดสอบโดยแบ่งออกเป็น 3 ชุด ตามตารางที่ 7 มีรายละเอียดดังนี้

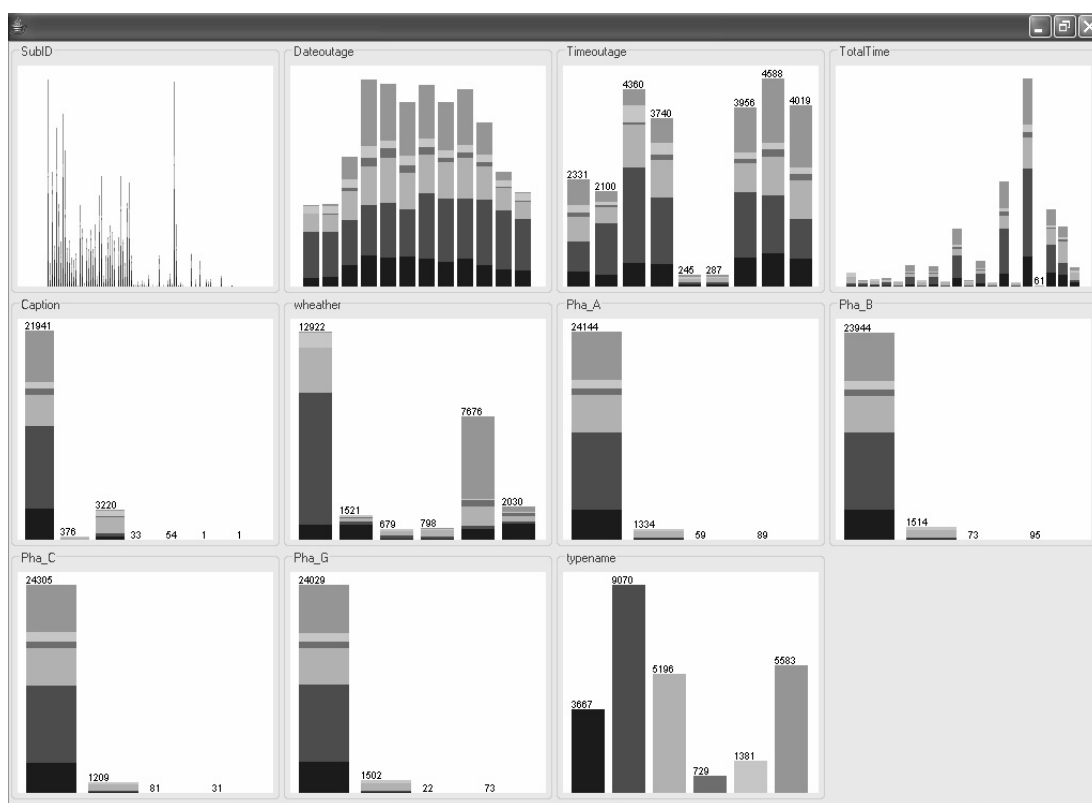
ชุดที่ 1 ใช้ชุดข้อมูลที่ทราบสาเหตุจำนวน 7,328 ข้อมูล สุ่มเลือกมาจำนวน 1,500 ข้อมูล แบ่งออกเป็น 3 ชุดๆละ 500 ข้อมูล เพื่อทดสอบความถูกต้องของผลจากการทดสอบกับแบบจำลอง โดยเปรียบเทียบกับสาเหตุจริงที่ระบุโดยทราบสาเหตุ

ชุดที่ 2 ใช้ชุดข้อมูลที่ไม่ทราบสาเหตุจำนวน 1,785 ข้อมูล สุ่มเลือกมาจำนวน 100 ข้อมูล เพื่อแสดงผลจากการทดสอบกับแบบจำลองเปรียบเทียบกับสาเหตุที่มีจากข้อมูลที่ไม่ทราบสาเหตุ

ชุดที่ 3 ใช้ชุดข้อมูลที่ไม่ทราบสาเหตุจำนวน 1,785 ข้อมูล สุ่มเลือกมาจำนวน 5 ข้อมูล เพื่อจำลองการเกิดเหตุการณ์กระแสไฟฟ้าขัดข้องในปัจจุบันและแสดงตัวอย่างของผลจากการทดสอบกับแบบจำลองโดยละเอียด

ตารางที่ 7 จำนวนการแบ่งชุดข้อมูล

ชุดข้อมูล		จำนวน ทั้งหมด	จำนวนข้อมูล			
ปี	ประเภท		ชุดข้อมูล เรียนรู้	ชุดข้อมูลทดสอบ		
				ชุด 1	ชุด 2	ชุด 3
2547-2550	ทราบสาเหตุ	25,626	25,626	-	-	-
	ไม่ทราบสาเหตุ	12,520	-	-	-	-
2551	ทราบสาเหตุ	7,328	-	1,500	-	-
	ไม่ทราบสาเหตุ	1,785	-	-	100	5



ภาพที่ 16 ลักษณะข้อมูลทั้งหมดที่เตรียมพร้อมสำหรับวิเคราะห์

Viewer

Relation: testset 51 cut4968-weka.filters.supervised.instance.Resample-B0.0-S4-Z10.07

No.	SubID Nominal	Dateoutage Nominal	Timeoutage Nominal	TotalTime Nominal	Caption Nominal	wheather Nominal	Pha_A Nominal	Pha_B Nominal	Pha_C Nominal	Pha_G Nominal	typename Nominal
467	NYB	ส.ค.	To3	To16	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สภาพสิ่งแวดลอม
468	NYB	พ.ย.	To3	To8	D/F	ลมแรง	none	none	none	none	ต้นไม้
469	NYB	ก.ค.	To3	To8	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สภาพสิ่งแวดลอม
470	OKA	เม.ย.	To1	To16	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	ภัยธรรมชาติ
471	KYE	พ.ย.	To3	To14	D/F	อากาศปกติ	none	none	none	none	สัตว์
472	BMA	พ.ย.	To1	To16	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สภาพสิ่งแวดลอม
473	BSG	ก.ค.	To7	To16	D/F	อากาศปกติ	none	none	none	none	สัตว์
474	WYA	ม.ค.	To7	To16	D/F	อากาศปกติ	none	none	none	none	สัตว์
475	NYB	ก.ค.	To3	To10	D/F	อากาศปกติ	none	none	none	none	สัตว์
476	NYB	ส.ค.	To9	To6	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สภาพสิ่งแวดลอม
477	PCK	มี.ค.	To3	To1	D/F	อากาศปกติ	none	none	none	none	สัตว์
478	BIA	เม.ย.	To1	To18	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สภาพสิ่งแวดลอม
479	PCK	พ.ย.	To9	To1	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สัตว์
480	NSI	พ.ย.	To7	To10	D/F	ลมแรง	none	none	none	none	อุปกรณ์
481	WAB	เม.ย.	To9	To14	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สภาพสิ่งแวดลอม
482	KLO	ก.ย.	To8	To16	D/F	ฝนตก_ลมแรง	none	none	none	none	ภัยธรรมชาติ
483	NYB	มี.ค.	To8	To14	D/F	อากาศปกติ	none	none	none	none	สัตว์
484	AYA	ก.ค.	To4	To14	D/F	อากาศปกติ	none	none	none	none	สัตว์
485	AYA	พ.ย.	To7	To14	D/F	อากาศปกติ	none	none	none	none	สัตว์
486	WAA	เม.ย.	To9	To16	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สภาพสิ่งแวดลอม
487	NSI	มี.ย.	To4	To10	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สภาพสิ่งแวดลอม
488	PKE	มี.ย.	To4	To14	D/F	อากาศปกติ	none	none	none	none	สัตว์
489	NSI	ส.ค.	To8	To16	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	อุปกรณ์
490	NYB	ส.ค.	To9	To7	D/F	ลมแรง	none	none	none	none	ต้นไม้
491	NYB	พ.ค.	To1	To16	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สภาพสิ่งแวดลอม
492	BSG	ต.ค.	To7	To16	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สภาพสิ่งแวดลอม
493	TDA	ก.ค.	To4	To16	D/F	อากาศปกติ	none	none	none	none	สัตว์
494	BSA	ก.ค.	To3	To16	D/F	อากาศปกติ	none	none	none	none	สัตว์
495	ATB	เม.ย.	To4	To16	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สภาพสิ่งแวดลอม
496	LBL	ก.ค.	To3	To16	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สภาพสิ่งแวดลอม
497	PCK	มี.ย.	To9	To1	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สภาพสิ่งแวดลอม
498	PAA	ธ.ค.	To8	To18	D/F	อากาศปกติ	none	none	none	none	สัตว์
499	KLJ	ต.ค.	To8	To16	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สภาพสิ่งแวดลอม
500	PCK	ส.ค.	To7	To1	T/L(Function)	ฝนตก	none	none	none	Time	อุปกรณ์

Undo OK Cancel

ภาพที่ 17 ลักษณะชุดข้อมูลทดสอบโดยนำข้อมูล ปี 2551 ที่ทราบสาเหตุ 500 ข้อมูล จำนวน 3 ชุด

Viewer

Relation: tes 2t-weka.filters.supervised.instance.Resample-B0.0-S1-Z89.12-weka.filters.supervised.instance.Resample-B0.0-S3-Z10.0

No.	SubID Nominal	Dateoutage Nominal	Timeoutage Nominal	TotalTime Nominal	Caption Nominal	wheather Nominal	Pha_A Nominal	Pha_B Nominal	Pha_C Nominal	Pha_G Nominal	typename Nominal
67	PQA	ฉ.ย.	To8	Tto10	D/F	อากาศปกติ	none	none	none	none	สัตว์
68	WYA	พ.ศ.	To4	Tto16	D/F	ฝนตก	none	none	none	none	สภาพสิ่งแวดล้อม
69	AYB	เม.ย.	To9	Tto16	D/F	ฝนตก_ลมแรง	none	none	none	none	ต้นไม้
70	PAB	ก.ค.	To9	Tto18	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สภาพสิ่งแวดล้อม
71	PQB	ก.ค.	To8	Tto18	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สภาพสิ่งแวดล้อม
72	OKA	พ.ย.	To2	Tto12	T/L(Function)	อากาศเย็น_หมอก	Time	none	none	Time	อุปกรณ์
73	BDA	พ.ค.	To9	Tto1	T/L(Instante...	อากาศเย็น_หมอก	none	none	none	Time	อุปกรณ์
74	BIA	เม.ย.	To8	Tto18	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สภาพสิ่งแวดล้อม
75	BPH	ฉ.ค.	To3	Tto14	D/F	อากาศปกติ	none	none	none	none	สัตว์
76	KYE	ส.ค.	To7	Tto14	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สภาพสิ่งแวดล้อม
77	NSI	ฉ.ค.	To4	Tto19	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	อุปกรณ์
78	PJA	ก.ค.	To1	Tto18	D/F	อากาศปกติ	none	none	none	none	สัตว์
79	WYA	ส.ค.	To7	Tto8	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สภาพสิ่งแวดล้อม
80	SMP	ต.ค.	To3	Tto4	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สภาพสิ่งแวดล้อม
81	SRB	เม.ย.	To4	Tto16	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	ภัยธรรมชาติ
82	ARA	เม.ย.	To8	Tto16	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สภาพสิ่งแวดล้อม
83	NYB	ก.ค.	To7	Tto10	T/L(Instante...	อากาศปกติ	Time	Time	Time	Time	อื่นๆ
84	MUA	ธ.ค.	To4	Tto10	D/F	อากาศปกติ	none	none	none	none	สัตว์
85	AYB	พ.ค.	To7	Tto14	D/F	อากาศปกติ	none	none	none	none	สัตว์
86	WNA	พ.ย.	To8	Tto8	D/F	ฝนตก_ลมแรง	none	none	none	none	ต้นไม้
87	PAB	ส.ค.	To9	Tto16	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สภาพสิ่งแวดล้อม
88	MUA	ธ.ค.	To8	Tto14	D/F	อากาศปกติ	none	none	none	none	ต้นไม้
89	SRD	ต.ค.	To8	Tto14	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สภาพสิ่งแวดล้อม
90	BPH	ก.พ.	To7	Tto14	D/F	อากาศปกติ	none	none	none	none	สัตว์
91	BIA	ก.ย.	To4	Tto9	D/F	อากาศปกติ	none	none	none	none	สัตว์
92	PAB	ส.ค.	To8	Tto18	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สภาพสิ่งแวดล้อม
93	WYA	เม.ย.	To1	Tto6	D/F	ฝนตก	none	none	none	none	สภาพสิ่งแวดล้อม
94	SRB	พ.ค.	To8	Tto16	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	ภัยธรรมชาติ
95	BMA	พ.ค.	To7	Tto16	D/F	อากาศปกติ	none	none	none	none	อุปกรณ์
96	BKA	ส.ค.	To7	Tto19	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สภาพสิ่งแวดล้อม
97	SKW	เม.ย.	To8	Tto6	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สภาพสิ่งแวดล้อม
98	NSI	เม.ย.	To8	Tto16	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สภาพสิ่งแวดล้อม
99	KYE	ส.ค.	To8	Tto14	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	สภาพสิ่งแวดล้อม
100	LUA	ฉ.ค.	To3	Tto14	D/F	อากาศปกติ	none	none	none	none	สัตว์

Undo OK Cancel

ภาพที่ 18 ลักษณะชุดข้อมูลทดสอบโดยนำข้อมูล ปี 2551 ที่ไม่ทราบสาเหตุ 100 ข้อมูล

Viewer

Relation: test 5 non

No.	SubID Nominal	Dateoutage Nominal	Timeoutage Nominal	TotalTime Nominal	Caption Nominal	wheather Nominal	Pha_A Nominal	Pha_B Nominal	Pha_C Nominal	Pha_G Nominal	typename Nominal
1	NYA	มิ.ย.	To3	Tto10	D/F	ฝนตก_ล...	none	none	none	none	
2	OKA	พ.ค.	To3	Tto3	T/L(Ins...	อากาศปกติ	none	Time	Time	Time	
3	ATB	พ.ค.	To4	Tto16	D/F	อากาศปกติ	none	Ins	Ins	none	
4	NSI	ก.ย.	To4	Tto19	D/F	ฝนตก_ฟิ...	none	none	none	none	
5	RSA	ต.ค.	To3	Tto16	D/F	อากาศปกติ	none	none	Time	Time	

Undo OK Cancel

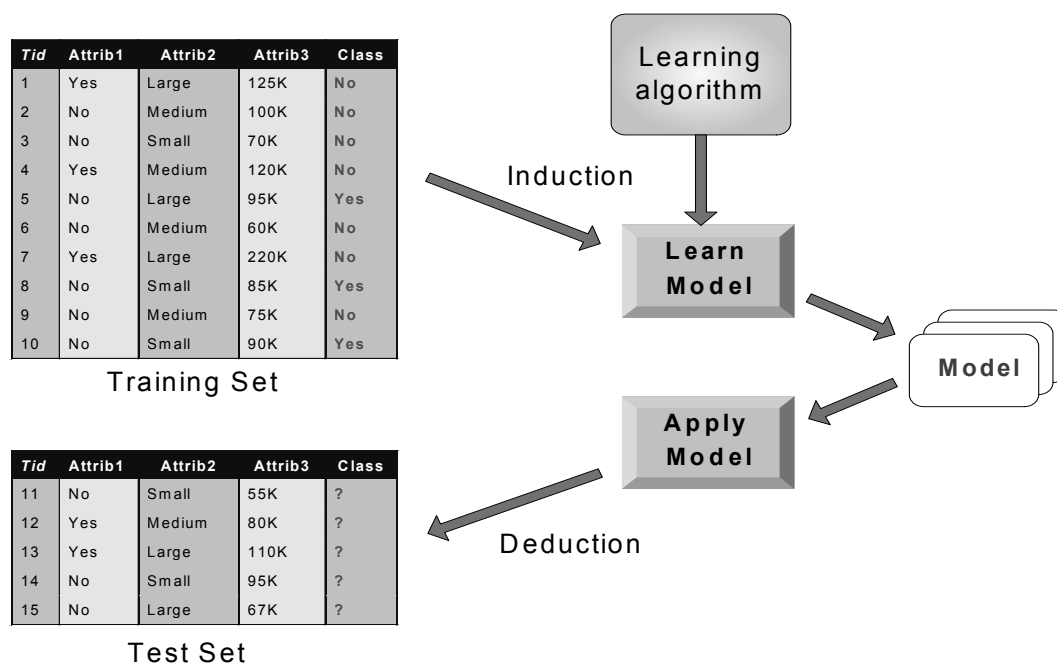
ภาพที่ 19 ลักษณะชุดข้อมูลทดสอบโดยนำข้อมูล ปี 2551 ที่ไม่ทราบสาเหตุ 5 ข้อมูล

3. วิเคราะห์สร้างแบบจำลอง

ขั้นตอนการวิเคราะห์ในกระบวนการจำแนกข้อมูล (Classification) ตามภาพที่ 20 ซึ่งในวิทยานิพนธ์นี้มีขั้นตอนดังนี้

- หาแบบจำลองโดยเลือกใช้อัลกอริทึมที่เหมาะสมที่สุดโดยพิจารณาจากค่าความถูกต้องและค่าความคลาดเคลื่อนของผลการวิเคราะห์

- นำแบบจำลองที่ได้ทดสอบกับชุดข้อมูลทดสอบเพื่อดูค่าความถูกต้องในการทำนายผลโดยใช้ข้อมูลที่ทราบผล และไม่ทราบผล เพื่อพิจารณาผลและการนำไปใช้งาน



ภาพที่ 20 กระบวนการของการจำแนกข้อมูล (Classification)

การสร้างแบบจำลอง ได้แบ่งการพิจารณาจากการวิเคราะห์แต่ละวิธีจากอัลกอริทึม 3 แบบ ได้แก่

- เทคนิคแบบนาอิวเบย์เซียน (Naïve Bayesian)
- กฎของริปเปอร์ (RIPPER) สำหรับในโปรแกรม Weka ใช้ JRip ซึ่งเป็น กฎของริปเปอร์ ที่สร้างขึ้นโดย Java
- ต้นไม้ตัดสินใจ (Decision Tree) ในการจำแนกเราใช้วิธี n-fold Cross Validation

จากชุดข้อมูลเรียนรู้ที่ทราบสาเหตุในตารางที่ 8 นำมาเพื่อหาอัลกอริทึมที่เหมาะสม โดยให้แบบจำลองที่มีประสิทธิภาพที่สุด โดยการนำชุดข้อมูลเรียนรู้มาทดสอบทีละชุดกับ อัลกอริทึมทั้ง 3 แบบในโปรแกรม Weka

ตารางที่ 8 ชุดข้อมูลเรียนรู้

สาเหตุ	ข้อมูลทั้งหมด	จำนวนข้อมูล		
		ชุดที่ 1	ชุดที่ 2	ชุดที่ 3
ต้นไม้	3,667	2,058	2,063	1,996
สัตว์	9,070	7,373	7,519	7,470
อุปกรณ์	5,196	2,637	2,601	2,619
ภัยธรรมชาติ	729	77	71	55
อื่นๆ	1,381	412	408	446
สภาพแวดล้อม	5,583	4,099	3,994	4,070

4. ทดสอบแบบจำลองกับชุดข้อมูลทดสอบ

จากขั้นตอนการหาแบบจำลอง เมื่อได้แบบจำลองที่มีประสิทธิภาพแล้ว ขั้นตอนต่อไปจะนำแบบจำลองที่ได้มาทดสอบกับชุดข้อมูลทดสอบโดยใช้ชุดทดสอบดังนี้

- ใช้ข้อมูลทดสอบในปี พ.ศ. 2551 ซึ่งมีข้อมูลที่ทราบสาเหตุจำนวน 7,328 ข้อมูล นำมาสุ่มเลือกมา 5 ชุดๆ ละ 500 ข้อมูล ในตารางที่ 9 เปรียบเทียบหาประสิทธิภาพของแบบจำลองในการทำนายเหตุการณ์กระแสไฟฟ้าขัดข้องที่ได้กับค่าจริงที่มีอยู่

ตารางที่ 9 ชุดข้อมูลทดสอบจำนวน 5 ชุดๆ ละ 500 ข้อมูล

สาเหตุ	ข้อมูลทั้งหมด	จำนวนข้อมูล				
		ชุดที่ 1	ชุดที่ 2	ชุดที่ 3	ชุดที่ 4	ชุดที่ 5
ต้นไม้	587	26	33	38	31	26
สัตว์	2,742	194	217	196	190	211
อุปกรณ์	1,239	69	70	79	77	68
ภัยธรรมชาติ	524	47	29	31	31	47
อื่นๆ	116	2	5	6	6	7
สภาพแวดล้อม	1,926	162	146	150	165	141

- ใช้ข้อมูลทดสอบในปี พ.ศ. 2551 ซึ่งมีข้อมูลที่ไม่ทราบสาเหตุจำนวน 1,745 ข้อมูล นำมาสุ่มเลือกมา 100 ข้อมูล ในตารางที่ 10 เพื่อผลการทำนายของแบบจำลองกับข้อมูลประเภทที่ระบุว่าไม่ทราบสาเหตุและได้สรุปสาเหตุจากการสันนิษฐาน

ตารางที่ 10 ชุดข้อมูลทดสอบที่ไม่ทราบสาเหตุจำนวน 100 ข้อมูล

สาเหตุ	จำนวนข้อมูล
ต้นไม้	5
สัตว์	28
อุปกรณ์	10
ภัยธรรมชาติ	9
อื่นๆ	6
สภาพแวดล้อม	42

- ใช้ข้อมูลทดสอบในปี พ.ศ. 2551 ซึ่งมีข้อมูลที่ไม่ทราบสาเหตุจำนวน 1,745 ข้อมูล นำมาสุ่มเลือกมา 5 ข้อมูล ในตารางที่ 11 โดยจะตัดการระบุสาเหตุในแอททริบิวต์เป้าหมายออกเพื่อจำลองการเกิดเหตุการณ์กระแสไฟฟ้าขัดข้องในปัจจุบัน และให้แบบจำลองทำนายสาเหตุจากชุดข้อมูลพร้อมแสดงผลอย่างละเอียด

ตารางที่ 11 ชุดข้อมูลทดสอบที่ไม่ทราบสาเหตุทั้งหมด 1,745 ข้อมูล สุ่มเลือกมา 5 ข้อมูล

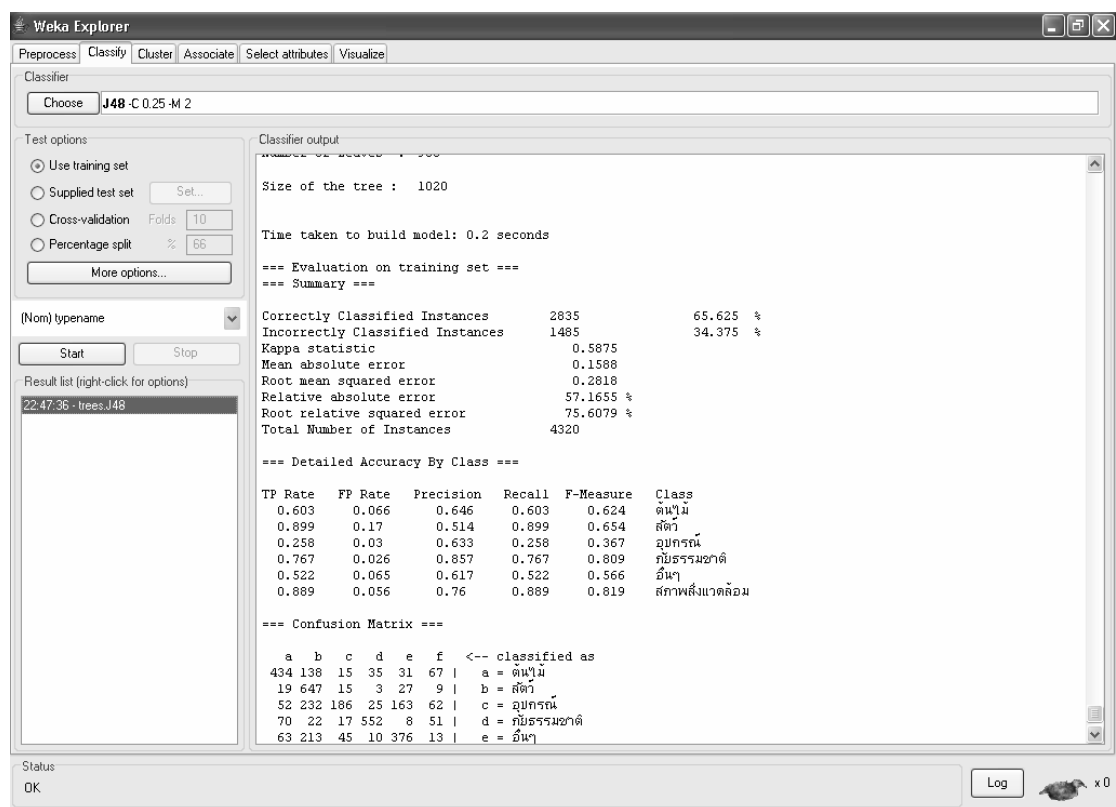
SubID	Date outage	Time outage	Total Time	Caption	Weather	Pha_A	Pha_B	Pha_C	Pha_G	Type name
NYA	มิ.ย.	To3	Tto10	D/F	ฝนตก_ลมแรง	none	none	none	none	?
OKA	พ.ค.	To3	Tto3	T/L	อากาศปกติ	none	time	time	time	?
(Instantaneous)										
ATB	ม.ค.	To4	Tto16	D/F	อากาศปกติ	none	Ins	Ins	none	?
NSI	ก.ย.	To4	Tto19	D/F	ฝนตก_ฟ้าคะนอง	none	none	none	none	?
RSA	ต.ค.	To3	Tto16	D/F	อากาศปกติ	none	none	time	time	?

ผลและวิจารณ์

ผล

1. ผลจากการสร้างแบบจำลอง

จากชุดข้อมูลเรียนรู้ที่ได้เตรียมไว้ชนิดที่ทราบสาเหตุ 3 ชุดๆละ 16,656 ข้อมูล นำมาวิเคราะห์กับแบบจำลอง โดยการนำชุดข้อมูลเรียนรู้มาทดสอบที่ละชุดกับอัลกอริทึมทั้ง 3 แบบ คือ อัลกอริทึมกฎของริบเปอร์ อัลกอริทึมนาอิวเบย์เซียน และอัลกอริทึมต้นไม้ตัดสินใจ ผลจากการวิเคราะห์ในโปรแกรม ได้แสดงในภาพที่ 21



ภาพที่ 21 ลักษณะผลที่ได้ในโปรแกรม จากการหาแบบจำลอง

จากผลในตารางที่ 12 เมื่อพิจารณาค่าความถูกต้อง Correctly Classified Instances และค่าความคลาดเคลื่อน Root Mean Squared Error (RMSE) แบบจำลองในการเรียนรู้จากอัลกอริทึมต้นไม้ไม่มีค่าที่ดีที่สุด

พิจารณาแบบจำลองในการเรียนรู้จากอัลกอริทึมต้นไม้ในข้อมูลชุดที่ 3 สามารถพิจารณาผลได้ดังนี้

- ค่าความถูกต้องสูงที่สุด 97.96%
- ค่าเฉลี่ยความคลาดเคลื่อน ต่ำที่สุด 7.30%

เมื่อเปรียบเทียบกับผลของอัลกอริทึมทั้งสอง จึงเหมาะกับการนำมาเป็นแบบจำลองในการทำนายสาเหตุการเกิดกระแสไฟฟ้าขัดข้องต่อไป

ตารางที่ 12 ค่าความถูกต้องและค่าเฉลี่ยความคลาดเคลื่อน จากการสร้างแบบจำลองในการเรียนรู้จากชุดข้อมูลเรียนรู้ ที่ 1-3

ผลเปรียบเทียบ		JRip'Rule	Naïve Bayesian	Decision Tree
		(%)	(%)	(%)
ข้อมูลชุดที่ 1	Correctly Instances	96.57	94.30	97.58
	Incorrectly Instances	3.43	5.70	2.42
	RMSE	10.16	12.39	7.94
ข้อมูลชุดที่ 2	Correctly Instances	96.60	94.45	97.80
	Incorrectly Instances	3.40	5.55	2.20
	RMSE	10.22	12.26	7.54
ข้อมูลชุดที่ 3	Correctly Instances	96.72	94.29	97.96
	Incorrectly Instances	3.28	5.71	2.04
	RMSE	10.04	12.36	7.30

การพิจารณาผลจากค่า Confusion Matrix ในตารางที่ 13-21 พบว่าในตารางที่ 19-21 ซึ่งเป็นผลจากการใช้อัลกอริทึมพบว่ามีการกระจายตัวของข้อมูลผลลัพธ์ที่มีลักษณะคล้ายกันและอัลกอริทึมต้นไม้ตัดสินใจมีจำนวนข้อมูลค่าจริง กับจำนวนข้อมูลจากการทำนายมีผลลัพธ์ตรงกันมากกว่าผลที่ได้จากอัลกอริทึมกฎของริบเปอร์ และอัลกอริทึมนาอีฟเบย์เซียน

ตารางที่ 13 ค่า Confusion Matrix ของการหาแบบจำลองจากข้อมูลชุดที่ 1 โดยอัลกอริทึมกฎของริบเปอร์

PREDICTED CLASS						ACTUAL CLASS	
a	b	c	d	e	f		
1974	34	23	12	1	14		a = ต้นไม้
15	7276	45	9	11	17		b = สัตว์
74	63	2405	21	10	64		c = อุปกรณ์
16	4	8	372	0	12		d = ภัยธรรมชาติ
3	7	7	0	59	1		e = อื่นๆ
56	11	26	8	0	3998		f = สภาพสิ่งแวดล้อม

ตารางที่ 14 ค่า Confusion Matrix ของการหาแบบจำลองจากข้อมูลชุดที่ 2 โดยอัลกอริทึมกฎของริบเปอร์

PREDICTED CLASS						ACTUAL CLASS	
a	b	c	d	e	f		
1983	39	18	5	1	17		a = ต้นไม้
9	7446	43	6	1	14		b = สัตว์
92	78	2335	12	8	76		c = อุปกรณ์
18	6	4	368	0	12		d = ภัยธรรมชาติ
1	13	9	0	45	3		e = อื่นๆ
54	11	13	3	0	3913		f = สภาพสิ่งแวดล้อม

ตารางที่ 15 ค่า Confusion Matrix ของการหาแบบจำลองจากข้อมูลชุดที่ 3 โดยอัลกอริทึมกฎของ
ริบเปอร์

PREDICTED CLASS						ACTUAL CLASS	
a	b	c	d	e	f		
1926	29	24	10	0	7		a = ต้นไม้
12	7380	57	8	3	10		b = สัตว์
55	75	2396	13	2	78		c = อุปกรณ์
12	4	7	411	0	12		d = ภัยธรรมชาติ
2	8	10	0	35	0		e = อื่นๆ
70	11	23	5	0	3961		f = สภาพสิ่งแวดล้อม

ตารางที่ 16 ค่า Confusion Matrix ของการหาแบบจำลองจากข้อมูลชุดที่ 1 โดยอัลกอริทึม
นาอีฟเบย์เซียน

PREDICTED CLASS						ACTUAL CLASS	
a	b	c	d	e	f	<-- classified as	
1874	40	67	46	1	30		a = ต้นไม้
18	7272	36	24	8	15		b = สัตว์
89	152	2148	37	15	196		c = อุปกรณ์
27	7	4	358	0	16		d = ภัยธรรมชาติ
1	19	11	1	45	0		e = อื่นๆ
52	16	7	14	0	4010		f = สภาพสิ่งแวดล้อม

ตารางที่ 17 ค่า Confusion Matrix ของการหาแบบจำลองจากข้อมูลชุดที่ 2 โดยอัลกอริทึม
 นาอีฟเบย์เขียน

PREDICTED CLASS						ACTUAL CLASS	
a	b	c	d	e	f	<-- classified as	
1904	39	62	36	1	21		a = ต้นไม้
18	7433	32	22	3	11		b = สัตว์
97	180	2097	39	15	173		c = อุปกรณ์
15	8	6	354	0	25		d = ภัยธรรมชาติ
1	22	15	5	28	0		e = อื่นๆ
52	11	8	7	0	3916		f = สภาพสิ่งแวดล้อม

ตารางที่ 18 ค่า Confusion Matrix ของการหาแบบจำลองจากข้อมูลชุดที่ 3 โดยอัลกอริทึม
 นาอีฟเบย์เขียน

PREDICTED CLASS						ACTUAL CLASS	
a	b	c	d	e	f	<-- classified as	
1811	41	75	55	2	12		a = ต้นไม้
13	7382	37	19	0	19		b = สัตว์
94	167	2098	28	20	212		c = อุปกรณ์
9	4	3	413	0	17		d = ภัยธรรมชาติ
1	14	10	0	29	1		e = อื่นๆ
62	17	9	9	0	3973		f = สภาพสิ่งแวดล้อม

ตารางที่ 19 ค่า Confusion Matrix ของการหาแบบจำลองจากข้อมูลชุดที่ 1 โดยอัลกอริทึมต้นไม้ตัดสินใจ

PREDICTED CLASS						ACTUAL CLASS	
a	b	c	d	e	f	<-- classified as	
2003	14	16	13	2	10		a = ต้นไม้
17	7294	44	6	9	3		b = สัตว์
36	49	2500	8	8	36		c = อุปกรณ์
16	11	13	369	0	3		d = ภัยธรรมชาติ
4	6	10	1	56	0		e = อื่นๆ
15	10	33	10	0	4031		f = สภาพสิ่งแวดล้อม

ตารางที่ 20 ค่า Confusion Matrix ของการหาแบบจำลองจากข้อมูลชุดที่ 2 โดยอัลกอริทึมต้นไม้ตัดสินใจ

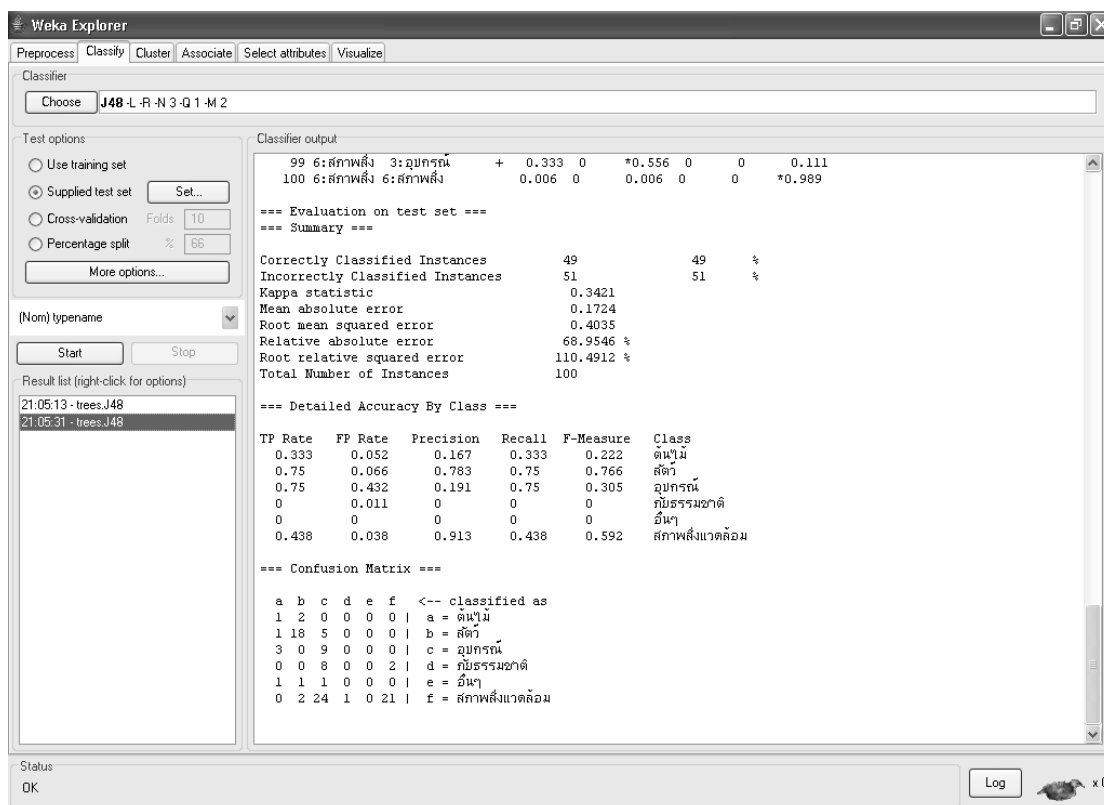
PREDICTED CLASS						ACTUAL CLASS	
a	b	c	d	e	f	<-- classified as	
2003	18	24	9	3	6		a = ต้นไม้
13	7454	36	10	3	3		b = สัตว์
35	36	2483	3	12	32		c = อุปกรณ์
12	9	15	368	3	1		d = ภัยธรรมชาติ
1	8	16	2	44	0		e = อื่นๆ
10	8	32	5	0	3939		f = สภาพสิ่งแวดล้อม

ตารางที่ 21 ค่า Confusion Matrix ของการหาแบบจำลองจากข้อมูลชุดที่ 3 โดยอัลกอริทึมต้นไม้ตัดสินใจ

PREDICTED CLASS						ACTUAL CLASS	
a	b	c	d	e	f	<-- classified as	
1941	22	11	15	0	7		a = ต้นไม้
18	7409	24	6	5	8		b = สัตว์
17	42	2509	11	2	38		c = อุปกรณ์
11	6	12	415	0	2		d = ภัยธรรมชาติ
2	6	10	0	36	1		e = อื่นๆ
12	10	32	9	0	4007		f = สภาพสิ่งแวดล้อม

2. ผลจากการทดสอบแบบจำลองกับชุดข้อมูลทดสอบที่ทราบสาเหตุ

นำข้อมูลที่ทราบสาเหตุทั้งหมด 7,328 ข้อมูลเหตุการณ์นำมาสุ่มเลือกมา 5 ชุดๆ ละ 500 ทดสอบกับแบบจำลองที่เลือกมาจากการใช้อัลกอริทึมต้นไม้ตัดสินใจ เพราะให้ค่าที่ถูกต้องมากที่สุดและค่าความคลาดเคลื่อนน้อยที่สุด ผลการใช้แบบจำลองจากอัลกอริทึมต้นไม้เพื่อการตัดสินใจ ทดสอบกับชุดข้อมูลทดสอบจำนวน 5 ชุด แสดงผลจากการวิเคราะห์ในโปรแกรม Weka ตามรูปที่ 22



ภาพที่ 22 ลักษณะการแสดงผลในโปรแกรม Weka ที่ได้จากการนำแบบจำลองทดสอบกับชุดข้อมูลทดสอบ

ผลการใช้แบบจำลองจากอัลกอริทึมต้นไม้เพื่อการตัดสินใจ ทดสอบกับชุดข้อมูลทดสอบจำนวน 5 ชุด ให้ผลการทดสอบแสดงในรูป Confusion Matrix ในตารางที่ 24-28 และในตารางที่ 22 มีความถูกต้องโดยเฉลี่ยมีค่ามากกว่า 82.32% ซึ่งเป็นค่าความถูกต้องที่ค่อนข้างสูง สามารถนำไปใช้ประกอบการพิจารณาสาเหตุได้

วิเคราะห์ข้อมูลที่ทำนายผิดพลาดเฉลี่ยประมาณ 17.68 % เมื่อดูผลที่ได้จากการทดสอบจากข้อมูลทดสอบทั้ง 5 ชุด จากค่า True Positive rate (TP rate) ในตารางที่ 23 เห็นได้ว่าการทำนายสาเหตุภัยธรรมชาติมีค่าเฉลี่ย 0.2960 และสาเหตุอื่นๆ มีค่าเฉลี่ย 0.0334

ตารางที่ 22 ผลการทดสอบชุดข้อมูลทดสอบ 5 ชุดๆ ละ 500 ข้อมูล

ผลการทดสอบ(%)	Testset 1	Testset 2	Testset 3	Testset 4	Testset 5
Correctly Instances	82.00	84.80	80.60	81.40	82.80
Incorrectly Instances	18.00	15.20	19.40	18.60	17.20
RMSE	24.08	21.40	24.45	24.26	23.25

ตารางที่ 23 ค่า True Positive rate (TP rate) จากผลทดสอบชุดข้อมูลทดสอบ 5 ชุดๆ ละ 500 ข้อมูล

ข้อมูลชุดที่	ค่า True Positive rate (TP rate)					
	ต้นไม้	สัตว์	อุปกรณ์	ภัยธรรมชาติ	อื่นๆ	สิ่งแวดล้อม
1	0.7310	0.9790	0.7100	0.3190	0	0.8460
2	0.7880	0.9680	0.7710	0.2410	0	0.8700
3	0.6320	0.9690	0.6840	0.2900	0	0.8400
4	0.8060	0.9680	0.6490	0.2900	0.1670	0.8360
5	0.7690	0.9860	0.6620	0.3400	0	0.8870
เฉลี่ย	0.7452	0.9740	0.6952	0.2960	0.0334	0.8558

ตารางที่ 24 ค่า Confusion Matrix จากผลทดสอบชุดข้อมูลทดสอบ 5 ชุดๆ ละ 500 ข้อมูล ของชุดที่ 1

PREDICTED CLASS						ACTUAL CLASS	
a	b	c	d	e	f	<-- classified as	
19	1	4	0	0	2		a = ต้นไม้
2	190	2	0	0	0		b = สัตว์
8	4	49	1	0	7		c = อุปกรณ์
16	1	4	15	0	11		d = ภัยธรรมชาติ
0	0	2	0	0	0		e = อื่นๆ
5	1	11	8	0	137		f = สภาพสิ่งแวดล้อม

ตารางที่ 25 ค่า Confusion Matrix จากผลทดสอบชุดข้อมูลทดสอบ 5 ชุดๆละ 500 ข้อมูล
ของชุดที่ 2

PREDICTED CLASS						ACTUAL CLASS
a	b	c	d	e	f	<-- classified as
26	1	3	0	0	3	a = ต้นไม้
1	210	5	0	0	1	b = สัตว์
3	2	54	0	0	11	c = อุปกรณ์
15	2	2	7	0	3	d = ภัยธรรมชาติ
1	2	2	0	0	0	e = อื่นๆ
5	0	5	9	0	127	f = สภาพสิ่งแวดล้อม

ตารางที่ 26 ค่า Confusion Matrix จากผลทดสอบชุดข้อมูลทดสอบ 5 ชุดๆ ละ 500 ข้อมูล
ของชุดที่ 3

PREDICTED CLASS						ACTUAL CLASS
a	b	c	d	e	f	<-- classified as
24	1	7	0	0	6	a = ต้นไม้
4	190	2	0	0	0	b = สัตว์
6	10	54	0	0	9	c = อุปกรณ์
12	0	6	9	0	4	d = ภัยธรรมชาติ
0	1	5	0	0	0	e = อื่นๆ
6	4	5	9	0	126	f = สภาพสิ่งแวดล้อม

ตารางที่ 27 ค่า Confusion Matrix จากผลทดสอบชุดข้อมูลทดสอบ 5 ชุดๆ ละ 500 ข้อมูล
ของชุดที่ 4

PREDICTED CLASS						ACTUAL CLASS
a	b	c	d	e	f	<-- classified as
25	0	5	0	0	1	a = ต้นไม้
3	184	2	0	0	1	b = สัตว์
9	7	50	0	0	11	c = อุปกรณ์
17	0	3	9	0	2	d = ภัยธรรมชาติ
1	1	3	0	1	0	e = อื่นๆ
6	4	7	10	0	138	f = สภาพสิ่งแวดล้อม

ตารางที่ 28 ค่า Confusion Matrix จากผลทดสอบชุดข้อมูลทดสอบ 5 ชุดๆ ละ 500 ข้อมูล
ของชุดที่ 5

PREDICTED CLASS						ACTUAL CLASS
a	b	c	d	e	f	<-- classified as
20	4	0	0	0	2	a = ต้นไม้
2	208	1	0	0	0	b = สัตว์
6	9	45	0	0	8	c = อุปกรณ์
20	0	3	16	0	8	d = ภัยธรรมชาติ
1	1	5	0	0	0	e = อื่นๆ
1	2	9	4	0	125	f = สภาพสิ่งแวดล้อม

ตารางที่ 29 ผลรวมในการทำนายหาสาเหตุอื่นๆ จาก Confusion Matrix ของชุดข้อมูลทดสอบ 5 ชุดๆ ละ 500 ข้อมูล

ข้อมูลชุดที่	จำนวนข้อมูลเหตุการณ์ทำนายสาเหตุอื่นๆ					
	ต้นไม้	สัตว์	อุปกรณ์	ภัยธรรมชาติ	อื่นๆ	สิ่งแวดล้อม
1	1	1	5	0	0	0
2	1	2	2	0	0	0
3	0	1	5	0	0	0
4	1	1	3	0	1	0
5	0	0	2	0	0	0

ผลจากตารางที่ 29 พบว่าการทำนายสาเหตุอุปกรณ์ของแบบจำลองมักทำนายเป็นสาเหตุเกิดจากอุปกรณ์ เป็นส่วนใหญ่ และมีสาเหตุต้นไม้ สัตว์ เป็นส่วนน้อย และไม่สามารถทำนายเป็นสาเหตุอื่นๆ ได้เลยและผลจากตารางที่ 30 พบว่าการทำนายสาเหตุภัยธรรมชาติของแบบจำลองจะทำนายเป็นสาเหตุเกิดจากต้นไม้ สิ่งแวดล้อม และอุปกรณ์ เป็นส่วนใหญ่

ตารางที่ 30 ผลรวมในการทำนายหาสาเหตุ ภัยธรรมชาติ จาก Confusion Matrix ของชุดข้อมูลทดสอบ 5 ชุดๆ ละ 500 ข้อมูล

ข้อมูลชุดที่	จำนวนข้อมูลเหตุการณ์ทำนายสาเหตุภัยธรรมชาติ					
	ต้นไม้	สัตว์	อุปกรณ์	ภัยธรรมชาติ	อื่นๆ	สิ่งแวดล้อม
1	16	1	4	15	0	11
2	15	2	2	7	0	3
3	12	0	6	9	0	4
4	17	0	3	9	0	2
5	20	0	3	16	0	8

3. ผลจากการทดสอบจำลองกับชุดทดสอบที่ไม่ทราบสาเหตุ

ใช้ข้อมูลทดสอบในปี พ.ศ. 2551 ซึ่งเป็นชุดข้อมูลที่ไม่ทราบสาเหตุจำนวน 1,745 ข้อมูล นำมาสุ่มเลือกมา 100 ข้อมูล เพื่อผลการทำนายของแบบจำลองกับข้อมูลประเภทไม่ทราบสาเหตุ และได้สรุปสาเหตุจากการสันนิษฐาน ผลทดสอบข้อมูลที่ไม่ทราบสาเหตุจำนวน 1,745 ข้อมูล นำมาสุ่มเลือกมา 100 ข้อมูล ได้ผลดังนี้

- ค่าความถูกต้อง 58 %
- ค่าเฉลี่ยความผิดพลาด 42 % (โดยค่าความถูกต้องและค่าเฉลี่ยความผิดพลาดเทียบกับสาเหตุจากการสันนิษฐานที่ได้ระบุจากข้อมูล)
- ค่า False Positive rate (FP rate) คือ ข้อมูลที่ไม่ได้อยู่ในคลาสนั้นแต่คำตอบจากโมเดลบอกว่าอยู่ในคลาสนั้น และค่า True Positive rate (TP rate) คือข้อมูลที่เป็นผลจากการทำนายด้วยโมเดลต้นไม่ตัดสินใจ ที่สร้างขึ้นมาแล้วให้คำตอบตรงกับความเป็นจริง มีค่าตามตารางที่ 17 เห็นได้ว่าสาเหตุที่เกิดจาก ภัยธรรมชาติ มีค่า FT rate = 0 และ TP rate = 0.011 และเกิดจาก อื่นๆ มีค่า FT rate = 0 และ TP rate = 0 แสดงให้เห็นว่า การทำนายสองสาเหตุนี้มีความผิดพลาดสูง

ตารางที่ 31 ผลทดสอบข้อมูลที่ไม่ทราบสาเหตุจำนวน 100 ข้อมูล

สาเหตุที่ทำให้เกิดเหตุการณ์	ค่า TP Rate	ค่า FP Rate
ต้นไม้	0.600	0.021
สัตว์	0.821	0.097
อุปกรณ์	0.700	0.344
ภัยธรรมชาติ	0	0.011
อื่นๆ	0	0
สภาพสิ่งแวดล้อม	0.595	0.017

เมื่อพิจารณาจากค่า Confusion Matrix ที่ได้จากการทดสอบจากโปรแกรมพบว่าสาเหตุภัยธรรมชาติ จะตอบเป็นเกิดจากอุปกรณ์ และสาเหตุอื่นๆ จะตอบว่าเกิดจากต้นไม้ สัตว์ และอุปกรณ์

ตารางที่ 32 ค่า Confusion Matrix ของผลทดสอบข้อมูลที่ไม่ทราบสาเหตุจำนวน 100 ข้อมูล

PREDICTED CLASS						ACTUAL CLASS
a	b	c	d	e	f	-- classified as
3	2	0	0	0	0	a = ต้นไม้
1	23	4	0	0	0	b = สัตว์
0	2	7	0	0	1	c = อุปกรณ์
0	0	9	0	0	0	d = ภัยธรรมชาติ
1	3	2	0	0	0	e = อื่นๆ
0	0	16	1	0	25	f = สภาพสิ่งแวดล้อม

4. ผลจากการทดสอบชุดข้อมูลไม่ทราบสาเหตุจำนวน 5 ข้อมูล

จากข้อมูลที่ไม่ทราบสาเหตุจำนวน 1,745 ข้อมูล นำมาสุ่มเลือกมา 5 ข้อมูล โดยจะตัดการระบุสาเหตุออกเพื่อให้แบบจำลองทำนายสาเหตุพร้อมแสดงผลอย่างละเอียด ผลทดสอบข้อมูลที่ไม่ทราบสาเหตุจำนวน 5 ข้อมูล หรือ 5 เหตุการณ์ ในตารางที่ 33 แสดงผลทำนายโดยละเอียด การพิจารณาเลือกสาเหตุ มาเป็นผลการทำนาย ทำโดยเลือกสาเหตุที่มีความน่าจะเป็นที่มีค่าสูงที่สุด โดยแต่ละสาเหตุจะแสดงค่าความน่าจะเป็น ซึ่งสามารถนำมาช่วยพิจารณาได้ กรณีการทำนายไม่ถูกต้อง สาเหตุที่มีความน่าจะเป็นรองลงมาเป็นสาเหตุที่จะพิจารณาต่อมา เพื่อนำสาเหตุที่ได้ไปตรวจสอบต่อไป

ในข้อมูลที่ 1 แสดงผลการทำนายให้เกิดจากสาเหตุ ต้นไม้ โดยให้ค่าความน่าจะเป็นที่จะเกิดจากสาเหตุจากต้นไม้ ถึง 100 %

ในข้อมูลที่ 2 แสดงผลการทำนายให้เกิดจากสาเหตุ อื่นๆ โดยให้ค่าความน่าจะเป็นที่จะเกิดจากสองสาเหตุเท่ากัน คือ 4.4 % และให้โดยให้ค่าความน่าจะเป็นที่จะเกิดจากสาเหตุอุปกรณ์ 95.6%

ในข้อมูลที่ 4 แสดงผลการทำนายที่เกิดจากสาเหตุ ต้นไม้และอุปกรณ์ โดยให้ค่าความน่าจะเป็นที่จะเกิดจากสองสาเหตุเท่ากัน คือ 0.6 % และให้โดยให้ค่าความน่าจะเป็นที่จะเกิดจากสาเหตุสภาพสิ่งแวดล้อม 98.8%

ตารางที่ 33 ผลการทดสอบชุดข้อมูล 5 ข้อมูล

ข้อมูลที่	ความน่าจะเป็นของแต่ละสาเหตุ						การทำนาย
	ต้นไม้	สัตว์	อุปกรณ์	ภัยธรรมชาติ	อื่นๆ	สภาพสิ่งแวดล้อม	
1	1	0	0	0	0	0	ต้นไม้
2	0	0	0.956	0	0.044	0	อุปกรณ์
3	0	1	0	0	0	0	สัตว์
4	0.006	0	0.006	0	0	0.989	สิ่งแวดล้อม
5	0	1	0	0	0	0	สัตว์

วิจารณ์

1. ผลจากการสร้างแบบจำลอง

การจำแนกข้อมูลสามารถช่วยในการวิเคราะห์แนวโน้มของข้อมูลได้โดยวิเคราะห์จากแอท-ทริบิวต์ในที่นี้ ได้เลือกแอททริบิวต์สาเหตุการเกิดกระแสไฟฟ้าขัดข้องเป็นแอททริบิวต์เป้าหมายเพื่อใช้ทำนาย แล้วนำแอททริบิวต์อื่นๆ มาใช้ในการวิเคราะห์

ขั้นตอนแรกคือการสร้างแบบจำลองโดยเรียนรู้จากชุดข้อมูลเรียนรู้ที่ได้กำหนดคลาสไว้เรียบร้อยแล้ว จากทั้งสามอัลกอริทึม ต้นไม้ตัดสินใจ กฎของริบเปอร์ และนาอีฟเบย์เซียน การเลือกใช้แบบจำลองจะพิจารณาการตรวจสอบความถูกต้องโดยใช้ชุดข้อมูลทดสอบเป็นข้อมูลที่ทราบอยู่แล้วว่าอยู่กลุ่มใดไปเปรียบเทียบกับกัน เพื่อทดสอบว่าโมเดลนี้สามารถจัดกลุ่มประเภทข้อมูลได้อย่างถูกต้องหรือไม่ โดยจะนำคลาสที่แท้จริงของข้อมูลเปรียบเทียบกับคลาสที่ได้จากโมเดล พบว่าความถูกต้องของแบบจำลองที่สร้างขึ้นโดยอัลกอริทึมต้นไม้ตัดสินใจมีค่าสูงสุด กล่าวคือ

แบบจำลองสามารถจัดกลุ่มหรือทำนายสาเหตุกระแสไฟฟ้าขัดข้องได้ใกล้เคียงกับสาเหตุจริงของข้อมูลที่นำมาทดสอบ

อัลกอริทึมต้นไม้ตัดสินใจเป็นที่นิยมกันมากเนื่องจากเป็นลักษณะที่คนจำนวนมากคุ้นเคย ทำให้เข้าใจได้ง่าย มีลักษณะเหมือนแผนภูมิองค์กร โดยที่แต่ละโหนดแสดงแอททริบิวต์ แต่ละกิ่งแสดงผลในการทดสอบ และลิฟโหนดแสดงคลาสที่กำหนดไว้ ในแต่ละแอททริบิวต์สามารถแบ่งได้มากกว่าสองคลาส อีกทั้งสามารถรองรับชุดข้อมูลแบบต่อเนื่องได้ด้วย จึงทำให้การวิเคราะห์ใช้เวลาที่สั้นและเข้าใจผลลัพธ์ง่าย ข้อมูลจากโปรแกรม จฟ.3 ประกอบไปด้วยแอททริบิวต์ที่มีความเกี่ยวข้องของข้อมูลน้อย และมีการแบ่งหลายคลาส อีกทั้งบางแอททริบิวต์เป็นข้อมูลแบบต่อเนื่อง เช่น ช่วงเวลาที่ไฟฟ้าดับ วันที่เกิดเหตุการณ์ เป็นต้น การใช้อัลกอริทึมต้นไม้ตัดสินใจ จึงสะดวกในการวิเคราะห์ และได้ผลการวิเคราะห์ที่ถูกต้องใกล้เคียงข้อมูลเหตุการณ์กระแสไฟฟ้าขัดข้องมากที่สุด ส่วนอัลกอริทึมนาอิวเบย์เซียนเป็นการใช้ความรู้ก่อนหน้าเข้ามาช่วยในการสร้างเงื่อนไข ซึ่งหากข้อมูลไม่ถูกต้องแล้ว ก็จะส่งผลถึงการสร้างแบบจำลองรวมไปถึงความถูกต้องแม่นยำในการจัดกลุ่มข้อมูลด้วย อัลกอริทึมกฎของริบเปอร์จะใช้หลักการจำแนกข้อมูลที่ละสองคลาสซึ่งไม่เหมาะสมกับข้อมูลของ จฟ.3 เนื่องจากแต่ละแอททริบิวต์มีหลายคลาสจะส่งผลให้ต้องใช้เวลาในการประมวลผลมากกว่าวิธีอื่นๆ อีกทั้งยังเป็นการลดความถูกต้องแม่นยำของกฎที่ถูกสร้างขึ้นจากข้อมูล

2. ผลจากการทดสอบแบบจำลองกับชุดข้อมูลทดสอบที่ทราบสาเหตุ

เมื่อนำแบบจำลองที่ได้มาทดสอบกับชุดข้อมูลทดสอบ เพื่อหาความแม่นยำในการทำนายข้อมูล เนื่องจาก ค่าความคลาดเคลื่อนที่ได้จากการสร้างแบบจำลองไม่สามารถบอกได้ว่าแบบจำลองนั้นมีความแม่นยำเพียงใด ผลการทดสอบค่าความแม่นยำ (Correctly Instances) จะมีค่าน้อยกว่าค่าความแม่นยำที่ได้จากการสร้างแบบจำลอง

3. ผลจากการทดสอบแบบจำลองกับชุดทดสอบที่ไม่ทราบสาเหตุ

การนำแบบจำลองมาทดลองใช้กับข้อมูลที่ไม่ทราบสาเหตุ เป็นการแสดงการประยุกต์ใช้แบบจำลองเพื่อทำนายสาเหตุการเกิดกระแสไฟฟ้าขัดข้อง ซึ่งจะทดสอบกับข้อมูลที่มีการกรอกข้อมูลจากการสันนิษฐานเพื่อเปรียบเทียบผลที่ได้จากการทำนาย กับการสันนิษฐาน

จากการทดสอบแบบจำลองกับชุดข้อมูลทดสอบที่ทราบสาเหตุกับชุดข้อมูลทดสอบที่ไม่ทราบสาเหตุพบว่า ผลที่ได้มีจำนวนข้อมูลที่แบบจำลองทำนายสาเหตุผิดและมีค่าความคลาดเคลื่อนสูง ซึ่งจากการวิเคราะห์พบว่าเกิดจากข้อมูลที่ระบุสาเหตุอื่นๆ และสาเหตุภัยธรรมชาติจากผลทดสอบที่ได้สามารถสรุปได้ดังนี้

- พบว่า สาเหตุอื่นๆ มีจำนวนเหตุการณ์ที่เกิดขึ้นน้อย ส่งผลให้ไม่สามารถทำนายได้เนื่องจากเป็นเหตุการณ์ที่เกิดขึ้นไม่บ่อยและการเกิดเหตุการณ์แต่ละครั้งมักจะเกิดอย่างอิสระไม่ขึ้นกับเวลา สถานที่ สภาพอากาศ เช่น อุบัติเหตุจากคน สิ่งของ รถยนต์ เป็นต้น

- ในส่วนของข้อมูลเหตุการณ์ที่ระบุสาเหตุจากภัยธรรมชาติ อาจเนื่องจากขณะเกิดภัยธรรมชาติ มีสาเหตุอื่นเกิดขึ้นซ้อน เช่น สาเหตุจากต้นไม้ อุปกรณ์ เป็นต้น ทำให้การทำงานของอุปกรณ์ป้องกันเป็นไปตามสาเหตุซ้อนที่เกิดขึ้น จึงอาจทำให้ผลการวิเคราะห์คลาดเคลื่อนจากความเป็นจริง

- การบันทึกข้อมูลโดยอาศัยประสบการณ์ในการสันนิษฐานสาเหตุของผู้กรอกขณะเกิดเหตุการณ์แต่ละครั้ง และการเข้าใจสภาพแวดล้อมต่างๆ อาจส่งผลให้เกิดข้อมูลที่บ่งบอกคุณลักษณะแตกต่างกันในส่วนของบางแอททริบิวต์ได้

4. ผลจากการทดสอบชุดข้อมูลไม่ทราบสาเหตุจำนวน 5 ข้อมูล

แบบจำลองสามารถนำมาประยุกต์ใช้ในการทำนายสาเหตุที่ไม่ทราบสาเหตุ ซึ่งในการทดสอบนี้ใช้ข้อมูลที่ไม่ทราบสาเหตุจำนวน 5 ข้อมูลจำลองการเกิดกระแสไฟฟ้าขัดข้องในเหตุการณ์ปัจจุบัน สามารถนำข้อมูลแวดล้อมตามแอททริบิวต์ที่ได้กำหนดไว้มาทดสอบกับแบบจำลองที่มีอยู่ สามารถแสดงผลพร้อมอย่างละเอียดออกมาเป็นค่าเปอร์เซ็นต์ความน่าจะเป็นของสาเหตุ สามารถใช้คำดังกล่าวช่วยในการพิจารณาในการสันนิษฐานเบื้องต้นของสาเหตุที่มีค่าความน่าจะเป็นสูงและรองลงมาให้เป็นสาเหตุรอง

สรุปและข้อเสนอแนะ

สรุป

การหาแบบจำลองเพื่อวิเคราะห์หาสาเหตุที่ทำให้เกิดเหตุการณ์กระแสไฟฟ้าขัดข้องที่ไม่ทราบสาเหตุโดยใช้อัลกอริทึมที่ต่างกัน ได้แก่ ต้นไม้ตัดสินใจ เทคนิคแบบนาอ็อฟเบย์เซียน และกฎของริปเปอร์ พบว่าการใช้อัลกอริทึมต้นไม้ตัดสินใจให้ค่าความคลาดเคลื่อนน้อยที่สุดจึงเหมาะสำหรับใช้วิเคราะห์หาแบบจำลอง

การทดสอบแบบจำลองโดยการทดสอบกับชุดข้อมูลทดสอบทราบสาเหตุ พบว่าให้ผลที่ถูกต้องเกิน 80 เปอร์เซ็นต์

การทดสอบแบบจำลองโดยการทดสอบกับชุดข้อมูลที่ไม่ทราบสาเหตุ ซึ่งมีการระบุสาเหตุต้นนิษฐาน ผลการทดสอบยังมีค่าความคลาดเคลื่อนจากสาเหตุที่ต้นนิษฐานไว้ 42 เปอร์เซ็นต์ ซึ่งสามารถพิจารณาเลือกสาเหตุที่แท้จริงได้ทั้งจากการต้นนิษฐานหรือจากผลการทำนาย

ผลจากการทดสอบแบบจำลองโดยการทดสอบกับชุดข้อมูลทดสอบที่ทราบสาเหตุและไม่ทราบสาเหตุพบว่าความคลาดเคลื่อนส่วนใหญ่เกิดจาก ผลการทำนายของสาเหตุอื่นๆ กับสาเหตุจากภัยธรรมชาติ

การทดสอบแบบจำลองโดยการทดสอบกับชุดข้อมูลที่ไม่ทราบสาเหตุ โดยตัดสาเหตุที่ต้นนิษฐานออก ผลที่ได้สามารถแสดงค่าความน่าจะเป็นของแต่ละสาเหตุ ทำให้สะดวกในการนำไปใช้ประกอบการพิจารณาเพื่อบ่งชี้เหตุการณ์ที่มีความเป็นไปได้อย่างรวดเร็วในขณะเกิดเหตุการณ์ โดยการนำข้อมูลแวดล้อมที่เกี่ยวข้องมาใช้กับแบบจำลองที่ได้ ส่งผลให้แก้ไขได้ตรงสาเหตุและสามารถนำระบบคืนสู่สภาพเดิมในระยะเวลาอันรวดเร็ว

จากผลการวิจัยจะเห็นได้ว่าการใช้ต้นไม้ตัดสินใจเหมาะสมกับการนำมาใช้ทำแบบจำลองเนื่องจากให้ค่าความถูกต้องสูงและมีค่าความคลาดเคลื่อนน้อย ดังนั้นการนำหลักการการทำเหมืองข้อมูลและผลที่ได้จากงานวิจัยมาประยุกต์ใช้กับการทำงานจะมีส่วนช่วย กฟผ. ในการบันทึกข้อมูลกระแสไฟฟ้าขัดข้องมีความแม่นยำถูกต้อง และมีหลักเกณฑ์การระบุสาเหตุมากขึ้น จึงส่งผลให้

ข้อมูลของ กฟภ. ที่จะนำไปใช้งานมีความน่าเชื่อถือมากขึ้น หลักการดังกล่าวสามารถใช้เป็นเครื่องมือในการสนับสนุนด้านงานวิจัย และงานวิเคราะห์วางแผนต่อไป

ข้อเสนอแนะ

จากการผลการวิเคราะห์และทดสอบในงานวิจัยนี้พบว่ายังมีชุดข้อมูลอีกหลายส่วนที่มีความสำคัญเหมาะสำหรับนำมาใช้ในการวิเคราะห์และทดสอบนี้ ความถูกต้องของข้อมูลมีผลมากต่อรูปแบบการทำนาย การกรอกข้อมูลอย่างถูกต้องและตรงกับความเป็นจริงมากที่สุดจึงเป็นสิ่งสำคัญในการสร้างแบบจำลองที่มีประสิทธิภาพสูง ดังนั้นการกรอกข้อมูล ผู้ที่รับผิดชอบที่กรอกข้อมูลขณะที่เกิดเหตุการณ์กระแสไฟฟ้าขัดข้องจึงจำเป็นต้องมีความรู้และประสบการณ์สูง เข้าใจสภาพแวดล้อม และมีการวิเคราะห์เบื้องต้นจากข้อมูลที่มีความจำเป็นที่จะต้องพิจารณาข้อมูลหลายๆ ส่วนประกอบกัน ตลอดจนการสันนิษฐานเหตุการณ์ในขณะนั้น ทั้งนี้เพื่อให้สาเหตุที่ถูกต้อง อาจรวมไปถึงการจ่ายไฟให้กลับสู่สภาพปกติเป็นไปได้โดยรวดเร็วส่งผลให้มีระยะเวลาในการดับไฟที่น้อยลงเพิ่มระดับความเชื่อถือได้

เอกสารและสิ่งอ้างอิง

กรุง สีนอภิมย์สรานู. ม.ป.ป. **CSC662 Data Mining, Data Warehouse and Visualization.**
ชุดเอกสารประกอบการสอน ภาควิชาคณิตศาสตร์. คณะวิทยาศาสตร์ จุฬาลงกรณ์
มหาวิทยาลัย.

นฤพนธ์ ว่องประชาณุกุล. 2548. **วิธีที่เหมาะสมสำหรับการตัดกิ่งต้นไม้ตัดสินใจของการทำเหมือง
ข้อมูล ทางด้านวิทยาศาสตร์.** วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยเทคโนโลยีสุรนารี.

บุญเสริม กิจศิริกุล. 2546. **รายงานวิจัยฉบับสมบูรณ์ โครงการย่อยที่ 7 อัลกอริทึมการทำเหมือง
ข้อมูล.** ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

Breslow, L. A. and D. W. Aha. 1997. Simplifying decision trees: A survey. **Knowledge
Engineering Review.** 12(1): 1-40.

Fayyad, U. M. and K. B. Irani. 1993. **Multi-interval discretization of continuous-valued
attributes for classification learning**, pp. 1022-1027. In Bajcsy, R. (ed.), Proc. of 13th
IJCAI, vol.2, Chambery, France, Morgan-Kaufmann, San Francisco.

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

Quinlan, J. R. 1993. **C4.5: Programs for Machine Learning.** CA: Morgan Kaufmann,
San Mateo.

Quinlan, J. R. and R. L. Rivest. 1989. Inferring decision trees using the minimum description
length principle. **Information and Computation.** 80(3): 227-248.

Weka. n.d. **Data Mining Software in Java.** Available Source: [http://www.cs.waikato.
ac.nz/~ml/weka/](http://www.cs.waikato.ac.nz/~ml/weka/), April 29, 2008.

William W. C. 1995. **Fast effective rule induction**, pp. 115-123. In A. Prieditis & S. Russell (Eds.), Proceedings of the 12th International Conference on Machine Learning (ML-95), CA: Morgan Kaufmann, Lake Tahoe.

Witten, I. H. and E. Frank. 2005. **Data Mining Practical Machine Learning Tools and Technique**. 2nd edition, Elsevier, San Francisco.

Wu, C., D. Landgrebe and P. Swain. 1975. **The decision tree approach to classification**. School Elec. Eng, Purdue Univ., West Lafayette, IN, Rep. RE-EE 75-17.

ประวัติการศึกษา และการทำงาน

ชื่อ – นามสกุล	นายจรัสศักดิ์ ศรีอ่อน
วัน เดือน ปี ที่เกิด	วันที่ 19 เมษายน 2518
สถานที่เกิด	จังหวัดสกลนคร
ประวัติการศึกษา	ปี 2541 วิศวกรรมศาสตรบัณฑิต (วิศวกรรมไฟฟ้า) มหาวิทยาลัยภาคตะวันออกเฉียงเหนือ
ตำแหน่งหน้าที่การงานปัจจุบัน	วิศวกรระดับ 6
สถานที่ทำงานปัจจุบัน	การไฟฟ้าส่วนภูมิภาค เขต 1 ภาคกลาง (อยุธยา)