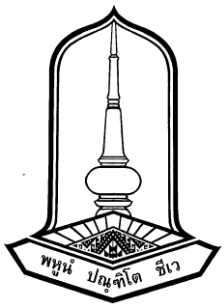


การคัดกรองผู้ป่วยโรคพาร์กินสันโดยใช้วิธีซัพพอร์ตเวคเตอร์แมชชีน และการค้นหาแบบกริด

ณัฐพล แสนคำ

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์
ธันวาคม 2559

ลิขสิทธิ์เป็นของมหาวิทยาลัยมหาสารคาม

Mahasarakham University

การคัดกรองผู้ป่วยโรคพาร์กินสันโดยใช้วิธีซัพพอร์ตเวคเตอร์แมชชีน และการค้นหาแบบกริด

ณัฐพล แสนคำ

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์
ธันวาคม 2559
ลิขสิทธิ์เป็นของมหาวิทยาลัยมหาสารคาม

Mahasarakham University

คณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาวิทยานิพนธ์ของนายณัฐพล แสนคำ แล้วเห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ ของมหาวิทยาลัยมหาสารคาม

คณะกรรมการสอบวิทยานิพนธ์

.....	ประธานกรรมการ
(ผศ. ดร.อนันต์ เครือทรัพย์ถาวร)	(ผู้ทรงคุณวุฒิภายนอก)
.....	กรรมการ
(ผศ. ดร.นิวัตร์ อังควิศิษฐพันธ์)	(อาจารย์บัณฑิตศึกษาประจำคณะ)
.....	กรรมการ
(ผศ. ดร.ณัฐวุฒิ สุวรรณทา)	(อาจารย์บัณฑิตศึกษาประจำคณะ)
.....	กรรมการ
(รศ. ดร.วรวัฒน์ เสงียมวิบูล)	(อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก)

มหาวิทยาลัยอนุมัติให้รับวิทยานิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ ของมหาวิทยาลัยมหาสารคาม

.....	.....
(รศ.ดร.อนงค์ฤทธิ์ แข็งแรง)	(ศ.ดร.ประดิษฐ์ เทอดทูล)
คณบดีคณะวิศวกรรมศาสตร์	คณบดีบัณฑิตวิทยาลัย

วันที่ 30 เดือน ธ.ค. พ.ศ. 2559

Mahasarakham University

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้ ได้รับทุนจากสำนักงานคณะกรรมการการอุดมศึกษา

วิทยานิพนธ์ฉบับนี้สำเร็จสมบูรณ์ได้ด้วยความกรุณาและความช่วยเหลืออย่างสูงยิ่งจาก รองศาสตราจารย์ ดร.วรวัฒน์ เสงี่ยมวิบูล อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก ผู้ช่วยศาสตราจารย์ ดร.ณัฐวุฒิ สุวรรณทา ประธานกรรมการสอบ และผู้ช่วยศาสตราจารย์ ดร.นิวัตร์ อังควิศิษฐพันธ์ กรรมการสอบ

ขอขอบพระคุณ รองศาสตราจารย์ ดร.วรวัฒน์ เสงี่ยมวิบูล ผู้เชี่ยวชาญที่ช่วยตรวจเครื่องมือการวิจัย

ขอขอบพระคุณอาจารย์ท่านอื่นๆ ในสาขาวิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ ผู้ให้คำแนะนำและให้การช่วยเหลือสนับสนุนการวิจัย

ณัฐพล แสนคำ

Mahasarakham University

ชื่อเรื่อง การคัดกรองผู้ป่วยโรคพาร์กินสันโดยใช้วิธีซัพพอร์ตเวคเตอร์แมชชีนและการค้นหาแบบกริด
ผู้วิจัย นายณัฐพล แสนคำ
ปริญญา ปรัชญาดุษฎีบัณฑิต **สาขาวิชา** วิศวกรรมไฟฟ้าและคอมพิวเตอร์
อาจารย์ที่ปรึกษา รองศาสตราจารย์ ดร.วรวัฒน์ เสงี่ยมวิบูล
มหาวิทยาลัย มหาวิทยาลัยมหาสารคาม **ปีที่พิมพ์** 2559

บทคัดย่อ

โรคพาร์กินสันเป็นโรคภาวะเสื่อมของเซลสมอง จากการสูญเสียของเซลล์สมองส่วนที่ผลิตสารโดพามีน ซึ่งมีหน้าที่ช่วยประสานการเคลื่อนไหวของร่ายกาย โดยทั่วไปผู้ป่วยโรคพาร์กินสันจะมีอาการสั่น การเคลื่อนไหวร่างกายจะช้าและการทรงตัวไม่ดี ซึ่งส่งผลกระทบในการใช้ชีวิตประจำวันต่อผู้ป่วยรวมทั้งคนใกล้ชิดเป็นอย่างมาก งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาการคัดกรองผู้ป่วยโรคพาร์กินสันจากข้อมูลการวิเคราะห์เสียงที่ไม่ปรกติของผู้ป่วยโรคพาร์กินสัน ข้อมูลกลุ่มตัวอย่างเป็นข้อมูลที่ได้จากการวิเคราะห์เสียงทางชีวการแพทย์ จำนวน 31 คน แยกเป็นผู้ป่วยโรคพาร์กินสันจำนวน 23 คน และคนปกติจำนวน 8 คน มี 23 คุณลักษณะ 195 ระเบียน วิเคราะห์ข้อมูลด้วยวิธีซัพพอร์ตเวคเตอร์แมชชีนและการค้นหาแบบกริด ร่วมกับวิธีลดมิติของข้อมูลด้วยขั้นตอนวิธีรีลีฟ ผลจากการทดลองปรากฏว่าความถูกต้องในการคัดกรองร้อยละ 100 สำหรับข้อมูลที่ใช้ในการสอนระบบ และความถูกต้องในการคัดกรองร้อยละ 89.74 สำหรับข้อมูลทดสอบ ซึ่งเพียงพอต่อการนำไปใช้ในการคัดกรองผู้ป่วยโรคพาร์กินสันโดยใช้วิธีซัพพอร์ตเวคเตอร์แมชชีนและการค้นหาแบบกริด

คำสำคัญ : พาร์กินสัน; ซัพพอร์ตเวคเตอร์แมชชีน; การหาค่าที่เหมาะสมที่สุด; การคัดเลือกคุณลักษณะ

Mahasarakham University

TITLE Classification of Parkinson's Disease Using Support Vector Machine and Grid Search
AUTHOR Mr. Nuttapol Saenkham
DEGREE Doctor of Philosophy **MAJOR** Electrical and Computer Engineering
ADVISOR Assoc. Prof. Worawat Sa-ngiamvibool, PH.D.
UNIVERSITY Mahasarakham University **YEAR** 2016

ABSTRACT

Parkinson's disease is a neurodegenerative brain disorder, which are the result of the loss of dopamine – producing brain cells and affecting the body movement. The typical symptoms are tremor, bradykinesia, and postural instability, which affects the everyday life of the patients and surrounded people very much. The objective of this research was to study the screening of patients with Parkinson's disease by using the data of voice-disorder analysis of the patients. The sample data was collected with biomedical voice analysis from 31 people in total- including of 23 Parkinson's disease patients and 8 normal people with 23 features and 195 records. The data obtained were analyzed by Support Vector Machine technique and Grid Search together with Dimensionality Reduction with Relief Methodology. The results revealed that the accuracy for prediction was 100% for the data used in machine teaching system, and was 89.74 % for testing data, which is sufficiency for using with the screening of Parkinson's disease by Support Vector Machine and Grid Search Techniques.

Keyword : Parkinson; Support vector machine; Optimization; Feature selection

Mahasarakham University

สารบัญ

Mahasarakham University

หน้า

Mahasarakham University

สารบัญตาราง

หน้า

Mahasarakham University

สารบัญภาพประกอบ

หน้า

Mahasarakham University

บทที่ 1

บทนำ

1.1 ภูมิหลัง

โรคพาร์กินสัน (Parkinson's Disease) เป็นโรคเสื่อมของสมอง มักพบในผู้อายุสูงอายุ พบได้ในประชากรทั่วโลกรวมทั้งในประเทศไทย สาเหตุของโรคเกิดจากการเสื่อมของเซลล์ประสาทในสมองส่วน substantia nigra ซึ่งมีหน้าที่ในการผลิตสารโดบามีน (dopamine) ในปัจจุบันยังไม่เป็นที่ทราบแน่ชัดถึงสาเหตุที่ทำให้เกิดการเสื่อมของเซลล์ประสาทดังกล่าว เนื่องจากการเสื่อมของเซลล์ประสาทส่วนนี้จะเกิดขึ้นอย่างต่อเนื่องในผู้ป่วยโรคพาร์กินสัน ส่งผลให้ร่างกายขาดโดพามีนซึ่งเป็นสารสื่อประสาทที่สำคัญช่วยในเรื่องการประสานการเคลื่อนไหวของร่างกายให้เป็นไปอย่างราบรื่น ผู้ที่ป่วยเป็นโรคพาร์กินสันจึงมีอาการสั่น (tremor) เกร็ง (rigidity) การเคลื่อนไหวช้า (bradykinesia) และการทรงตัวไม่ดี (postural instability) แม้จะมีอาการที่มีลักษณะเด่น แต่อาการดังกล่าวนี้ยังไม่สามารถใช้วินิจฉัยได้แน่ชัดว่าป่วยเป็นโรคพาร์กินสันหรือไม่ เนื่องจากยังมีอีกหลายโรคที่มีอาการคล้ายๆ กัน เช่น โรคพากินสันเทียม โรคเอแอลเอส (Amyotrophic lateral sclerosis : ALS) โรคฮันติงตัน (Huntington's disease : HD) เป็นต้น

ในการวินิจฉัยโรคพาร์กินสันทางคลินิกโดยทั่วไปแพทย์จะวินิจฉัยจากการซักประวัติผู้ป่วย เช่น การนอนละเมอ ดมกลิ่นไม่ค่อยได้ การท้องผูก เคลื่อนไหวช้า นอกจากนี้แล้วอาจสังเกตและการตรวจลักษณะทางร่างกาย การตรวจโดยทั่วไปจะพบอาการสั่นเกร็ง การเคลื่อนไหวช้า การเคลื่อนไหวของกล้ามเนื้อไม่ปกติ ในการตรวจอาจต้องอาศัยผู้เชี่ยวชาญด้านกายภาพมาช่วยในการวินิจฉัย การวินิจฉัยเบื้องต้นนี้จะมีความถูกต้องประมาณ 60-70% เท่านั้น เพื่อผลการวินิจฉัยที่แม่นยำยิ่งขึ้นจึงมีการตรวจผู้ป่วยขั้นสูงทางห้องปฏิบัติการเพื่อทำการจำแนกโรคใกล้เคียงออก ได้แก่ การตัดชิ้นเนื้อไปตรวจ การเอ็กซ์เรย์คอมพิวเตอร์สมอง การวัดกระแสไฟฟ้าที่กล้ามเนื้อ และการปล่อยสารเคมีเพื่อเข้าไปจับกับเซลล์ในสมอง การวินิจฉัยในขั้นสูงจะมีความยุ่งยากซับซ้อน เสียเวลามีค่าใช้จ่ายสูงและอาจมีผลข้างเคียงอีกด้วยการทำเหมืองข้อมูลทางทางการแพทย์ คือการประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลกับข้อมูลทางการแพทย์ โดยอาศัยเทคนิคการวิเคราะห์ที่โดดเด่นซับซ้อนกว่าการวิเคราะห์ทางสถิติและการสืบค้นฐานข้อมูลแบบเดิมๆ ปัจจุบันมีการพัฒนาเทคนิคใหม่ๆ มาช่วยในการตัดสินใจวินิจฉัยทางการแพทย์ได้ เรียกว่าเทคนิคการเรียนรู้ของเครื่องจักร (Machine Learning Technique) เป็นเทคโนโลยีอย่างหนึ่งที่มีที่มาจากงานวิจัยด้านปัญญาประดิษฐ์ เช่น อัลกอริทึม ต้นไม้ตัดสินใจ ซัพพอร์ตเวคเตอร์แมชชีน เบย์เซียนเน็ตเวิร์ค โครงข่ายประสาทเทียม เป็นต้น เทคนิคการเรียนรู้ของเครื่องจักรเป็นการสร้างขั้นตอนวิธี (Algorithms) จากการให้ข้อมูลสอน (Training data) สำหรับสอน

Mahasarakham University

ให้คอมพิวเตอร์เรียนรู้ เพื่อให้ได้สมมติฐาน (Hypothesis) ในการนำมาใช้จำแนกและพยากรณ์ ซึ่งเป็นวิธีการที่ถูกนำมาประยุกต์ใช้อย่างแพร่หลาย ในการวิเคราะห์รูปแบบของข้อมูล วิเคราะห์แนวโน้มของข้อมูล และศึกษาองค์ความรู้จากข้อมูล มีการประยุกต์ใช้ในการวิเคราะห์ข้อมูลอย่างแพร่หลาย ในส่วนการประยุกต์ใช้ทางการแพทย์มีบทบาทหลายด้าน ทั้งการวินิจฉัยโรค การพยากรณ์โรค การรักษาโรค และการศึกษากลไกของโรค หากมีระบบตรวจคัดกรอง ที่เชื่อถือได้ จะช่วยประหยัดเวลา ค่าใช้จ่ายในการเดินทางไปที่คลินิก ลดผลข้างเคียงที่อาจเกิดขึ้นจากการตรวจวินิจฉัยในบางขั้นตอนได้ นอกจากนี้ยังช่วยเพิ่มความถูกต้องของการประเมินผลทางการแพทย์ได้อีกด้วย เทคนิคการจำแนกประเภทข้อมูลเป็นกระบวนการสร้างโมเดลจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ มี 2 ขั้นตอนหลักคือ ขั้นตอนแรกเป็นการสร้างโมเดลในการจำแนกประเภทข้อมูล เป็นการนำกลุ่มตัวอย่างข้อมูลที่เรียกว่าข้อมูลสอน (training data) ไปผ่านกระบวนการของอัลกอริทึมการจำแนกประเภทข้อมูล ซึ่งผลลัพธ์ที่ได้จะอยู่ในรูปของโมเดลการจำแนก และขั้นตอนที่ 2 เป็นการใช้โมเดลเพื่อพยากรณ์ เมื่อมีข้อมูลใหม่จะสามารถทำนายได้ โดยการนำข้อมูลมาทำการเปรียบเทียบกับโมเดลที่ได้จากการจำแนกประเภทข้อมูลจะทำให้สามารถพิจารณาคลาสในข้อมูลที่ยัง มิได้แบ่งกลุ่มในอนาคต เทคนิค การจำแนกประเภทข้อมูลนี้ได้นำไปประยุกต์ใช้ในหลายด้าน เช่น การจัดกลุ่มลูกค้าทางการตลาด, การตรวจสอบความผิดปกติ และการวิเคราะห์ทางการแพทย์ เป็นต้น

จากการศึกษาหลายครั้งก่อนหน้านี้ [1, 2] แสดงให้เห็นว่าข้อบกพร่องในการพูดและความผิดปกติของเสียงที่เปล่งออกมาของผู้ป่วยพาร์กินสัน อาจจะเป็นหนึ่งในตัวชี้วัดของโรคพาร์กินสันได้ ผู้วิจัยจึงมีความสนใจที่ประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่องจักรด้วยวิธีการ SVM มาใช้ในจำแนกข้อมูลโรคพาร์กินสัน นอกจากนี้ยังได้นำขั้นตอนการคัดเลือกคุณสมบัติของข้อมูล (Feature Selection Algorithm) และขั้นตอนการหาค่าที่เหมาะสมที่สุดของพารามิเตอร์ (Parameter Optimization) มาร่วมด้วย เพื่อเพิ่มความแม่นยำต่อการฝึกการเรียนรู้ให้แก่ข้อมูล และนำโมเดลที่ไปได้พัฒนาระบบ Clinical Diagnosis of Parkinson Diseases System เพื่อทำให้ผลการจำแนกโรคมีความถูกต้องมากยิ่งขึ้น

1.2 ความมุ่งหมายของการวิจัย

1.2.1 เพื่อศึกษาเกี่ยวกับการเรียนรู้ของเครื่องจักรและการขุดค้นข้อมูล

1.2.2 เพื่อศึกษาการจำแนกผู้ป่วยโรคพาร์กินสันจากผู้มีสุขภาพปกติ

1.2.3 เพื่อปรับปรุงประสิทธิภาพของการจำแนกผู้ป่วยโรคพาร์กินสันโดยใช้หลักการซัพพอร์ตเวคเตอร์แมชชีนร่วมกับขั้นตอนการคัดเลือกคุณสมบัติของข้อมูล (Feature Selection Algorithm) และขั้นตอนการหาค่าที่เหมาะสมที่สุดของพารามิเตอร์ (Parameter Optimization)

Mahasarakham University

1.3 ขอบเขตของการวิจัย

1.3.1 ประชากรประกอบด้วยกลุ่มผู้ป่วยโรคพาร์กินสัน และกลุ่มคนที่มีสุขภาพปกติ

1.3.2 กลุ่มตัวอย่าง จำนวน 31 คน แบ่งเป็นกลุ่มตัวอย่างผู้ป่วยโรคพาร์กินสัน 23 คน และกลุ่มของคนสุขภาพดีจำนวน 8 คน โดยแต่ละคนแบ่งช่วงของการวัดเสียงจากผู้เชี่ยวชาญเป็น 6 ช่วง และมีข้อมูลต่างๆ ของกลุ่มตัวอย่าง ทั้งหมดเป็น 23 แอตทริบิว

1.4 ความสำคัญของการวิจัย

1.4.1 มีระบบสารสนเทศเพื่อสนับสนุนการพยากรณ์และการวินิจฉัยพาร์กินสัน

1.4.2 เกิดการพัฒนาองค์ความรู้เชิงวิชาการที่เกี่ยวกับระบบสารสนเทศทางการแพทย์เพื่อสนับสนุนการพยากรณ์และการวินิจฉัยโรคพาร์กินสันโดยโมเดลใหม่ที่เกิดจากการผสมผสานหลายเทคนิคและวิธีการ ซึ่งสามารถนำไปใช้พยากรณ์ได้และเป็นประโยชน์เชิงวิชาการ

Mahasarakham University

บทที่ 2

ปริทัศน์เอกสารข้อมูล

ในการศึกษาการคัดกรองผู้ป่วยโรคพาร์กินสันโดยใช้วิธีซัพพอร์ตเวคเตอร์แมชชีนและการค้นหาแบบ กริดครั้งนี้ ได้กำหนดกรอบในแนวคิดในการศึกษา เอกสาร ทฤษฎีและงานวิจัยที่เกี่ยวข้อง ดังต่อไปนี้

2.1 โรคพาร์กินสัน

โรคพาร์กินสัน (Parkinson's Disease) มีที่มาจากชื่อ นพ.เจมส์ พาร์กินสัน (Dr. James Parkinson) นายแพทย์ชาวอังกฤษผู้ค้นพบโรคพาร์กินสัน ซึ่งเขาได้เขียนอธิบายอาการของโรคพาร์กินสันเป็นครั้งแรกลงในบทความที่เรียกว่า "Shaking Palsy" ในปี ค.ศ. 1817 ในปัจจุบันโรคพาร์กินสันจัดเป็นโรคความเสื่อมของระบบประสาท (Neurodegenerative Diseases) ที่พบได้บ่อยเป็นอันดับ 2 รองมาจากโรคอัลไซเมอร์ (Alzheimer's Disease) ในอัตราส่วนประมาณร้อยละ 1 ในคนที่มีอายุมากกว่า 65 ปี ขึ้นไป และเพิ่มเป็นร้อยละ 1-3 ในผู้มีอายุมากกว่า 80 ปี มองดูแล้วตัวเลขอาจจะดูไม่มาก แต่ถ้าเป็นอัตราส่วนเทียบกับคนปกติเป็นจำนวนไม่น้อยเลยที่เดียว โรคพาร์กินสัน (Parkinson's disease) เป็นโรคทางสมองที่เกิดจากเซลล์ประสาทในบางตำแหน่งเกิดมีการตายโดยไม่ทราบสาเหตุที่แน่ชัด ทำให้สารสื่อประสาทในสมองที่ชื่อว่า โดปามีน (Dopamine) มีปริมาณลดลง [3] ส่งผลให้ผู้ป่วยมีอาการที่สำคัญ คือ อาการสั่นขณะช่วงการพัก (Resting tremor) เคลื่อนไหวร่างกายช้าลง (Bradykinesia) ร่างกายมีสภาพแข็งเกร็ง (Rigidity) และการทรงตัวขาดความสมดุล (Postural instability)

2.1.1 อาการแสดงหลักทางคลินิกของผู้ป่วยโรคพาร์กินสัน

Jean Martin Charcot ชาวฝรั่งเศส เป็นคนแรกที่กล่าวถึงอาการแสดงทางคลินิกของโรคพาร์กินสันได้อย่างครบถ้วนทั้ง 4 อย่างที่เรียกว่า cardinal sign อันได้แก่ อาการสั่น (tremor) อาการแข็งเกร็ง (rigidity) อาการเคลื่อนไหวช้า (bradykinesia) และอาการเสียการทรงตัว (postural imbalance)

1) อาการสั่น (tremor) มักเป็นอาการแรกที่สังเกตเห็นได้และในระยะเริ่มแรกจะมีอาการสั่นเพียงด้านเดียวของร่างกาย (unilateral) โดยเริ่มต้นที่มือหรือปลายนิ้ว pill-rolling คือมือสั่นเหมือนกับปั้นเม็ดยา จะเกิดขึ้นเมื่ออยู่นิ่ง ถ้าเคลื่อนไหวจะไม่ค่อยสั่นและจะมีอาการมากขึ้นเมื่ออ่อนเพลียหรือเกิดอาการเครียด แต่จะหายไปเมื่อมีการเคลื่อนไหวหรือนอนหลับ เมื่ออาการของโรคดำเนินไปมากขึ้นจะพบอาการสั่นทั้งสองด้านของร่างกาย (bilateral) และอาการสั่นจะรุนแรงมากขึ้น อย่างไรก็ตามผู้ป่วยอาจไม่มีอาการสั่นที่มือให้เห็น แต่มีอาการสั่นบริเวณอื่น เช่น คางหรือริมฝีปาก

Mahasarakham University

2) กล้ามเนื้อแข็งเกรง (rigidity) อาการกล้ามเนื้อแข็งเกร็งเกิดจากการที่กล้ามเนื้อมีความตึงตัวมากขึ้น ผู้ป่วยจะมีอาการกล้ามเนื้อแข็งเกร็งด้านเดียวกับด้านที่มีอาการสั่น ทำให้ผู้ป่วยมีการเคลื่อนไหวแบบล้อเฟือง โดยจะมีการขยับทีละนิดคล้ายกับการหมุนของเฟือง (Cogwheel)

3) การเคลื่อนไหวช้าลง (bradykinesia) ผู้ป่วยมักจะเคลื่อนไหวหรือทำอะไรช้าลง โดยมักจะเกิดด้านเดียวกับด้านที่มีอาการสั่น รวมทั้งมีการแสดงออกทางสีหน้าลดลงที่เรียกว่า masked facies หรือมองแบบเลื่อนลอยโดยกระพริบตาน้อยลง ผู้ป่วยจะใช้เวลานานก่อนที่จะเคลื่อนไหว เมื่ออาการดำเนินไปมากขึ้นจะทำให้ผู้ป่วยเริ่มก้าวขาและหยุดก้าวขาได้ยากขึ้นจนเกิดการเดินแบบก้าวสั้นๆ แต่ซอยเท้าถี่ (festinating gait)

4) สูญเสียการทรงตัว (postural imbalance) มักเกิดหลังมีอาการผ่านไปแล้ว 2-5 ปี โดยผู้ป่วยจะมีลักษณะของลำตัวโน้มไปข้างหน้า (stooped posture) สูญเสียการทรงตัว เมื่อมีการผลักไปข้างหน้าหรือดึงไปด้านหลัง และการเกิด postural reflex บกพร่อง อย่างไรก็ตามผู้ป่วยไม่จำเป็นที่จะมีอาการแสดงทั้ง 4 อาการ เพื่อที่จะถูกวินิจฉัยว่าเป็นโรคพาร์กินสัน เนื่องจากผู้ป่วยโรคพาร์กินสันหลายรายไม่มีอาการสั่นให้เห็นชัดเจน นอกจากนั้นอาจมีอาการดังต่อไปนี้ เช่น เขียนตัวหนังสือเล็กลง (micrographic) ผู้ป่วยโรคพาร์กินสัน มักมีอาการของการทำงานของระบบประสาทอัตโนมัติบกพร่อง เช่น น้ำลายไหลยืด (drooling) อาการผื่นผิวหนังอักเสบจากการมีผิวมัน (seborrhea) และท้องผูก (constipation) อาจเกิดสมรรถภาพทางเพศบกพร่อง เช่น อวัยวะเพศไม่แข็งตัว หรือไม่สามารถถึงจุดสุดยอดได้ ผู้ป่วยโรคพาร์กินสันมักมีเสียงพูดที่เบาลงและมีโทนเสียงในการพูดโทนเดียว และอาจมีอาการผิดปกติทางจิต เช่น กระวนกระวาย วิตกกังวล และซึมเศร้า นอกจากนี้การเรียนรู้และความจำของผู้ป่วยจะลดลง รวมทั้งเกิดโรคสมองเสื่อม (dementia) ได้ประมาณร้อยละ 10-30 เป็นต้น

2.1.2 ระยะของโรคพาร์กินสัน

ในการวัดระดับความรุนแรงของโรคพาร์กินสัน จะมีการใช้เกณฑ์ที่เรียกว่า The Hoehn and Yahr scale (1997) ซึ่งมีการกำหนดเกณฑ์ไว้ดังต่อไปนี้ (ตาราง 2.1)

ตาราง 2.1 เกณฑ์การพิจารณาระดับความรุนแรงและการดำเนินไปของโรค

ระยะที่ 1	ร่างกายเกิดความผิดปกติในการเคลื่อนไหวไม่มาก เช่น แขน ขา ข้างใดข้างหนึ่งของลำตัว
ระยะที่ 2	ร่างกายเกิดความผิดปกติในการเคลื่อนไหวไม่มาก เช่น บริเวณแขน ขา แต่เป็นทั้งสองข้างของลำตัว
ระยะที่ 3	ร่างกายเกิดความผิดปกติในการเคลื่อนไหวทั้งสองข้าง อีกทั้งเกิดสภาวะการสูญเสียความสมดุลในการทรงตัว
ระยะที่ 4	ร่างกายเกิดความผิดปกติในการเคลื่อนไหวเป็นอย่างมาก และในการใช้ชีวิตปกติต้องได้รับความช่วยเหลือจากผู้อื่นพอประมาณ
ระยะที่ 5	ร่างการความผิดปกติในการเคลื่อนไหวเป็นอย่างมาก และส่วนใหญ่มักจะอยู่บนเตียงหรือบนรถเข็น บางรายอาจเดินได้ ถ้าได้รับการดูแลอย่างใกล้ชิด

Mahasarakham University

โดยทั่วไปผู้ป่วยโรคพาร์กินสันที่อยู่ในระยะที่ 1 และ 2 (Stage 1, 2) อาการจะไม่รุนแรง และไม่มีผลกระทบต่อการทำงานและการใช้ชีวิตประจำวัน อาจไม่จำเป็นต้องทำการรักษา ในระยะที่ 3 (Stage 3) หากผู้ป่วยไม่ได้รับการรักษา ผู้ป่วยจะทำงานและใช้ชีวิตประจำวันได้ลำบากมากขึ้น ระยะที่ 4 (Stage 4) การทำกิจกรรมต่างๆ จะลำบากมากขึ้นอีก และต้องการความช่วยเหลือผู้อื่นมากขึ้น จะไม่สามารถยืนหรือเดินด้วยตัวเองได้ ซึ่งจำเป็นต้องได้รับการรักษามากขึ้น และระยะที่ 5 (Stage 5) เป็นระยะสุดท้ายของโรค โดยชีวิตของผู้ป่วยส่วนใหญ่จะอยู่บนเตียงหรือรถเข็นเพียงอย่างเดียว ต้องได้รับความช่วยเหลือจากคนอื่นอยู่ตลอดเวลา ผู้ป่วยไม่สามารถทำกิจกรรมต่างๆ เองได้ ผู้ป่วยจะไม่ค่อยตอบสนองต่อการรักษาด้วยยา เพราะที่อาการของโรคมีการดำเนินไปอย่างมาก

2.1.3 การวิเคราะห์เสียง

การวินิจฉัยโรคพาร์กินสันนั้นโดยปกติใช้ลักษณะอาการและอาการแลดงทางคลินิกเป็นหลัก แต่จากการศึกษาหลายครั้งก่อนหน้านี้ [4] แสดงให้เห็นว่าการวินิจฉัยภาวะความผิดปกติทางเสียงพูดของผู้ป่วยพาร์กินสัน เป็นหนึ่งในตัวชี้วัดของโรคพาร์กินสันได้ ดังนั้นในส่วนนี้จึงจะเป็นนำเสนอการวินิจฉัยภาวะความผิดปกติทางเสียงพูด ที่ใช้ในทางการแพทย์

เสียงพูด (Voice) เป็นสิ่งสำคัญที่มนุษย์ใช้ติดต่อสื่อสารกัน กิจวัตรต่างๆ ของมนุษย์ล้วนใช้เสียงเพื่อติดต่อสื่อสารเป็นหลัก หากเกิดปัญหาในการใช้เสียงก็จะทำให้มีปัญหาในการทำงานติดต่อสื่อสารในสังคม ดังนั้นการวินิจฉัยภาวะความผิดปกติทางเสียงพูด (Voice Disorder) จึงเป็นสิ่งสำคัญ เสียงพูดเกิดจากการเคลื่อนที่ของลมจากปอดผ่านกล่องเสียง จนทำให้เส้นเสียง (Vocal folds) สั่นและเกิดเสียง โดยเสียงที่เกิดขึ้นจะเดินทางผ่านคอหอย จมูก และปาก จนเกิดเป็น “เสียงพูด” ดังนั้นความผิดปกติทางเสียงพูด จึงได้แก่ปัญหาเกี่ยวกับ ระดับสูงต่ำของเสียง ความก้อง ความดัง และคุณภาพของเสียง ปัจจัยด้าน การตรวจร่างกาย อายุ น้ำหนัก ส่วนสูง การแสดงออกทางสีหน้า ผิวหนัง เส้นผม เล็บ สุขอนามัยส่วนบุคคล ศีรษะและคอ สามารถช่วยผู้ตรวจบอกตำแหน่ง ความผิดปกติ รูปร่างของผู้ป่วย เช่น น้ำหนักมากหรือน้อยเกินไป หรือ มีประวัติน้ำหนักเปลี่ยนแปลงเร็ว มีโรคประจำตัวหรือโรคจิตเวช อาจทำให้เกิดปัญหาต่อคุณภาพของเสียงได้

ในปัจจุบันได้มีการสร้างการทดสอบต่างๆ ขึ้นมามากมายเพื่อใช้ในการประเมินการใช้เสียง จุดมุ่งหมายหลักคือ การวัดเสียงให้เป็นวัตถุ ผลของฟังก์ชันการทดสอบเสียง ควรจะได้รับการแปลผลตามปัจจัยต่างๆ ต่อไปนี้ เช่น ในรูปของการกระตุ้น (stimulus) สัญลักษณ์ (token) มาตรวัด (measure) และ อุปกรณ์ (equipment) เพื่อให้การประเมินน่าเชื่อถือจึงต้องมีการควบคุมการทดสอบโดย ใช้โปรโตคอลมาตรฐาน ขั้นตอนวิธีในการบันทึกเสียง การชี้แนะผู้ป่วย และสภาพแวดล้อม ที่เป็นมาตรฐาน อย่างไรก็ตามคุณภาพของเสียงเป็นข้อมูลหลายมิติ โดยทั่วไปแล้วการวัดเสียงจึงมักจะใช้ตัววัดหลายชนิด เช่น คุณภาพ (quality) ความดัง (loudness) และ ระดับ (pitch) องค์ประกอบของสัญญาณเสียงนี้สามารถแยกได้โดยใช้ได้หลายเทคนิค องค์ประกอบที่น่าสนใจ ได้แก่

Mahasarakham University

ความถี่ (Frequency) เป็นการวัดที่เป็นพื้นฐานเกี่ยวกับการวิเคราะห์เสียง Frequency คือความถี่ของการสั่นสะเทือนของสัญญาณเสียง เป็นจำนวนรอบช้า ๆ โดยจะวัดเป็นจำนวนรอบต่อวินาที หรือ เรียกว่า Hertz (Hz) การรับรู้ของความถี่นี้เรียกว่าระดับเสียงหรือ pitch ในความเป็นจริงมี หลายๆปัจจัยที่มีผลต่อการรับรู้ของระดับเสียง ถ้าความยาวคลื่นมากขึ้นของแต่ละรอบ (lower frequency) จะเป็น lower pitch และถ้าความยาวคลื่นสั้นขึ้นของแต่ละรอบ (higher frequency) ก็จะเป็น higher pitch การที่มีระดับเสียงที่เปลี่ยนแปลงไป หรือ ช่วงระดับเสียงถูกจำกัด ก็จะเกิดอาการที่พบได้บ่อยในภาวะความผิดปกติทางเสียงพูด และทำให้คนไข้เป็นกังวลกับสิ่งที่เกิดขึ้นได้ ความถี่มูลฐานเฉลี่ยในการพูด(fundamental frequency) และช่วงความถี่สูงสุดในกระบวนการออกเสียงพูด (maximum phonational frequency range) เป็นสิ่งที่สำคัญที่สุดที่เรา จะต้องตรวจหาความถี่นั้นอิทธิพลต่างๆ ที่มีผลต่อค่าความถี่ปกติ ได้แก่ ความเข้ม (intensity) ตัวอย่างการพูด (speech sample) ประเภทของสระ (vowel type) อายุ และ เพศ

ความเข้ม (Intensity) มีหลายปัจจัยมีผลต่อความดัง ความเข้มของเสียงเป็นปัจจัยหนึ่งที่เกี่ยวกับความดัง จะวัดเป็นระดับความดันเสียง (sound pressure level) ใน หน่วย decibels (dB) โดยใช้เครื่องมือ sound level meter หรือ acoustic analysis equipment การวัดดังกล่าวมีปัจจัยอื่นๆ เข้ามาเกี่ยวข้อง ได้แก่ ความถี่ (frequency) สระ (vowel) ตัวอย่างการพูด (speech sample) อุปกรณ์ (equipment) ระยะห่างระหว่างไมค์กับปาก (หรือ sound level meter) และ เสียงรบกวน การวัดที่ใช้บ่อยคือ average speaking intensity และความดังสุด และเบาสุดของความเข้ม ค่าเฉลี่ยของในผู้ชายและผู้หญิงขณะพูดอยู่ที่ 70 dB ถึงแม้จะมีความหลากหลายใน การสนทนาก็ตาม ค่าเฉลี่ยส่วนใหญ่จะน้อยกว่า 60 dB จนถึงอย่างน้อย 110 dB การวัด ความเข้มของเสียงจะมีประโยชน์ในการบอกให้คนไข้และคนในครอบครัวใส่ใจกับการลดความดังลง ซึ่งพบบ่อยในคนไข้ Parkinson's disease หรือ vocal fold motion impairment และช่วยระบุคนไข้ที่มีปัญหาการพูดเบาๆ ซึ่งพบบ่อยใน vocal fold scarring หรือ lesions

Jitter คือ ความผันแปรของความถี่ (cycle-to-cycle variation in frequency) ขณะที่ shimmer คือ ความผันแปรของความดัง (cycle-to-cycle variation in intensity) การวัดความผันแปรของ harmonics หรือสัญญาณรบกวน (noise) ได้ถูกนำมาเผยแพร่ และรวมถึงวิธีการวัดอื่นๆ ในทางทฤษฎีแล้วการวัดความผันแปรของคลื่นเสียงจะสอดคล้องกับเสียงที่แหบ (roughness or hoarseness) แต่ในทางปฏิบัติเราวัดคุณภาพเสียงได้ไม่แน่นอนและยังไม่มีเครื่องมือในการวัดที่แม่นยำในการที่จะวินิจฉัยภาวะความผิดปกติทางเสียงพูดปัจจุบันมีโปรแกรมที่นิยมใช้ในการแปรผลค่า Fo, jitter, shimmer, NHR อยู่ 2 โปรแกรม คือ Multi-Dimensional Voice Program (MDVP) และ Praat จากการศึกษาที่ผ่านมาได้มีการเปรียบเทียบทั้งสองโปรแกรม ผลที่ได้พบว่าผลค่า Fo ทั้งสองโปรแกรมได้ค่าใกล้เคียงกัน ส่วนผลค่าของ jitter, shimmer, NHR นั้นโปรแกรม MDVP จะมีค่าสูงกว่า

Mahasarakham University

การประเมินและวัดคุณภาพเสียง ควรจะมองในหลายๆ แง่มุมเพราะเสียงเป็นสิ่งที่ซับซ้อน และใช้สำหรับติดต่อสื่อสาร ในการวิเคราะห์เสียงแบ่งเป็นการประเมิน 3 อย่างคือ patient scales perceptual evaluation, and measures ซึ่งแต่ละวิธีวัดก็มีปัจจัยต่างๆ ที่มาเกี่ยวข้อง ซึ่งจำเป็นอย่างยิ่งที่แพทย์ต้องเข้าใจก่อนที่จะนำผลตรวจมาแปลผล วิธีวัดต่างๆ เปรียบเสมือนตัวช่วยแพทย์ไม่มีการวัดวิธีไหนวิธีเดียวที่ดีที่สุด จำเป็นต้องใช้การประเมินหลายๆ วิธีร่วมกันโดยการเลือกวิธีการวัดอย่างเหมาะสม การพัฒนามาตรฐานการวัดเสียงจึงเป็นสิ่งสำคัญที่จะนำไปสู่ การพัฒนาความสามารถของแพทย์และนักวิจัยในการเข้าใจเสียงของมนุษย์อย่างถ่องแท้

2.1.4 การออกแบบกลไกวินิจฉัย

กลไกวินิจฉัยเป็นส่วนของระบบผู้เชี่ยวชาญที่ใช้ข้อมูลในฐานความรู้เพื่อการวินิจฉัยตามที่ต้องการจนกว่าจะพบคำตอบ หรือจนกว่าจะหาคำตอบไม่ได้อันเนื่องมาจากฐานความรู้มีไม่เพียงพอ เราอาจจะแบ่งกลไกวินิจฉัยออกเป็นประเภทใหญ่ๆ ได้ 2 ประเภท คือ

1) ประเภทที่ให้คำตอบที่น่าจะเป็นไปได้ หรือ Probabilistic มักจะไม่เกี่ยวข้องกับความเป็นจริงทางธรรมชาติ ส่วนใหญ่จะขึ้นอยู่กับตัวประกอบหลายอย่างซึ่งอาจจะแปรเปลี่ยนไปตามสังคมหรือวัฒนธรรมได้ ตัวอย่างของการให้คำปรึกษาอาชีพจัดอยู่ในประเภท Probabilistic นั่นคือระบบเพียงแต่ให้ความเห็นว่า ผู้มีคุณสมบัติอย่างนี้ควรจะมีอาชีพอย่างไร ซึ่งมิได้หมายความว่าจะต้องเป็นจริงเสมอไป แต่ได้มีตัวอย่างข้อมูลในอดีตมาแล้วว่า ถ้าได้ผู้ที่มีลักษณะดังกล่าวมักจะเป็นผู้ที่ประสบความสำเร็จในอาชีพนั้น

2) ประเภทที่ให้คำตอบที่แน่นอน หรือ Deterministic เป็นประเภทที่ให้คำตอบได้แน่นอนหรือค่อนข้างจะแน่นอน ส่วนใหญ่มักจะเป็นปัญหาที่เกี่ยวข้องกับกฎธรรมชาติหรือความเป็นจริงที่สามารถพิสูจน์ได้แน่นอนเช่น รถยนต์วิ่งไม่ได้เพราะน้ำมันหมด เป็นต้น ในที่นี้ “น้ำมันหมด” คือคุณลักษณะหรือAttribute ส่วน “รถยนต์วิ่งไม่ได้” เป็นเป้าหมายหรือ Object ในระบบผู้เชี่ยวชาญ อย่างไรก็ตามกลไกวินิจฉัยทั้ง 2 ประเภทนี้สามารถสร้างขึ้นได้หลายวิธี สำหรับในระบบผู้เชี่ยวชาญมีวิธีที่เป็นพื้นฐานอยู่ 2 วิธีคือ

(1) Backward-Chaining Method หรือ Object-Driven Method กลไกสำหรับการวินิจฉัยแบบ Backward-Chaining Method จะเริ่มต้นที่เป้าหมายหรือ Object แล้วจึงพยายามที่จะค้นหาข้อมูลที่จะสนับสนุนให้เป้าหมายนี้เป็นจริง (ค้นหา Attribute ที่จะสนับสนุนเป้าหมาย) โดยระบบจะเริ่มที่ส่วนหัวของกฎ โดยมุ่งเป้าหมายไปที่ปัญหา จากนั้นก็จะทำการถามโต้ตอบกับผู้ใช้เพื่อที่จะหาข้อมูลสนับสนุนให้ตรงกับที่ได้ระบุไว้ในส่วนหางของกฎถ้าข้อมูลที่สนับสนุนข้อใดข้อหนึ่งไม่ตรงกับส่วนหางของกฎ ระบบก็จะเลื่อนไปที่กฎต่อไป นั่นคือ ระบบก็จะมุ่งเป้าหมายใหม่ของปัญหาซึ่งระบบก็จะพยายามถามข้อมูลจากผู้ใช้ระบบเพื่อสนับสนุนส่วนหางของกฎนี้อีกเช่นเดิม เราจะเห็นว่าการทำงานของกลไกนี้เป็นการทำงานย้อนหลังดังชื่อที่เรียกว่า Backward นั่นเอง

Mahasarakham University

(2) Forward-Chaining Method หรือ Data-Driven Method กลไกการวินิจฉัย วิธีนี้ตรงข้ามกับวิธีที่ได้กล่าวมาแล้วคือ แทนที่จะเริ่มสมมติเป้าหมายแล้วพยายามค้นหาข้อมูลเพื่อสนับสนุนเป้าหมายนั้นวิธีการ Forward-Chaining จะถามคำถามจากผู้ใช้แล้วใช้ประโยชน์จากคำถามนี้ไปในการหาทางเดินเข้าสู่เป้าหมาย ดังนั้นกลไกวินิจฉัยจึงเริ่มจากการหาข้อมูลแล้วจึงพยายามที่จะค้นหาเป้าหมายที่คล้องจองกับข้อมูลที่ได้มาเช่น ถ้ามีอาการไข้ มีน้ำมูกและปวดเมื่อยตามกล้ามเนื้อ สรุปคืออาจเป็นไปได้ว่าเป็นไข้หวัดใหญ่ เป็นต้น

2.2 การทำเหมืองข้อมูล (Data miming)

แนวโน้มของการนำสารสนเทศมาช่วยประกอบการตัดสินใจในงานสาขาต่าง ๆ มีมากขึ้น แต่บางครั้งไม่สามารถสร้างสารสนเทศที่ตรงกับความต้องการขององค์กรได้ ซึ่งในองค์กรต่าง ๆ ส่วนใหญ่ได้มีการเก็บข้อมูลไว้เป็นจำนวนมากโดยที่ข้อมูลเหล่านี้สามารถนำมาใช้ประโยชน์ได้มากแต่ไม่ค่อยได้ถูกนำมาใช้อย่างจริงจัง การทำเหมืองข้อมูล (Data Mining) เป็นวิธีการหนึ่งที่สามารถนำมาข้อมูลเหล่านั้นมาใช้ให้เกิดประโยชน์

2.2.1 นิยามของเหมืองข้อมูล

Data Mining [5] ศัพท์ที่ราชบัณฑิตยสถานกำหนดไว้คือ การทำเหมืองข้อมูล ซึ่งหมายถึง การสกัดหรือวิเคราะห์ ค้นหาข้อมูลที่ต้องการจากข้อมูลจำนวนมากได้ หรือกล่าวอีกนัยหนึ่ง Data Mining คือ ชุด Software วิเคราะห์ข้อมูลที่ได้ถูกออกแบบมาเพื่อระบบสนับสนุนความต้องการของผู้ใช้ในการค้นหาข้อมูลที่ต้องการจากข้อมูลจำนวนมากได้สำหรับ Philippe Nieuwbourg (CXP Information) กล่าวไว้ว่า “Data Mining คือ เทคนิคที่ผู้ใช้สามารถปฏิบัติการได้โดยอัตโนมัติกับข้อมูลที่ไม่รู้จัก ซึ่งเป็นการเพิ่มคุณค่า ให้กับข้อมูลที่มี” จากประโยคข้างต้นมีคำอยู่สามคำที่สำคัญ คือ คำแรก “อัตโนมัติ” คือกระบวนการทำงานของ Data Mining จะทำงานเอง คำที่สอง “ข้อมูลที่ไม่รู้จัก” คือ Data Mining จะไม่ประมวลแต่ข้อมูลปัจจุบันหรือข้อมูลที่ผู้ใช้ป้อนให้เท่านั้นแต่จะประมวลผลข้อมูลทั้งหมดที่มี และสุดท้าย “เพิ่มคุณค่า” นั่นหมายถึง ข้อมูลที่ได้ไม่ได้เป็นแค่ข้อมูลทางสถิติ แต่เป็นข้อมูลที่ช่วยในระดับตัดสินใจ

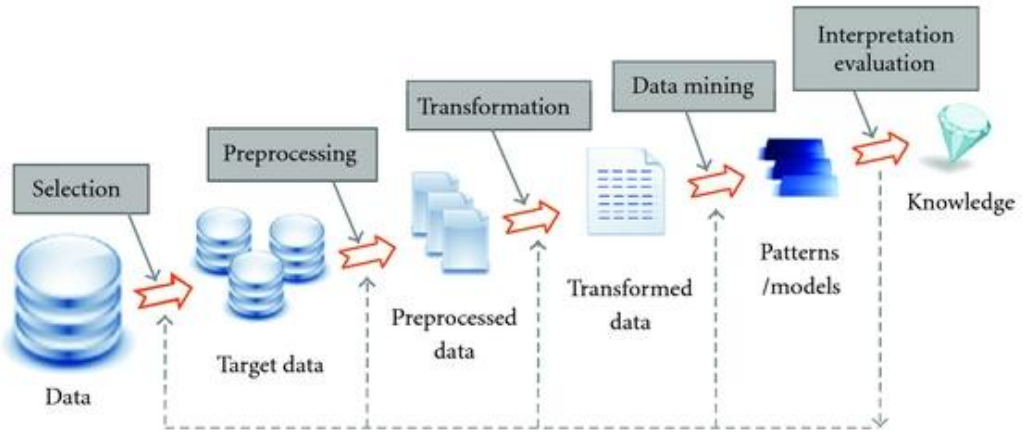
ดังนั้นเหมืองข้อมูลจึงเป็นกระบวนการในการค้นหาความสัมพันธ์ของรูปแบบ และแนวโน้มจากข้อมูลที่เก็บรวบรวมไว้เป็นจำนวนมาก โดยจะใช้เทคนิคทางสถิติหรือเทคนิคทางคณิตศาสตร์ นอกจากนี้ยังมีผู้นิยามเหมืองข้อมูลไว้อีกหลายนิยาม เช่น เหมืองข้อมูลเป็นการค้นหา วิเคราะห์ หรือสร้างองค์ความรู้ใหม่ จากข้อมูลขนาดใหญ่ซึ่งอาจเป็นการค้นหารูปแบบหรือ กฎ โดยการใช้เทคนิคทางสถิติ ทางคณิตสาสตร์หรือเทคนิคทางวิทยาการคอมพิวเตอร์ หรือเหมืองข้อมูล คือการกลั่นกรองสารสนเทศที่อยู่ในฐานข้อมูลขนาดใหญ่ โดยใช้เทคนิคการเรียนรู้ของเครื่องจักร กล่าวคือวิธีทางสถิติ วิธีทางคณิตศาสตร์ วิธีทางฐานข้อมูล และการแสดงข้อมูลในรูปแบบรายงานต่าง ๆ [6] จากนิยาม

Mahasarakham University

การทำเหมืองข้อมูลจะช่วยให้ได้สารสนเทศ กฎ หรือองค์ความรู้ใหม่ แต่การทำเหมืองข้อมูลค่อนข้างยุ่งยากและซับซ้อน จึงมีความจำเป็นต้องมีโปรแกรมคอมพิวเตอร์มาช่วยในการวิเคราะห์ข้อมูลเช่น Matlab, SPSS, Microsoft SQL Server หรือโปรแกรมทางสถิติและทางคณิตศาสตร์ต่างๆ เป็นต้น ทั้งนี้ผู้ทำเหมืองข้อมูลยังสามารถพัฒนาโปรแกรมขึ้นเอง เพื่อนำมาใช้ในการวิเคราะห์เองได้ด้วย เมื่อเปรียบเทียบกับการวิเคราะห์ข้อมูลด้วยโปรแกรมสถิติอื่นๆ เช่น SPSS, SAS นักวิเคราะห์ข้อมูลจะต้องเป็นผู้กำหนดว่าจะศึกษาลักษณะใดจากข้อมูล และจะใช้ข้อมูลส่วนใดบ้าง แต่ Data Mining จะกระทำขั้นตอนต่างๆ เหล่านี้ให้โดยอัตโนมัติ โปรแกรม Data Mining มีความสามารถที่จะค้นหาแนวโน้ม รูปแบบร่วม หรือลักษณะอื่นๆ ที่น่าสนใจ โดยไม่ต้องพึ่งพาการสั่งงานทุกขั้นตอนจากนักวิเคราะห์ข้อมูล และอาจจะสามารถค้นพบลักษณะที่น่าสนใจจากข้อมูลซึ่งนักวิเคราะห์ข้อมูลไม่ได้คาดหมายมาก่อน นอกจากนี้ Data Mining ยังมีความแตกต่างจากระบบผู้เชี่ยวชาญ (Expert System) ตรงที่ฐานความรู้ของ Data Mining ได้จากการสังเคราะห์ขึ้นจากข้อมูลโดยตรง สามารถปรับปรุงฐานความรู้ของตัวเองได้อัตโนมัติตามข้อมูลใหม่ที่ได้รับเพิ่มขึ้น ซึ่งต่างจากระบบผู้เชี่ยวชาญที่ฐานความรู้ถูกป้อนเข้ามาในระบบและจะคงตัวอยู่เช่นนั้นตลอดการใช้งาน

2.2.2 ตัวแบบการดำเนินการของเหมืองข้อมูล

เนื่องจากการทำเหมืองข้อมูลเป็นขั้นตอนหนึ่งที่สำคัญของ Knowledge Discovery data (KDD) ซึ่ง KDD คือกระบวนการในการกำหนดและแสวงหารูปแบบที่ชัดเจน เป็นองค์ความรู้ใหม่ที่มีประโยชน์และเข้าใจได้จากสิ่งที่ซ่อนอยู่ในข้อมูลดังในภาพประกอบ 2.1 งานวิจัยนี้ได้นำเสนอตัวแบบการดำเนินการที่มีการดำเนินการเพียง 7 ขั้นตอน [6] ดังต่อไปนี้



ภาพประกอบ 2.1 Knowledge Discovery Data (KDD)

Mahasarakham University

1) กำหนดลักษณะของจุดมุ่งหมาย (Goals identification) ประกอบด้วย การตั้งวัตถุประสงค์ของการทำเหมืองข้อมูล (Determine objective) ตั้งเกณฑ์วัดความสำเร็จ (Define success criteria) ประเมินสถานการณ์ในด้านต่างๆ (Assess situation) ระบุเป้าหมายที่ใช้ในการตัดสินใจทำเหมืองข้อมูล (Determine data mining goals) วางแผนการทำเหมืองข้อมูล (Produce a project plan) ว่าจะเก็บข้อมูลด้วยวิธีใด และใช้อัลกอริทึมไหน ต้นทุนของการดำเนินการ การเลือกเครื่องมือที่ใช้ในการทำเหมืองข้อมูล การรับคำปรึกษาของผู้เชี่ยวชาญ การวางแผนการจัดการทรัพยากรมนุษย์และทรัพยากรขององค์กร รวมทั้งการวางแผนการบำรุงรักษาและการปรับปรุงระบบหลังการดำเนินการเสร็จแล้ว โดยเป็นการวางแผนระยะยาว

2) การสร้างเซตข้อมูลเป้าหมาย (Creating a target data set) ประกอบด้วย การกำหนดคุณสมบัติของข้อมูล (Define success criteria) อธิบายรายละเอียดของข้อมูล (Describe data) การสำรวจข้อมูล (Explore data) การตรวจสอบความถูกต้องและความสมบูรณ์ของข้อมูล (Verify data quality) เป็นการกำหนดตัวแปรที่จะใช้จากแหล่งข้อมูล โดยแหล่งข้อมูลนั้นอาจได้มาจากคลังข้อมูลฐาน ข้อมูลธุรการ หรือจากแฟ้มงานต่าง ๆ เช่น แฟ้มงานสปรีตชีต (Spreadsheet)

3) การเตรียมข้อมูล (Data preprocessing) เป็นการตรวจสอบข้อมูลให้ถูกต้อง พร้อมทั้งแจ้งเตือนข้อมูลที่ไม่ถูกต้องให้ทราบก่อนที่ข้อมูลจะถูกนำไปใช้ขั้นตอนการเตรียมข้อมูลประกอบด้วย การคัดเลือกข้อมูลที่จะนำมาใช้ (Select data), การทำความสะอาดข้อมูล (Clean data) ซึ่งเป็นกระบวนการเตรียมข้อมูลให้เหมาะสมที่สุดเพื่อนำไปใช้ในขั้นตอนต่อไป เช่น

3.1) แก้ไขข้อมูลให้ถูกต้องสมบูรณ์ เช่น การแก้ไขค่าว่างของข้อมูลโดยใส่ค่า 9

3.2) ปรับเปลี่ยนข้อมูลให้มีค่าที่เหมาะสมในการตัดสินใจ เช่น ข้อมูลที่มีค่า “มาม่า” และ “ไวไว” อาจเปลี่ยนค่าเป็น “บะหมี่กึ่งสำเร็จรูป”

3.3) เลือกข้อมูลเฉพาะที่สนใจ เช่น ต้องการหาลักษณะลูกค้าที่ซื้อรถเก๋ง ไม่ควรนำรายชื่อพนักงานขายเข้ามาเกี่ยวข้อง

3.4) คอลัมน์ที่มีค่าสำหรับทุกแถวเป็นค่าเดียว เช่น “สัญชาติไทย” หรือ คอลัมน์ที่มีค่าไม่ซ้ำกันเลย เช่น “หมายเลขสมาชิก” ไม่ควรนำมาใช้เพราะไม่สามารถบอกรูปแบบของข้อมูลได้

4) การแปลงข้อมูล (Data transformation) เป็นการทำให้ข้อมูลอยู่รูปแบบตามความจำเป็นต่างๆ ซึ่งมีหลายเหตุผลด้วยกัน การปรับเปลี่ยนรูปแบบข้อมูล (Transform data) เช่น นำตารางในฐานข้อมูลมาเชื่อมต่อกัน ขั้นตอนนี้เป็นขั้นตอนที่สำคัญมาก เนื่องจากความถูกต้อง และสมบูรณ์ของผลลัพธ์สุดท้ายซึ่งขึ้นอยู่กับว่านักวิเคราะห์ข้อมูลนั้นตัดสินใจกำหนดโครงสร้างและเสนอลักษณะของข้อมูลที่จะใช้ในการประมวลผลอย่างไรอย่างไร กรรมวิธีนี้ รวมไปถึงการทำ Data Recording (การจัดเก็บข้อมูล) และ Data Format Conversion (รูปแบบในการแปลงข้อมูล) เช่น การแปลงเวลา Unix timestamp เป็นเวลาปัจจุบันเป็นต้น ขอยกตัวอย่างการแปลงข้อมูลดังนี้

Mahasarakham University

4.1) Data normalization เป็นการแปลงข้อมูลโดยการเปลี่ยนค่าข้อมูลให้อยู่ในรูปแบบมาตรฐานเดียวกัน เนื่องจากข้อมูลที่ได้มีค่าน้ำหนัก (Weight) ไม่เท่ากัน จึงมีความจำเป็นต้องทำการปรับข้อมูลให้เป็นมาตรฐานเดียวกัน ซึ่งเป็นวิธีการช่วยการเพิ่มความเร็วระยะที่เรียนรู้และเพื่อให้อยู่ในรูปมาตรฐาน เมื่อปรับข้อมูลเสร็จเรียบร้อยแล้วจึงสามารถนำข้อมูลที่ได้ไปใช้ในการวินิจฉัยต่อไป วิธีการ Data normalization มีหลายวิธีเช่น

1) Decimal scaling เป็นการหารข้อมูลด้วยตัวเลข โดยส่วนมากตัวเลขที่เรานำมาหารนั้นจะใช้ตัวเลขยกกำลัง เช่น ถ้าเรารู้ว่าข้อมูลมากและน้อยอยู่ในช่วง - 10,000 ถึง 10,000 ถ้าเราต้องการให้ข้อมูลอยู่ในช่วง -10 ถึง 10 จะต้องนำ 1,000 มาหารตลอดทุกข้อมูล

2) Min Max normalization เป็นเทคนิคที่ต้องรู้ค่าสูงสุดและค่าต่ำสุดของข้อมูล จะทำให้ข้อมูลที่ได้อยู่ในช่วง 0 และ 1 โดยมีสูตรการคำนวณดังนี้

$$newValue = \frac{originalvalue - oldMax}{oldMax - oldMin} \quad (2.1)$$

3) การทำให้อยู่ในรูปคะแนนมาตรฐาน Z (Z-score) โดยต้องทราบค่าเฉลี่ย (μ) และค่าส่วนเบี่ยงเบนมาตรฐาน(σ) โดยมีสูตรการคำนวณดังนี้

$$newValue = \frac{originalvalue - \text{ค่าเฉลี่ย}}{\text{ค่าส่วนเบี่ยงเบนมาตรฐาน}} \quad (2.2)$$

4.2) การเปลี่ยนชนิดของข้อมูล (Data type conversion) เทคนิคเหมืองข้อมูลบางชนิดไม่สามารถวิเคราะห์ข้อมูลเป็นข้อมูลเชิงกลุ่มได้ จึงต้องทำการแปลงข้อมูลเชิงกลุ่มให้เป็นตัวเลขก่อนการนำข้อมูลไปวิเคราะห์

5) การทำเหมืองข้อมูล (Data mining) หรือขั้นตอนการสร้างแบบจำลอง เป็นการเลือกเทคนิคที่เหมาะสมในการทำเหมืองข้อมูล (Select modeling technique) บางครั้งสามารถเลือกอัลกอริทึมได้หลายวิธี การกำหนดว่าข้อมูลใดเป็นข้อมูลที่สร้างและข้อมูลใดเป็นข้อมูลที่ใช้ในการทดสอบผลลัพธ์ รวมทั้งการวิเคราะห์ข้อมูล กำหนดรูปแบบการทดสอบผลลัพธ์ (Generate test design) สร้างแบบจำลองตามเทคนิคที่เลือก (Model Building) ทดสอบความถูกต้องและความน่าเชื่อถือของแบบจำลองที่สร้างขึ้น (Model Assassin)

6) การแปลผลและการประเมินผล (Interpretation and evaluation) ประกอบด้วยการประเมินผลที่ได้จากการทดลอง (evaluate Results) อาจจะเป็นการประเมินแบบจำลองที่สร้างขึ้น

Mahasarakham University

ด้วยการลองนำไปใช้กับสถานการณ์จริงเพื่อตรวจสอบประสิทธิภาพของแบบจำลอง การทบทวนกระบวนการ (review process) ใช้เป็นขั้นตอนถัดไปในการตัดสินใจ (Determine next steps) เป็นการพิจารณาผลลัพธ์ จากขั้นตอนที่ 5 ว่าสามารถตอบคำถามหรือแก้ปัญหาจากขั้นตอนที่ 1 หรือไม่ ตัดสินใจว่าจะกระทำขั้นตอนที่ 5 ซ้ำหรือไม่ และรวมถึงการแปลผลไปให้ผู้ใช้ข้อมูลเข้าใจ

7) การนำไปใช้ (Taking action) ประกอบด้วย แผนการในการนำไปใช้ (Plan the deployment) และ สรุปผลของการทดลอง เมื่อลงความเห็นว่าจะนำองค์ความรู้ที่ได้ไปใช้ องค์ความรู้นั้นจะถูกรวมเข้ากับระบบที่ใช้อยู่เช่น การสร้างระบบรายงานเกี่ยวกับองค์ความรู้ที่ได้

2.2.3 แบบชนิดของข้อมูล

ในการทำเหมืองข้อมูลนั้นก่อนอื่นจะต้องทำการศึกษาว่าข้อมูลที่จะทำการศึกษานั้นเป็นข้อมูลชนิดใด จึงจะสามารถนำข้อมูลนั้นไปทำการวิเคราะห์ได้อย่างถูกต้อง ดังนี้

แบบชนิดของข้อมูล (Data Type) คือ ชนิดของตัวแปรที่ใช้อธิบายข้อมูล การใช้ตัวแปรอธิบายข้อมูลสามารถใช้ตัวแปรได้ตั้งแต่ 1 ตัวขึ้นไป ตัวแปรมีหลายชนิดได้แก่

1) อินเตอร์วอล – สเกลวาเรียเบิล (Interval – Scaled Variable) คือ ตัวแปรที่ค่าของมันมีความต่อเนื่องกันตัวอย่าง เช่น ส่วนสูง น้ำหนัก อุณหภูมิ เป็นต้น

2) ไบนารีวาเรียเบิล (Binary Variable) มีค่าของตัวแปรเพียง 2 สถานะ คือ 0 หรือ 1 ({0, 1}) โดยสถานะ 0 หมายถึง ข้อมูลที่ไม่ได้แสดงตัวแปร นั้นๆ หรือ ตัวแปรไม่มีค่า และสถานะ 1 หมายถึง ข้อมูลมีค่าของตัวแปรนั้นๆ อยู่ ตัวอย่างเช่น ให้ตัวแปร “smoker” อธิบายถึง ข้อมูลผู้ป่วยแต่ละคนถ้าตัวแปรนี้มีค่าเป็น 1 จะบ่งชี้ว่าผู้ป่วยคนนั้นสูบบุหรี่ แต่ถ้าเป็น 0 จะบ่งชี้ว่าผู้ป่วยนั้นไม่ได้สูบบุหรี่

3) โนมินอลวาเรียเบิล (Nominal Variable) โนมินอลวาเรียเบิลจะคล้ายกันกับไบนารีวาเรียเบิลตรงที่ค่าของตัวแปรจะเป็นสถานะ แต่ต่างกันตรงที่โนมินอลวาเรียเบิลสามารถมีสถานะได้มากกว่า 2 สถานะ ตัวอย่างเช่น ตัวแปร “color” สามารถกำหนดให้มี 5 สถานะ ได้แก่ {red, yellow, green, pink, blue}

4) ออดินอลวาเรียเบิล (Ordinal Variable) ออดินอลวาเรียเบิลคล้ายกับโนมินอลวาเรียเบิล เพียงแต่สถานะของตัวแปรแบบออดินอลวาเรียเบิลจะมีการจัดลำดับด้วย (ranking) ตัวอย่างเช่น ตัวแปร “professional” มีสถานะดังนี้ ผู้ช่วยศาสตราจารย์ (assistant) รองศาสตราจารย์ (associate) และศาสตราจารย์ (full) ซึ่งกำหนดสถานะให้ตัวแปรจะต้องเป็นไป ตามลำดับ

5) เรโชสเกลวาเรียเบิล (Ratio Scaled Variable) เป็นตัวแปรที่ค่าของมันเป็นค่ามากซึ่งได้จากการวัดบนสเกลไม่เชิงเส้น (Nonlinear Scale) เช่น สเกลเอกซ์โปแนนเธียล (Exponential Scale)

6) ตัวแปรหลายชนิดผสมกัน (Variable of Mixed Type) ในบางครั้งข้อมูลหนึ่งถูกอธิบายข้อมูลโดยใช้ตัวแปรหลายตัวหลายชนิดผสมกัน ซึ่งทำให้แบบชนิดข้อมูลเป็นแบบผสม

Mahasarakham University

2.3 เทคนิคการเรียนรู้ของเครื่อง (Learning Machine)

วิทยาการทางด้านการเรียนรู้ของเครื่องเติบโตไปพร้อมๆ กับศาสตร์ด้านปัญญาประดิษฐ์ อาจกล่าวได้ว่าการเรียนรู้ของเครื่องมีมาตั้งแต่ยุคต้นๆ ของปัญญาประดิษฐ์ เหล่านักวิจัยต่างให้ความสนใจในการสร้างเครื่องจักรที่สามารถเรียนรู้จากข้อมูลได้ จึงเริ่มศึกษาการเรียนรู้ของเครื่องหลายๆ วิธีการ ที่ได้รับการยอมรับและความนิยมมาก เช่น โครงข่ายประสาทเทียม และในเวลาต่อมา ได้มีการคิดค้นโมเดลเชิงเส้นทั่วไปจากหลักการทางสถิติศาสตร์ ไปจนถึงการพัฒนาวิธีการให้เหตุผลตามหลักความน่าจะเป็น โดยเฉพาะในการประยุกต์ด้านการวินิจฉัยโรคอัตโนมัติ อย่างไรก็ตามนักวิจัยในสายปัญญาประดิษฐ์ยุคต่อมาเริ่มหันมาให้ความสำคัญกับตรรกศาสตร์และใช้วิธีการทางการแทนความรู้มากขึ้น จนทำให้ปัญญาประดิษฐ์เริ่มแยกตัวออกจากกับศาสตร์การเรียนรู้ของเครื่อง หลังจากนั้นเริ่มมีการใช้หลักการความน่าจะเป็นมากขึ้นในการดึงและการแทนข้อมูล ต่อมาในปี 1980 ระบบผู้เชี่ยวชาญเริ่มโดดเด่นในสายของปัญญาประดิษฐ์จนหมดยุคของการใช้หลักสถิติ มีงานวิจัยด้านการเรียนรู้เชิงสัญลักษณ์และบนพื้นฐานของฐานความรู้ออกมา เรื่อยๆ จนกลายศาสตร์ด้านการโปรแกรมตรรกะเชิงอุปนัยได้ถือกำเนิดขึ้นมา แต่งานด้านสถิติก็ยังถือว่ามีบทบาทมากนอกสาขาของปัญญาประดิษฐ์ เช่น การรู้จำแบบและการค้นคืนสารสนเทศ นักวิจัยสายปัญญาประดิษฐ์และนักวิทยาศาสตร์คอมพิวเตอร์ได้ทิ้งงานวิจัยด้านโครงข่ายประสาทเทียมไปในเวลาเดียวกัน แต่ก็ยังมีนักคณิตศาสตร์บางคน เช่น จอห์น ฮอปฟิลด์ เดวิด โรเมลฮาร์ต และเจฟฟรีย์ ฮินตันที่ยังพัฒนาโครงข่ายประสาทเทียมต่อไป จนกระทั่งได้ค้นพบหลักการการแพร่คืนย้อนกลับของโครงข่ายประสาทเทียม ที่ประสบความสำเร็จมากมายในเวลาต่อมา การเรียนรู้ของเครื่องกับการทำเหมืองข้อมูลมักจะใช้วิธีการเหมือนๆ กัน และมีส่วนคาบเกี่ยวกันอย่างเห็นได้ชัด สิ่งที่แตกต่างระหว่างสองศาสตร์นี้คือ

การเรียนรู้ของเครื่อง เน้นเรื่องการพยากรณ์ข้อมูลจากคุณสมบัติที่"รู้"แล้วที่ได้เรียนรู้มาจากข้อมูลชุดสอน

การทำเหมืองข้อมูล เน้นเรื่องการค้นหาคุณสมบัติที่"ไม่รู้"จากข้อมูลที่ได้มา กล่าวได้ว่าเป็นขั้นตอนการวิเคราะห์เพื่อค้นหา"ความรู้"ในฐานข้อมูล

สองศาสตร์นี้มีส่วนคาบเกี่ยวกันไม่น้อย คือ การทำเหมืองข้อมูลใช้วิธีการทางการเรียนรู้ของเครื่อง แต่มักจะมีเป้าหมายในใจที่แตกต่างออกไปเล็กน้อย ส่วนการเรียนรู้ของเครื่องก็ใช้วิธีการของการทำเหมืองข้อมูลบางอย่าง เช่น การเรียนรู้แบบไม่มีผู้สอน หรือขั้นตอนการเตรียมข้อมูลเพื่อปรับปรุงความถูกต้องของการเรียนรู้ บ่อยครั้งที่นักวิทยาศาสตร์ผสมสองสาขานี้เข้าด้วยกันด้วยเหตุผลที่ว่าประสิทธิภาพของการเรียนรู้ของเครื่องมักจะดีขึ้นหากมีความสามารถในการรู้ ความรู้บางอย่างในขณะที่การค้นหาความรู้และการทำเหมืองข้อมูลนั้นกุญแจสำคัญคือการค้นหาความรู้ที่ไม่รู้มาก่อนหากมีการวัดประสิทธิภาพจากสิ่งที่ไม่รู้มาก่อน วิธีการเรียนรู้แบบมีผู้สอนของการเรียนรู้ของเครื่อง

Mahasarakham University

ก็มักจะให้ผลได้ดีกว่าการใช้วิธีการเรียนรู้แบบไม่มีผู้สอนอย่างเดียว การเรียนรู้ของเครื่องยังมีความคล้ายคลึงกับการหาค่าเหมาะที่สุด (optimization) นั่นคือ การเรียนรู้หลายอย่างมักจะถูกจัดให้อยู่ในรูปแบบของการหาค่าที่น้อยที่สุด ของฟังก์ชันการสูญเสียบางอย่างจากข้อมูลชุดสอน ฟังก์ชันการสูญเสียหมายถึงความแตกต่างระหว่างสิ่งที่พยากรณ์ไว้กับสิ่งที่เป็นจริง

หลักสำคัญของการเรียนรู้ของเครื่องคือ การทำให้โมเดลมีความ "เป็นธรรมชาติ" (general) มากขึ้นจากประสบการณ์ที่ได้มา การทำให้เป็นธรรมชาติมากขึ้นนี้จะทำให้เครื่องสามารถพยากรณ์หรือทำงานกับตัวอย่างข้อมูลที่ไม่เคยเห็นมาก่อนได้อย่างแม่นยำมากขึ้น บางครั้งข้อมูลชุดสอนก็มาจากการสุ่มและผู้เรียนรู้จะต้องทำให้โมเดลมีความธรรมชาติขึ้น เพื่อจะได้ทำการพยากรณ์ข้อมูลใหม่ๆ ได้อย่างถูกต้องเพียงพอ การวิเคราะห์เชิงคำนวณของการเรียนรู้ของเครื่อง และการวัดประสิทธิภาพการเรียนรู้ เป็นอีกสาขาหนึ่งทางวิทยาการคอมพิวเตอร์สายทฤษฎี ที่รู้จักกันในชื่อ ทฤษฎีการเรียนรู้เชิงคำนวณ อย่างไรก็ตามทฤษฎีก็ไม่สามารถรับประกันประสิทธิภาพของอัลกอริทึมได้เพราะข้อมูลนั้นมีจำกัดและอนาคตมีความไม่แน่นอน แต่ทฤษฎีก็สามารถบอกขอบเขตบนความน่าจะเป็นได้ว่าประสิทธิภาพน่าจะอยู่ในช่วงใด นอกจากนี้นักวิทยาศาสตร์ด้านนี้ยังได้ศึกษาดูต้นทุนทางเวลาและความเป็นไปได้ของการเรียนรู้ของเครื่องด้วย โดยการคำนวณที่ถือว่าเป็นไปได้ในการเรียนรู้นั้นจะต้องสามารถเรียนรู้ได้ในเวลาโพลิโนเมียล

2.3.1 การจำแนกข้อมูลและการทำนาย (Classification and Prediction)

เป็นกระบวนการสร้างโมเดลจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ ประกอบด้วยการตรวจสอบลักษณะของสิ่งที่เราสนใจ และนำไปสู่คลาส (Class) ที่ได้ถูกกำหนดไว้ก่อนแล้ว การจำแนกประเภทข้อมูล เป็นการจัดประเภทของข้อมูลจากค่าของคุณลักษณะ (Attribute) ตัวอย่างเช่น การจัดกลุ่มนักเรียนว่า ดีมาก ดี ปานกลาง ไม่ดีโดยพิจารณาจากประวัติและผลการเรียน หรือแบ่งประเภทของลูกค้าว่าเชื่อถือได้หรือไม่โดยพิจารณาจากข้อมูลที่มีอยู่

โมเดลที่ใช้จำแนกข้อมูลออกเป็นกลุ่มตามที่ได้กำหนดไว้ จะขึ้นอยู่กับการวิเคราะห์เซตของข้อมูลทดลอง (Training Data) โดยนำข้อมูลทดลอง (Training Data) มาสอนให้ระบบเรียนรู้ว่ามีข้อมูลใดอยู่ในคลาส (Class) เดียว ผลลัพธ์ที่ได้จากการเรียนรู้ คือ โมเดลจัดประเภทข้อมูล(Classifier Model) โมเดลนี้ สามารถแทนได้ในหลายรูปแบบเช่น Classification (IF-THEN) Rules, ต้นไม้ตัดสินใจ (Decision Tree), Mathematical Formulae หรือโครงข่ายประสาทเทียม (Neural Networks) และจะนำข้อมูลส่วนที่เหลือจากข้อมูลทดลอง (Training Data) เป็นข้อมูลที่ใช้ทดสอบ (Testing Data) ซึ่งเป็นกลุ่มที่แท้จริงของข้อมูลที่ใช้ทดสอบนี้จะถูกนำมาเปรียบเทียบกับกลุ่มที่หามาได้จากโมเดลเพื่อทดสอบความถูกต้อง โดยจะปรับปรุงโมเดลจนกว่าจะได้ค่าความถูกต้องในระดับที่น่าพอใจ หลังจากนั้นเมื่อมีข้อมูลใหม่เข้ามา จะนำข้อมูลผ่านโมเดล โดยโมเดลจะสามารถทำนายกลุ่มของข้อมูลนี้ได้ ซึ่งเทคนิค

Mahasarakham University

ที่ใช้ในการทำเหมืองข้อมูลแบบการจำแนกข้อมูล (Classification) ที่ใช้กันในปัจจุบัน ได้แก่ต้นไม้ตัดสินใจ (Decision Tree), โครงข่ายประสาทเทียม (Neural Networks), ขั้นตอนวิธีการหาเพื่อนบ้านที่ใกล้ที่สุด (K-Nearest Neighbor : K-NN)

กระบวนการจำแนกประเภทข้อมูลนี้แบ่งออกเป็น 3 ขั้นตอน

ขั้นตอนที่ 1 การสร้างโมเดล (Model Construction: Learning) เป็นขั้นการสร้างโมเดล โดยการเรียนรู้จากข้อมูลที่ได้กำหนดคลาสไว้เรียบร้อยแล้ว (Training Data) ซึ่งโมเดลที่ได้อาจแสดงในต้นไม้ตัดสินใจ (Decision Tree) คล้ายโครงสร้างต้นไม้ ที่แต่ละโหนด (Node) แสดงคุณลักษณะแต่ละกิ่งแสดงผลในการทดสอบ และใบของต้นไม้ตัดสินใจ (Leaf Node) แสดงคลาสที่กำหนดไว้ ซึ่งต้นไม้ตัดสินใจง่ายต่อการปรับเปลี่ยนเป็นกฎของการทำการจำแนกประเภทข้อมูล

ขั้นตอนที่ 2 การประมาณความถูกต้องของโมเดล (Model Evaluation : Accuracy) เป็นขั้นการประมาณความถูกต้องโดยอาศัยข้อมูลที่ใช้ทดสอบ (Testing Data) ซึ่งคลาสที่แท้จริงของข้อมูลที่ใช้ทดสอบนี้จะถูกนำมาเปรียบเทียบกับคลาสที่หามาได้จาก Model เพื่อทดสอบความถูกต้อง

ขั้นตอนที่ 3 การใช้งานโมเดล (Model Usage : Classification) เป็นโมเดลสำหรับใช้กับข้อมูลที่ไม่เคยเห็นมาก่อน (Unseen Data) โดย จะทำการกำหนดคลาสให้กับวัตถุ (Object) ใหม่ที่ได้มา หรือทำนายค่าออกมาตามต้องการ

ลักษณะของการจำแนกข้อมูล (Classification) นั้นคำนึงถึงผลกำหนดที่ออกมาชัดเจนว่าคุณสมบัติดังกล่าวจะอยู่ในชั้นใด แต่การประมาณการ (Estimation) เป็นการประเมินที่ไม่สามารถกำหนดค่าหรือคุณสมบัติดังกล่าวให้ชัดเจนเป็นการจัดการกับค่าที่มีผลในการวัดที่ต่อเนื่อง เช่น การประเมินรายได้ของครอบครัว การประเมินความสูงของบุคคลในครอบครัว เป็นต้น

การทำนาย (Prediction) เหมือนกับการจำแนกประเภทข้อมูล (Classification) และการประเมิน (Estimation) ยกเว้นว่ารายการข้อมูลที่ถูกแยกจัดลำดับนั้นเกิดขึ้นตามการ ทำนายพฤติกรรมในอนาคตหรือการทำนายค่าที่จะเกิดขึ้นในอนาคต ข้อมูลในอดีตจะถูกสร้างเป็นโมเดลขึ้นมาเพื่อทำนายหรืออธิบายสิ่งที่จะเกิดขึ้นในอนาคตเช่น การทำนายว่ายอดซื้อของลูกค้าจะเป็นเท่าใดถ้าบริษัทลดราคาสินค้า 10 %

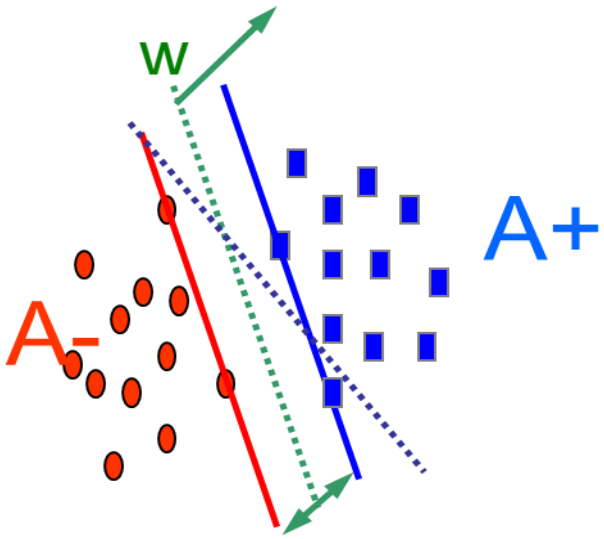
2.3.2 ซัพพอร์ตเวคเตอร์แมชชีน (Support Vector Machine : SVM)

ซัพพอร์ตเวคเตอร์แมชชีน (Support Vector Machine) [7] เป็นเครื่องมือที่ใช้ในการจำแนกข้อมูล ซึ่งเทคนิคหนึ่งที่สำคัญของการสืบค้นความรู้บนข้อมูลขนาดใหญ่ โดยเป็นกระบวนการสร้างโมเดลจัดประเภทข้อมูล (Classifier Model) เพื่อช่วยในการตัดสินใจสำหรับการทำนายแนวโน้มข้อมูลที่อาจเกิดขึ้นในอนาคต ซึ่งโมเดลจัดประเภทข้อมูลดังกล่าวเกิดขึ้นจากการสอนกลุ่มข้อมูลตัวอย่างที่เรียกว่า ข้อมูลสอน (Training Data) ให้อยู่ในกลุ่มที่กำหนดให้เพื่อแสดงให้เห็นความแตกต่างระหว่างกลุ่มของข้อมูล (Class) กล่าวคือ เมื่อมีข้อมูลใหม่เข้ามาในระบบ โมเดลจัดประเภทข้อมูลดังกล่าวจะทำการประมวลผลและสามารถทำนายกลุ่มของข้อมูลนี้ได้ เทคนิคการจำแนกประเภทข้อมูลนี้ได้นำไป

Mahasarakham University

ประยุกต์ใช้ในหลายด้าน เช่น การจัดกลุ่มลูกค้าทางการตลาดการตรวจสอบความผิดปกติ และการวิเคราะห์ทางการแพทย์ เป็นต้น [6]

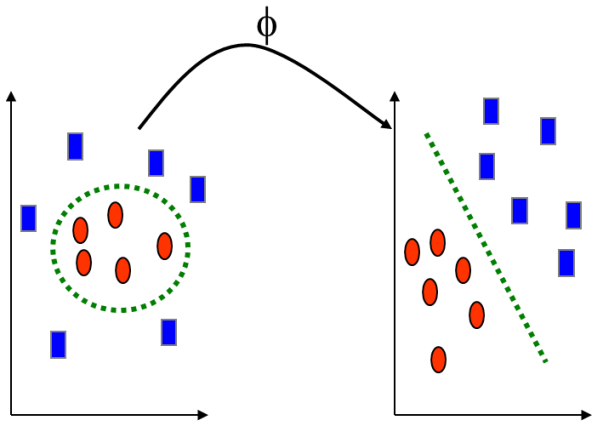
เป้าหมายของวิธีซัพพอร์ตเวคเตอร์แมชชีน คือ กระบวนการสอนเครื่องแบบมีผู้สอน (Supervise Learning) เพื่อให้สามารถสร้างตัวจัดประเภทข้อมูล (Classifier) ที่มีความเป็นทั่วไป (Generalize) สูง นั่นคือสามารถทำงานได้ดีกับตัวอย่างที่ไม่รู้จัก (Unknown Dataset) ด้วยกระบวนการปรับรูปแบบข้อมูลจากข้อมูลที่มีมิติต่ำ (Low Dimension Dataset) บนพื้นที่ข้อมูลนำเข้า (Input Space) ให้อยู่ในรูปแบบของข้อมูลที่มีมิติสูง (High Dimension Dataset) บนพื้นที่ข้อมูลคุณลักษณะ (Feature Space) โดยใช้ฟังก์ชันในการปรับรูปแบบข้อมูลที่เรียกว่าฟังก์ชันเคอร์เนล (Kernel Function) ซึ่งความสามารถดังกล่าวช่วยให้การสร้างตัวจัดประเภทข้อมูลด้วยสมการกำลังสอง (Quadratic Equation) บนพื้นที่ข้อมูลคุณลักษณะเป็นไปได้ง่ายขึ้นและ มีความชัดเจนในการจัดประเภทมากยิ่งขึ้นด้วย นอกจากนี้ ตัวจัดประเภทข้อมูลที่ดีควรมีโครงสร้างแบบเส้นตรง (Linear Classifier) และสามารถสร้างพื้นที่ระยะห่างระหว่างตัวจัดประเภทข้อมูลเองกับค่าที่ใกล้ที่สุดของแต่ละกลุ่มข้อมูลได้มากที่สุดเพื่อประสิทธิภาพในการแยกแยะประเภทของชุดข้อมูลแต่ละประเภทออกจากกันอย่างชัดเจน ซึ่งเส้นที่เหมาะสมดังกล่าว ถูกเรียกว่า ระนาบแบ่งเขตข้อมูลที่เหมาะสม (The Optimal Separating Hyperplane) โดยหลักการในการทำงานเพื่อจำแนกประเภทข้อมูลของวิธีซัพพอร์ตเวคเตอร์แมชชีน สามารถแสดงตามภาพประกอบ 2.2



ภาพประกอบ 2.2 อัลกอริทึมของ Support Vector Machines

Mahasarakham University

ในกรณีการจัดแบ่งกลุ่มข้อมูลโดยใช้ระนาบแบบไม่เป็นเส้นตรง ซัพพอร์ตเวคเตอร์แมชชีนจะอาศัยหลักการของการแปลงข้อมูลจากพื้นที่ข้อมูลนำเข้า (Input Space) ให้เป็นพื้นที่คุณลักษณะ (Feature Space) ที่มีมิติสูงขึ้น จากภาพประกอบ 2.3 แสดงให้เห็นถึงแนวคิดของซัพพอร์ตเวคเตอร์แมชชีน ซึ่งทำการแปลงข้อมูลที่ใช้ในการเรียนรู้แบบไม่เชิงเส้นไปเป็นขนาดพื้นที่คุณลักษณะที่ใหญ่ขึ้นผ่านฟังก์ชันเคอร์เนล (Kernel Function : Φ) และสร้างระนาบซึ่งแบ่งข้อมูลสองกลุ่มได้ดีที่สุด ทำให้เกิดเป็นขอบเขตการตัดสินใจ (Decision Surface) แบบไม่เชิงเส้นในพื้นที่ข้อมูลนำเข้า ในขณะที่ซัพพอร์ตเวคเตอร์แมชชีนแบบเชิงเส้นจะสร้างระนาบในพื้นที่คุณลักษณะที่ใหญ่ขึ้นภายใต้ทฤษฎีของ Mercer ซึ่งต้องการการคำนวณที่สิ้นเปลืองในส่วนของตัวอย่างเพื่อให้ได้ขนาดพื้นที่คุณลักษณะที่ใหญ่ขึ้น ปัญหาดังกล่าวสามารถแก้ไขได้โดยการใช้ฟังก์ชันเคอร์เนล เพื่อให้ได้ผลลัพธ์ที่น่าพึงพอใจเช่นเดียวกัน ซึ่งการใช้ฟังก์ชันเคอร์เนลจะทำให้สามารถคำนวณระนาบได้โดยไม่ต้องอาศัยการแปลงไปเป็นพื้นที่คุณลักษณะ



ภาพประกอบ 2.3 แนวความคิดการจำแนกข้อมูลของวิธีซัพพอร์ตเวคเตอร์แมชชีน

ฟังก์ชันเคอร์เนล $K(x_i, x_j)$ เป็นฟังก์ชันที่แก้ปัญหาภายใต้เงื่อนไขของ Mercer's ซึ่งมีค่าเท่ากับการคูณกันของสองเวคเตอร์ x_i, x_j ในพื้นที่คุณลักษณะ $\Phi(x_i)$ และ $\Phi(x_j)$

$$k(x_i + x_j) = \Phi(x_i).\Phi(x_j) \tag{2.3}$$

โดยที่ Φ คือ ฟังก์ชันการแปลงแบบไม่เป็นเชิงเส้น (Nonlinear Projection Function) ซึ่งฟังก์ชันเคอร์เนลหลายตัวได้ถูกนำมาใช้กับซัพพอร์ตเวคเตอร์แมชชีนแบบไม่เป็นเชิงเส้น

Mahasarakham University

อย่างประสบความสำเร็จ การฟังก์ชันเคอร์เนลที่แตกต่างกันของซัพพอร์ตเวคเตอร์แมชชีนสามารถนำมาซึ่งวิธีการเรียนรู้ที่หลากหลาย ซึ่งตัวอย่างของฟังก์ชันเคอร์เนล มีดังนี้ [8]

1) โพลิโนเมียลเคอร์เนล (Polynomial Kernel)

$$k(x_i + x_j) = [(x_i, x_j + 1]^p \quad (2.4)$$

2) เกาเซียนเรเดียลเบสิสฟังก์ชัน (Gaussian Radial Basis Function)

$$k(x_i + x_j) = exp - \left|\frac{|x_i - x_j|^2}{2\sigma^2}\right| \quad (2.5)$$

3) ไฮเปอร์โบลิกแทนเจนต์เคอร์เนล (Hyperbolic Tangent Kernel)

$$k(x_i + x_j) = tanh(x_i * x_j + c) \quad (2.6)$$

เมื่อ k คือ ฟังก์ชันเคอร์เนล เช่น รูปแบบฟังก์ชันเคอร์เนลของเกาเซียนเรเดียลเบสิสฟังก์ชัน (Gaussian Radial Basis Function) มีรูปแบบดังสมการที่ (2.5) เมื่อค่าของชุดข้อมูล x_i = 2 และ x_j = 1 และกำหนดค่า σ = 10 ทำให้ค่า

$k(x_i + x_j) = exp - \left|\frac{|2-1|^2}{2(10)^2}\right| = 0.995$ หลังจากถูกแทนค่าเข้าไปในสมการที่ (2.5)

กล่าวโดยสรุปได้ว่าขั้นตอนวิธีการซัพพอร์ตเวคเตอร์แมชชีน เป็นขั้นตอนทางคณิตศาสตร์ที่ใช้ในการแก้ไขข้อบกพร่องขั้นตอนวิธีของโครงข่ายประสาทเทียมแบบเดิมๆ และช่วยให้ผู้ใช้สามารถออกแบบอัลกอรึทึมได้สะดวกขึ้นกว่าเดิม เนื่องจากขั้นตอนวิธีซัพพอร์ตเวคเตอร์แมชชีนไม่จำเป็นต้องกำหนดจำนวนชั้นซ่อน จำนวนนิวรอล ค่าอัตราการเรียนรู้ ค่าน้ำหนักเริ่มต้น ฯลฯ วิธีการซัพพอร์ตเวคเตอร์แมชชีน มีชั้นซ่อนเพียงชั้นเดียว ซึ่งจะสามารถสร้างระนาบเกินที่เหมาะสมที่สุดในการแบ่งกลุ่ม หรือใช้แทนกลุ่มข้อมูลในการวิเคราะห์แบบถดถอย ซึ่งสามารถเลือกใช้งานได้ตามความเหมาะสม นอกจากนี้วิธีการซัพพอร์ตเวคเตอร์แมชชีนยังมีการใช้ซัพพอร์ตเวคเตอร์น้ำหนักที่ใช้ในโครงข่ายประสาทเทียมแบบดั้งเดิม เป็นการช่วยลดเวลาที่ใช้ในการเรียนรู้ และมีผลลัพธ์ที่ถูกต้องมากขึ้นด้วย ในด้านการออกแบบโมเดลด้วยวิธีการซัพพอร์ตเวคเตอร์แมชชีน ยังมีการออกแบบอย่างมีหลักการเพิ่มมากขึ้นโดยอาศัยหลักกระบวนการลดความเสี่ยงเชิงโครงสร้างให้ต่ำที่สุด ทำให้ผลที่ได้มีความน่าเชื่อถือมากขึ้นและใช้เวลาในการเรียนรู้ที่น้อยกว่าอีกด้วย อย่างไรก็ตามการเลือกฟังก์ชันเคอร์เนลและปัจจัยที่เกี่ยวข้องที่มีความเหมาะสมยังคงเป็นปัญหาที่จะต้องอาศัยการทดสอบและแก้ไขสำหรับการเลือกแบบจำลอง

Mahasarakham University

2.3.3 ขั้นตอนวิธีการเบย์ (Bayes)

การเรียนรู้แบบเบย์ (Bayesian Learning) เป็นการเรียนรู้โดยอาศัยความน่าจะเป็น (Probability) เพื่ออนุมานคำตอบที่ต้องการ เช่น คลาสหรือประเภทของตัวอย่าง บนสมมติฐานว่าปริมาณที่สนใจจะอยู่ภายใต้การแจกแจงความน่าจะเป็น (Probability Distribution) ดังนั้นความน่าจะเป็นของตัวอย่างจึงสามารถใช้ประกอบการตัดสินใจอย่างมีเหตุผลได้ ในงานจำแนกประเภท (Classification) ตัวแบบเบย์เป็นตัวจำแนกประเภท (Classifier) ค่อนข้างจะแตกต่างจากตัวจำแนกประเภท เช่น โครงข่ายประสารทเทียม หรือ SVM เหล่านี้ทำการสร้างสมมติฐานหรือพยายามสร้างระนาบเกิน (Hyperplane) แบ่งคลาสแต่ละคลาสออกจากกัน แต่การเรียนรู้แบบเบย์ไม่ได้ทำการสร้างระนาบเกินเพื่อทำการแบ่งคลาสแต่อาศัยทฤษฎีของเบย์แทน

ทฤษฎีของเบย์ (Bayes1 theorem) หรือกฎของเบย์ (Bayes' law) ตั้งชื่อตาม โทมัส เบย์ (Thomas Bayes) นักสถิติและนักปราชญ์ชาวอังกฤษ กล่าวถึงความสัมพันธ์ระหว่างเหตุการณ์ในปัจจุบันและสิ่งที่เกิดก่อนหน้าโดยมีความน่าจะเป็นแบบมีเงื่อนไขเป็นประเด็นสำคัญในทฤษฎีนี้ ความน่าจะเป็นแบบมีเงื่อนไข (Conditional Probability) หมายถึง ความน่าจะเป็นของการเกิดเหตุการณ์ A เมื่อกำหนดว่าเหตุการณ์ B เกิดขึ้นแล้ว (Conditional Probability of A given B) โดยสามารถเขียนสัญลักษณ์ได้ดังนี้ P(A|B) ซึ่งสามารถคำนวณได้จากความน่าจะเป็นร่วม (Joint probability) หรือความน่าจะเป็นที่เหตุการณ์ A และเหตุการณ์ B เกิดขึ้นร่วมกัน การจำแนกประเภทด้วยทฤษฎีของเบย์ เพื่อใช้ในการจำแนกประเภท (Classification) จากกฎของเบย์ สามารถกำหนดได้ดังนี้

$$P(C|A) = \frac{P(A|C)xP(C)}{P(A)} \qquad (2.7)$$

P(C) หมายถึง ความน่าจะเป็นของคลาส C (Prior probability)

P(C|A) หมายถึง ค่าความน่าจะเป็นที่ข้อมูลที่มีคุณลักษณะเป็น A จะมีคลาส C หรือ ความน่าจะเป็นภายหลัง (Posterior probability) ของสมมติฐาน A เมื่อกำหนดชุดข้อมูลที่ใช้สอน C หรือเรียกว่า Posterior

P(A|C) หมายถึง ค่าความน่าจะเป็นที่ข้อมูลที่ใช้สอนที่มีคลาส C และมีคุณลักษณะ A โดยที่ A = a_1, a_2 ..., a_M โดยที่ M คือจำนวนคุณลักษณะในข้อมูลที่ใช้สอนหรือ ความน่าจะเป็นภายหลัง (Posterior probability) ของชุดข้อมูลที่ใช้สอน C เมื่อกำหนดสมมติฐาน A หรือเรียกว่า Likelihood

การคำนวณค่าต่างๆ จากชุดข้อมูลที่ใช้สอน เพื่อสร้างเป็นโมเดล Naive Bayes โดยใช้ข้อมูล weather ดังในตาราง 2.2

Mahasarakham University

ตาราง 2.2 ชุดข้อมูล Weather

0	outlook	temperature	humidity	windy	play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	mild	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

จากข้อมูลในตาราง 2.2 สามารถคำนวณค่าความน่าจะเป็นจากตารางได้ดังนี้

P(play = yes) = 9/14 – 0.64

P(play = no) = 5/14 - 0.36

Mahasarakham University

ตาราง 2.3 ชุดข้อมูล Weather

attribute	play = Yes	play = No
outlook = sunny	2/9 = 0.22	3/5 = 0.60
outlook - overcast	4/9 = 0.45	0/5 = 0.00
outlook = rainy	3/9 - 0.33	2/5 = 0.40
temperature - hot	2/9 = 0.22	2/5 = 0.40
temperature = mild	4/9 = 0.45	2/5 = 0.40
temperature - cool	3/9 = 0.33	1/5 = 0.20
humidity = high	3/9 = 0.33	4/5 = 0.80
humidity = normal	6/9 = 0.67	1/5 = 0.20
windy = TRUE	3/9 = 0.33	3/5 = 0.60
windy - FALSE	6/9 = 0.67	2/5 = 0.40

จากตาราง 2.3 คือโมเดลของ Naive Bayes ที่สร้างได้จากข้อมูลที่ใช้สอน หากทดลองนำเอาข้อมูลรายการแรกจากในตาราง 2.3 มาทำนายด้วยโมเดล Naive Bayes ข้อมูลในรายการแรกประกอบด้วย

คุณลักษณะ outlook = sunny

คุณลักษณะ temperature = hot

คุณลักษณะ humidity = high

คุณลักษณะ windy = FALSE

เราจะต้องคำนวณค่าความน่าจะเป็นที่มีคุณลักษณะเหล่านี้แล้วตอบคลาส play = yes ได้ดังนี้

$$P(play=yes|A) = P(outlook=sunny|play=yes) \times P(temperature=hot|play=yes) \times P(humidity=high|play=yes) \times P(windy=FALSE|play=yes) \times P(play=yes)$$

$$= 0.22 \times 0.22 \times 0.33 \times 0.67 \times 0.64$$

$$= 0.0068$$

Mahasarakham University

หลังจากนั้นจะคำนวณค่าความน่าจะเป็นที่มีคุณลักษณะเหล่านี้แล้วตอบคลาส play = no ได้ดังนี้

P(play=no|A) = P(outlook=sunny|play=no) x P(temperature=hot|play=no) x
P(humidity=high|play=no) x P(windy=FALSE|play=no) x
P(play=no)
= 0.60x0.40x0.80x0.40x0.36
= 0.0276

เมื่อเปรียบเทียบค่าความน่าจะเป็นที่ได้จาก 2 คลาสแล้วพบว่าค่า P(play = no|A) (=0.0276) มีค่ามากกว่า P(play = yes|A) (=0.0068) ดังนั้นโมเดลของเราจึงทำนายว่าข้อมูล instance นี้มีค่าคลาส play = no

ถ้าสังเกตตารางโมเดล Naive Bayes จะพบว่ามีค่าความน่าจะเป็นของบางคุณลักษณะเป็น 0 นั่นคือไม่มีรูปแบบของคุณลักษณะนี้เกิดขึ้นในข้อมูลที่ใช้สอนเลย ดังนั้นการใช้งานโมเดลที่มีค่าความน่าจะเป็นมีค่าเท่ากับ 0 เช่นนี้จะทำให้ค่าที่จะทำนายมีค่าเป็น 0 ไปด้วย จึงมีการเพิ่มค่าความถี่ของข้อมูลเข้าไปอีกครั้งละ 1 เช่น จะได้เป็น P(outlook = overcast | play = yes) มีค่าเท่ากับ 5/12 = 0.42 และ P(outlook = overcast | play = no) = 1/8 = 0.13 วิธีการนี้เรียกว่า Laplace smoothing

ตัวอย่างการประยุกต์ใช้กฎของเบย์กับปัญหาทางการแพทย์ ในทางการแพทย์นั้นผลการตรวจทางห้องปฏิบัติการมีความผิดพลาดได้ สมมุติให้ผลการตรวจโรคชนิดหนึ่งจากห้องปฏิบัติการมีความแม่นยำในการตรวจว่าเป็นโรค (Correct positive) ร้อยละ 98 และมีความแม่นยำในการตรวจว่าไม่เป็นโรค (Correct negative) ร้อยละ 97 โดยมีสถิติการเป็นโรคนี้ของประชากรทั่วโลกเป็นร้อยละ 0.8 จากข้อมูลในตัวอย่างนี้ จะมีความน่าจะเป็นเท่าใดที่ผู้ป่วยรายนี้จะเป็นโรคจริงหากมีผลตรวจจากห้องปฏิบัติการของผู้ป่วยรายหนึ่งว่าเป็นโรค หรือ P(ill|+)

กำหนดให้ ชุดของสมมติฐาน H={A,-A} หรือผู้ป่วยเป็นโรค (A) หรือโม่เป็นโรค (-A) และ
ชุดข้อมูลตัวอย่าง D= {+,-} หรือผลตรวจจากห้องปฏิบัติการว่าเป็น (+)
หรือไม่เป็น (-)

จากกฎของเบย์เราสามารถหา P(A|+) ได้จากสมการที่ 8

$$P(A|+) = \frac{P(+|A)P(A)}{P(+)} \quad (2.8)$$

จากข้อมูลในตัวอย่างเราทราบว่า P(A) = 0.008 และ P(+|A) = 0.98 ดังนั้นต้องทำการคำนวณหา P(+) ได้ดังนี้

Mahasarakham University

$$P(+) = P(+|A)\,P(A) + P(+ \mid -A)\,P(-A) \quad (2.9)$$
$$= (0.98)(0.008) + (0.03)(1- 0.008)$$
$$= 0.0078+0.0298$$
$$= 0.21$$

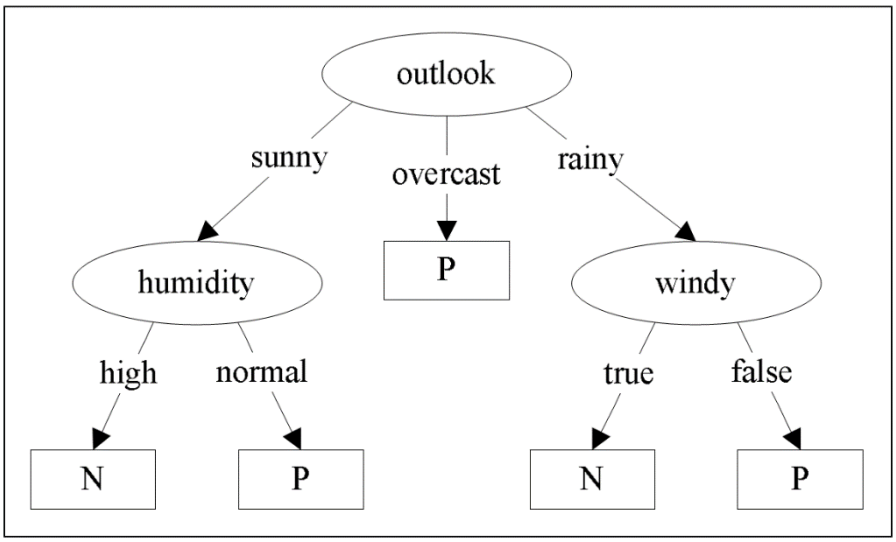
จากผลการคำนวณในตัวอย่างนี้พบว่าหากผลการตรวจจากห้องปฏิบัติการออกมาเป็นบวก จะมีความน่าจะเป็นเพียงร้อยละ 21 เท่านั้นที่จะเป็นโรคจริง

จุดหมายสำคัญของการเรียนรู้แบบเบย์ คือ การเลือกสมมติฐานที่เหมาะสม หากความน่าจะเป็นของแต่ละสมมติฐาน P(h) ทั้งหมดมีค่าเท่ากันจะมีลักษณะการแจกแจงเป็นรูปแบบเดียวกัน (Uniform) เมื่อถูกอนุมาน (Inference) ด้วยชุดข้อมูลตัวอย่างที่ใช้สอนค่า Posterior หรือ P(h|D) ของสมมติฐานที่สอดคล้องกับข้อมูลตัวอย่างที่ใช้สอนนั้นจะมีค่ามากขึ้นและในทางกลับกันค่า Posterior ของสมมติฐานที่ไม่สอดคล้องกับข้อมูลตัวอย่างที่ใช้สอนจะมีค่าลดลง และเมื่อจำนวนชุดข้อมูลตัวอย่างที่ใช้สอน D มีจำนวนมากขึ้น ความแตกต่างของค่า Posterior จะมากขึ้นกล่าวคือจะมีสมมติฐานจำนวนน้อยที่มีค่า Posterior สูง และจะเป็นลักษณะนี้ไปเรื่อยๆ เมื่อจำนวนตัวอย่างเพิ่มขึ้น โดยทั่วไปแล้วสมมติฐานที่เราต้องการคือสมมติฐานที่ทำให้ค่า Posterior หรือ P(h|D) มีค่ามากที่สุดเมื่อกำหนดชุดข้อมูลที่ใช้สอน D หรือเรียกสมมติฐานนี้ว่า สมมติฐาน Maximum a posteriori (MAP)

2.3.4 ขั้นตอนวิธีต้นไม้ตัดสินใจ (decision tree)

ต้นไม้ตัดสินใจ (decision tree) คือขั้นตอนวิธีที่มีลักษณะเหมือนต้นไม้กลับหัวโดยมีรากอยู่ด้านบนและมีใบอยู่ทางด้านล่าง โครงสร้างของต้นไม้ประกอบด้วยโหนด (node) ซึ่งแต่ละโหนดเป็นการตัดสินใจจากข้อมูลในคุณลักษณะต่าง ๆ กิ่งจะเป็นผลที่ได้จากตรวจสอบ และใบจะเป็นผลลัพธ์ที่ได้จากการพยากรณ์ ส่วนรูทโหนด (root node) จะอยู่บนสุด ดังแสดงองค์ประกอบของต้นไม้ตัดสินใจได้ดังภาพประกอบ 2.4 ซึ่งเป็นตัวอย่างจากปัญหาการตัดสินใจว่าจะออกไปเล่นกอล์ฟหรือไม่ [9] โดยอาศัยปัจจัยสภาพอากาศต่าง ๆ เพื่อใช้สนับสนุนการตัดสินใจ โดยรูปวงรีแสดงถึงการตรวจสอบค่าที่เป็นไปได้ของคุณลักษณะนั้นๆ และรูปสี่เหลี่ยมคือใบจะแสดงการจำแนกกลุ่มของข้อมูล ซึ่งเป็นผลลัพธ์ของการทำนายว่าจะออกไปเล่นกอล์ฟ (P) หรือไม่ออกไปเล่น (N) จากการตรวจสอบตามเส้นทางของต้นไม้ตัดสินใจ ในการแบ่งกลุ่มข้อมูลใหม่นั้น ค่าของคุณลักษณะต่าง ๆ จะถูกตรวจสอบด้วยต้นไม้ตัดสินใจโดยจะเริ่มจากรูทโหนดไปถึงใบ โดยใบคือคำตอบจากการทำนายว่าข้อมูลนั้นอยู่ในกลุ่มใด

Mahasarakham University



ภาพประกอบ 2.4 ซึ่งแสดงถึงต้นไม้ที่ใช้ในการตัดสินใจว่าจะออกไปเล่นกอล์ฟ

โดยทั่วไปสร้างต้นไม้ตัดสินใจ จะดำเนินการในลักษณะจากบนลงล่าง (top-down) คือเริ่มสร้างรูทหรือส่วนรากของต้นไม้ก่อน แล้วค่อยทำต่อไปกิ่งและใบ โดยสามารถแสดงกระบวนการสร้างต้นไม้ตัดสินใจได้ดังต่อไปนี้ [6]

1) เริ่มจากมีเพียงโหนดเดียว คือชุดข้อมูลสอน (training set)

2) ถ้าข้อมูลอยู่กลุ่มเดียวกันหมด ให้เป็นโหนดใบและตั้งชื่อตามกลุ่มของข้อมูลนั้น

3) หาค่าเกน (gain) ของแต่ละคุณลักษณะ ถ้าโหนดนั้นมีข้อมูลหลายกลุ่มอยู่ด้วยกัน เพื่อใช้เป็นเงื่อนไขในการคัดเลือกคุณลักษณะ ที่ประสิทธิภาพในการจำแนกข้อมูลได้ดีที่สุด โดยเลือกคุณลักษณะที่เมื่อคำนวณแล้วได้ค่าเกนมากที่สุด ให้เป็นตัวทดสอบเพื่อใช้ในการตัดสินใจ โดยอยู่ในรูปของโหนดบนต้นไม้ตัดสินใจ

4) กิ่งถูกสร้างจากค่าที่เป็นไปได้ของโหนดที่ใช้ทดสอบ และข้อมูลจะถูกจำแนกออกตามกิ่งต่าง ๆ ที่ได้สร้างขึ้น

4) วนรอบเพื่อหาค่าเกนที่มีค่ามากที่สุดของแต่ละคุณลักษณะ จากข้อมูลที่ถูกจำแนกในแต่ละกิ่ง นำคุณลักษณะที่ได้มาสร้างเป็นโหนดตัดสินใจต่อไป โดยมีเงื่อนไขว่าจะไม่เลือกคุณลักษณะที่เคยถูกเลือกมาเป็นโหนดแล้ว ในโหนดในระดับต่อ ๆ ไป

5) วนรอบเพื่อจำแนกข้อมูลและแตกกิ่งของต้นไม้ไปเรื่อย ๆ โดยจะหยุดวนรอบเมื่อเงื่อนไขต่อไปนี้เป็นจริง ในข้อใดข้อหนึ่ง

Mahasarakham University

(1) ถ้าทุกข้อมูลในโหนดอยู่ในกลุ่มเดียวกัน ให้สร้างใบตามกลุ่มของข้อมูลนั้น

(2) ถ้าไม่เหลือคุณลักษณะใดที่ใช้ในการจำแนกข้อมูลแล้ว ซึ่งจะใช้กลุ่มที่มีค่าข้อมูลสนับสนุนมากที่สุดมาเป็นใบในกรณีนี้

(3) ถ้าไม่มีข้อมูลสนับสนุนในกิ่งนั้นๆ แล้ว ให้ทำการสร้างใบตามกลุ่มที่มีข้อมูลสนับสนุนมากที่สุด

กระบวนการสร้างต้นไม้ตัดสินใจด้วยขั้นตอนวิธี C4.5 เป็นขั้นตอนวิธีที่ได้รับความนิยมและมีการใช้อย่างแพร่หลาย พัฒนามาจากขั้นตอนวิธี ID3 โดย Quinlan [10] ที่เขาได้พัฒนาขึ้นเป็นวิธีการเรียนรู้จาก ชุดข้อมูลสอน (training set) โดยใช้เทคนิคการจัดหมวดหมู่เพื่อสร้างต้นไม้ตัดสินใจ ชุดข้อมูลสอนจะมีลักษณะคล้ายกับข้อมูลในฐานข้อมูลเชิงสัมพันธ์ (relational database) อยู่ในรูปของตารางที่มีแถวแสดงข้อมูล และคอลัมน์แสดงคุณลักษณะของข้อมูล ซึ่งแบ่งออกเป็น 2 ชนิดคือ

1) คุณลักษณะที่เป็นจุดมุ่งหมาย (goal attribute) ของการจำแนกกลุ่มข้อมูล เป็นคุณลักษณะที่กำหนดว่าตัวอย่างนั้นๆ ถูกจัดอยู่ในกลุ่มไหน โดยจะมีเพียงคุณลักษณะเดียวในแต่ละชุดข้อมูล และข้อมูลจะเป็นชนิดข้อความเท่านั้น

2) คุณลักษณะประกอบการทำนาย (predicting attribute) เป็นคุณลักษณะที่บ่งบอกถึงคุณสมบัติต่าง ๆ ของตัวอย่างแต่ละตัวอย่าง โดยแต่ละคุณลักษณะอาจมีข้อมูลเป็นชนิดข้อความหรือตัวเลขก็ได้

2.3.4.1 การคัดเลือกคุณลักษณะเพื่อจำแนกกลุ่มของข้อมูล

ในการสร้างต้นไม้ตัดสินใจ ปัญหาสำคัญที่ต้องพิจารณาคือ ควรจะตัดสินใจเลือกคุณลักษณะใดมาทำหน้าที่เป็นโหนดราก ในแต่ละขั้นตอนของการสร้างต้นไม้และต้นไม้ย่อย (subtree) ของต้นไม้ตัดสินใจ เกณฑ์ที่ใช้ช่วยประกอบการเลือกคุณลักษณะคือการคำนวณค่าเกน (gain) ซึ่งเป็นค่าที่บ่งบอกว่าคุณลักษณะนั้นจะสามารถจำแนกกลุ่มของข้อมูลได้ดีเพียงใด โดยทดลองเลือกแต่ละคุณลักษณะที่เป็นไปได้จากชุดข้อมูลมาทำหน้าที่เป็นโหนดราก ถ้าคุณลักษณะใดให้ค่าเกนที่สูงที่สุดแสดงว่าคุณลักษณะนั้นสามารถจำแนกกลุ่มของข้อมูลได้ดีที่สุด หรือเป็นคุณลักษณะที่จัดกลุ่มของข้อมูลแล้ว ได้ข้อมูลในแต่ละใบของต้นไม้เป็นกลุ่มเดียวกันทั้งหมด หรือมีข้อมูลต่างกลุ่มปะปนมาบ้างเพียงเล็กน้อยเท่านั้นโดยค่าเกนสำหรับการเลือกคุณลักษณะที่สำคัญแสดงได้ดังนี้

1) ค่าบรรทัดฐานเกน (Gain criterion)

วิธีการสร้างต้นไม้ตัดสินใจโดยใช้อัลกอริทึม ID3 จะใช้ค่าบรรทัดฐานในการตัดสินใจเลือกคุณลักษณะที่จะใช้เป็นโหนดรากของต้นไม้หรือของต้นไม้ย่อย โดยการคำนวณค่าเกนของแต่ละคุณลักษณะเมื่อใช้แบ่งกลุ่มตัวอย่าง และเลือกคุณลักษณะที่มีค่าเกนสูงที่สุดมาเป็นโหนดรากซึ่งคุณลักษณะนี้จะมีความสามารถในการจำแนกกลุ่มข้อมูลสูง โดยที่ต้องการข้อมูลจำนวนน้อยที่สุดในการที่จะระบุว่าข้อมูลนั้นอยู่นกลุ่มใด และการคัดเลือกคุณลักษณะนี้ทำให้สามารถแบ่งข้อมูลออกมาโดยที่มี

Mahasarakham University

การปะปนกันของกลุ่มที่ต่างกันเกิดขึ้นน้อยอีกด้วย ค่าเกนนี้คำนวณได้โดยใช้ความรู้จากทฤษฎีสารสนเทศ (information theory) ซึ่งมีสาระสำคัญคือ ค่าสารสนเทศของข้อมูลจะขึ้นอยู่กับความน่าจะเป็นของข้อมูล ซึ่งสามารถวัดอยู่ในรูปของบิต (bits) เขียนเป็นสมการได้ดังนี้

ค่าสารสนเทศของข้อมูล = $-\log_2$ (ความน่าจะเป็นของข้อมูล) (2.10)

การใช้ค่าสารสนเทศเกน จะช่วยลดจำนวนการทดสอบในการจำแนกข้อมูล และยังช่วยให้มั่นใจว่าไม่มีเกิดความซับซ้อนมากเกินไปในต้นไม้ตัดสินใจที่ได ซึ่งค่าสารสนเทศเกนสามารถหาได้จากสมการดังต่อไปนี้ [6] โดยที่ S เป็นชุดของข้อมูล ซึ่งประกอบไปด้วยข้อมูล s รายการ m เป็นจำนวนกลุ่มทั้งหมดที่ต่างกันของข้อมูลชุดนั้น ให้ C_i แทนกลุ่มในลำดับที่ i โดยที่ i มีค่าระหว่าง 1 ถึง m ให้ s_i แทนจำนวนข้อมูลที่เป็นสมาชิกของ S และอยู่ในกลุ่ม C_i ให้ s_{ij} แทนจำนวนข้อมูลที่เป็นสมาชิกของ S ในกลุ่ม C_i จากการแบ่งข้อมูลด้วยค่าที่เป็นไปได้ j ของคุณลักษณะ A โดยที่ j มีค่าระหว่าง 1 ถึง vs โดย S_j/s แทนค่าความน่าจะเป็นที่ข้อมูลจะอยู่ในกลุ่ม C_i ค่าสารสนเทศที่ต้องการสำหรับการแบ่งข้อมูลออกเป็นแต่ละกลุ่มหาได้โดย

$$I(S_1, S_{2,} \dots, S_m) = -\sum_{j=1}^{m} \frac{S_i}{s} log2 \frac{S_i}{s} \quad (2.11)$$

ค่า entropy ของคุณลักษณะ A ซึ่งมีค่าของคุณลักษณะเป็น (a_1, a_2, a_3,..., a_v) หาได้โดย

$$E(A) = \sum_{j=1}^{v} \frac{S_{1j}+\dots+S_{mj}}{S} I\,(S_{1j}, \dots, S_{mj}) \quad (2.12)$$

ค่าบรรทัดฐานเกนที่จะใช้ในการเลือกคุณลักษณะ A มาเป็นโหนดของต้นไม้มีค่าเท่ากับ ปริมาณข้อมูลที่ต้องการเพื่อให้สามารถจำแนกกลุ่มของข้อมูลได้ ลบด้วยปริมาณข้อมูลที่ต้องการเพื่อการจำแนกกลุ่มของข้อมูลโดยใช้คุณลักษณะ A เป็นตัวตรวจสอบเพื่อจำแนกกลุ่มของข้อมูล เขียนเป็น

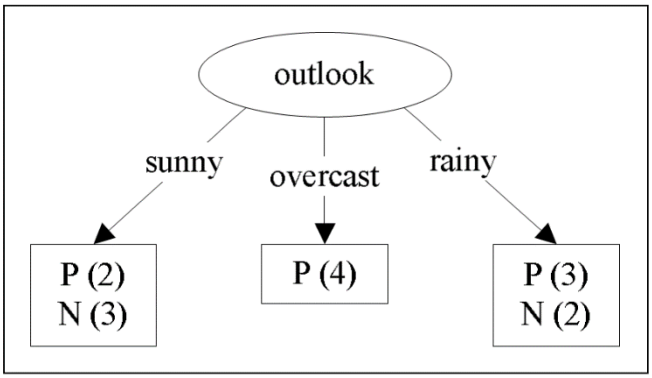
$$Gain(A) = I(S_1, S_{2,} \dots, S_m) - E(A) \quad (2.13)$$

Mahasarakham University

จากตัวอย่างเรื่องการตัดสินใจออกไปเล่นกอล์ฟโดยใช้ข้อมูลสภาพอากาศ ชุดข้อมูลที่ใช้สอน T ประกอบไปด้วยข้อมูลทั้งหมด 14 ระเบียน แยกออกเป็น 2 ชุดคือ ชุดข้อมูลที่ตอบว่าออกไปเล่น (Label = P) จำนวน 9 ระเบียน และชุดที่ตอบว่าไม่ออกไปเล่น (Label = N) จำนวน 5 ระเบียน การจะกำหนดว่าข้อมูลหนึ่งระเบียนอยู่ในกลุ่ม P หรือ N หาได้จาก

$$I(T) = -(9/14) \times \log_2(9/14) - (5/14) \times \log_2(5/14) \quad (2.14)$$
$$= 0.940 \text{ บิต}$$

การทำนายว่ารายการข้อมูลใดเป็นคำตอบว่าจะออกไปเล่นหรือไม่นั้น จำเป็นต้องใช้คุณลักษณะอื่น เพื่อใช้ในการทำนายหาคำตอบ ภาพประกอบ 2.5 เป็นการแสดงผลว่าหากใช้คุณลักษณะ outlook เป็นตัวแบ่งกลุ่มของข้อมูล โดยในวงเล็บคือจำนวนรายการข้อมูลของแต่ละกลุ่ม เมื่อแบ่งตามค่าที่เป็นไปได้จะต้องการปริมาณข้อมูลเพิ่มเพื่อประกอบการเลือกกลุ่ม และสามารถคำนวณค่า entropy ของคุณลักษณะได้



ภาพประกอบ 2.5 การจำแนกกลุ่มของข้อมูลโดยใช้คุณลักษณะ outlook

$$E(\text{outlook}) = (5/14) \times (-(2/5) \times \log_2(2/5) - (3/5) \times \log_2(3/5)) + (4/14) \times (-(4/4) \times \log_2(4/4) - (0/4) \times \log_2(0/4)) + (5/14) \times (-(3/5) \times \log_2(3/5) - (2/5) \times \log_2(2/5))$$
$$= 0.693 \text{ บิต}$$

ดังนั้นหากต้องการแบ่งกลุ่มของข้อมูลใหม่ โดยการใช้คุณลักษณะ outlook ทำหน้าที่ในการทดสอบเพื่อใช้ในการแบ่งกลุ่มข้อมูล การใช้ค่าคุณลักษณะ outlook ของข้อมูลใหม่นี้ต้องมีการใช้ข้อมูลเพิ่มอีก 0.693 บิต จึงจะสามารถหาค่าคำตอบที่ถูกต้องได การเลือกคุณลักษณะ outlook เพื่อใช้ในการจำแนกข้อมูลสามารถคำนวณหาค่าเกนได้จากสมการ ดังนี้

Mahasarakham University

$$\text{Gain (outlook)} = \text{I(T)} - \text{E(outlook)} \quad (2.15)$$

= 0.940 – 0.693

= 0.247 บิต

เราสามารถคำนวณหาค่าเกนจากคุณลักษณะอื่นๆ ที่เหลือ ประกอบด้วย คุณลักษณะ temperature, humidity และ windy ได้ดังนี้

$$\text{Gain(temperature)} = \text{I(T)} - \text{E(temperature)} \quad (2.16)$$

= 0.940 – 0.911

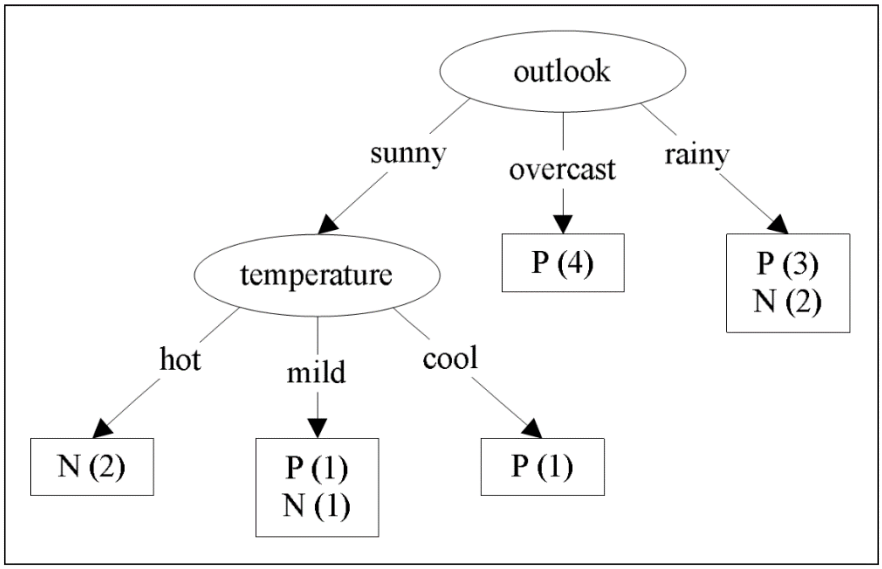
= 0.029 บิต

$$\text{Gain(humidity)} = \text{I(T)} - \text{E(humidity)} \quad (2.17)$$

= 0.940 – 0.788

จากการคำนวณสรุปได้ว่าคุณลักษณะที่มีค่าเกนมากที่สุดคือ outlook ดังนั้น outlook จึงถูกเลือกให้เป็นรูทของต้นไม้ตัดสินใจ แต่ยังคงต้องสร้างต้นไม้ตัดสินใจต่อไปอีก เพราะว่าแค่นี้ยังไม่สามารถแบ่งกลุ่มของข้อมูลให้เป็นกลุ่มเดียวกันได้ทั้งหมด ดังนั้นจึงทำการเลือกคุณลักษณะที่จะมาเป็นโหนดในระดับต่อไปเพื่อแบ่งกลุ่มข้อมูล แต่ถ้า outlook = overcast ไม่ต้องสร้างต้นไม้ตัดในใจเพิ่มอีกแล้ว เพราะว่าแบ่งกลุ่มของข้อมูลที่เป็นกลุ่ม P ได้หมดแล้ว คุณลักษณะที่ถูกเลือกเป็นโหนดระดับที่ 2 คือ temperature สามารถแบ่งกลุ่มของข้อมูลได้ดังภาพประกอบ 2.6 และสามารถหาค่าเกนได้ดังต่อไปนี้

Mahasarakham University



ภาพประกอบ 2.6 การจำแนกกลุ่มของข้อมูลโดยโหนดระดับที่ 2 (temperature)

$$I(\text{outlook} = \text{sunny}) = -(2/5) \times \log2(2/5) - (3/5) \times \log2(3/5)$$

= 0.971 บิต

$$E_{\text{temperature}}(\text{outlook} = \text{sunny}) = (2/5) \times (-(0/2) \times \log_2(0/2) - (2/2) \times \log_2(2/2))$$
$$+ (2/5) \times (-(1/2) \times \log_2(1/2) - (1/2) \times \log_2(1/2))$$
$$+ (1/5) \times (-(1/1) \times \log_2(1/1) - (0/1) \times \log_2(0/1))$$

= 0.4 บิต

$$\text{Gain}(\text{temperature}) = I(\text{outlook} = \text{sunny}) - E_{\text{temperature}}(\text{outlook} = \text{sunny})$$

= 0.971 – 0.4

= 0.571 บิต

คุณลักษณะอื่นๆ คือ humidity และ windy สามารถใช้เป็นตัวทดสอบเพื่อแบ่งกลุ่มของข้อมูลสอน สามารถคำนวณค่าเกนจากการเลือกแต่ละคุณลักษณะได้ดังนี้

$$\text{Gain}(\text{humidity}) = I(\text{outlook} = \text{sunny}) - E_{\text{humidity}}(\text{outlook} = \text{sunny})$$

= 0.971 – 0

= 0.971 บิต

$$\text{Gain}(\text{windy}) = I(\text{outlook} = \text{sunny}) - E_{\text{windy}}(\text{outlook} = \text{sunny})$$

= 0.971 – 0.951

= 0.020 บิต

Mahasarakham University

ดังนั้นคุณลักษณะ humidity ให้ค่าเกนมากที่สุด จึงถูกเลือกเป็นโหนดระดับที่ 2 ต่อจาก outlook = sunny และยังคงเหลือโหนดลูกทางขวาของโหนด outlook (outlook = rainy) ที่ต้องพิจารณาเลือกคุณลักษณะและจากวิธีการคำนวณค่าเกนที่แสดงด้วยตัวอย่างก่อนหน้านี้ สามารถเลือกได้ว่าคุณลักษณะ windy จะให้ค่าเกนสูงที่สุด จึงถูกเลือกเป็นโหนดระดับที่ 2 ต่อจาก outlook = rainy กระบวนการสร้างต้นไม้ตัดสินใจจะสิ้นสุดเมื่อโหนดใบเป็นกลุ่มของข้อมูลเดียวกันทั้งหมด และจะได้โครงสร้างของต้นไม้ตัดสินใจเป็นภาพประกอบ 2.4

2) ค่าบรรทัดฐานอัตราส่วนเกน (Gain ratio criterion)

ในอัลกอริทึม ID3 จะใช้ค่าบรรทัดฐานเกนเป็นหลักในการเลือกคุณลักษณะที่จะใช้เป็นโหนดรากของต้นไม้ตัดสินใจหรือของต้นไม้ย่อย แต่ในอัลกอริทึม C4.5 ได้เพิ่มการใช้ค่าบรรทัดฐานอัตราส่วนเกนในการตัดสินใจเลือกคุณลักษณะที่จะใช้เป็นโหนดรากเข้ามาด้วย เนื่องจากค่าบรรทัดฐานจะมีความลำเอียงอย่างมาก กับข้อมูลที่ประกอบด้วยคุณลักษณะที่มีค่าที่เป็นไปได้จำนวนมาก ๆ เช่น ในชุดข้อมูลที่มีคุณลักษณะหมายเลขประจำตัว ซึ่งมีค่าไม่ซ้ำกัน ถ้าจำแนกข้อมูลตามคุณลักษณะนี้จะทำให้ได้เพียง 1 ตัวอย่างต่อ 1 กิ่ง และเมื่อหาค่าเอนโทรปีจากการจำแนกตัวอย่างบนคุณลักษณะนี้จะเป็น 0 ทำให้คุณลักษณะนี้มีค่าเกนค่าสูงที่สุด [11]

จากข้อมูลตัวอย่างการตัดสินใจเล่นกอล์ฟในตารางที่ 2.1 ถ้าใช้คุณลักษณะ ID ในการจัดกลุ่มข้อมูลจะต้องการปริมาณข้อมูลประกอบการตัดสินใจเพื่อจำแนกกลุ่มดังนี้

$$E(ID) = (1/14) \times (-(0/1) \times \log2(0/1) - (1/1) \times \log2(1/1)) + \ldots +$$
$$(1/14) \times (-(0/1) \times \log2(0/1) - (1/1) \times \log2(1/1))$$
$$= 0 \text{ บิต}$$

เมื่อแบ่งตัวอย่างบนคุณลักษณะนี้จะได้ค่า entropy เท่ากับ 0 ดังนั้น ค่าบรรทัดฐานของคุณลักษณะนี้จะเท่ากับปริมาณข้อมูลที่ต้องการจะระบุว่าข้อมูลหนึ่งรายการอยู่ในกลุ่ม P หรือ N ที่โหนดรากซึ่งมีค่าเท่ากับ 0.940 บิต ทำให้ค่าบรรทัดฐานนี้มีค่าสูงกว่าคุณลักษณะอื่น ๆ ดังนั้น คุณลักษณะ ID นี้จะถูกเลือกมาเป็นตัวทดสอบเพื่อจัดกลุ่มของข้อมูลสอน ดังนั้นจะเห็นว่า การวัดค่าบรรทัดฐานจะได้ค่ามากเมื่อคุณลักษณะนั้น มีค่าที่เป็นไปได้จำนวนมาก ๆ ซึ่งไม่สามารถนำมาใช้เป็นโหนดของต้นไม้ เพื่อแบ่งกลุ่มของข้อมูลใหม่ที่ไม่เคยเห็นได้อย่างแม่นยำ จึงต้องแก้ไขความบกพร่องนี้โดยวิธีการปรับค่าเกนให้ถูกต้อง โดยใช้ค่าสารสนเทศการแบ่งแยก (split information) ของแต่ละคุณลักษณะ [12] เพื่อใช้คำนวณค่าบรรทัดฐานอัตราส่วนเกน

ถ้ากำหนดให้ T แทนชุดของข้อมูลสอน เมื่อแบ่งตัวอย่างโดยใช้คุณลักษณะ A จะได้ชุดของตัวอย่างย่อยในแต่ละกิ่งเป็น $\{t_1, t_2, \ldots, t_v\}$ จำนวน v ชุด ตามค่าที่เป็นไปได้ของคุณลักษณะ A และสามารถคำนวณค่าสารสนเทศการแบ่งแยกได้ดังนี้

Mahasarakham University

$$\text{ค่าสารสนเทศการแบ่งแยก} = -\sum_{j=1}^{m} \frac{|t_i|}{|T|} log2 \frac{|t_i|}{|T|} \quad (2.18)$$

ค่าสารสนเทศการแบ่งแยกนี้จะแสดงถึงระดับการกระจายของข้อมูล เมื่อแบ่งข้อมูลตัวอย่าง T เป็น v ชุดย่อยตามค่าที่เป็นไปได้ของคุณลักษณะ A โดยค่านี้จะมีค่าสูงสุดเมื่อ $|t_i|$ เป็น 1 เท่ากันในทุกกิ่ง และจะลดลงเมื่อค่า $|t_i|$ เพิ่มขึ้น เมื่อนำค่านี้ไปหารค่าบรรทัดฐานจะได้ค่าบรรทัดฐานอัตราส่วนเกน ซึ่งช่วยแก้ไขความลำเอียงที่เกิดขึ้นของค่าบรรทัดฐานได้ โดยทำให้ค่าบรรทัดฐานอัตราส่วนเกนของคุณลักษณะที่มีค่าที่เป็นไปได้จำนวนมากถูกปรับลดลง [11]

$$\text{ค่าบรรทัดฐานอัตราส่วนเกน} = \text{ค่าบรรทัดฐานเกน} \,/\, \text{ค่าสารสนเทศการแบ่งแยก} \quad (2.19)$$

จากตัวอย่างข้อมูลการตัดสินใจเล่นกอล์ฟในตารางที่ 2.1 สามารถคำนวณค่าอัตราส่วนเกนของคุณลักษณะ outlook ได้ดังนี้

ค่าสารสนเทศการแบ่งแยก (outlook) = $-(5/14) \times \log_2(5/14) - (4/14) \times \log_2(4/14) - (5/14) \times \log_2(5/14)$

= 1.577 บิต

อัตราส่วนเกน (outlook) = 0.247 / 1.577

= 0.156

เมื่อทำการแบ่งข้อมูลตัวอย่างด้วยคุณลักษณะ temperature, humidity และ windy สามารถหาค่าอัตราส่วนเกนได้ดังนี้

อัตราส่วนเกน (temperature) = 0.029 / 1.362

= 0.021

อัตราส่วนเกน (humidity) = 0.152 / 1.000

= 0.152

อัตราส่วนเกน (windy) = 0.048 / 0.985

= 0.049

Mahasarakham University

จากข้อมูลสรุปได้ว่าว่าคุณลักษณะที่ให้ค่าอัตราส่วนเกนมากที่สุดคือ outlook สอดคล้องกับการคำนวณค่าสารสนเทศการแบ่งแยก ด้วยเหตุนี้คุณลักษณะ outlook จึงถูกเลือกเป็นโหนดรูท และจะสร้างต้นไม้ตัดสินใจไปเรื่อยๆ จนสามารถจำแนกกลุ่มของข้อมูลให้เป็นกลุ่มเดียวกันได้ทั้งหมด

2.3.4.2 การตัดกิ่งต้นไม้ตัดสินใจ

ในช่วงที่กำลังสร้างต้นไม้ตัดสินใจ อาจมีการการสร้างต้นไม้อย่างผิดปกติในแต่ละกิ่ง เนื่องจากมีข้อมูลรบกวน (noise) ในข้อมูลสอน ซึ่งอาจมีสาเหตุจากผิดพลาดในการเก็บข้อมูลหรือความผิดพลาดที่เกิดจากตัวระบบเอง หรืออาจเป็นไปได้ว่าในชุดข้อมูลมีข้อมูลที่ผิดปกติจากข้อมูลส่วนใหญ่ (outlier) ปะปนมาด้วย การตัดกิ่งต้นไม้ตัดสินใจเป็นวิธีที่ใช้ในการแกปัญหานี้ และเป็นวิธีที่ช่วยลดการเกิดปัญหาการเจาะจงโมเดลกับข้อมูลเกินไป (overfitting) ได้ โดยปัญหานี้ทำให้ได้โครงสร้างต้นไม้ที่สามารถแบ่งข้อมูลได้ดีกับชุดข้อมูลที่ใช้สร้างต้นไม้ตัดสินใจเท่านั้น แต่เมื่อนำไปใช้กับข้อมูลใหม่ประสิทธิภาพในการจำแนกกลุ่มข้อมูลจะลดลง เปรียบเหมือนคนรู้ข้อสอบและจำข้อสอบไว้ ทำให้เวลาทำข้อสอบจะได้คะแนนเยอะ แต่เวลาเจอข้อสอบใหม่จะไม่สามารถทำได้ เทคนิคการตัดกิ่งต้นไม้ตัดสินใจจะใช้ค่าข้อมูลทางสถิติในการตัดกิ่งที่มีความน่าเชื่อถือน้อยที่สุดออกไป เพื่อทำให้ต้นไม้ใหม่ที่ได้สามารถทำงานได้รวดเร็วขึ้น นอกจากนี้ยังเป็นการเพิ่มความสามารถของต้นไม้เพื่อทำนายข้อมูลใหม่ ๆ ได้อย่างแม่นยำมากยิ่งขึ้นอีกด้วย โดยการตัดกิ่งต้นไม้ตัดสินใจที่นิยมมีอยู่ 2 ประเภทดังต่อไปนี้

1) การตัดกิ่งขณะที่เรียนรู้ (pre-pruning)

คือการไม่แตกกิ่งในกระบวนการสร้างต้นไม้ตัดสินใจ โดยใช้เทคนิคการเปลี่ยนโหนดที่ถูกตัดให้กลายเป็นใบ และให้ใบนั้นทำการรายงานกลุ่มที่มีความเป็นเป็นไปได้ที่ข้อมูลจะอยู่ในกลุ่มนั้นมากที่สุด [13]

ในระหว่างที่ทำการสร้างต้นไม้ตัดสินใจนั้น จะต้องมีการหาหรือวัดค่าทางสถิติที่จำเป็น เช่น X2, สารสนเทศเกน เพื่อใช้พิจารณาว่าควรที่จะสร้างหรือแตกกิ่งของต้นไม้อย่างไรถ้าค่าที่วัดได้ไม่ถึงเกณฑ์ที่กำหนดไวก็จะถือว่าไม่สมควรที่จะทำการแตกกิ่งในโหนดนั้นต่อไป ซึ่งนับว่าเป็นเรื่องยากในการหาค่าที่เหมาะสมในการหาค่าเกณฑ์ที่กำหนด หากตั้งค่ามากเกินไปต้นไม้ที่ได้จะมีความซับซ้อนสูง หากตั้งค่าน้อยเกินไปต้นไม้ที่ได้ก็จะมีขนาดเล็กจนใช้การไม่ได้

2) การตัดกิ่งหลังการเรียนรู้ (post-pruning)

คือการตัดกิ่งของต้นไม้ตัดสินใจที่ถูกสร้างขึ้นสมบูรณแล้ว โดยเทคนิคการหาค่าความซับซ้อนของแต่ละโหนด เมื่อทำการตัดกิ่งของต้นไม้เสร็จแล้วโหนดที่ที่ไม่ได้ถูกตัดซึ่งอยู่ล่างสุดจะถูกทำให้กลายใบและจะรายงานกลุ่มที่มีจำนวนข้อมูลสนับสนุนมากที่สุด [13]

Mahasarakham University

กรณีโหนดต่างๆ ที่ไม่ใช่ใบ จะมีการหาค่าอัตราความผิดพลาดที่คาดหวังไว้ ซึ่งเป็นค่าที่ใช้อธิบายถึงความผิดพลาดที่จะปรากฏ หากมีการตัดโหนดของต้นไม้ย่อย โดยที่ค่าความผิดพลาดของโหนดที่ไม่ถูกตัดจะหาได้จากค่าความผิดพลาดโดยรวมของแต่ละกิ่ง และให้ค่าความสำคัญตามอัตราของกิ่งนั้น ๆ ถ้าเกิดค่าความผิดพลาดที่สูงขึ้นจากการตัดโหนดนั้น ก็จะต้องยังคงโหนดนั้นไว้ แต่ถ้าได้ค่าความผิดพลาดอยู่ในเกณฑ์ที่ยอมรับได้ก็จะตัดโหนดนั้นออกไป เมื่อตัดกิ่งต้นไม้ตัดสินใจเสร็จแล้ว ต้องวัดค่าความแม่นยำ (accuracy) ของต้นไม้ที่ทำการตัดกิ่งแล้วด้วย โดยจะเลือกต้นไมที่มีค่าความผิดพลาดน้อยที่สุด

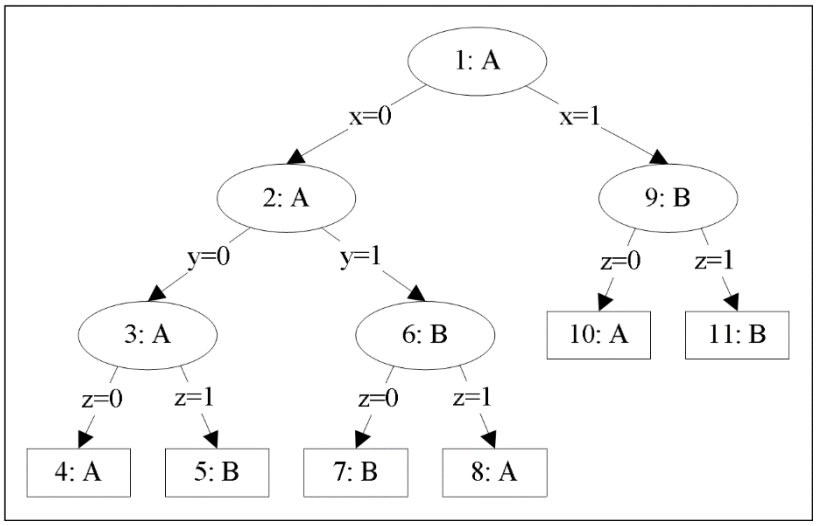
นอกจากนี้แล้วยังมีวิธีการอื่น ๆ ที่ใช้ในการตัดกิ่งของต้นไม้เช่น การเข้ารหัสหรือ encode ด้วยการใช้แนวทางของ Minimum Description Length (MDL) ด้วย [9] เป็นต้น

โดยที่จากการศึกษาที่ผ่านมา [15] แสดงให้เห็นว่าว่าการตัดกิ่งแบบนี้มีความแน่นอนกว่าและมีความสามารถที่ทำให้เกิดผลในการงานสูงกว่าการตัดกิ่งขณะที่เรียนรู้ (pre-pruning) เนื่องจากสามารถคัดเลือกโหนดที่ไม่มีประโยชนจากต้นไม้ที่สร้างขึ้น และใช้เทคนิคต่าง ๆ ในการคำนวณค่าความผิดพลาดของโหนดเพื่อใช้ในการตัดสินใจว่าจะตัดกิ่งของต้นไม้หรือไม่ สามารถแสดงกระบวนการทำงานของเทคนิคการตัดกิ่งต้นไม้ตัดสินใจแบบการตัดกิ่งหลังการเรียนรู้ ได้ดังนี้

1) การตัดกิ่งแบบลดความผิดพลาด (Reduced-error pruning)

การตัดกิ่งแบบลดความผิดพลาด [9] ถือเป็นวิธีที่มีแนวคิดง่ายที่สุดในบรรดาเทคนิคในการตัดกิ่งต้นไม้ตัดสินใจ โดยการแยกชุดข้อมูลการตัดกิ่ง (pruning set) ออกจากชุดข้อมูลสอน(training set) เพื่อหาค่าความถูกต้องของโหนดและใบของต้นไม้ที่ได้จากขั้นตอนการสร้างต้นไม้ตัดสินใจ ทำให้มีความโน้มเอียงของค่าอัตราความผิดพลาดลดลงเมื่อมีการนำไปใช้กับข้อมูลใหม่ที่ไม่เคยเห็น กระบวนการในการทำงานของเทคนิคนี้ จะใช้วิธีการตรวจสอบโหนดจากด้านล่างไปด้านบนของต้นไม้ (bottom-up strategy) โดยใช้วิธีการท่องไปแบบโพสออร์เดอร์ ทำการเปลี่ยนโหนดให้กลายเป็นใบที่มีกลุ่มของข้อมูลเป็นกลุ่มหลักของกลุ่มตัวอย่าง จากการแบ่งกลุ่มของชุดข้อมูลสอนที่โหนดนั้นของต้นไม้ นับจำนวนตัวอย่างที่ไม่ถูกต้องหรือไม่ใช่พวกเดียวกันกับตัวอย่างที่ใบนี้ เมื่อทดสอบด้วยชุดข้อมูลการตัดกิ่งเปรียบเทียบกับจำนวนตัวอย่างที่ไม่ถูกต้องของโหนดลูกของมัน ถ้าค่าความผิดพลาดในการจำแนกข้อมูลของโหนดที่เปลี่ยนเป็นใบ มีค่าน้อยกว่าหรือเท่ากับค่าความผิดพลาดในการจำแนกข้อมูลของโหนดลูกแล้วจะเลือกตัดโหนดนั้นออกไปแล้วเปลี่ยนเป็นใบ การตรวจสอบนี้จะทำซ้ำในแต่ละโหนด ถ้าไม่ทำให้ผลรวมทั้งหมดของความผิดพลาดในการจำแนกมีค่าเพิ่มขึ้น ผลที่ได้จากวิธีการนี้จะได้ต้นไม้ตัดสินใจที่มีขนาดเล็ก และให้ค่าความผิดพลาดต่ำที่สุดเมื่อทดสอบกับชุดข้อมูลการตัดกิ่ง

Mahasarakham University



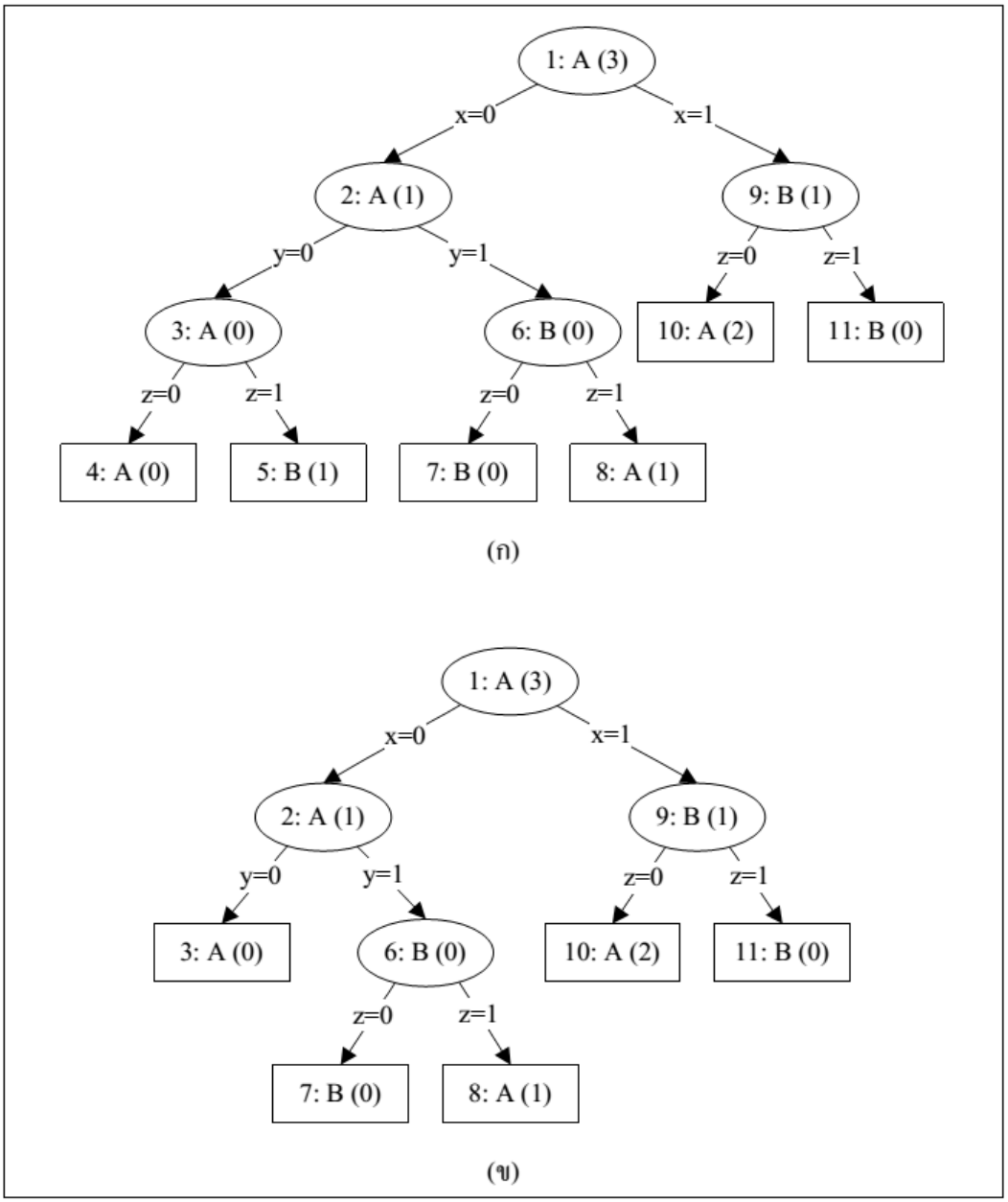
ภาพประกอบ 2.7 ต้นไม้ตัดสินใจที่สร้างขึ้นโดยมีข้อมูล 2 กลุ่ม (A และ B)

ตาราง 2.4 ตัวอย่างของชุดข้อมูลการตัดกิ่ง

x	y	z	class
0	0	1	A
0	1	1	B
1	1	0	B
1	0	0	B
1	1	1	A

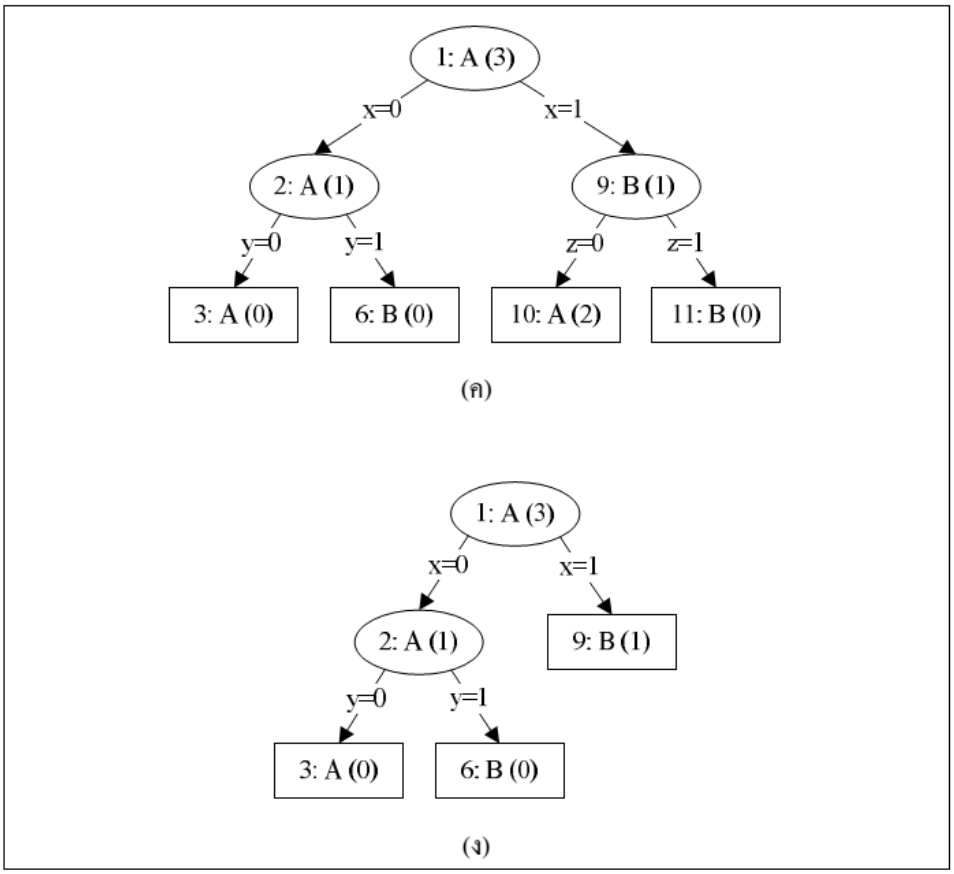
ต้นไม้ตัดสินใจที่แสดงในภาพประกอบ 2.7 [15] ได้จากกระบวนการในการสร้างต้นไม้ ซึ่งยังไม่มีการตัดกิ่งใด ๆ ออกไป โดยแสดงกลุ่มหลักของข้อมูลจากการแบ่งกลุ่มของชุดข้อมูลสอนที่แต่ละโหนดของต้นไม้และแสดงหมายเลขประจำแต่ละโหนดด้วย การตัดกิ่งต้นไม้ตัดสินใจด้วยวิธีการตัดกิ่งแบบลดความผิดพลาด ต้องใช้ชุดข้อมูลการตัดกิ่งวัดค่าความถูกต้องของต้นไม้ ซึ่งแสดงตัวอย่างชุดข้อมูลการตัดกิ่งสำหรับต้นไม้ตัดสินใจภาพประกอบ 7 ได้ดังตาราง 2.4

Mahasarakham University



ภาพประกอบ 2.8 ตัวอย่างการตัดกิ่งต้นไม้ตัดสินใจด้วยวิธีการตัดกิ่งแบบลดความผิดพลาด

Mahasarakham University



ภาพประกอบ 2.8 ตัวอย่างการตัดกิ่งต้นไม้ตัดสินใจด้วยวิธีการตัดกิ่งแบบลดความผิดพลาด (ต่อ)

ขั้นตอนการทำงานของการตัดกิ่งด้วยวิธีการตัดกิ่งแบบลดความผิดพลาด แสดงได้ดังภาพประกอบ 2.8 [15] โดยจะแสดงจำนวนตัวอย่างที่ไม่ถูกต้องของต้นไม้แต่ละต้น เมื่อทดสอบแต่ละโหนดด้วยชุดข้อมูลการตัดกิ่งจากตาราง 2.2 ไวในวงเล็บด้วย สามารถแสดงกระบวนการในการทำงานได้ดังนี้ [5]

1.1) พิจารณาที่โหนด 3 ของต้นไม้จากภาพประกอบ 2.8(ก) จะได้จำนวนตัวอย่างที่ไม่ถูกต้องเท่ากับ 0 ซึ่งมีค่าน้อยกว่าผลที่ได้จากการบวกของจำนวนตัวอย่างที่ไม่ถูกต้องที่ใบของมันซึ่งเท่ากับ 1 ดังนั้นจะตัดโหนดลูกของโหนด 3 ทิ้งแล้วเปลี่ยนโหนด 3 เป็นโหนดใบ แสดงได้ดังภาพประกอบ 2.8(ข)

1.2) พิจารณาที่โหนด 6 ของต้นไม้จากรูปที่ ภาพประกอบ 2.8(ข) จะไดจำนวนตัวอย่างที่ไม่ถูกต้องเท่ากับ 0 ซึ่งมีค่าน้อยกว่าผลที่ได้จากการบวกของจำนวนตัวอย่างที่ไมถูกต้องที่ใบของมันซึ่งเท่ากับ 1 ดังนั้นจะตัดโหนดลูกของโหนด 6 ทิ้งแล้วเปลี่ยนโหนด 6 เป็นโหนดใบ แสดงได้ดังภาพประกอบ 2.8(ค)

Mahasarakham University

1.3) พิจารณาที่โหนด 2 ของต้นไม้จากรูปที่ ภาพประกอบ 2.8(ค) จะได้จำนวนตัวอย่างที่ไม่ถูกต้องเท่ากับ 1 ซึ่งมีค่ามากกว่าผลที่ได้จากการบวกของจำนวนตัวอย่างที่ไม่ถูกต้องที่ใบของมันซึ่งเท่ากับ 0 จะเห็นว่าความผิดพลาดที่โหนดเมื่อตัดกิ่งแล้วมีค่ามากกว่า ดังนั้นจึงไม่ตัดกิ่งของโหนดนี้ แสดงได้ดังภาพประกอบ 2.8(ง)

1.4) พิจารณาที่โหนด 9 ของต้นไม้จากรูปที่ ภาพประกอบ 2.8(ค) จะได้จำนวนตัวอย่างที่ไม่ถูกต้องเท่ากับ 1 ซึ่งมีค่าน้อยกว่าผลรวมของจำนวนตัวอย่างที่ไม่ถูกต้องที่ใบของมันซึ่งเท่ากับ 2 ดังนั้นจะตัดโหนดลูกของโหนด 9 ทิ้งแล้วเปลี่ยนโหนด 9 เป็นโหนดใบ แสดงได้ดังภาพประกอบ 2.8(ง)

1.5) พิจารณาที่โหนด 1 ของต้นไม้จากรูปที่ ภาพประกอบ 2.8(ง) จะได้จำนวนตัวอย่างที่ไม่ถูกต้องเท่ากับ 3 ซึ่งมีค่ามากกว่าผลที่ได้จากการบวกของจำนวนตัวอย่างที่ไม่ถูกต้องจากโหนดลูกของโหนด 1 ซึ่งเท่ากับ 2 จะเห็นว่าความผิดพลาดที่โหนดเมื่อตัดกิ่งแล้วมีค่ามากกว่า ดังนั้นจึงไม่ตัดกิ่งของโหนดนี้ ดังนั้นต้นไม้ตัดสินใจที่ตัดกิ่งด้วยวิธีลดความผิดพลาดอย่างสมบูรณแล้วแสดงได้ดังภาพประกอบ 2.8(ง) การตัดกิ่งด้วยวิธีลดความผิดพลาด จะเข้าถึงแต่ละโหนดเพียงครั้งเดียวเพื่อประเมินโอกาสที่จะตัดโหนดนั้นออกไป ทำให้ค่าความซับซ้อนเชิงคำนวณ (computational complexity) เป็นแบบเชิงเสน ตามจำนวนโหนดของต้นไม้ตัดสินใจ O(n) เมื่อ n เป็นจำนวนโหนดของต้นไม้ตัดสินใจ แต่วิธีนี้ก็มีข้อเสียคือจำเป็นต้องใช้ชุดข้อมูลการตัดกิ่งแยกจากชุดข้อมูลสอนซึ่งอาจมีปัญหากับบางชุดข้อมูลที่มีกลุ่มตัวอย่างที่มีจำนวนไม่มาก และอาจจะไม่สามารถแบ่งแยกข้อมูลชนิดพิเศษที่อยู่นอกเหนือจากกลุ่มตัวอย่างส่วนใหญ่ได้อย่างถูกต้อง ถ้าข้อมูลนั้นไม่มีอยู่ในชุดข้อมูลการตัดกิ่ง เนื่องจากส่วนนั้น ของต้นไม้ตัดสินใจจะถูกตัดออกไปด้วย [16]

2) การตัดกิ่งแบบความผิดพลาดในแง่ร้าย (Pessimistic error pruning)

การตัดกิ่งแบบความผิดพลาดในแง่ร้าย [9] เป็นเทคนิคการตัดกิ่งที่ตรงกันข้ามกับวิธีการตัดกิ่งแบบลดความผิดพลาด คือในการตัดกิ่งต้นไม้ตัดสินใจไม่ต้องใช้ชุดข้อมูลที่แยกต่างหาก แต่ชุดข้อมูลสอนจะใช้ในกระบวนการสร้างและตัดกิ่งที่ไม่สำคัญออกไปด้วยกำหนดให้ต้นไม้ตัดสินใจ T สร้างขึ้นจากชุดข้อมูลสอนที่มีจำนวนตัวอย่างเป็น N ถ้ามีตัวอย่าง K ตัวอย่างที่ใบและมีตัวอย่างที่ไม่ถูกต้องหรือไม่ใช่พวกเดียวกันกับตัวอย่างที่ใบนี้เป็น J ดังนั้นค่าอัตราความผิดพลาดที่ใบนี้จะเท่ากับ J/K จากการทดสอบด้วยชุดข้อมูลสอนค่าอัตราความผิดพลาดนี้จะเป็นค่าสมมุติยังไม่ใช่ค่าที่แท้จริงเมื่อนำไปทำนายข้อมูลใหม่ที่ไม่เคยเห็น ดังนั้นจึงต้องมีการระบุค่าคงที่เพิ่มเข้าไปด้วย โดยให้ค่าอัตราความผิดพลาดเกิดจากการใช้ค่าปรับแก่ความตอเนื่องของการแจกแจงทวินาม (binomial distribution) โดยให้ค่า J สามารถแทนด้วย J + 1/2

ถ้าเราพิจารณาต้นไม้ย่อย T’ ของต้นไม้ T ซึ่งมีจำนวนใบของต้นไม้ย่อยเป็น L(T’) จะได้จำนวนตัวอย่างที่ไม่ถูกต้องทั้งหมดเท่ากับ $\sum_J$ เมื่อแทนค่า J ด้วย J + 1/2 จะได้จำนวนตัวอย่างทีไม่ถูกต้องเมื่อทดสอบกับข้อมูลใหม่ที่ไม่เคยเห็นเท่ากับ $\sum_J$ + L(T’)/2 วิธีนี้จะตัดกิ่งต้นไม้ย่อย

Mahasarakham University

ออกไป แล้วเปลี่ยนเป็นโหนดใบที่มีกลุ่มของข้อมูลเป็นกลุ่มหลักของกลุ่มตัวอย่างจากการแบ่งกลุ่มของชุดข้อมูลสอนที่โหนดนั้นของต้นไม้ ถ้าจำนวนตัวอย่างที่ไม่ถูกต้องของโหนดใบนี้ มีค่าน้อยกว่าหรือเท่ากับจำนวนตัวอย่างที่ไม่ถูกต้องของต้นไม้ย่อยบวกด้วยค่าความผิดพลาดมาตรฐาน (standard error) ของมัน [9]

การใช้ชุดข้อมูลสอนเพียงชุดข้อมูลเดียวสำหรับสร้างและตัดกิ่งต้นไม้ตัดสินใจ เป็นข้อดีสำหรับวิธีนี้ และสามารถทำงานได้เร็ว เพราะการท่องไปในแต่ละโหนดจะทำงานจากบนลงล่าง ทำการตรวจสอบเพียงครั้งเดียวเริ่มจากบนสุดไปยังใบของต้นไม้ โดยเมื่อต้นไม้ย่อยถูกตัดออกไปแล้วก็ไม่จำเป็นต้องตรวจสอบต้นไม้ย่อยที่อยู่ด้านล่างอีก แต่การนำค่าคงที่ 1/2 มาใช้และประมาณค่า

3) การตัดกิ่งโดยใช้ค่าความผิดพลาด (Error-based pruning)

วิธีนี้เป็นเทคนิคในการตัดกิ่งต้นไม้ตัดสินใจที่ได้นำมาใช้ในขั้นตอนวิธีในการสร้างต้นไม้ตัดสินใจที่เรียกว่า C4.5 [10] โดยปรับปรุงรูปแบบมาจากเทคนิคการตัดกิ่งแบบความผิดพลาดในแง่ร้าย โดยได้ปรับปรุงในส่วนวิธีการหาค่าความน่าจะเป็นของอัตราความผิดพลาด วิธีนี้ใช้ชุดข้อมูลสอนในการสร้างและตัดกิ่งต้นไม้ตัดสินใจ โดยไม่ต้องใช้ชุดข้อมูลที่แยกออกต่างหากในการตัดกิ่ง โดยเฉพาะขั้นตอนการทำงานจะตรวจสอบโหนดจากล่างสุดขึ้น ไปยังรากของต้นไม้ โดยใช้วิธีการท่องไปในแต่ละโหนดแบบโพสออร์เดอร์ t มีทางเลือก 2 แนวทางในการเพิ่มประสิทธิภาพต้นไม้ตัดสินใจคือ ตัดกิ่งต้นไม้ T ตรงตำแหน่งของโหนด t แล้วเปลี่ยนเป็นโหนดใบ ที่มีกลุ่มของข้อมูลเป็นกลุ่มหลักของกลุ่มตัวอย่างจากการแบ่งกลุ่มของชุดข้อมูลสอนที่โหนดนั้น หรือตัดกิ่ง T_t' ที่เป็นต้นไม้ย่อยของ T_t ที่มีผลรวมของจำนวนตัวอย่างมากที่สุดขึ้นมาแทนที่ตรงตำแหน่งของโหนด t โดยที่ไม่ทำให้ค่าความผิดพลาดเมื่อปรับปรุงต้นไม้ตัดสินใจแล้วมีค่าเพิ่มขึ้น [16]

ในการวัดความถูกต้องของต้นไม้ตัดสินใจ ไม่จำเป็นต้องวัดจากประชากรทั้งหมด จะวัดค่าจากชุดข้อมูลสอนเท่านั้น เพื่อใช้แทนค่าความน่าจะเป็นของความผิดพลาด เมื่อใช้ทำนายข้อมูลใหม่ที่ไม่เคยเห็น จึงใช้ค่าจำกัดบนของการแจกแจงแบบทวินาม (binomial distribution) ที่ระดับความเชื่อมั่นเท่ากับ CF (confidence factor) เป็นตัวแทนความผิดพลาดของประชากรแทนด้วย UCF(E,N) โดย E แทนจำนวนตัวอย่างที่แบ่งกลุ่มไม่ถูกต้องจาก N ตัวอย่าง [11]

จากตัวอย่างของต้นไม้ตัดสินใจก่อนการตัดกิ่งในภาพประกอบ 2.8 เป็นต้นไม้ตัดสินใจที่สร้างขึ้นเพื่อใช้จำแนกลักษณะของกลุ่มคนที่เลือกพรรคการเมืองในการเลือกตั้ง โดยจำแนกกลุ่มของข้อมูลออกเป็น 2 กลุ่ม {democrat, republican} ต้นไม้ตัดสินใจนี้สร้างขึ้นจากชุดข้อมูลสอนจำนวน 300 ตัวอย่าง โดยตัวเลขที่อยู่ด้านหลังโหนดใบ จะแสดงจำนวนตัวอย่างที่สามารถแบ่งกลุ่มตัวอย่างได้จากใบนั้น ต้นไม้นี้สร้างจากอัลกอริทึม C4.5 โดยแสดงในรูปแบบข้อความ ในโหนดหนึ่งของต้นไม้ที่ประกอบด้วยกิ่งและใบดังนี้

education spending = n: democrat (6.0)
education spending = y: democrat (9.0)
education spending = u: republican (1.0)

Mahasarakham University

เทคนิคการตัดกิ่งด้วยค่าความผิดพลาดใช้ในการคำนวณหาระดับหรือช่วงความเชื่อมั่น (confidence level) เพื่อเป็นการลดการโน้มเอียงที่เกิดจากการใช้ชุดข้อมูลสอนหาค่าความผิดพลาดเพียงชุดข้อมูลเดียว โดยกำหนดโครงสร้างการแจกแจงทวินามกำหนดได้เป็นการแจกแจงปกติ (normal distribution) ในชุดข้อมูลที่มีกลุ่มตัวอย่างที่มีสัดส่วนมากส่งผลให้การตัดกิ่งต้นไม้ตัดสินใจด้วยวิธีนี้จะให้ค่าความแม่นยำที่น้อยลงเมื่อใช้กับชุดข้อมูลที่มีกลุ่มตัวอย่างจำนวนน้อยกว่า 100 รายการ [15]

เทคนิคการตัดกิ่งโดยใช้ค่าความผิดพลาดมีแนวทางในการทำงาน 2 วิธี ในการแก้ปัญหาและเพิ่มประสิทธิภาพต้นไม้ตัดสินใจคือ ตัดกิ่งต้นไม้ T ที่ตำแหน่งของโหนด t แล้วปรับเป็นโหนดใบแทน ที่มีกลุ่มของข้อมูลเป็นกลุ่มหลักของกลุ่มตัวอย่างจากการแบ่งกลุ่มของชุดข้อมูลสอนที่โหนดนั้น หรือตัดกิ่ง $T_{t'}$ ที่เป็นต้นไม้ย่อยของ T_t ที่มีผลรวมของจำนวนตัวอย่างมากที่สุดขึ้นมาแทนที่ตรงตำแหน่งของโหนด t โดยที่ไม่ทำให้ค่าความผิดพลาดเมื่อปรับปรุงต้นไม้ตัดสินใจแล้วมีค่าเพิ่มขึ้น ดังนั้นค่าความซับซ้อนเชิงคำนวณของการตัดกิ่งด้วยวิธีการตัดกิ่งโดยใช้ค่าความผิดพลาดจึงมีค่าเท่ากับ O(n(log n)2) เมื่อ n เป็นจำนวนโหนดของต้นไม้ตัดสินใจ [12]

4) การตัดกิ่งแบบค่าความซับซ้อน (Cost-complexity pruning)

วิธีการนี้เป็นเทคนิคการตัดกิ่งต้นไม้ตัดสินใจที่ใช้ในขั้นตอนวิธีในการสร้างต้นไม้ตัดสินใจที่ชื่อ CART [15] โดยสามารถแบ่งกระบวนการทำงานออกเป็น 2 ขั้นตอนคือ [17]

(1) การคัดเลือกเซตของต้นไม้ย่อย จากต้นไม้ที่ได้จากขั้นตอนการสร้าง

ต้นไม้ตัดสินใจ T_{max} ได้เป็น $\{ T_0, T_1, T_2, \dots , T_L \}$

โดยที่ $T_0 = T_{max}$ และ T_L คือรากของต้นไม้ตัดสินใจ

(2) การเลือกต้นไม้ที่ดีที่สุด T_i จากเซตที่ได้ โดยใช้การประเมินความถูกต้องของต้นไม้ตัดสินใจในขั้นตอนแรกต้นไม้ T_i+1 ได้รับมาจาก T_i โดยการตัดกิ่งที่ทำให้ค่าอัตราความผิดพลาดในการจำแนก (resubstitution errors) เพิ่มขึ้นน้อยที่สุด โดยเมื่อต้นไม้ T ถูกตัดกิ่งที่โหนด t จะได้ค่าอัตราความผิดพลาดเพิ่มขึ้นเท่ากับ $R(t) - R(T_t)$ และทำให้จำนวนของใบลดลงเท่ากับ $L(T_t) - 1$

$$\frac{R(t)-R(T_t)}{L(T_t)-1} = \alpha T \tag{2.20}$$

ค่าอัตราส่วนการเพิ่มขึ้นของอัตราความผิดพลาด ต่อจำนวนใบของต้นไม้ที่ถูกตัดกิ่งออกไปเป็นค่าความซับซ้อน (cost-complexity) ของต้นไม้ T' ดังนั้น T_{i+1} จะได้รับมาจาก T_i โดย การตัดกิ่งต้นไม้ตัดสินใจที่ทำให้ค่าความซับซ้อนมีค่าน้อยที่สุด โดยถ้าต้นไม้มีค่าความซับซ้อนเท่ากันจะเลือกต้นไม้ที่มีจำนวนโหนดน้อยกว่า [9]

Mahasarakham University

ค่าความซับซ้อนที่ได้มีค่าน้อยที่สุด ดังนั้นต้นไม้ T_1 ซึ่งตกทอดมาจากต้นไม้ที่ได้จากกระบวนการในการสร้าง T_0 โดยสับเปลี่ยนต้นไม้ย่อยนี้ด้วยใบจะถูกคัดเลือกไวในชุด เพื่อนำไปคำนวณหาค่าความถูกต้อง เพื่อให้ได้ต้นไม้ตัดสินใจที่สามารถแบ่งแยกข้อมูลใหม่ได้อย่างมีประสิทธิภา และต้นไม้ตัดสินใจที่ใช้เทคนิคการตัดกิ่งแบบค่าความซับซ้อนแล้ว ในขั้นตอนต่อมาคือขั้นตอนที่2สองเป็นกระบวนการในการพิจารณาคัดสรรค์ต้นไม้ที่เหมาะสมที่สุด โดยใช้เทคนิคการวัดความถูกต้องในแบ่งแยกเพื่อหาค่าอัตราความผิดพลาดของต้นไม้แต่ละต้น โดยทั่วไปนิยมใช้เทคนิคที่ชื่อว่าการตรวจสอบแบบไขว้ (cross-validation) หรือใช้ชุดข้อมูลการตัดกิ่งในการตรวจสอบการวิธีการตัดกิ่งแบบค่าความซับซ้อนแต่จะมีข้อด้อยกว่าการตัดกิ่งแบบลดความผิดพลาด ในเรื่องการใช้ชุดข้อมูลการตัดกิ่ง เพราะว่าสามารถพิจารณาต้นไม้เฉพาะภายในชุดเท่านั้น จะไม่สามารถพิจารณาจากต้นไม้ย่อยที่มีโอกาสเป็นไปได้ทั้งหมดของต้นไม้ที่สร้างขึ้นอย่างสมบูรณ ดังนั้นหากต้นไม้ย่อยที่สร้างขึ้นมีความถูกต้องมากที่สุดเมื่อใช้กับชุดข้อมูลการตัดกิ่งไม่อยู่ในชุดแล้ว เทคนิคนี้จะไม่สามารถเลือกต้นไม้นั้นได้ [17] และการใช้วิธีตรวจสอบแบบไขว้เพื่อวัดค่าความถูกต้องของต้นไม้ตัดสินใจ จะต้องใช้เวลาในการประมวลผลมากขึ้นด้วย [9] เนื่องจากต้องมีการแบ่งเป็นกลุ่มย่อย และมีการวนรอบในการตรวจสอบมากขึ้น

2.3.4 การหาค่าเหมาะที่สุดด้วยกริด (Grid Search Optimization)

ในงานด้านต่างๆ เช่น งานวิศวกรรมศาสตร์ วิทยาศาสตร์ การจัดการ และอื่น ๆ บ่อยครั้งที่เราจะเกี่ยวข้องกับการหาจุดที่เหมาะสมที่สุด หรือจุดที่ดีที่สุด ตัวอย่างเช่น การออกแบบเครื่องบินให้มีน้ำหนักต่ำสุดและมีความแข็งแกร่งสูงสุด การออกแบบโครงสร้างของอาคารให้มีค่าใช้จ่ายที่ต่ำที่สุด หาจุดที่เหมาะสมที่สุดสำหรับการวางแผนและจัดการเวลา เป็นต้น มีวิธีการทางคณิตศาสตร์มากมายที่สามารถประยุกต์ใช้เพื่อหาจุดที่ดีที่สุด การหาค่าเหมาะที่สุด (optimization or mathematical programming) ก็คือ “การหาค่า x ซึ่งทำให้ f(x) มีค่าต่ำสุด หรือสูงสุด”

หลักการทำงานโดยทั่วไปของการหาค่าเหมาะที่สุดด้วยกริด [18] จะอาศัยวิธีการสร้างกริด (grid) ขึ้นมา และทำการหาค่าฟังก์ชันที่ทุกๆ จุดตามกริดที่กำหนด และหาค่าช่วงที่จุดต่ำสุดจะถูกบรรจุอยู่ ถ้ารู้ในเบื้องต้นว่าจุดต่ำสุดต้องอยู่ในช่วง [a,b] อย่างแน่นอน ดัง วิธีการนี้เหมาะที่จะใช้ในกรณีที่ช่วงดังกล่าวมีค่าไม่กว้างนัก มิฉะนั้นจะทำให้ใช้เวลาในการคำนวณนานเกินไป วิธีการนี้จะทำการกำหนดช่วงการค้นหาไปเรื่อยๆ จนกว่าช่วงการค้นหาจะมีค่าน้อยกว่าความคลาดเคลื่อนสูงสุดที่ยอมรับได้ กำหนดให้แบ่งช่วงการค้นหา [a,b] ออกเป็น n - 1 ช่วงเท่า ๆ กัน จะได้จำนวนจุดทั้งสิ้นที่จุดการค้นหาจะดำเนินการโดยคำนวณค่าฟังก์ชันที่ตำแหน่งกริดทั้ง n จุด หลักการค้นหาจะพิจารณาลดช่วงการค้นหาจากช่วง [a,b] ให้มีขนาดเล็กลง โดยพิจารณาคู่ของจุดที่อยู่ติดกันที่มี โอกาสบรรจุค่าต่ำสุด

Mahasarakham University

จากหลายการศึกษาก่อนหน้านี้พบกว่า วิธีการซัพพอร์ตเวคเตอร์แมชชีนจำเป็นต้องทำการกำหนดค่าพารามิเตอร์ของซัพพอร์ตเวคเตอร์แมชชีนให้เหมาะสมถึงจะทำงานได้อย่างมีประสิทธิภาพ เพื่อให้การหาค่าพารามิเตอร์นี้ได้ง่ายขึ้น จึงนิยมใช้วิธีหาค่าที่เหมาะสมที่สุด ให้พารามิเตอร์ของซัพพอร์ตเวคเตอร์แมชชีน อยู่ 2 ค่า C, γ ในการหาค่าที่เหมาะสมที่สุดของพารามิเตอร์มีอยู่หลายวิธี เช่น วิธีการเชิงวิวัฒนาการต่างๆ งานวิจัยนี้ได้เลือกวิธีการหาค่าที่เหมาะสมด้วยวิธี Grid search เนื่องจากวิธีการที่เข้าใจง่าย และมีประสิทธิ์ภาพดี โดย Grid search มีขั้นตอนการทำงานดังนี้ [18]

ขั้นที่ 1: เตรียมข้อมูล กำหนดตัวนับและค่าเริ่มต้นช่วงการค้นหา [a,b] ใด ๆ

iter_no = 1 ; $f_{min} = \infty$;

ขั้นที่ 2: แบ่งช่วงการค้นหา [a,b] ออกเป็น n - 1 ช่วงย่อย กำหนดตัวนับ i = 1

ขั้นที่ 3: ตรวจสอบค่าฟังก์ชันของจุด w_i

ถ้า $f(w_i) < f_{min}$ ให้ $f_{min} = f(w_i)$ และ idmin = i

ขั้นที่ 4: ถ้า $i \leq n - 1$ เพิ่มตัวนับ i = i +1 ทำซ้ำขั้นที่ 3

ขั้นที่ 5: ให้ k = idmin กำหนดช่วงการค้นหาในรอบต่อไปเป็น [a,b] = $[w_{k-1}, w_{k+i}]$

ขั้นที่ 6: ถ้า $S = b - a > S_{allow}$ ทำซ้ำขั้นที่ 2 และ iter_no = iter_no + 1,

ถ้าไม่ใช่ จุดคำตอบมีค่าเป็น $x_{opt} = (a+b) / 2$

2.3.5 การคัดเลือกคุณลักษณะ

การคัดเลือกคุณลักษณะมีชื่อเรียกหลายอย่าง ในงานด้านสถิติมักเรียกว่า การเลือกตัวแปร (variable selection) เพราะมองว่าคุณลักษณะแต่ละอันคือตัวแปรแบบสุ่ม (random variable) การลดคุณลักษณะ (Feature Reduction) การคัดเลือกตามลักษณะประจำ (Attribute Selection) การคัดเลือกเซตย่อยของตัวแปร (Variable Subset Selection) หรือใช้ชื่อที่เฉพาะเจาะจงกับตัวแปร เช่น การคัดเลือกยีน (Gene Selection) การคัดเลือกความยาวคลื่น (Wavelength Selection) บางทีเรียกว่า subset selection เพราะการเลือกของจำนวนหนึ่งออกจากของทั้งหมดก็คือการเลือก subset นั่นเอง และอย่างที่ทราบกันว่าถ้าเรามี คุณลักษณะ d ตัวแล้ว จำนวน subset ทั้งหมดที่เป็นไปได้คือ 2d ซึ่งใหญ่มาก (แค่มี คุณลักษณะ 20 ตัวก็สามารถเลือกได้มากกว่า 1 ล้านแบบแล้ว) เราไม่สามารถไล่เช็คที่ละ subset แล้วหาว่าอันไหนดีที่สุดได้ (brute-force search)

การคัดเลือกคุณลักษณะได้รับการนิยามจาก ผู้เขียนหลายท่าน (Guyon, 2008; John, Kohavi, & Pfleger, 1994; Kira & Rendell, 1992; Koller & Sahami, n.d.; Narendra & Fukunaga, 1977) ซึ่งมีความหมายที่ครอบคลุมในประเด็น ต่าง ๆ 4 ประเด็น สรุปได้ดังนี้

Mahasarakham University

คำนิยามตามความหมายอุดมคติ การคัดเลือกคุณลักษณะเป็นการหาเซตย่อยของคุณลักษณะที่มีขนาดเล็กที่สุดที่จำเป็นและเพียงพอ สำหรับการจำแนกประเภทข้อมูล

คำนิยามตามความหมายดั้งเดิม การคัดเลือกคุณลักษณะเป็นการเลือกเซตย่อยของคุณลักษณะขนาด m จากกลุ่มคุณลักษณะที่มีขนาด p โดยที่มีค่าฟังก์ชันเกณฑ์ที่เหมาะสมที่สุดในบรรดาเซตย่อยขนาด m เมื่อ m และ p เป็นจำนวนเต็มบวก และ $m < p$

คำนิยามในแง่ของการปรับปรุงความแม่นยำในการทำนาย การคัดเลือกคุณลักษณะมีเป้าหมายในการเลือกเซตย่อยของคุณลักษณะ เพื่อปรับปรุงความแม่นยำของการจำแนกประเภทข้อมูล หรือการลดขนาดของโครงสร้าง โดยไม่ลดความแม่นยำในการจำแนกประเภทอย่างมีนัยสำคัญ เมื่อใช้เฉพาะคุณลักษณะที่เลือกในการสร้างตัวแบบจำแนกประเภท

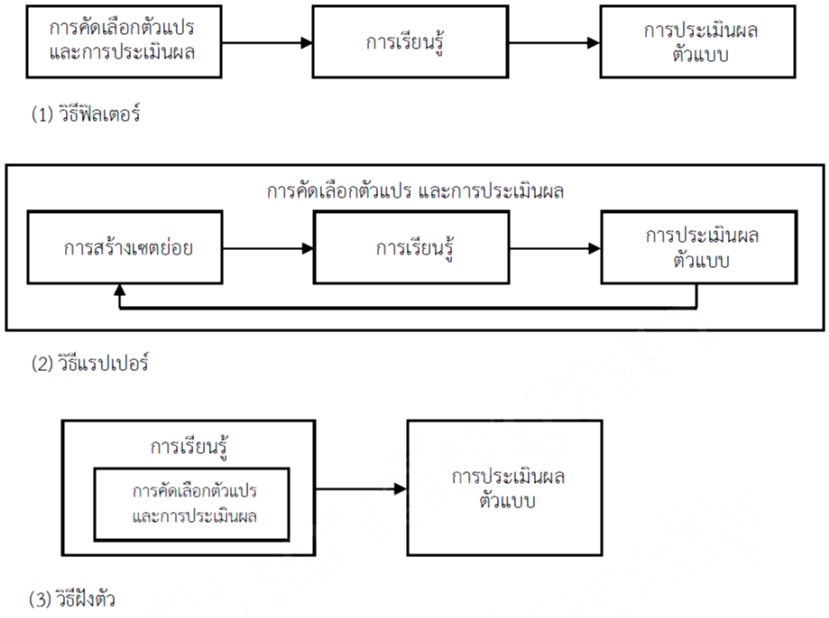
คำนิยามในแง่ของการประมาณการกระจายของกลุ่มเริ่มต้น การคัดเลือกคุณลักษณะมีเป้าหมายเพื่อเลือกเซตย่อยที่มีขนาดเล็กที่การกระจายของกลุ่ม (Class Distribution) เมื่อใช้เฉพาะคุณลักษณะที่ถูกเลือกมีลักษณะใกล้เคียงกับการกระจายของกลุ่มเริ่มต้นเมื่อมีคุณลักษณะครบ โดยสรุปแล้วการคัดเลือกคุณลักษณะ หมายถึงกระบวนการลดขนาดของคุณลักษณะ เพื่อให้ได้เซตย่อยของคุณลักษณะที่มีความเกี่ยวข้องกับการจำแนกประเภท เพื่อวัตถุประสงค์สำหรับการปรับปรุงประสิทธิภาพในการจำแนกประเภท การจัดเตรียมคุณลักษณะสำหรับการจำแนกประเภทที่สามารถประมวลผลได้อย่างรวดเร็วมีประสิทธิภาพ และเพิ่มความเข้าใจต่อตัวแบบที่ได้แนวทางของการคัดเลือกคุณลักษณะ

ดังนั้นการคัดเลือกคุณลักษณะ ก็คือการเลือกคุณลักษณะที่สามารถอธิบาย Y ได้ พูดอีกอย่างก็คือการตัดคุณลักษณะที่ไม่จำเป็นออก จุดต่างของการหาคุณลักษณะพิเศษและการคัดเลือกคุณลักษณะ คือการสร้างคุณลักษณะใหม่หรือไม่ การหาคุณลักษณะพิเศษมีการ “แปลง” แปลว่าเราได้คุณลักษณะชุดใหม่ แต่การคัดเลือกคุณลักษณะไม่มีการแปลง แค่ตัดออกหรือเลือกเก็บไว้ จึงได้คุณลักษณะชุดเดิมแต่จำนวนน้อยลง ทำให้ความหมายของคุณลักษณะยังคงเดิม ยกตัวอย่างเดิมคือเรื่องการจำแนกเอกสาร ถ้าเราใช้ถุงคำ(bag of words) เป็นคุณลักษณะแล้ว เราทำการคัดเลือกคุณลักษณะ เพื่อเลือก คุณลักษณะ 100 ตัวที่ดีที่สุด เราก็จะรู้ได้ว่าคำ 100 คำที่ดีที่สุดที่สามารถจำแนกประเภทเอกสารได้คือคำอะไรบ้าง เป็นต้น

การคัดเลือกคุณลักษณะ เป็นกระบวนการที่เลือกกลุ่มย่อยจากเซตของคุณลักษณะ (Feature set) ต้นฉบับ ซึ่งจะทำให้ได้คุณลักษณะที่เหมาะสมในการนำไปใช้ในการจำแนกข้อมูลทั้งหมด ซึ่งวิธีการตัดเลือกคุณลักษณะนี้จะช่วยปรับปรุงความถูกต้องในการจำแนกข้อมูลและหลีกเลี่ยงการเกิดปัญหาจำโมเดลมากเกินไป (overfitting) ได้

ในบริบทของการจำแนกประเภท การคัดเลือกคุณลักษณะแบ่งออกได้เป็น 3 วิธี ได้แก่ (1) วิธีฟิลเตอร์ (Filter Method) (2) วิธีแรปเปอร์ (Wrapper Method) และ (3) วิธีฝังตัว (Embedded Method) ซึ่งภาพประกอบ 2.9 แสดงการเปรียบเทียบแนวคิดของทั้ง 3 วิธีดังกล่าว

Mahasarakham University



ภาพประกอบ 2.9 การเปรียบเทียบแนวคิดของการคัดเลือกคุณลักษณะ (1) วิธีฟิลเตอร์ (2) วิธีแรปเปอร์ และ (3) วิธีฝังตัว

วิธีฟิลเตอร์คัดเลือกคุณลักษณะโดยประเมินความเกี่ยวข้องหรือวัดความสำคัญของคุณลักษณะต่อการจำแนกประเภทด้วยการพิจารณาคุณสมบัติในเนื้อแท้ของข้อมูลอย่างเป็นอิสระกับวิธีการจำแนกประเภท ซึ่งวิธีฟิลเตอร์จะคำนวณค่าความสำคัญของคุณลักษณะ หรือเซตย่อยของคุณลักษณะจากดัชนีวัดความสำคัญ และเลือกคุณลักษณะหรือเซตย่อยของคุณลักษณะที่ให้ค่าดัชนีสูง

วิธีแรปเปอร์อาศัยวิธีการจำแนกประเภท ในการวัดความสำคัญของเซตย่อยของคุณลักษณะ โดยเลือกเซตย่อยที่มีความแม่นยำในการจำแนกประเภทข้อมูลสูง หรือใช้ความแม่นยำในการจำแนกประเภทข้อมูลเมื่อใช้เซตย่อยนั้น ๆ ในการเป็นดัชนีวัดความสำคัญของเซตย่อย ซึ่งทำให้ได้เซตย่อยที่มีความแม่นยำในการจำแนกประเภทสูงกว่าการใช้ดัชนีวัดความสำคัญอื่น ๆ แต่เนื่องจากในการประเมินความเกี่ยวข้องของเซตย่อยแต่ละครั้งต้องเข้าสู่กระบวนการของวิธีการจำแนกประเภททำให้การคัดเลือกคุณลักษณะด้วยวิธีแรปเปอร์นั้นใช้เวลาในการประมวลผลค่อนข้างมากถ้าข้อมูลมีจำนวนคุณลักษณะมาก [19]

วิธีฝังตัวเป็นวิธีที่การคัดเลือกคุณลักษณะเป็นส่วนหนึ่งในกระบวนการจำแนกประเภทด้วย โดยทำการเลือกเซตย่อยที่มีคุณลักษณะที่เหมาะสมในแต่ละขั้นตอนวิธี พร้อมไปกับการสร้างตัวแบบสำหรับการจำแนกประเภท ซึ่งวิธีฝังตัวมีลักษณะคล้ายกับวิธีแรปเปอร์ที่ว่ามีการผูกติดกับวิธีการจำแนกประเภทที่เฉพาะเจาะจง แต่ใช้เวลาในการประมวลผลน้อยกว่าวิธีแรปเปอร์

Mahasarakham University

วิธีแรปเปอร์เป็นวิธีการคัดเลือกคุณลักษณะที่มีความซับซ้อนในการคำนวณสูงสุด ตามมาด้วยวิธีฝังตัว ทั้งสองวิธีดังกล่าวมีแนวทางในการคัดเลือกเซตย่อยของคุณลักษณะบนพื้นฐานของวิธีการจำแนกประเภทที่เฉพาะเจาะจง ซึ่งมีแนวโน้มที่จะได้ตัวแบบที่สามารถนำมาใช้ในการเรียนรู้ได้ดี แต่นำไปใช้

ทำนายข้อมูลอื่นได้ไม่ดี เนื่องจากทำให้เกิดปัญหาการจำโมเดล (Overfitting) โดยแนวโน้มการเกิดปัญหาดังกล่าวมีมากกว่าวิธีฟิลเตอร์ซึ่งเป็นอิสระจากวิธีการจำแนกประเภท นอกจากนี้ถึงแม้ว่าวิธีแรปเปอร์ และวิธีฝังตัวจะเป็นวิธีการคัดเลือกคุณลักษณะที่มีความแม่นยำในการจำแนกประเภทสำหรับปัญหาที่เฉพาะเจาะจง แต่สำหรับข้อมูลที่มีจำนวนมิติมากนั้น วิธีฟิลเตอร์มักเป็นวิธีที่ถูกเลือกใช้ในการคัดเลือกคุณลักษณะ เนื่องจากประมวลผลได้เร็วกว่าทั้งสองวิธี

วิธีฟิลเตอร์

วิธีฟิลเตอร์เป็นวิธีการคัดเลือกคุณลักษณะที่เร็วและง่ายต่อการตีความ โดยจะกำจัดคุณลักษณะที่ไม่เกี่ยวข้องต่อการจำแนกประเภทข้อมูล ด้วยคุณสมบัติในเนื้อแท้ของข้อมูล ซึ่งกระบวนการเป็นอิสระจากวิธีการจำแนกประเภทวิธีการคัดเลือกคุณลักษณะโดยวิธีฟิลเตอร์สามารถแบ่งได้เป็น 2 ประเภท ได้แก่

วิธีฟิลเตอร์แบบคุณลักษณะเดียว (Univariate Filter Method) และวิธีฟิลเตอร์แบบหลายคุณลักษณะ (Multivariate Filter Method) โดยวิธีฟิลเตอร์แบบคุณลักษณะเดียวจะพิจารณาความเกี่ยวข้องของคุณลักษณะทำนายต่อคุณลักษณะกลุ่ม (ตัวแปรตาม) โดยพิจารณาแต่ละคุณลักษณะทำนายแยกกัน ในขณะที่วิธีฟิลเตอร์แบบหลายคุณลักษณะจะรวมความสัมพันธ์ระหว่างคุณลักษณะทำนายด้วยกันให้มีผลต่อการพิจารณาคัดเลือกคุณลักษณะโดยส่วนใหญ่จะพิจารณาเซตย่อยของคุณลักษณะ เพื่อที่จะนำความสัมพันธ์ระหว่างคุณลักษณะเข้ามาพิจารณาด้วย ซึ่งทำให้มีโอกาสได้เซตย่อยของคุณลักษณะที่เหมาะสมมากกว่า แต่อย่างไรก็ดีกระบวนการค้นหาเซตย่อยจะทำให้ใช้เวลาในการประมวลผลมาก และลดความสามารถในการจัดการกับข้อมูลที่มีขนาดใหญ่ ได้กล่าวถึงการคัดเลือกคุณลักษณะที่มีลักษณะเดียวกับวิธีฟิลเตอร์แบบคุณลักษณะเดียวโดยเรียกว่าเป็นตัวจำแนกตัวแปรเชิงเดี่ยว (Single Variable Classifiers) ส่วนวิธีฟิลเตอร์แบบหลายคุณลักษณะซึ่งถูกรวมเข้ากับวิธีแรปเปอร์ และวิธีฝังตัวจะเรียกเป็นการคัดเลือกเซตย่อยของคุณลักษณะ

วิธีฟิลเตอร์แบบคุณลักษณะเดียว

ในบางครั้งอาจเรียกวิธีนี้ว่าการเรียงลำดับคุณลักษณะ (Feature Ranking) การถ่วงน้ำหนักคุณลักษณะ (Feature Weighting) หรือการประเมินคุณลักษณะเดียว (Individual Evaluation) วิธีฟิลเตอร์แบบคุณลักษณะเดียวจึงเป็นวิธีการคัดเลือกคุณลักษณะโดยประเมินผลทีละคุณลักษณะแยกจากกัน จากนั้นเรียงลำดับความสำคัญ หรือความเกี่ยวข้องของคุณลักษณะโดยอาศัยดัชนีวัดความสำคัญในการเรียงลำดับคุณลักษณะ ได้แก่ มาตรวัดระยะทาง มาตรวัดสารสนเทศ

Mahasarakham University

และมาตรวัดความไม่เป็นอิสระ คุณลักษณะที่มีค่าดัชนีสูงสุด m คุณลักษณะ จะถูกคัดเลือก หรือเลือกคุณลักษณะ ที่มีค่าสูงกว่าเกณฑ์จุดตัด (Cutoff Threshold Criterion) t โดยที่ m และ t เป็นเกณฑ์ที่ผู้ใช้เป็นผู้กำหนด วิธีนี้เป็นวิธีที่ทำได้ง่าย และประมวลผลได้รวดเร็ว แต่มีข้อเสียในเรื่องการละเลยความสัมพันธ์ระหว่างคุณลักษณะทำนายด้วยกัน คุณลักษณะที่ถูกเลือกถึงแม้เป็นคุณลักษณะที่มีความเกี่ยวข้องกับ การจำแนกประเภท แต่ในบรรดาคุณลักษณะเหล่านั้นบางคุณลักษณะอาจไม่ได้เพิ่มสารสนเทศที่ช่วยใน การจำแนกประเภท หรือเรียกว่าเป็นตัวแปรที่ซ้ำซ้อน (Redundant Variable) และนอกจากนี้การคัดเลือกคุณลักษณะโดยวิธีนี้อาจละเลยคุณลักษณะที่ไม่เกี่ยวข้องกับการจำแนกประเภทเมื่อพิจารณาเดี่ยว ๆ แต่จะมีความเกี่ยวข้องเมื่อพิจารณาร่วมกับคุณลักษณะอื่น

วิธีฟิลเตอร์แบบหลายคุณลักษณะ

วิธีนี้เป็นวิธีการคัดเลือกคุณลักษณะที่ได้พิจารณาความสัมพันธ์ระหว่างคุณลักษณะทำนายด้วยกันใน ระดับหนึ่งด้วย ซึ่งจะทำให้ได้คุณลักษณะที่ทำนายผลได้แม่นยำขึ้น ช่วยแก้ปัญหาที่เป็นข้อเสียของวิธี ฟิลเตอร์แบบคุณลักษณะเดียว

วิธีการคัดเลือกคุณลักษณะด้วย Relief Algorithm

การคัดเลือกคุณลักษณะ เป็นกระบวนการในการเลือกกลุ่มย่อยจากชุดของคุณลักษณะ (Feature set) ต้นฉบับ ซึ่งจะทำให้ได้คุณลักษณะที่เหมาะสมในการนำไปใช้ในการจำแนกข้อมูลทั้งหมด ซึ่งวิธีการตัดเลือกคุณลักษณะนี้จะช่วยปรับปรุงความถูกต้องในการจำแนกข้อมูลและหลีกเลี่ยงการเกิดปัญหาการเรียนรู้แบบจำคำตอบได้ (overfitting)

การเลือกคุณลักษณะของข้อมูลแบบรีลีฟเป็นอัลกอริทึม ซึ่งได้รับความนิยมเป็นอย่างมาก เนื่องจากสามารถเข้าใจและแปลความหมายได้ง่าย ทำงานได้เร็ว มีประสิทธิภาพ และยังสามารถที่จะพัฒนาโปรแกรมได้ไม่ยากนัก หลักการของรีลีฟได้รับการพัฒนามาจากการเรียนรู้ โดยใช้กลุ่มตัวอย่าง (Instance-based Learning) ซึ่งอาศัยการถ่วงน้ำหนักของคุณลักษณะ รีลีฟใช้การหาคุณลักษณะที่มีความเกี่ยวข้องกับค่าเป้าหมาย (Target Concept) โดยอาศัย หลักการทางสถิติ ซึ่งเป็นวิธีการที่มีประสิทธิภาพ เวลาที่ใช้ในการทำงานของรีลีฟขึ้นกับจำนวนของคุณลักษณะ และจำนวนตัวอย่างของข้อมูลที่ใช้ในการเรียนรู้ คุณลักษณะที่มีความเกี่ยวข้องกับค่าเป้าหมายจะถูกเลือก และจะละทิ้งคุณลักษณะที่ไม่มีความเกี่ยวข้องกับค่าเป้าหมายไปทำให้คุณลักษณะที่เหลืออยู่มีจำนวนลดลงได้ รีลีฟอัลกอริทึมมีขั้นตอนการทำงานดังนี้ [20]

Relief (S,m,t)

1. แยกตัวอย่าง S เป็นกลุ่มตัวอย่างบวก S+ และ กลุ่มตัวอย่างลบ S-
2. กำหนดเวกเตอร์น้ำหนักเริ่มต้น W = (0, 0, ... , 0)
3. ทำซ้ำจาก i = 1 ถึง m
   3.1 สุ่มเลือก X ที่เป็นสมาชิกของ S

Mahasarakham University

3.2 เลือกข้อมูลค่าบวกที่ใกล้เคียงกับ X มากที่สุด, Z+ ∈ S+

3.3 เลือกข้อมูลค่าลบที่ใกล้เคียงกับ X มากที่สุด, Z- ∈ S-

3.4 ตรวจสอบว่าถ้า X เป็นข้อมูลค่าบวกแล้ว NearHit = Z+; NearMiss = Z- มิฉะนั้นNearHit = Z-; NearMiss= Z+

3.5 ทำซ้ำจาก j=1 ถึง p เมื่อ p คือจำนวนคุณลักษณะ

คำนวณค่า W_j $W_j = W_j - diff(X_j - NearHit_j)2 + diff(X_j - NearMiss_j)^2$

โดยที่ diff(x,y) = 0 ถ้า x และ y เหมือนกัน

= 1 ถ้า x และ y ต่างกัน

4. ให้ R = (1/m) W

5. ทำซ้ำจาก j=1 ถึง p

ถ้า $R_j \geq t$ แล้ว f_j เป็น คุณลักษณะที่เกี่ยวข้อง มิฉะนั้น f_j เป็นคุณลักษณะที่ไม่เกี่ยวข้อง

จากอัลกอริทึมขั้นตอนการทำของขั้นตอนวิธีรีลีฟ เมื่อ S คือข้อมูลที่ใช้ในการเรียนรู้ m เป็นจำนวนรอบของการสุ่มตัว อย่าง คือ Threshold ของความเกี่ยวข้องกับ Target Concept และ p คือจำนวนคุณลักษณะ ขั้นตอนวิธีรีลีฟใช้ระยะทางของยูคลิด p มิติ ในการเลือก Near-Hit และ Near-Miss ในแต่ละรอบน้ำหนักของคุณลักษณะ W จะถูกปรับปรุง และสุดท้ายจะพิจารณาค่าเฉลี่ยน้ำหนักของแต่ละคุณลักษณะซึ่งจะดูว่าถึงค่าThreshold ที่กำหนดหรือไม่ ถ้าถึงระดับ Threshold จะเป็นคุณลักษณะที่มีความเกี่ยวข้องกับ Target Concept แต่ถ้าไม่ถึงระดับ Threshold จะเป็นคุณลักษณะที่ไม่มีความเกี่ยวข้องกับ Target Concept

รีลีฟจะทำการปรับค่าน้ำหนักของแต่ละคุณลักษณะ โดยใช้ระยะทางในการเลือกตัวอย่างใกล้สุดที่อยู่ในกลุ่มเดียวกัน (Near-Hit) และตัวอย่างใกล้สุดที่อยู่คนละกลุ่ม (Near-Miss) ผลลัพธ์ที่ได้จะเป็นค่าน้ำหนักที่อยู่ในช่วง -1 ถึง 1 ของแต่ละคุณลักษณะ ในแต่ละรอบน้ำหนักของคุณลักษณะจะถูกปรับค่า และเมื่อครบตามจำนวนรอบที่กำหนดไว้หรือตรงตาม เงื่อนไขที่ตั้งไว้ ก็จะพิจารณาค่าน้ำหนักของแต่ละคุณลักษณะว่าถึงค่าที่กำหนดหรือไม่ ถ้าถึงระดับที่กำหนดไว้ก็จะเลือกเก็บไว้เป็นคุณลักษณะที่มีความสำคัญต่อการจำแนกข้อมูลต่อไปจุดเด่นของรีลีฟ คือทนทานต่อสิ่งรบกวน แต่จะไม่ได้พิจารณาความสัมพันธ์ระหว่างกันของคุณลักษณะใช้การถ่วงน้ำหนักของคุณลักษณะ Relief ใช้การหาคุณลักษณะที่มีความเกี่ยวข้องกับ Target Concept โดยอาศัยหลักการทางสถิติและมีประสิทธิภาพของการทำงานที่ดีการทำงานของ Relief ใช้เวลาเป็นเชิงเส้นตรง (Linear Time) ซึ่งขึ้นกับจำนวนของคุณลักษณะและจำนวนตัวอย่างของข้อมูลที่ใช้ในการเรียนรู้จำนวนของคุณลักษณะที่ผ่าน Relief มักน้อยลงอันเนื่องมาจากอัลกอริทึมได้คัดเลือกคุณลักษณะที่มีความเกี่ยวข้องกับ Target Concept (Relevant Feature) และละทิ้งคุณลักษณะที่ไม่มีความเกี่ยวข้องกับ Target Concept

Mahasarakham University

จุดดีของขั้นตอนวิธีรีลีฟ คือ ไม่อ่อนไหวต่อสิ่งรบกวนรอบข้างง่ายๆ (Noise Tolerant) ทำให้ไม่ผลต่อคุณลักษณะมีความเกี่ยวข้องกัน (Feature Interaction) เพราะโดยทั่วไปแล้วคุณลักษณะมักมีเกี่ยวข้องและสัมพันธ์ระหว่างกัน

2.3.6 การลดจำนวนคุณลักษณะ (dimensionality reduction)

โดยทั่วไปในการจำแนกข้อมูลด้วยการเรียนรู้แบบมีผู้สอน (supervised learning) การใช้คุณลักษณะ (feature) ทั้งหมดที่มีอยู่ในการสอนระบบถือว่าเป็นเรื่องธรรมดา แต่ในบางครั้งหรือบางงานที่คุณลักษณะมีจำนวนมากเกินไปจนทำให้การสอนระบบทำได้ไม่ดี ดังนั้นวิธีการลดจำนวนคุณลักษณะจะช่วยให้สามารถสอนระบบได้อย่างมีประสิทธิภาพมากยิ่งขึ้น เช่น งานด้านการจำแนกไฟล์ข้อความ (text classification) ถือเป็นตัวอย่างของปัญหาที่มีตัวแปรมากเกินไป เนื่องจากงานด้านนี้นิยมใช้การสร้างคุณลักษณะด้วยวิธีการที่เรียกว่า ถุงคำ (bag of words) กล่าวคือ ให้หนึ่งคำเป็นหนึ่งคุณลักษณะ แล้วค่าของคุณลักษณะนั้นๆ คือจำนวนครั้งของคำนั้นที่เกิดขึ้นในเอกสาร การใช้ถุงคำแบบนี้จะทำให้จำนวนคุณลักษณะเท่ากับจำนวนคำทั้งหมดในคลังข้อความ ซึ่งโดยปกติจะพบได้มากกว่า 20,000 คำ นั่นแสดงว่าระบบต้องเรียนรู้ในพื้นที่ที่ใหญ่มากกว่า 20,000 มิติ

การที่โมเดลสอนระบบโดยใช้คุณลักษณะจำนวนมากๆ อาจส่งผลให้เกิดปัญหาดังต่อไปนี้

1. ถ้ามีจำนวนคุณลักษณะมาก สิ่งรบกวน (noise) ก็เยอะตามไปด้วย ซึ่งอาจส่งผลกระทบต่อการเรียนรู้
2. สิ้นเปลืองเนื้อที่ในการเก็บข้อมูลโดยเปล่าประโยชน์
3. ต้องใช้เวลามากขึ้นในการสอนระบบ
4. ตีความโมเดลได้ยาก เนื่องจากขึ้นอยู่กับตัวแปรจำนวนมาก

ด้วยเหตุนี้จึงมีงานวิจัยด้านการลดจำนวนคุณลักษณะ (dimensionality reduction) เกิดขึ้นมาเพื่อแก้ปัญหาดังกล่าว งานด้านนี้สามารถแบ่งได้เป็น 2 หมวดกว้างๆ คือ การหาคุณลักษณะพิเศษ (feature extraction) และ การคัดเลือกคุณลักษณะ (feature selection)

การหาคุณลักษณะพิเศษ (Feature Extraction) การหาคุณลักษณะพิเศษ คือ การแปลง (transform) คุณลักษณะให้อยู่ในปริภูมิที่มีมิติที่ต่ำกว่าเพื่อให้การเรียนรู้ทำได้ง่ายขึ้น การ “แปลง” ในที่นี้ก็คือการสร้างคุณลักษณะใหม่ขึ้นมานั่นเอง โดยที่ คุณลักษณะ ใหม่ที่ว่านี้เกิดจากการทำอะไรบางอย่างกับ คุณลักษณะ เก่าที่มี เขียนเป็นสมการได้แบบนี้ กำหนดให้ X=(X1,...,XD) แทน คุณลักษณะ ชุดเดิมจำนวน D อัน ให้ X′=(X′1,...,X′d) แทน คุณลักษณะ ชุดใหม่ซึ่งมี d อัน โดยที่ d<D และให้ Y แทน output ที่ต้องการเรียน เช่นหากพูดถึงงานจำแนกเอกสาร Y คือประเภทหรือหมวดหมู่ของเอกสารการทำการหาคุณลักษณะพิเศษ ก็คือการใช้ฟังก์ชัน g เพื่อให้ได้คุณลักษณะชุดใหม่

Mahasarakham University

โดยที่ฟังก์ชัน g จะทำอะไรก็ขึ้นอยู่กับขั้นตอนวิธี (algorithm) ของการหาคุณลักษณะพิเศษ ที่ใช้ ยกตัวอย่างเช่นใน Linear Discriminant Analysis (LDA) g จะเป็นฟังก์ชันเพื่อฉาย (project) ลงบนปริภูมิย่อยในลักษณะที่ ข้อมูลที่มาจาก class ต่างกัน (เช่นเอกสารมาจากต่างหมวดกัน) จะแยกจากการ และข้อมูลใน class เดียวกันจะเกาะกลุ่มกันเพื่อให้สามารถจำแนกประเภทได้ง่าย จะเห็นว่าการทำการหาคุณลักษณะพิเศษแบบนี้สามารถลดเวลาการฝึกสอนได้ เนื่องจาก d < D และเพิ่มความถูกต้องในการเรียนรู้ด้วยหากคุณลักษณะใหม่ที่ได้เหมาะกับปัญหานั้นๆ แต่การหาคุณลักษณะพิเศษไม่ได้ช่วยแก้ปัญหาเรื่องการตีความคุณลักษณะชุดเดิม หากต้องการหาว่า คุณลักษณะตัวไหนมีส่วนช่วยในการเรียนรู้ได้มาก เราต้องทำการหาคุณลักษณะพิเศษ

2.3.7 ข้อมูลที่ไม่สมดุล (Imbalanced Datasets)

ข้อมูลไม่สมดุล หมายถึง ข้อมูลมีลักษณะการกระจายตัวที่ไม่เท่ากัน หรือ การที่ข้อมูลกลุ่มหนึ่งมีจำนวนมากกว่าจำนวนข้อมูลของกลุ่มที่เหลือเป็นจำนวนมาก ยกตัวอย่าง เช่น ข้อมูลการฉ้อโกง (fraud) มีคำตอบที่เราต้องการ คือ คำตอบ normal จะมีจำนวนเยอะมากๆ ส่วนคำตอบที่เป็น fraud จะมีจำนวนน้อยมาก หรือข้อมูลทางด้านการแพทย์ ข้อมูลผู้ป่วยโรคต่างๆ ปัญหานี้มีคำตอบที่ต้องการ คือ ผู้ป่วยรายนั้นเป็นโรค (Positive) หรือ ไม่เป็นโรค (Negative) ซึ่งข้อมูลผู้ป่วยที่ไม่เป็นโรค (กลุ่มหลัก) อาจจะมีข้อมูลหลายร้อยคน แต่ข้อมูลผู้ที่เป็นโรคอาจมีแค่หลักสิบคนเท่านั้น (กลุ่มรอง) ดังนั้น ถ้าเรานำข้อมูลทั้งสองกลุ่มมาสอนเพื่อแบ่งกลุ่มข้อมูลพร้อมกันทั้งหมด ผลลัพธ์ที่ได้พบว่า ความถูกต้องของการจำแนกประเภทข้อมูลมีความเอนเอียง นั่นคือสามารถจำแนกประเภทข้อมูลกลุ่มที่เป็นข้อมูลที่อยู่ในคลาสส่วนมากได้อย่างถูกต้องแม่นยำ ในขณะเดียวกันข้อมูลที่อยู่ในกลุ่มที่เป็นข้อมูลที่อยู่ในคลาสส่วนน้อยจะไม่สามารถจำแนกประเภทข้อมูลได้หรือจำแนกประเภทข้อมูลได้น้อย ทั้งนี้เนื่องจากในขั้นตอนการเรียนรู้ของโมเดลนั้นจะให้ความสำคัญกับข้อมูลที่อยู่ในคลาสส่วนมากเมื่อนำข้อมูลที่ไม่เคยผ่านขั้นตอนการเรียนรู้เข้าไปทดสอบ ความน่าจะเป็นของการจำแนกประเภทข้อมูลก็จะเกิดความเอนเอียงไปยังกลุ่มของคลาสส่วนมากส่งผลให้ข้อมูลกลุ่มที่เป็นข้อมูลที่อยู่ในคลาสส่วนน้อยเกิดการจำแนกประเภทผิดกลุ่มโดยทั่วไปข้อมูลกลุ่มที่มีจำนวนมากจะถูกเรียกว่า คลาสส่วนมาก(Majority Class หรือ Negative Class) และข้อมูลกลุ่มที่มีจำนวนน้อยจะถูกเรียกว่า คลาสส่วนน้อย(Minority Classหรือ Positive Class) ซึ่งข้อมูลที่อยู่ในคลาสส่วนน้อยจะเป็นข้อมูลที่งานวิจัยนี้ให้ความสำคัญมากกว่าข้อมูลที่อยู่ในคลาสส่วนมาก

วิธีในการแบ่งกลุ่มข้อมูลที่ไม่สมดุล จะมี 3 วิธีการหลัก ๆ คือ Sampling Methods, Cost-Sensitive Methods และ Kernel-Based Methods

Sampling Methods สำหรับวิธีการนี้จะเป็นการประยุกต์เอาวิธีสุ่มตัวอย่างซึ่งเป็นวิธีการทางสถิติ เพื่อสร้างข้อมูลสำหรับการสอน โดยมีจุดประสงค์เพื่อให้จำนวนสมาชิกในข้อมูลทั้งสองกลุ่มมีความสมดุลกัน ซึ่งประกอบด้วย 2 วิธีการใหญ่ ๆ คือ Oversampling และ Undersampling

Mahasarakham University

โดยวิธีการ Oversampling จะทำการสุ่มข้อมูลในกลุ่มรองเพื่อสร้างข้อมูลใหม่ของกลุ่มรองให้มีจำนวนเพิ่มมากขึ้น ให้ใกล้เคียงหรือเท่ากับจำนวนข้อมูลในกลุ่มหลัก และในทางตรงข้ามวิธีการ Undersampling จะทำการสุ่มเลือกข้อมูลสำหรับการสอนจากข้อมูลในกลุ่มหลัก ให้ได้จำนวนที่ใกล้เคียงกับจำนวนข้อมูลในกลุ่มรอง

Cost-Sensitive Methods วิธีการนี้จะต่างจากวิธีการแรกที่กล่าวมา โดยวิธีการนี้จะพิจารณาขั้นตอนการเรียนรู้ โดยการสร้างสมมติฐานของการแบ่งกลุ่มข้อมูลที่ไม่สมดุลซึ่งให้ค่าความผิดพลาดจากการสอน (Misclassifying examples) ในการแบ่งกลุ่มข้อมูลให้น้อยที่สุดเป็นวิธีการแก้ปัญหาที่นำทั้งการแก้ปัญหาที่ระดับข้อมูล และระดับอัลกอริทึมมาทำงานร่วมกัน โดยที่ระดับข้อมูลจะทำการเพิ่มค่าน้ำหนัก (Cost) ที่พิเศษสำหรับกรณีที่มีการจำแนกประเภทผิดพลาด และที่ระดับอัลกอริทึมจะทำการปรับการเรียนรู้ของอัลกอริทึมมาตรฐานให้สอดคล้องกับการจำแนกประเภทข้อมูลผิดพลาด

Kernel-based Methods วิธีการนี้เป็นวิธีการใหม่ที่กำลังได้รับความนิยมในการดำเนินการกับกลุ่มข้อมูลที่ไม่สมดุล โดยหลักการแล้วสำหรับวิธีการนี้จะทำการย้ายตำแหน่งของข้อมูล (Map) ที่ไม่สามารถแบ่งกลุ่มได้ในระนาบปกติ โดยการเพิ่มมิติข้อมูลให้สูงขึ้นจนทำให้สามารถแบ่งข้อมูลทั้งสองกลุ่มออกจากกันได้การแก้ปัญหาระดับขั้นตอนวิธีการ เป็นการแก้ปัญหาโดยการปรับการเรียนรู้ของอัลกอริทึมมาตรฐานสำหรับการจำแนกประเภทข้อมูลที่มีอยู่เดิมให้สามารถเรียนรู้ข้อมูลไม่สมดุลโดยให้มีการเอนเอียงไปทางข้อมูลของคลาสกลุ่มน้อย

2.3.6 การวัดประสิทธิภาพของการจำแนก (Measurement)

การวัดประสิทธิภาพของการจัดกลุ่มข้อมูลมีอยู่หลายวิธี แต่มี 2 วิธีที่นิยมตามมาตรฐานของระบบค้นคืนสารสนเทศ [12] ก็คือการใช้การวัดค่าความแม่นยำ (Precision) และค่าความระลึก (Recall)

ค่าความแม่นยำ (Precision: P) เป็นอัตราส่วนของการค้นพบข้อมูลที่ถูกต้องจากจำนวนข้อมูลทั้งหมดที่ทำการค้นหามาได้

$$\text{ค่าความแม่นยำ} = \frac{\text{จำนวนข้อมูลที่ถูกนำมาจัดกลุ่มและถูกต้อง}}{\text{จำนวนข้อมูลทั้งหมดที่จัดอยู่ในกลุ่ม}} \quad (2.21)$$

ค่าความระลึก (Recall: R) เป็นอัตราส่วนของการค้นพบข้อมูลที่ถูกต้องจากจำนวนข้อมูลที่ถูกต้องทั้งหมด

$$\text{ค่าความระลึก} = \frac{\text{จำนวนข้อมูลที่จัดอยู่ในกลุ่ม}}{\text{จำนวนข้อมูลทั้งหมด}} \quad (2.22)$$

Mahasarakham University

โดยทั่วไปแล้วสำหรับฐานข้อมูลสารสนเทศที่มีขนาดใหญ่มาก ๆ มักจะไม่ทราบว่าข้อมูลที่ถูกต้องทั้งหมดมีอยู่เท่าใด ทำให้ต้องทำการประมาณโดยใช้การสุ่มตัวอย่าง (Sampling) ตามหลักทางสถิติหรือด้วยวิธีอื่น ๆ ด้วย โดยทั่วไปจะเป็นการหาค่า F-measure ซึ่งเป็นการวัดค่าความสัมพันธ์ระหว่างค่าความระลึกและค่าความแม่นยำในเชิงฮาร์โมนิค (Harmonic) โดยที่ค่า F-measure จะมีค่าระหว่าง 0 ถึง 1 ซึ่งถ้า F การให้ผลในการจัดกลุ่มข้อมูลมีประสิทธิภาพมากขึ้นเท่านั้นแสดงถึงค่าความแม่นยำ การค้นพบข้อมูลที่ถูกต้องจากจำนวนข้อมูลทั้งหมดที่ทำการค้นหามาได้ (ค่า P) และค่าความระลึกการค้นพบข้อมูลที่ถูกต้องจากจำนวนข้อมูลที่ถูกต้องทั้งหมด (ค่า R) ทั้งสองค่ามีค่ามากเท่าไรจะทำให้ค่าของการวัดประสิทธิภาพของการจัดกลุ่มข้อมูลมากขึ้น ซึ่งแสดงได้ดังสมการ

$$\mathrm{F-measure} = \frac{2 \times \text{ค่าความแม่นยำ} \times \text{ค่าความระลึก}}{\text{ค่าความแม่นยำ} + \text{ค่าความระลึก}} \qquad (2.23)$$

ขั้นตอนการวัดประสิทธิภาพเป็นขั้นตอนของการนำเอากลุ่มของข้อมูลที่จัดได้มาทำการประเมินประสิทธิภาพ โดยจะตรวจสอบคู่ว่ากลุ่มของข้อมูลที่จัดได้มีค่าเป็นอย่างไร เมื่อเทียบกับกลุ่มของข้อมูลที่ถูกต้องซึ่งวัดจากค่าความระลึก (Recall) และค่าความแม่นยำ (Precision) ค่าความแม่นยำจะเป็นค่าที่แสดงว่า การค้นพบข้อมูลได้ตรงกับความต้องการเพียงใด ส่วนค่าความระลึกจะเป็นค่าที่แสดงถึงความครอบคลุมในการจัดกลุ่มข้อมูล ในงานวิจัยนี้ได้จัดประสิทธิภาพการจัดกลุ่มข้อมูลออกเป็น 3 กลุ่มคือ

1) กลุ่มวัดประสิทธิภาพการจัดกลุ่มข้อมูลได้ ดีที่สุด คือกลุ่มที่มีค่าความแม่นยำ และค่าความระลึกสูง แสดงว่าการจัดกลุ่มข้อมูลได้ตรงกับกลุ่มข้อมูลและถูกต้องมากที่สุด

2) กลุ่มวัดประสิทธิภาพการจัดกลุ่มข้อมูลได้ปานกลาง คือกลุ่มที่ค่าความแม่นยำสูงแต่ค่าความระลึกต่ำ แสดงว่าการจัดกลุ่มข้อมูลได้ตรงกับกลุ่มข้อมูลแค่มีข้อมูลบางส่วนมีความคล้ายคลึงกับกลุ่มข้อมูลอื่น

3) กลุ่มวัดประสิทธิภาพการจัดกลุ่มข้อมูลได้ต่ำ คือ กลุ่มที่ค่าความแม่นยำต่ำ แต่ค่าความระลึกสูง แสดงว่าการจัดกลุ่มข้อมูลได้ไม่ตรงกับกลุ่มข้อมูลและมีข้อมูลที่ความคล้ายคลึงกับกลุ่มข้อมูลอื่น เนื่องจากข้อมูลมีการใช้คำสำคัญ ข้อมูลที่ให้ความหมายที่ต่างกัน สมมุติตัวอย่าง ถ้ามีข้อมูล 100 ข้อมูล และมีข้อมูลที่จัดอยู่ในกลุ่มค้นออกมาได้ 60 ข้อมูล ซึ่งเป็นข้อมูลที่เกี่ยวข้องและถูกต้อง 30 ข้อมูล แต่ข้อมูลที่จัดอยู่ในกลุ่มค้นออกมาได้และเป็นข้อมูลที่ถูกต้องมี 20 ข้อมูลสามารถคำนวณค่าความแม่นยำ และค่าความระลึกได้ดังนี้

$$\text{ค่าความแม่นยำ} = \frac{\text{จำนวนข้อมูลที่ลูกนำมาจัดกลุ่มและถูกต้อง}}{\text{จำนวนข้อมูลทั้งหมดที่จัดอยู่ในกลุ่ม}}$$

$$= \frac{20}{60}$$

Mahasarakham University

$$= 0.34 \quad (2.24)$$

$$\text{ค่าความระลึก} = \frac{\text{จำนวนข้อมูลที่จัดอยู่ในกลุ่ม}}{\text{จำนวนข้อมูลทั้งหมด}}$$

$$= \frac{60}{100}$$

$$= 0.60 \quad (2.25)$$

และจาก

$$\text{F} - \text{measure} = \frac{2 \times \text{ค่าความแม่นยำ} \times \text{ค่าความระลึก}}{\text{ค่าความแม่นยำ} + \text{ค่าความระลึก}}$$

$$= \frac{2 \times 0.34 \times 0.6}{0.34 + 0.6}$$

$$= 0.4286 \quad (2.26)$$

นั่นหมายความว่าระบบให้ประสิทธิภาพของการจัดกลุ่มข้อมูลคิดเป็นร้อยละ 42.86 แสดงให้ว่าการวัดประสิทธิภาพจัดอยู่ในกลุ่มวัดประสิทธิภาพการจัดกลุ่มข้อมูลได้ต่ำ

2.4 งานวิจัยที่เกี่ยวข้อง

ปัจจุบันมีงานวิจัยเกี่ยวกับการวินิจฉัยโรคต่างๆ มากมาย ซึ่งแต่ละงานวิจัยล้วนแต่มีเทคนิคที่น่าสนใจ และมีประสิทธิภาพในการวินิจฉัยโรคที่แตกต่างกันออกไป ในงานการศึกษาในครั้ง นี้ได้มีการศึกษาผลงานวิจัยต่างๆ ที่เกี่ยวข้องดังนี้

David และ Magnus [20] ได้ทำการวินิจฉัยโรคพาร์กินสันโดยใช้โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ (Backpropagation) และซัพพอร์ตเวคเตอร์แมชชีน (Support Vector Machines) ในการวินิจฉัยซึ่งประกอบด้วย ข้อมูลเสียงของกลุ่มตัวอย่างจำนวน 31 คน แบ่งเป็นผู้ป่วยโรคพาร์กินสันจำนวน 23 คนและผู้ที่ร่างกายปกติจำนวน 8 คน ผลที่ได้จากการศึกษาคือโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับให้ความถูกต้องในการวินิจฉัยร้อยละ 92.31 ซัพพอร์ตเวคเตอร์แมชีน (linear kernel) ให้ความถูกต้องในการวินิจฉัยร้อยละ 91.79 และซัพพอร์ตเวคเตอร์แมชีน(Universal PearsonVII function based kernel) ให้ความถูกต้องในการวินิจฉัยร้อยละ 93.33 สรุปได้ว่า ความถูกต้องของวิธีการโครงข่ายประสาทเทียม และซัพพอร์ตเวคเตอร์แมชีนค่อนข้างสูง และความถูกต้องของการวินิจฉัยโดยซัพพอร์ตเวคเตอร์แมชชีนขึ้นอยู่กับ kernel function ด้วย

Mahasarakham University

Engin และคณะ [21] ได้ศึกษาการจำแนกประเภทสัญญาณการสั่นของร่างกาย เพื่อการวินิจฉัยทางการแพทย์ โดยใช้วิธีการโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับด้วยการเรียนรู้แบบ scaleconjugate (SCG) และ Brodyen-Fletcher-Goldfarb-Shanno (BFGS) ข้อมูลที่ใช้ในการศึกษาคือ ข้อมูลกระแสไฟฟ้าที่กล้ามเนื้อส่งออกมา(EMG) ของผู้ป่วยโรคพาร์กินสัน โรค Essential Tremor และผู้ที่ร่างกายปกติ ซึ่งโครงสร้างของนิวรอนที่ใช้ในการทดลองคือ ในชั้นอินพุตใช้นิวรอน 15 นิวรอน มีชั้นซ่อน 1 ชั้น ซึ่งมี 2 นิวรอน และชั้นเอาท์พุตมี 2 นิวรอน ใช้ฟังก์ชันซิกมอยด์ (Sigmoid function) เป็นฟังก์ชันกระตุ้น ผลที่ได้จากการศึกษาคือ การเรียนรู้ด้วยขั้นตอนของ BFGS (Brodyen-Fletcher-Goldfarb-Shanno) ให้ความถูกต้องร้อยละ 91.02 มีผลที่ดีกว่าการเรียนรู้ด้วย ขั้นตอนของ scale-conjugate(SCG) ซึ่งให้ความถูกต้องร้อยละ 88.48 ซึ่งผลจากการทดลองพบว่า สามารถนำไปใช้จำแนกประเภทของโรคจากสัญญาณของการสั่นของร่างกายได้

Das [13] ได้ทำการเปรียบเทียบการจัดแบ่งประเภทโรค ด้วยวิธีการต่างๆ สำหรับการวินิจฉัยโรคพาร์กินสัน ซึ่งวิธีการที่ใช้ได้แก่ วิธีการโครงข่ายประสาทเทียม (Neural Network) แบบแพร่ย้อนกลับ DMneural Regression และ Decision Tree ข้อมูลนำเข้าที่ใช้ในการศึกษาคือ เสียงจากการพูดจากผู้ที่ร่างกายปกติจำนวน 8 คนและผู้ป่วยโรคพาร์กินสันจำนวน 23 คน ซึ่งความถูกต้องของข้อมูลในการทดสอบด้วยโครงข่ายประสาทเทียมร้อยละ 92.90 วิธีการ DMneural มีความถูกต้องร้อยละ 84.30 วิธีการ Regression มีความถูกต้องร้อยละ 88.60 วิธีการ Decision Tree มีความถูกต้องร้อยละ 84.30

Ene [22] ได้ศึกษาการจำแนกผู้ป่วยโรคพาร์กินสัน โดยใช้โครงข่ายประสาทเทียมชนิดความน่าจะเป็น (Probabilistic Neural Network , PNN) ซึ่งมีระเบียบวิธีที่ใช้ในการทดลองใช้ประเภทการ search 3 แบบ คือ Incremental search (IS), Monte Carlo search (MCS) และ Hybrid search(HS) ข้อมูลนำเข้าที่ใช้ในการศึกษาคือ เสียงจากการพูดจากผู้ที่ร่างกายปกติ จำนวน 8 คน และผู้ป่วยโรคพาร์กินสันจำนวน 23 คน ซึ่งผลที่ได้จากการคำนวณด้วยวิธีการ ต่างๆ พบว่าการจำแนกด้วยโครงข่ายประสาทเทียมชนิดความน่าจะเป็นด้วย hybrid search(HS) ให้ค่าความถูกต้องของข้อมูลที่ใช้ในการทดสอบร้อยละ 81.28 ซึ่งให้ค่าความถูกต้อง มากกว่าวิธีการอื่นๆ

Mahasarakham University

บทที่ 3

วิธีดำเนินการวิจัย

งานวิจัยนี้มีจุดมุ่งหมายเพื่อพัฒนาระบบและปรับปรุงผลการจำแนกข้อมูลโรคพาร์กินสัน เพื่อใช้ในการคัดกรองผู้ป่วยโรคพาร์กินสัน โดยเน้นการพัฒนาด้วยขั้นตอนวิธีซัพพอร์ตเวคเตอร์แมชชีน รายละเอียดเนื้อหาในบทนี้ประกอบด้วยหัวข้อ 3.1 การศึกษาและรวบรวมข้อมูล ในหัวข้อ 3.2 เป็นการอธิบายการออกแบบสถาปัตยกรรมของระบบ 3.3 เป็นการกลาวถึงโมดูลการสร้างองค์ความรู้ ในหัวขอ 3.4 เป็นการกลาวถึงโมดูลการอนุมานความรู้ และหัวข้อ 3.5 กล่าวถึงรายละเอียดของการทดสอบเปรียบเทียบวิธีการอื่นๆ กับโมเดลที่ได้ออกแบบไว้

3.1 การศึกษาและรวบรวมข้อมูล

ในการดำเนินการวิจัยได้อาศัยหลักการ Knowledge Discovery data[7] ทั้ง 7 ขั้นตอน มาใช้ในการดำเนินงาน ซึ่งประกอบด้วย

1) กำหนดลักษณะของจุดมุ่งหมาย(Goals identification) ประกอบด้วยการตั้งวัตถุประสงค์ ตั้งเกณฑ์วัดความสำเร็จ วางแผนแนวทางการศึกษาวิจัย โดยใช้ข้อมูลจาก UCI Machine Learning Repository และใช้ขั้นตอนวิธีซัพพอร์ตเวคเตอร์แมชชีนเป็นอัลกอริทึมหลักที่ใช้ในการศึกษาและใช้อัลกอริทึมอื่นๆ มาเปรียบเทียบผล

2) การสร้างเซตข้อมูลเป้าหมาย (Creating a target data set) ประกอบด้วยการกำหนดคุณสมบัติของข้อมูล (Define success criteria) อธิบายรายละเอียดของข้อมูล (Describe data) การสำรวจข้อมูล (Explore data) การตรวจสอบความถูกต้องและความสมบูรณ์ของข้อมูล (Verify data quality) เป็นการกำหนดตัวแปรที่จะใช้จากแหล่งข้อมูล โดยแหล่งข้อมูลในการวิจัยครั้งนี้ได้มาจากนำมาจาก UCI Machine Learning Repository เพื่อให้สำหรับสอนระบบและทดลอง

3) การเตรียมข้อมูล (Data preprocessing) เป็นการตรวจสอบข้อมูลให้ถูกต้อง พร้อมทั้งแจ้งเตือนข้อมูลที่ไม่ถูกต้องให้ทราบก่อนที่ข้อมูลจะถูกนำไปใช้ขั้นตอนการเตรียมข้อมูลประกอบด้วยการการคัดเลือกข้อมูลที่จะนำมาใช้ (Select data) การทำความสะอาดข้อมูล (Clean data) ซึ่งเป็นกระบวนการเตรียมข้อมูลให้เหมาะสมที่สุดเพื่อนำไปใช้ในขั้นตอนต่อไป

4) การแปลงข้อมูล (Data transformation) เป็นการทำให้ข้อมูลอยู่รูปแบบตามความจำเป็นต่างๆ การปรับข้อมูลให้อยู่ในรูปที่ง่ายต่อการประมวลผล ในงานวิจัยนี้ได้การปรับเปลี่ยนรูปแบบ

Mahasarakham University

ข้อมูล โดยใช้เทคนิค Min Max normalization ซึ่งเป็นเทคนิคที่ต้องรู้ค่าสูงสุดและค่าต่ำสุดของข้อมูล จะทำให้ข้อมูลที่ได้อยู่ในช่วง 0 และ 1

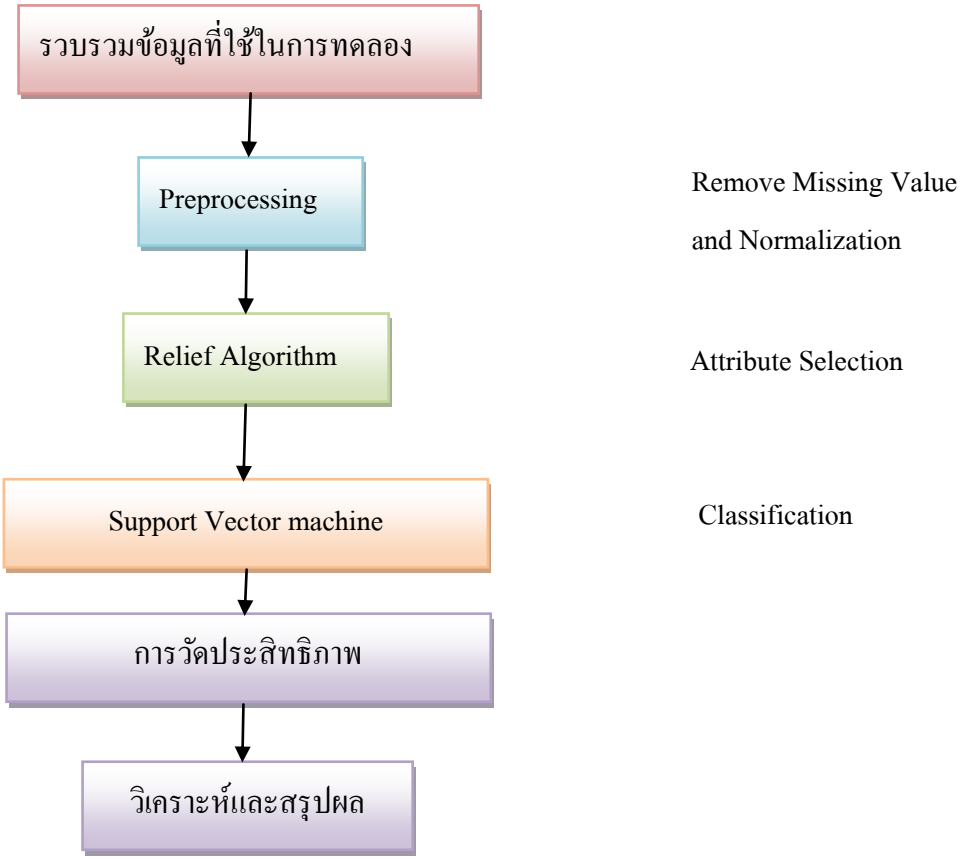
5) การทำเหมืองข้อมูล (Data mining) หรือขั้นตอนการสร้างโมเดลในการทดลอง โดยการเลือกเทคนิคซัพพอร์ตเวคเตอร์แมชชีนในการทำเหมืองข้อมูล และใช้วิธี K-fold Cross validation ในการกำหนดว่าข้อมูลใดเป็นข้อมูลที่สร้างและข้อมูลใดเป็นข้อมูลที่ใช้ในการทดสอบผลลัพธ์ รวมทั้งการวิเคราะห์ข้อมูล กำหนดรูปแบบการทดสอบผลลัพธ์ สร้างโมเดลตามเทคนิคที่เลือก ทดสอบความถูกต้องและความน่าเชื่อถือของโมเดลที่สร้างขึ้น

6) การแปลผลและการประเมินผล (Interpretation and evaluation) ประกอบด้วย การประเมินผลที่ได้จากการทดลอง โดยการประเมินแบบจำลองที่สร้างขึ้นด้วยการลองนำไปใช้กับสถานการณ์จริงเพื่อตรวจสอบประสิทธิภาพของแบบจำลอง การทบทวนกระบวนการ ใช้เป็นขั้นตอนถัดไปในการตัดสินใจ เป็นการพิจารณาผลลัพธ์ จากขั้นตอนที่ 5 ว่าสามารถตอบคำถามหรือแก้ปัญหาจากขั้นตอนที่ 1 หรือไม่ ตัดสินใจว่าจะกระทำขั้นตอนที่ 5 ซ้ำหรือไม่ และรวมถึงการแปลผลไปให้ผู้ใช้ข้อมูลเข้าใจ

7) การนำไปใช้ (Taking action) ประกอบด้วยแผนการในการนำไปใช้ และสรุปผลของการทดลอง เมื่อลงความเห็นว่าจะนำองค์ความรู้ที่ได้ไปใช้ องค์ความรู้นั้นจะถูกรวมเข้ากับระบบที่ใช้อยู่

จากหลักการของ Knowledge Discovery data ทั้ง 7 ขั้นตอน สามารถนำมาเขียนเป็นแผนภาพวิธีดำเนินการวิจัยของระบบวินิจฉัยโรคพาร์กินสันโดยใช้วิธีการซัพพอร์ตเวคเตอร์แมชชีน ได้ดังภาพประกอบ 3.1

Mahasarakham University



ภาพประกอบ 3.1 แนวทางในการดำเนินการ

จากภาพประกอบ 3.1 เมื่อได้กำหนดเป้าหมายและวัตถุประสงค์ จึงทำการรวบรวมข้อมูลจาก UCI Machine Learning Repository นำข้อมูลเข้าสู่กระบวนการเตรียมข้อมูลและแปลงข้อมูล จากนั้นจึงเป็นขั้นตอนการสร้างโมเดล โดยการเลือกขั้นตอนวิธีซัพพอร์ตเวคเตอร์แมชชีน และใช้วิธี K-fold Cross validation ในการวัดประสิทธิ์ภาพโมเดลที่ใช้ในการสอน และวัดประสิทธิ์ภาพความแม่นยำกับข้อมูลทดสอบที่ระบบยังไม่เคยเห็น จากนั้นนำผลที่ได้ไปเปรียบเทียบกับวิธีการอื่นๆ และสรุปผล

เพื่อความสะดวกในการใช้งานระบบที่พัฒนาขึ้นต้องตอบสนองต่อการใช้งาน และสามารถเรียกดูข้อมูลจากหลายๆ แหล่งพร้อมๆ กันได้ ตามแต่ความต้องการของผู้ใช้ ดังนั้นการใช้งานผ่านทางเว็บไซต์จึงเป็นทางเลือกที่เหมาะสมกับงานนี้

Mahasarakham University

ชุดข้อมูลที่ใช้ในการศึกษา (Dataset)

ชุดข้อมูลเสียงจากการพูดของผู้ป่วยโรคพาร์กินสันชุดนี้ได้มาจาก UCI Machine Learning Repository [4] ซึ่งประกอบด้วยกลุ่มตัวอย่างทั้งหมด 31 คน แบ่งเป็นกลุ่มตัวอย่างผู้ป่วยโรคพาร์กินสัน จำนวน 23 คน และกลุ่มของคนสุขภาพดี จำนวน 8 คน โดยแต่ละคนจะถูกแบ่งข้อมูลเป็นช่วงของการวัดเสียงและข้อมูลต่างๆ จากผู้เชี่ยวชาญเฉพาะด้าน และรายการข้อมูลทั้งหมด 195 รายการ ประกอบไปด้วยข้อมูลผู้ป่วยโรคพาร์กินสันจำนวน 48 รายการ และข้อมูลผู้ไม่ป่วยโรคพาร์กินสันจำนวน 147 รายการ ในชุดข้อมูลนี้ประกอบด้วยคุณลักษณะทั้งหมด 23 คุณลักษณะ ซึ่งคุณลักษณะต่างๆ ของข้อมูลชุดนี้ ดังแสดงในตารางที่ 2

ในศึกษานี้จะแบ่งข้อมูลสำหรับใช้สอนระบบร้อยละ 80 เป็นข้อมูลจำนวน 156 รายการ และที่เหลือร้อยละ 20 เป็นข้อมูลสำหรับทดสอบโมเดล โดยในการแบ่งข้อมูลครั้งนี้จะใช้วิธีการสุ่มตัวอย่างแบบชั้นภูมิ (Stratified sampling) เป็นการสุ่มตัวอย่างโดยแยกประชากรออกเป็นกลุ่มประชากรย่อยๆ หรือแบ่งเป็นชั้นภูมิก่อน โดยหน่วยประชากรในแต่ละชั้นภูมิจะมีลักษณะเหมือนกันแล้วสุ่มอย่างง่ายเพื่อให้ได้จำนวนกลุ่มตัวอย่างตามสัดส่วนของขนาดกลุ่มตัวอย่างและกลุ่มประชากร จากการสำรวจชุดข้อมูล พบว่าชุดข้อมูลนี้มีลักษณะเป็นข้อมูลที่ไม่สมดุล (Imbalanced Datasets) ข้อมูลมีลักษณะการกระจายตัวที่ไม่เท่ากัน เนื่องจากมีจำนวนคำตอบในคลาสคำตอบที่ต่างกันมากคือ ข้อมูลผู้ป่วยที่ไม่เป็นโรคพาร์กินสัน (majority class) มีข้อมูล 147 รายการ แต่มีข้อมูลคำตอบผู้ที่เป็นโรคพาร์กินสัน(minority class) เพียง 48 รายการเท่านั้น ดังนั้นถ้าเรานำข้อมูลทั้งสองกลุ่มมาสอนเพื่อแบ่งกลุ่มข้อมูลพร้อมกันทั้งหมด ผลลัพธ์ที่ได้พบว่า ความถูกต้องของการจำแนกประเภทข้อมูลมีความเอนเอียง นั่นคือสามารถจำแนกประเภทข้อมูลกลุ่มที่เป็นข้อมูลที่อยู่ในกลุ่มหลักได้อย่างถูกต้องแม่นยำ ในขณะเดียวกันข้อมูลที่อยู่ในกลุ่มรองจะจำแนกประเภทข้อมูลได้น้อยมากหรือไม่ได้เลย ดังนั้นในการศึกษานี้จะเลือกใช้วิธีการ undersampling โดยการใช้เทคนิคการสุ่มลดแบบ resampling ทำให้ข้อมูลมีจำนวนใกล้เคียงกัน

Mahasarakham University

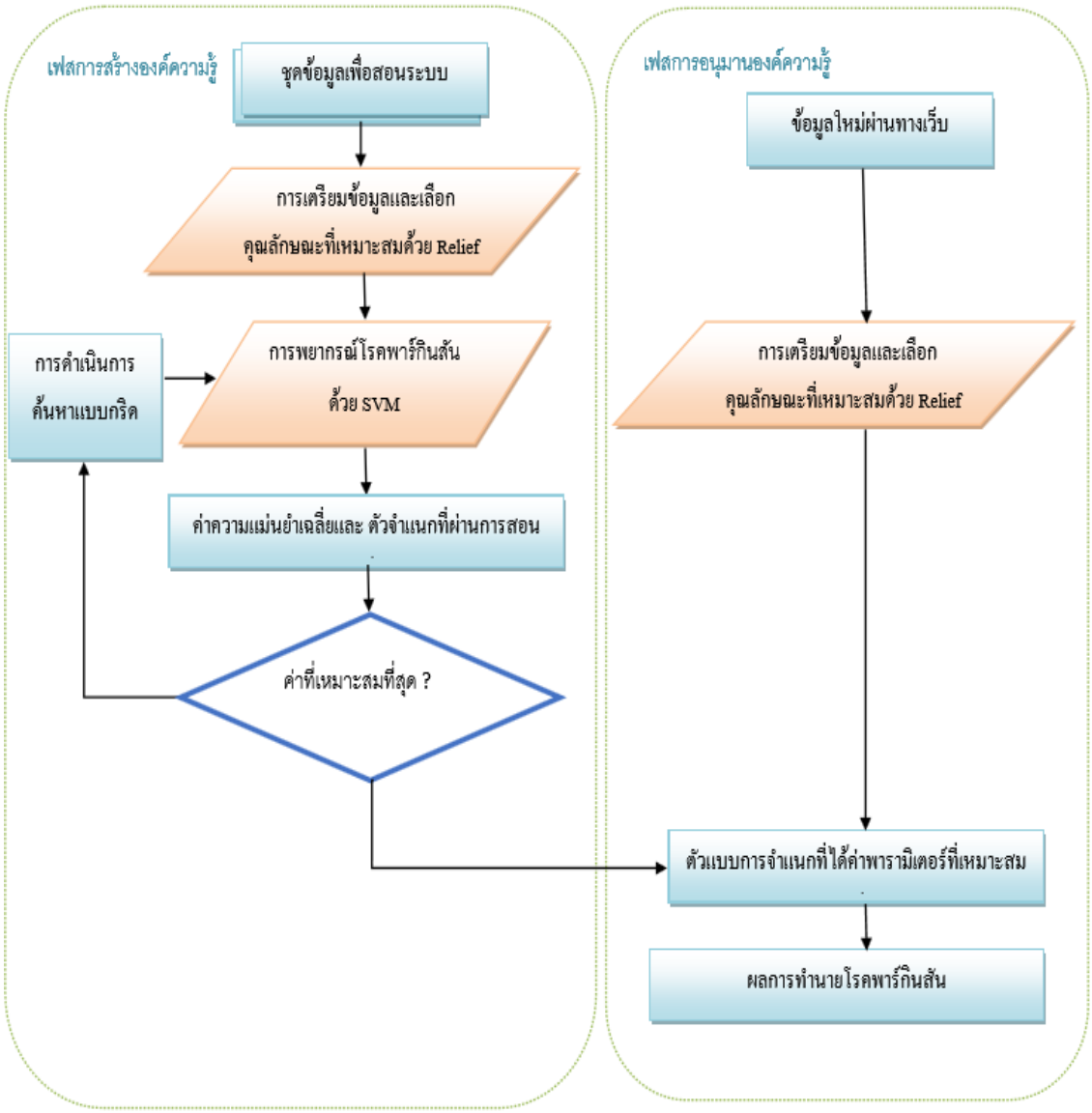
ตาราง 3.1 คุณลักษณะของข้อมูล

คุณลักษณะ	รายละเอียด
Name	- ชื่อและหมายเลขเรคอร์ด
MDVP:Fo(Hz)	- ค่าเฉลี่ยความถี่มูลฐาน
MDVP:Fhi(Hz)	- ค่าสูงสุดความถี่มูลฐาน
MDVP:Flo(Hz)	- ค่าต่ำสุดความถี่มูลฐาน
MDVP:Jitter(%)	- มาตรวัดการแปรผันในความถี่พื้นฐาน
MDVP:Jitter(Abs)	- มาตรวัดการแปรผันในความถี่พื้นฐาน
MDVP:RAP	- มาตรวัดการแปรผันในความถี่พื้นฐาน
MDVP:PPQ	- มาตรวัดการแปรผันในความถี่พื้นฐาน
Jitter:DDP	- มาตรวัดการแปรผันในความถี่พื้นฐาน
MDVP:Shimmer	- ความผันแปรของความดัง
MDVP:Shimmer(dB)	- ความผันแปรของความดัง
Shimmer:APQ3	- มาตรวัดการแปรผันในแอมพลิจูด
Shimmer:APQ5	- มาตรวัดการแปรผันในแอมพลิจูด
MDVP:APQ	- มาตรวัดการแปรผันในแอมพลิจูด
Shimmer:DDA	- มาตรวัดการแปรผันในแอมพลิจูด
NHR (Noise to Harmonics Ratio)	- มาตรวัดอัตราของสัญญาณรบกวนในเสียง
HNR (Harmonics to Noise Ratio)	- มาตรวัดอัตราของสัญญาณรบกวนในเสียง
RPDE	- มาตรวัดความซับซ้อนพลศาสตร์ไม่เป็นเชิงเส้น
D2	- มาตรวัดความซับซ้อนพลศาสตร์ไม่เป็นเชิงเส้น
DFA	- มาตรวัดลักษณะเด่นของสัญญาณ
spread1	- ค่าวัดแปรผันความถี่พื้นฐานไม่เป็นเชิงเส้นการ
spread2	- ค่าวัดแปรผันความถี่พื้นฐานไม่เป็นเชิงเส้นการ
PPE	- ค่าวัดแปรผันความถี่พื้นฐานไม่เป็นเชิงเส้นการ
Status	– สถานะ (1) เป็นโรคพาร์กินสัน (0) ปกติ

Mahasarakham University

3.2 สถาปัตยกรรมของระบบ

ในการออกแบบสถาปัตยกรรมของระบบ งานวิจัยนี้ใช้กลไกการเรียนรู้ของเครื่องจักรด้วยวิธีซัพพอร์ตเวคเตอร์แมชชีนเพื่อการพยากรณ์และการวินิจฉัยโรคพาร์กินสัน ซึ่งสถาปัตยกรรมโดยรวมของระบบแสดงในภาพประกอบ 3.2 จากภาพจะเห็นว่ามีการแบ่งการจัดการองค์ความรู้เป็น 2 โมดูล คือ โมดูลแรกเป็นการสร้างองค์ความรู้ ซึ่งในโมดูลนี้จะเป็นการใช้วิธีการซัพพอร์ตเวคเตอร์แมชชีนหลักเพื่อใช้ในการสร้างองค์ความรู้ไว้ในรูปโมเดลการพยากรณ์ ส่วนโมดูลที่ 2 คือโมดูลที่ใช้ในการอนุมานองค์ความรู้ จะเป็นการนำโมเดลที่ได้ในโมดูลการสร้างความรู้ มาเป็นกลไกในการอนุมานการพยากรณ์โรค



ภาพประกอบ 3.2 สถาปัตยกรรมโดยรวมของระบบ

Mahasarakham University

จากภาพประกอบ 3.2 สถาปัตยกรรมของระบบที่แบ่งสองโมดูล คือการสร้างองค์ความรู้และการอนุมานองค์ความรู้ในขั้นตอนการสร้างองค์ความรู้ จะเป็นการนำชุดข้อมูลมาสอนระบบ ชุดข้อมูลนั้นจะถูกแบ่งเป็น 2 ส่วนคือส่วนแรกจำนวนร้อยละ 80 จะถูกนำมาเป็นข้อมูลที่ใช้ในการสอนให้ระบบ และส่วนที่ 2 จำนวนร้อยละ 20 จะถูกนำมาใช้ในการทดสอบระบบ โดยในการแบ่งข้อมูลนี้จะใช้วิธีการสุ่มตัวอย่างแบบชั้นภูมิ (Stratified sampling) ข้อมูลทั้งหมดจะต้องผ่านกระบวนการกรอง และเตรียมข้อมูลก่อนการประมวลผล เป็นการปรับปรุงคุณภาพของข้อมูล เพื่อลดความความขัดแย้งและความไม่สอดคล้องของข้อมูล จากนั้นจึงเป็นการเข้าสู่กระบวนการคัดเลือกคุณลักษณะที่สำคัญด้วยขั้นตอนวิธีรีลีฟอัลกอริทึม

ในขั้นตอนการอนุมานองค์ความรู้ จะนำข้อมูลทดสอบมาวัดประสิทธิภาพ ก่อนนำไปใช้กับข้อมูลจริง เมื่อมีการนำเข้าข้อมูลใหม่ที่ระบบยังไม่เคยเห็นผ่านทางเว็บเพจ เพื่อให้ระบบทำนายความน่าจะเป็นของโรคพาร์กินสัน โดยใช้ตัวแบบการจำแนกด้วยซัพพอร์ตเวคเตอร์แมชชีนที่ได้ค่าพารามิเตอร์ที่เหมาะสมแล้ว

3.3 โมดูลการสร้างองค์ความรู้ (The knowledge creating module)

ในโมดูลนี้จะเป็นการสร้างโมเดลเพื่อใช้ในการทำนายผู้ป่วยโรคพาร์กินสัน โดยกระบวนการทำงานจะเริ่มจากชุดข้อมูลที่เตรียมไว้ถูกดึงมาในระบบ เพื่อส่งต่อไปยังขั้นตอนการเตรียมข้อมูลก่อนการประมวลผล เช่น การทำความสะอาด การกรองและแปลงข้อมูลให้เป็นมาตรฐานเดียว โดยใช้เทคนิควิธี Min – Max normalization จะทำให้ข้อมูลที่ได้อยู่ในช่วง 0 และ 1 จากนั้นนำข้อมูลที่ผ่านกระบวนการข้างต้นไปเลือกคุณลักษณะด้วยวิธีรีลีฟอัลกอริทึมเพื่อถ่วงน้ำหนักให้แต่ละคุณลักษณะก่อนนำไปจำแนกด้วยวิธีการซัพพอร์ตเวคเตอร์แมชชีน (Support Vector Machine หรือ SVM) การศึกษาของ Huanga พบว่าวิธีการซัพพอร์ตเวคเตอร์แมชชีนให้ผลที่แม่นยำกว่าวิธีโครงข่ายประสาทเทียม อีกทั้งได้มีการนำเทคนิคซัพพอร์ตเวคเตอร์แมชชีน มาใช้จำแนกข้อมูลซึ่งผลที่ได้จากการวิจัยรายงานว่า เทคนิคซัพพอร์ตเวคเตอร์แมชชีน ให้ผลที่แม่นยำกว่าวิธีการจำแนกข้อมูลแบบโครงข่ายประสาทเทียม และการวิเคราะห์จำแนกประเภทหากแต่ปัญหาที่พบในเทคนิคซัพพอร์ตเวคเตอร์แมชชีน คือ การกำหนดค่าพารามิเตอร์หากกำหนดค่าไม่เหมาะสม จะมีผลทำให้โมเดลที่ได้มีประสิทธิภาพไม่ดีนัก อีกทั้งวิธีซัพพอร์ตเวคเตอร์แมชชีน ไม่คงทนต่อข้อมูลรบกวน (Noise) และ ข้อมูลที่มีลักษณะผิดแยกออกมาจากกลุ่ม (Outlier)

จากหลายการศึกษาก่อนนี้ พบว่าหากมีการเลือกคุณลักษณะข้อมูลที่เหมาะสม จะช่วยในการลดผลกระทบของ Noise และ Outlier ในส่วนของข้อมูลสำหรับโมเดลที่เรียนรู้ และยังเป็นการเพิ่มประสิทธิภาพให้กับโมเดลได้อีกด้วย ในโมดูลนี้จึงใช้รีลีฟอัลกอริทึมในการคัดเลือกคุณลักษณะข้อมูลที่เหมาะสมโดยการให้ค่าน้ำหนักของแต่ละแอททริบิวต์ดังแสดงในภาพประกอบ 3.2 ซึ่งมีวิธีการดังที่กล่าวไว้ในบทที่ 2

Mahasarakham University

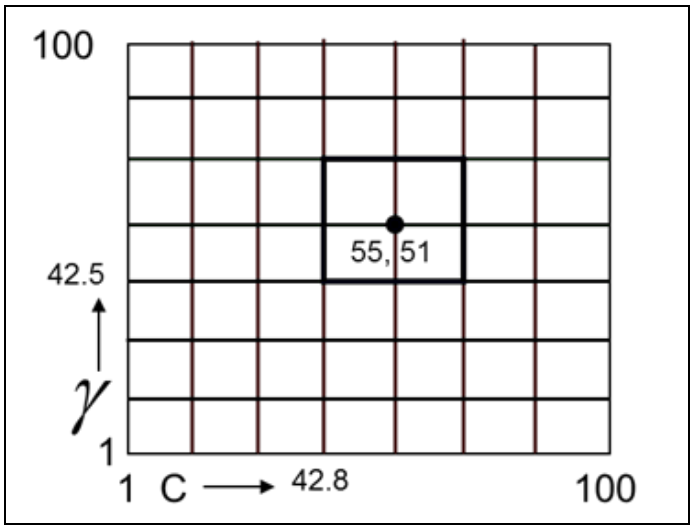
attribute	weight
MDVP:Fo(Hz)	0.312
MDVP:Fhi(Hz)	0.096
MDVP:Flo(Hz)	0.256
MDVP:Jitter(%)	0.088
MDVP:Jitter(Abs)	0.091
MDVP:RAP	0.088
MDVP:PPQ	0.106
Jitter:DDP	0.088
MDVP:Shimmer	0.177
MDVP:Shimmer(dB)	0.150
Shimmer:APQ3	0.167
Shimmer:APQ5	0.162
MDVP:APQ	0.143
Shimmer:DDA	0.167
NHR	0.029
HNR	0.361
RPDE	0.340
DFA	0.205
spread1	0.469
spread2	0.373
D2	0.233
PPE	0.426

ภาพประกอบ 3.3 ตัวอย่างการให้น้ำหนักด้วยขั้นตอนวิธีรีลีฟ

นอกจากนี้ยังมีผู้นำเสนอแนวทางในการกำหนดค่าพารามิเตอร์ที่เหมาะสมสำหรับเทคนิคซัพพอร์ตเวคเตอร์แมชชีน เพื่อปรับปรุงประสิทธิภาพของเทคนิคซัพพอร์ตเวคเตอร์แมชชีน ซึ่งหากกำหนดค่าพารามิเตอร์ที่เหมาะสมให้กับเทคนิคซัพพอร์ตเวคเตอร์แมชชีนจะสามารถทำให้โมเดลที่ได้มีประสิทธิภาพความแม่นยำในการจำแนกข้อมูลที่ดีขึ้น ดังนั้นโมดูลนี้จึงใช้วิธีหาค่าพารามิเตอร์ที่เหมาะสมสำหรับเทคนิคซัพพอร์ตเวคเตอร์แมชชีนด้วยวิธีการค้นหาแบบกริด (Grid Search) เนื่องจากเป็นวิธีการที่เรียบง่ายและมีประสิทธิ์ภาพดี โดยค้นหาแบบกริดใช้เวลาในการค้นหาค่าที่เหมาะสมไม่มากไปกว่าวิธีการอื่นๆ และใช้เวลาในการคำนวณเพื่อหาค่าพารามิเตอร์ได้ดี เนื่องจากมีเพียงสองพารามิเตอร์ นอกจากนี้ยังมีผลในเรื่องทางจิตวิทยา คือเราอาจไม่รู้สึกปลอดภัยที่จะใช้วิธีการในการค้นหาพารามิเตอร์ที่ละเอียดถี่ถ้วน เช่น วิธีทางการประมาณหรือการวิเคราะห์พฤติกรรม [] นอกจากนี้ค้นหาแบบกริดสามารถทำงานคู่ขนานได้อย่างง่ายดายเพราะแต่ละพารามิเตอร์ (C, γ) มีความเป็นอิสระต่อกัน ในขณะที่วิธีการอื่นบางวิธีทำไม่ได้

Mahasarakham University

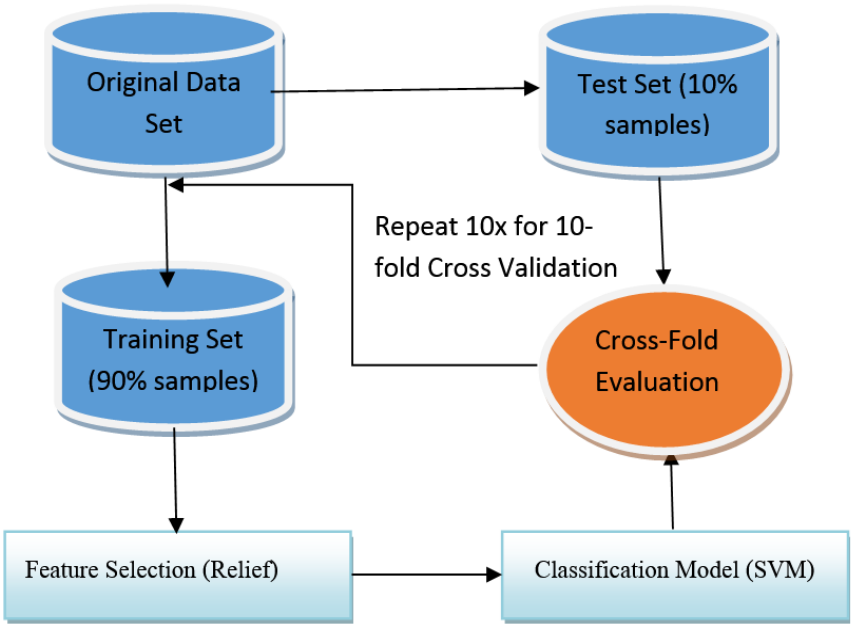
ในการศึกษานี้ในจะใช้กริด (grid) ขนาด 10x11 ทำการหาค่าฟังก์ชันแบบเชิงเส้นที่ทุกๆ จุดตามกริดที่กำหนด และหาค่าช่วงที่จุดต่ำสุดจะถูกบรรจุอยู่ โดยรู้ในเบื้องต้นว่าจุดต่ำสุดต้องอยู่ในช่วง [100,100] อย่างแน่นอน วิธีการนี้จะทำการกำหนดช่วงการค้นหาไปเรื่อยๆ จนกว่าช่วงการค้นหาจะมีค่าน้อยกว่าความคลาดเคลื่อนสูงสุดที่ยอมรับได้ ดังภาพประกอบ 3.3 กำหนดให้แบ่งช่วงการค้นหา [100,100] ออกเป็น n - 1 ช่วงเท่า ๆ กัน จะได้จำนวนจุดทั้งสิ้นที่จุดการค้นหาจะดำเนินการโดยคำนวณค่าฟังก์ชันที่ตำแหน่งกริดทั้ง n จุด หลักการค้นหาจะพิจารณาลดช่วงการค้นหาจากช่วง [a,b] ให้มีขนาดเล็กลง โดยพิจารณาคู่ของจุดที่อยู่ติดกันที่มี โอกาสบรรจุค่าต่ำสุด



ภาพประกอบ 3.4 ตัวอย่างวิธีการค้นหาค่าที่เหมาะสมด้วยกริด

ในงานวิจัยนี้ได้ใช้ 10-Fold Cross-Validation เพื่อเป็นวิธีการตรวจสอบค่าความผิดพลาดในการพยากรณ์ของโมเดล โดยพื้นฐานวิธีการ 10-Fold Cross-validation เป็นวิธีการที่แบ่งข้อมูลออกเป็นกลุ่มจำนวน 10 กลุ่ม) k-Fold) ในตอนแรกเลือกข้อมูลกลุ่มที่ 1 เป็นข้อมูลชุดทดสอบ และข้อมูลชุดที่เหลือจะเป็นข้อมูลชุดสอนนำข้อมูลไปจำแนกข้อมูล จากนั้นจะสลับข้อมูลกลุ่มที่ 2 มาเป็นชุดทดสอบและข้อมูลกลุ่มอื่นๆที่เหลือเป็นชุดทดสอบ สลับอย่างนี้ไปเรื่อย ๆ จนครบ 10 กลุ่ม ในขั้นตอนสุดท้ายจะหาค่าเฉลี่ยของค่าความถูกต้องในแต่ละกลุ่ม ดังแสดงในภาพประกอบ 3.5 วิธีการนี้ข้อมูลทุกตัวอย่างจะได้เป็นทั้งชุดทดสอบและชุดสอน

Mahasarakham University



ภาพประกอบ 3.5 กระบวนการทำงานของ 10-Fold Cross-Validation

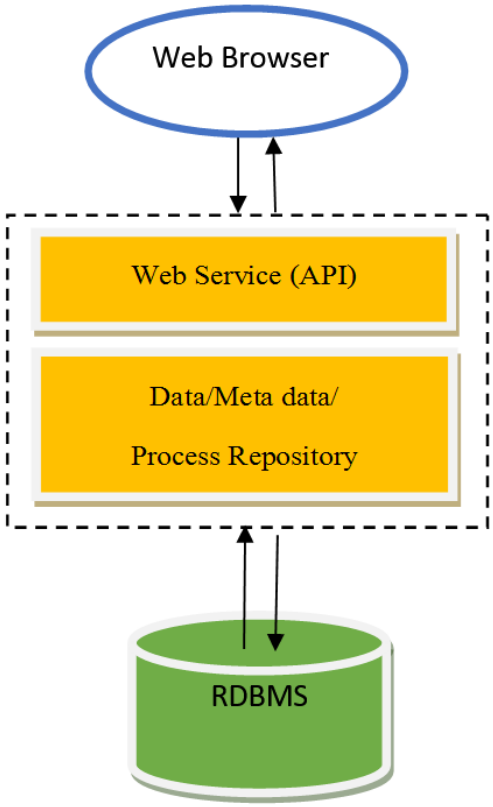
3.4 โมดูลการอนุมานองค์ความรู้ (The knowledge inferring module)

ในโมดูลการอนุมานองค์ความรู้ เป็นการนำโมเดลจากโมดูลการสร้างองค์ความรู้ไปใช้งานจริง โดยใช้ในการทำนายข้อมูลใหม่ โดยในการทดสอบการวัดประสิทธิภาพในการทำนาย จะนำข้อมูลทดสอบจำนวนร้อยละ 20 ที่ระบบไม่เคยเห็นมาใช้ในการทดสอบเพื่อหาความแม่นยำในการทำนาย

สำหรับการใช้งานในระบบเมื่อผู้ใช้นำข้อมูลที่ต้องการให้ระบบทำนายผ่านทางเว็บเพจ ข้อมูลชุดนี้จะถูกส่งผ่านไปยังเว็บเซิร์ฟเวอร์ที่อยู่ในเครือข่ายคอมพิวเตอร์ ที่เว็บเซิร์ฟเวอร์จะเชื่อมต่อกับ Process Repository ที่ใช้ในการเก็บโมเดลสำหรับทำนายผู้ป่วยโรคพาร์กินสันผ่านทางเว็บเซอร์วิส ดังแสดงในภาพประกอบ 3.6

เมื่อข้อมูลใหม่ที่ต้องการให้ระบบทำนายจะถูกส่งต่อไปยังขั้นตอนการเตรียมข้อมูลก่อนการประมวลผล แปลงข้อมูลให้เป็นมาตรฐานเดียว โดยใช้เทคนิควิธี Min – Max normalization จากนั้นข้อมูลที่ผ่านกระบวนข้างต้น จะถูกนำไปเลือกคุณลักษณะด้วยขั้นตอนวิธีรีลีฟอัลกอริทึมเพื่อถ่วงน้ำหนักหาคุณลักษณะที่จำเป็นต้องใช้ในการทำนาย จากนั้นระบบจะทำนายความน่าจะเป็นของโรคพาร์กินสันโดยการใช้โมเดลสำหรับพยากรณ์ด้วยซัพพอร์ตเวคเตอร์แมชชีนที่ได้ค่าพารามิเตอร์ที่เหมาะสมแล้ว ผลการทำนายที่ได้จะส่งกลับไปยังผู้ใช้ผ่านทางเว็บเพจอีกครั้ง

Mahasarakham University



ภาพประกอบ 3.6 กระบวนการวิเคราะห์ข้อมูลผ่านทางเว็บเพจ

เนื้อหาในบทนี้เป็นการแสดงวิธีดำเนินการวิจัย ประกอบไปด้วยขบวนการการศึกษาและรวบรวมข้อมูล โดยแบ่งข้อมูลออกเป็น 2 ส่วน คือข้อมูลสอบและข้อมูลทดสอบในอัตราส่วน 80:20 ในการออกแบบสถาปัตยกรรมของระบบ ประกอบไปด้วย 2 โมดูล คือ 1) โมดูลการสร้างองค์ความรู้ใช้ในการสร้างโมเดลการทำนายโรคพาร์กินสันด้วยวิธีการซัพพอร์ตเวคเตอร์แมชชีนร่วมกับวิธีเลือกคุณลักษณะและการหาค่าที่เหมาะสมให้กับซัพพอร์ตเวคเตอร์แมชชีน โดยการแบ่งข้อมูลที่ใช้ในการสอนและทดสอบด้วยวิธี 10-Fold Cross-Validation และ 2) โมดูลการอนุมานองค์ความรู้ เป็นการนำโมเดลที่สร้างจากโมดูลสร้างองค์ความรู้มาใช้ในการทำนายข้อมูลใหม่ที่ยังไม่เคยเห็นผ่านทางเว็บ

ในการศึกษานี้ได้เน้นศึกษาไปที่การปรับปรุงประสิทธิภาพวิธีการซัพพอร์ตเวคเตอร์แมชชีนและเปรียบเทียบผลการศึกษากับวิธีการอื่นๆ เช่น ต้นไม้ตัดสินใจ และเนอีพเบย์ โดยจะทำการสรุปผลโดยเปรียบเทียบแต่ละเทคนิคว่าประสิทธิภาพที่ได้จะเป็นอย่างไรซึ่งจะแสดงในบทที่ 4

Mahasarakham University

บทที่ 4

ผลการวิจัยและการอภิปรายผล

ในส่วนนี้จะเป็นการนำเสนอผลการวิจัยและเปรียบเทียบประสิทธิภาพกับขั้นตอนวิธีอื่นๆ โดยจะทำการเปรียบเทียบโมเดลที่ศึกษากับขั้นตอนวิธีอื่นๆ สำหรับการทดสอบระบบจะใช้เครื่องคอมพิวเตอร์ Intel Core2duo ความเร็ว 2.9 GHz หน่วยความจำหลัก 8.00 GB ฮาร์ดดิสก์ความจุ 2 TB และทำการทดสอบกับข้อมูลเสียงผู้ป่วยโรคพาร์กินสัน ปรากฏในรายละเอียดหัวข้อที่ 4.1 ผลของการลดคุณลักษณะ 4.2 ผลของการจำแนกข้อมูลด้วยอัลกอริทึมต่างๆ

การทดสอบประสิทธิภาพโมเดลเพื่อการจำแนกข้อมูล ได้ใช้ข้อมูลจาก UCI Machine Learning Repository (http://archive.ics.uci.edu/ml) ที่เป็นข้อมูลสำหรับใช้ในการศึกษาการจำแนกข้อมูล (Classification)

4.1 ผลการทดลองของการลดคุณลักษณะของข้อมูล

ในการกำหนดค่าน้ำหนักให้กับแต่คุณลักษณะจะใช้ขั้นตอนวิธีรีลีฟและขั้นตอนวิธีอื่น ๆ คือ Information Gain และ ขั้นตอนวิธี Chi squared เพื่อเปรียบเทียบกับขั้นตอนวิธีรีลีฟ ในการกำหนดค่าน้ำหนัก โดยจะนำข้อมูลตัวอย่างที่มีส่งให้รีลีฟเรียนรู้ค่าน้ำหนักจากนั้นจะนำคุณลักษณะมาจัดเรียงตามค่าน้ำหนักที่ได้จากขั้นตอนวิธีรีลีฟจากมากไปน้อย เพื่อทำการนอร์มอลไลเซชั่นให้น้ำหนักมีค่าอยู่ระหว่าง 0-1 ตามวิธี Min-Max Normalization ซึ่งได้ผลจากการทดลองดังต่อไปนี้

ตาราง 4.1 น้ำหนักของแต่ละคุณลักษณะด้วยรีลีฟ

คุณลักษณะ	น้ำหนัก
spread1	1.00
PPE	0.82
RPDE	0.81
HNR	0.79
spread2	0.75
DFA	0.58
MDVP:Flo(Hz)	0.58

Mahasarakham University

ตาราง 4.1 น้ำหนักของแต่ละคุณลักษณะด้วยรีลีฟ (ต่อ)

คุณลักษณะ	น้ำหนัก
MDVP:Fo(Hz)	0.54
D2	0.47
Shimmer:APQ3	0.34
Shimmer:DDA	0.34
MDVP:Shimmer	0.27
MDVP:Shimmer(dB)	0.21
Shimmer:APQ5	0.19
MDVP:Jitter(%)	0.19
MDVP:Jitter(Abs)	0.16
MDVP:RAP	0.16
Jitter:DDP	0.16
MDVP:PPQ	0.15
MDVP:APQ	0.10
MDVP:Fhi(Hz)	0.09
NHR	0.00

จากตาราง 4.1 จะพบว่ารีลีฟให้ค่าน้ำหนักกับคุณลักษณะ spread1 มากที่สุด แสดงว่าคุณลักษณะ spread1 มีผลต่อการทำนายมากที่สุด ส่วนคุณลักษณะ NHR มีค่าน้ำหนักน้อยที่สุด หรืออาจกล่าวได้ว่าเป็นคุณลักษณะที่มีผลต่อการทำนายน้อยมาก ในการศึกษาครั้งนี้จะเลือกเฉพาะคุณลักษณะที่ผ่านการนอร์มอลไลเซชันแล้วมีค่าน้ำหนักมากกว่า 0.2 เนื่องจากทำการทดสอบกับโมเดลที่เราสนใจแล้ว ให้ค่าผลลัพธ์ที่ดีที่สุด ทำให้เราได้คุณลักษณะทั้งหมด 13 คุณลักษณะ ประกอบด้วย spread1, PPE, RPDE, HNR, spread2, DFA, MDVP:Flo (Hz), MDVP:Fo (Hz), D2, Shimmer:APQ3, Shimmer:DDA, MDVP:Shimmer และ MDVP:Shimmer (dB)

Mahasarakham University

ตาราง 4.2 น้ำหนักของแต่ละคุณลักษณะด้วย Information Gain

คุณลักษณะ	น้ำหนัก
PPE	1.00
spread1	0.92
MDVP:APQ	0.74
Shimmer:APQ5	0.58
MDVP:Shimmer	0.58
MDVP:Flo(Hz)	0.58
MDVP:Fo(Hz)	0.55
MDVP:RAP	0.53
Jitter:DDP	0.53
MDVP:Shimmer(dB)	0.51
spread2	0.50
MDVP:Jitter(Abs)	0.48
Shimmer:APQ3	0.45
Shimmer:DDA	0.45
MDVP:PPQ	0.39
NHR	0.38
MDVP:Jitter(%)	0.31
MDVP:Fhi(Hz)	0.24
HNR	0.19
RPDE	0.09
D2	0.07
DFA	0.00

จากตาราง 4.2 จะพบว่า Information Gain ให้ค่าน้ำหนักกับคุณลักษณะ PPE มากที่สุด แสดงว่าคุณลักษณะ PPE มีผลต่อการทำนายข้อมูลชุดนี้ด้วยวิธี Information Gain มากที่สุด ส่วนคุณลักษณะ DFA มีค่าน้ำหนักน้อยที่สุด หมายความว่าแทบไม่มีผลต่อการทำนายเลย ในการศึกษาครั้งนี้สนใจเลือกเฉพาะคุณลักษณะที่ผ่านการนอร์มอลไลเซชันแล้วมีค่าน้ำหนักมากกว่า 0.2 เท่านั้น ทำให้เราได้คุณลักษณะทั้งหมด 18 คุณลักษณะ ประกอบด้วยคุณลักษณะ PPE, spread1, MDVP:APQ,

Mahasarakham University

Shimmer:APQ5, MDVP:Shimmer, MDVP:Flo(Hz), MDVP:Fo (Hz), MDVP:RAP, Jitter:DDP, MDVP:Shimmer (dB), spread2, MDVP:Jitter (Abs), Shimmer:APQ3, Shimmer:DDA, MDVP:PPQ, NHR, MDVP:Jitter (%) และ MDVP:Fhi (Hz)

ตาราง 4.3 น้ำหนักของแต่ละคุณลักษณะด้วย chi squared

คุณลักษณะ	น้ำหนัก
PPE	1.00
spread1	0.93
MDVP:Fo(Hz)	0.86
MDVP:Flo(Hz)	0.72
spread2	0.56
MDVP:APQ	0.48
Shimmer:APQ5	0.48
MDVP:Shimmer	0.44
MDVP:Fhi(Hz)	0.41
HNR	0.40
MDVP:Shimmer(dB)	0.39
DFA	0.34
Shimmer:APQ3	0.34
Shimmer:DDA	0.34
RPDE	0.32
D2	0.30
MDVP:PPQ	0.28
MDVP:Jitter(Abs)	0.27
MDVP:Jitter(%)	0.23
MDVP:RAP	0.19
Jitter:DDP	0.19
NHR	0.00

Mahasarakham University

จากตาราง 4.3 จะพบว่า chi squared ให้ค่าน้ำหนักกับคุณลักษณะ PPE มากที่สุด แสดงว่าคุณลักษณะ PPE มีผลต่อการทำนายมากที่สุด ส่วนคุณลักษณะ NHR มีค่าน้ำหนักน้อยที่สุด หรือกล่าวได้ว่าไม่มีผลต่อการทำนายเลย ในการศึกษาครั้งนี้สนใจเลือกเฉพาะคุณลักษณะที่ผ่านการนอร์มอลไลเซชันแล้วมีค่าน้ำหนักมากกว่า 0.2 เท่านั้น ทำให้เราได้คุณลักษณะทั้งหมด 20 คุณลักษณะ ประกอบด้วย PPE, spread1, MDVP:Fo (Hz), MDVP:Flo (Hz), spread2, MDVP:APQ, Shimmer:APQ5, MDVP:Shimmer, MDVP:Fhi (Hz), HNR, MDVP:Shimmer (dB), DFA, Shimmer:APQ3, Shimmer:DDA, RPDE, D2, MDVP:PPQ, MDVP:Jitter (Abs) และ MDVP:Jitter (%) ตามลำดับน้ำหนักจากมากไปน้อย

จากการทดลองการเลือกคุณลักษณะด้วยขั้นตอนวิธี Relief Information Gain และ Chi squared สามารถสรุปจำนวนคุณลักษณะได้ดังตารางที่ 9 จะเห็นได้ว่าขั้นตอนวิธี Relief สามารถลดคุณลักษณะของข้อมูลชุดนี้ได้มากที่สุดที่ 13 คุณลักษณะ รองลงมาคือขั้นตอนวิธี Information Gain และ ขั้นตอนวิธี Chi squared ตามลำดับ

ตาราง 4.4 สรุปผลการลดคุณลักษณะ

ขั้นตอนวิธี	จำนวนคุณลักษณะ
Relief	13
Information Gain	18
Chi squared	20

4.2 ผลของการประเมินประสิทธิภาพการจำแนกประเภทข้อมูลด้วยอัลกอริทึมต่างๆ

เพื่อให้การวัดผลของการประเมินประสิทธิภาพการจำแนกประเภทข้อมูลให้เป็นมาตรฐาน ในการศึกษานี้จึงเลือกใช้ค่าหรือวิธีที่นิยมใช้และผ่านการตีพิมพ์ในระดับสากลมาแล้วเป็นตัววัดผล ซึ่งประกอบไปด้วย

เมตริกซ์วัดประสิทธิภาพ (Confusion Matrix) แสดงผลสรุปการประเมินความสามารถในการจำแนกข้อมูลจากการทดสอบด้วยชุดทดสอบ

Mahasarakham University

ตาราง 4.5 แสดงเมตริกซ์วัดประสิทธิภาพสำหรับการจำแนกประเภทข้อมูล 2 กลุ่ม

True	0	1
0	True Positive (TP)	False Positive (FP)
1	False Negative (FN)	True Negative (TN)

จากตาราง 4.5 แถวของเมตริกซ์จะแสดงจำนวนของตัวอย่างจริงของแต่ละคลาส และคอลัมน์จะแสดงจำนวนที่ทำนายได้ของแต่ละคลาส โดยจะแบ่งออกเป็น 4 กรณี ดังนี้

ค่า TP หรือ True Positive คือ จำนวนข้อมูลที่อยู่ในคลาส Positive แล้วโมเดลทำนายได้ถูกต้องว่าเป็นคลาส Positive

ค่า FN หรือ False Negative คือ จำนวนข้อมูลที่อยู่ในคลาส Positive แล้วโมเดลทำนายผิดว่าเป็นคลาส Negative

ค่า FP หรือ False Positiveคือ จำนวนข้อมูลที่อยู่ในคลาส Negative แล้วโมเดลทำนายผิดว่าเป็นคลาส Positive

ค่า TNหรือ True Negativeคือ จำนวนข้อมูลที่อยู่ในคลาส Negativeแล้วโมเดลทำนายได้ถูกต้องว่าเป็นคลาส Negative

ความแม่น (Accuracy) เป็นการประเมินประสิทธิภาพการจำแนกประเภทข้อมูลโดยรวมทุกคลาสของแบบจำลอง เป็นตัวบ่งชี้ว่าผลการทดสอบมีค่าเข้าใกล้ค่าจริงหรือค่าอ้างอิงหรือค่าที่ยอมรับ เนื่องจากในทางปฏิบัติยากที่จะทราบค่าจริง จึงใช้วิธีเปรียบเทียบกับค่าที่ยอมรับแทน เป็นคุณลักษณะที่แสดงถึงความสอดคล้องกับค่าจริง

$$Accuracy = \frac{TP + TN}{(TP + FN + TN + FP)} \quad (4.1)$$

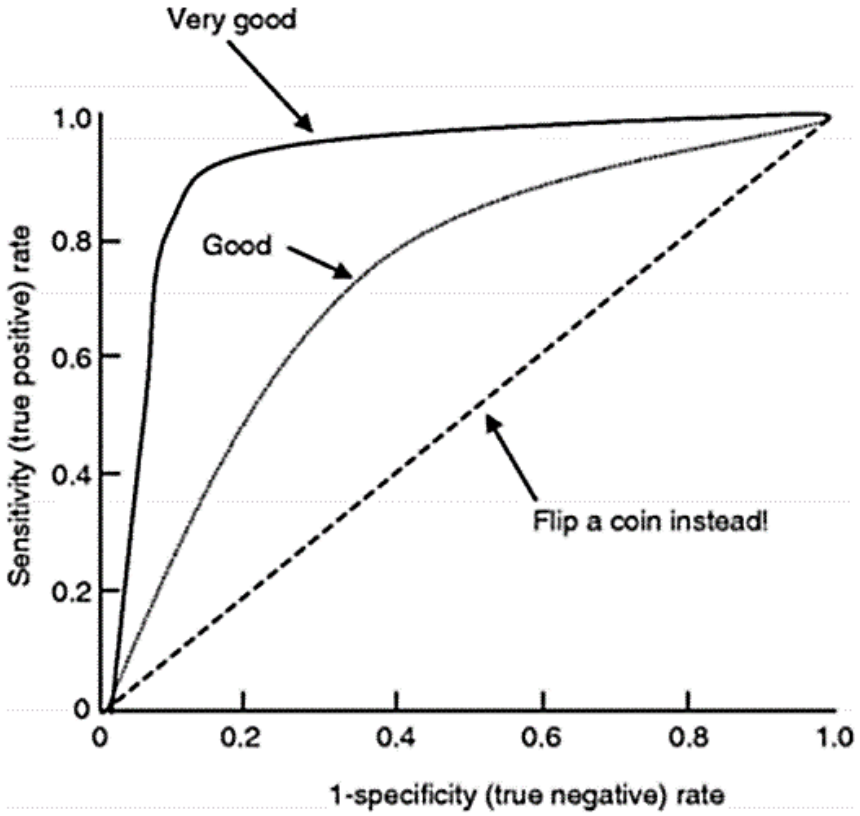
Classification error เป็นค่าที่ใช้ประเมินประสิทธิภาพการจำแนกประเภทข้อมูลที่ทำนายไม่ถูกในอัตราร้อยละ มีสมการในการหาค่า Classification error ดังนี้

Classification error = 100 – Accuracy (4.2)

Mahasarakham University

Roc curve คือกราฟความสัมพันธ์ระหว่าง true positive rate (Sensitivity) กับ false positive rate เครื่องมือในการตรวจวินิจฉัยควรมี Sensitivity สูง และ มี Specificity สูง ซึ่งประการหลังจะทำให้มี false positive rate ต่ำ ส่งผลให้ ROC curve เข้าชิดมุมซ้ายบนมากที่สุด

นอกจากนี้การสร้าง ROC curve ยังช่วยในการเปรียบเทียบประสิทธิภาพของการทำนายได้ด้วยโดยเปรียบเทียบพื้นที่ใต้เส้นโค้ง(AUC) ของการทำนายแต่ละชนิด พื้นที่ใต้ โค้งที่มากกว่าแสดงถึงประสิทธิภาพที่สูงกว่า



ภาพประกอบ 4.1 Roc curve

AUC คือ พื้นที่ใต้เส้นโค้ง ROC (Receiver Operating Characteristic curve) โดยที่เส้นโค้งROC จะเป็นเส้นกราฟที่พล็อตระหว่างค่า Sensitivityซึ่งเป็นค่าที่ทำนายได้ถูกต้องของการเกิดเหตุการณ์ที่สนใจซึ่งจะแทนด้วยแกน y และค่า 1- Specificity หรือค่าที่ทำนายผิดพลาดของการเกิดเหตุการณ์ที่สนใจซึ่งแทนด้วยแกน x ดังรูปที่ 16 โดยที่ AUC จะแสดงให้เห็นถึงความสามารถในการจำแนกกลุ่มของเหตุการณ์ที่สนใจออกจากกลุ่มของเหตุการณ์ที่ไม่สนใจ ค่า AUC สามารถคำนวณหาได้ดังสมการ

Mahasarakham University

$$AUC = \frac{1+TPrate-FPrate}{2} \quad (4.3)$$

ความเที่ยง (Precision) เป็นค่าที่ใช้วัดความใกล้เคียงของกลุ่มที่ทำการวัด นิยมใช้และแสดงความหมายใกล้เคียงกับความถูกต้องแม่นยำ (accuracy) ซึ่งในความเป็นจริงแล้วความเที่ยงตรงมีความหมายที่แตกต่างจากความแม่นยำ โดยความเที่ยงตรงเป็นค่าที่แสดงถึงความสามารถของเครื่องมือวัดในการแสดงค่าเดิมเมื่อทำการวัดหลาย ๆ ครั้ง ของการทำนายข้อมูลที่อยู่ในคลาส Positive โดยหาจากอัตราส่วนของการทำนายข้อมูลที่อยู่ในคลาส Positive ได้ถูกต้องเทียบกับจำนวนข้อมูลที่ทำนายว่าเป็นคลาส Positive ทั้งหมดดังสมการ

$$Precision = \frac{TP}{(TP+FP)} \quad (4.4)$$

ค่าระลึก (Recall) บางครั้งจะเรียกว่า True Positive rate (TPrate) หรือค่า Sensitive จะเป็นการวัดความสามารถในการค้นหาข้อมูลที่อยู่ในคลาส Positive โดยหาจากอัตราส่วนของการทำนายข้อมูลที่อยู่ในคลาส Positiveได้ถูกต้องเทียบกับข้อมูลจริงทั้งหมดของคลาส Positive ดังสมการ

$$Recall = \frac{TP}{(TP+FN)} \quad (4.5)$$

F_measure เป็นการวัดความแม่นยำโดยดูจากผลเฉลี่ยของ Precision และ Recall ซึ่งสามารถหาได้จากสมการ

$$F\_measure = \frac{(2\ x\ Precision\ x\ Recall)}{(Precision+\ Recall\ )} \quad (4.6)$$

4.2.1 ผลการทดลองด้วยขั้นตอนวิธี Relief + SVM + Grid ในส่วนข้อมูลการสอน

Accuracy: 100.00% +/- 0.00% จากการทดลองพบว่าการใช้ขั้นตอนวิธีรีลีฟร่วมกับซัพพอร์ตเวคเตอร์แมชชีนและการค้นหาแบบกริดในส่วนการใช้ชุดข้อมูลการสอนให้ค่าความถูกต้องเท่ากับร้อยละ 100 และค่าส่วนเบี่ยงเบนมาตรฐานร้อยละ 0

Confusion Matrix:

True:	1	0
1:	38	0
0:	0	38

สามารถคำนวณหาค่า accuracy ได้จากสมการ

Mahasarakham University

$$Accuracy = \frac{TP + TN}{(TP + FN + TN + FP)}$$
$$= \frac{38 + 38}{(38 + 0 + 38 + 0)}$$
$$= 1\ x\ 100$$
$$= 100\%$$

(4.7)

classification_error: 0.00% +/- 0.00%

Confusion Matrix:

True:	1	0
1:	38	0
0:	0	38

สามารถคำนวณหาค่า accuracy ได้จากสมการ

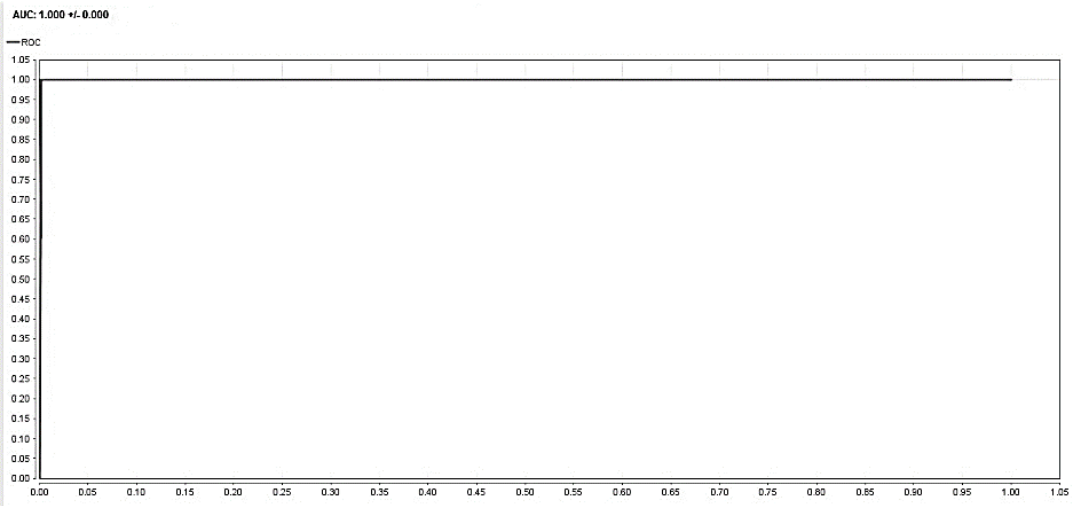
Classification_error = 100 – accuracy

= 100 - 100

= 0 %

(4.8)

AUC: 1.000 +/- 0.000



ภาพประกอบ 4.2 ค่าพื้นที่ใต้กราฟของขั้นตอนวิธี Relief + SVM + Grid ส่วนข้อมูลสอน

จากภาพประกอบ 4.2 พบว่าค่าพื้นที่ใต้กราฟเท่ากับ 1 แสดงให้เห็นถึงประสิทธิภาพในการจำแนกกลุ่มข้อมูลของผู้ที่ป่วยเป็นโรคพาร์กินสันและผู้ที่ไม่ได้เป็นผู้ป่วยเป็นโรคพาร์กินสันออกจากกันได้ และแสดงให้เห็นว่าขั้นตอนวิธีนี้นี้เป็นเทคนิคที่มีประสิทธิภาพสูงในการคัดกรองผู้ป่วยโรคพาร์กินสัน

Mahasarakham University

Precision: 100.00% +/- 0.00%

Confusion Matrix:

True: 1 0

1: 38 0

0: 0 38

สามารถคำนวณหาค่า precision ได้จากสมการ

$$Precision = \frac{TP}{(TP + FP)} = \frac{38}{(38 + 0)} = 1\ x\ 100 = 100\% \quad (4.9)$$

Recall: 100.00% +/- 0.00%

Confusion Matrix:

True: 1 0

1: 38 0

0: 0 38

สามารถคำนวณหาค่า Recall ได้จากสมการ

$$Recall = \frac{TP}{(TP + FN)} = \frac{38}{(38 + 0)} = 1\ x\ 100 = 100\% \quad (4.10)$$

F_measure: 100.00% +/- 0.00%

Confusion Matrix:

True: 1 0

1: 38 0

0: 0 38

สามารถคำนวณหาค่า Recall ได้จากสมการ

Mahasarakham University

$$F\_measure = \frac{(2\ x\ Precision\ x\ Recall)}{(Precision + Recall\ )}$$
$$= \frac{(2\ x\ 1\ x\ 1)}{(1 + 1\ )}$$
$$= 1\ x\ 100$$
$$= 100\%$$

(4.11)

4.2.2 ผลการทดลองขั้นตอนวิธี Relief + SVM + Grid ส่วนข้อมูลทดสอบ

Accuracy: 89.74%

Confusion Matrix:

True: 1 0

1: 25 0

0: 4 10

สามารถคำนวณหาค่า accuracy ได้จากสมการ

$$Accuracy = \frac{TP + TN}{(TP + FN + TN + FP)}$$
$$= \frac{25 + 10}{(25 + 4 + 10 + 0)}$$
$$= 0.8205\ x\ 100$$
$$= 89.74\%$$

(4.12)

classification_error: 17.95%

Confusion Matrix:

True: 1 0

1: 25 0

0: 4 10

สามารถคำนวณหาค่า accuracy ได้จากสมการ

classification_error = 100 – accuracy

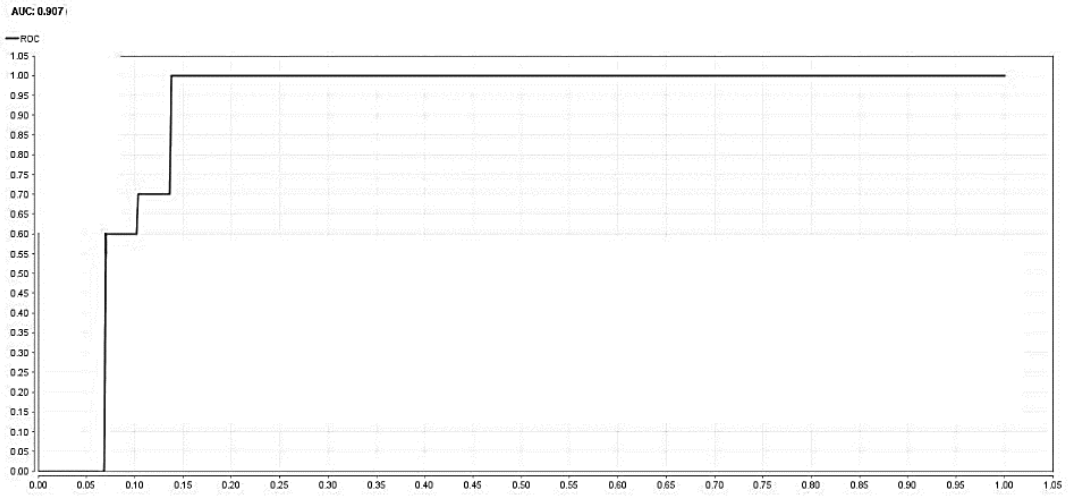
= 100 - 89.74

= 10.25 %

(4.13)

Mahasarakham University

AUC: 0.962



ภาพประกอบ 4.3 ค่าพื้นที่ใต้กราฟของขั้นตอนวิธี Relief + SVM + Grid ส่วนข้อมูลทดสอบ

จากภาพประกอบ 4.3 พบว่าค่าพื้นที่ใต้กราฟเท่ากับ 0.962 ซึ่งเข้าใกล้ 1 มาก แสดงว่าขั้นตอนวิธีนี้นี้เป็นเทคนิคที่มีประสิทธิภาพสูงในการคัดกรองผู้ป่วยโรคพาร์กินสัน

Precision: 100%

Confusion Matrix:

True: 1 0

1: 25 0

0: 4 10

สามารถคำนวณหาค่า precision ได้จากสมการ

$$Precision = \frac{TP}{(TP + FP)}$$

$$= \frac{25}{(25 + 0)}$$

$$= 1 \; x \; 100$$

$$= 100\%$$

(4.14)

Mahasarakham University

Recall: 86.20%

ConfusionMatrix:

True: 1 0

1: 25 0

0: 4 10

สามารถคำนวณหาค่า Recall ได้จากสมการ

$$Recall = \frac{TP}{(TP + FN)}$$
$$= \frac{25}{(25 + 4)}$$
$$= 0.8620\ x\ 100$$
$$= 86.20\%$$

(4.15)

f_measure: 92.59%

สามารถคำนวณหาค่า Recall ได้จากสมการ

$$F\_measure = \frac{(2\ x\ Precision\ x\ Recall)}{(Precision + Recall\ )}$$
$$= \frac{(2\ x\ 1\ x\ 0.8620)}{(1 + 0.8620\ )}$$
$$= 0.9259\ x\ 100$$
$$= 92.59\%$$

(4.16)

Mahasarakham University

4.2.3 ผลการทดลองขั้นตอนวิธีต้นไม้การตัดสินใจข้อมูลสอน

Accuracy: 80.12% +/- 10.19%

Confusion Matrix:

True: 1 0

1: 108 21

0: 10 17

สามารถคำนวณหาค่า accuracy ได้จากสมการ

$$Accuracy = \frac{TP + TN}{(TP + FN + TN + FP)}$$
$$= \frac{108 + 17}{(108 + 21 + 17 + 10)}$$
$$= 0.8012 * 100$$
$$= 80.12\%$$

(4.17)

classification_error: 19.87%

Confusion Matrix:

True: 1 0

1: 108 21

0: 10 17

สามารถคำนวณหาค่า accuracy ได้จากสมการ

classification_error = 100 – accuracy

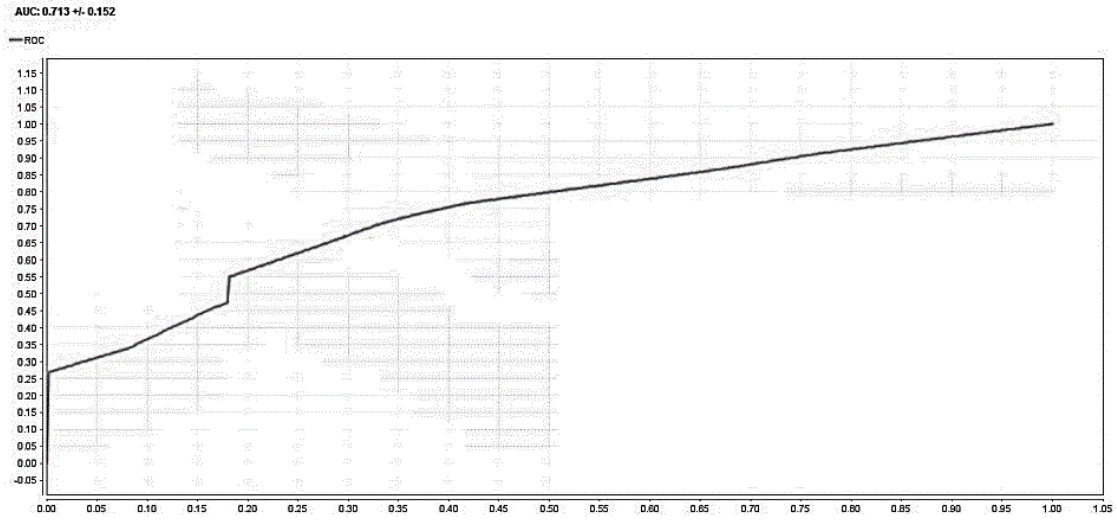
= 100 - 80.12

= 19.87%

(4.18)

Mahasarakham University

AUC: 0.713 +/- 0.152



ภาพประกอบ 4.4 ค่าพื้นที่ใต้กราฟของขั้นตอนวิธีต้นไม้การตัดสินใจส่วนข้อมูลสอน

จากภาพประกอบ 4.4 พบว่าค่าพื้นที่ใต้กราฟยังห่างจาก 1 พอสมควร แสดงว่าขั้นตอนวิธีนี้นี้เป็นเทคนิคที่ไม่ค่อยมีประสิทธิภาพในการคัดกรองผู้ป่วยโรคพาร์กินสัน

Precision: 83.72%

Confusion Matrix:

True:	1	0
1:	108	21
0:	10	17

สามารถคำนวณหาค่า precision ได้จากสมการ

$$\begin{aligned} Precision &= \frac{TP}{(TP+FP)} \\ &= \frac{108}{(108+21)} \\ &= 0.8372 \ x\, 100 \\ &= 83.72\ \% \end{aligned}$$

(4.19)

Mahasarakham University

Recall: 91.53%

Confusion Matrix:

True: 1 0

1: 108 21

0: 10 17

สามารถคำนวณหาค่า Recall ได้จากสมการ

$$Recall = \frac{TP}{(TP + FN)}$$
$$= \frac{108}{(108 + 10)}$$
$$= 0.9153 \ x \ 100$$
$$= 91.53\ \%$$

(4.20)

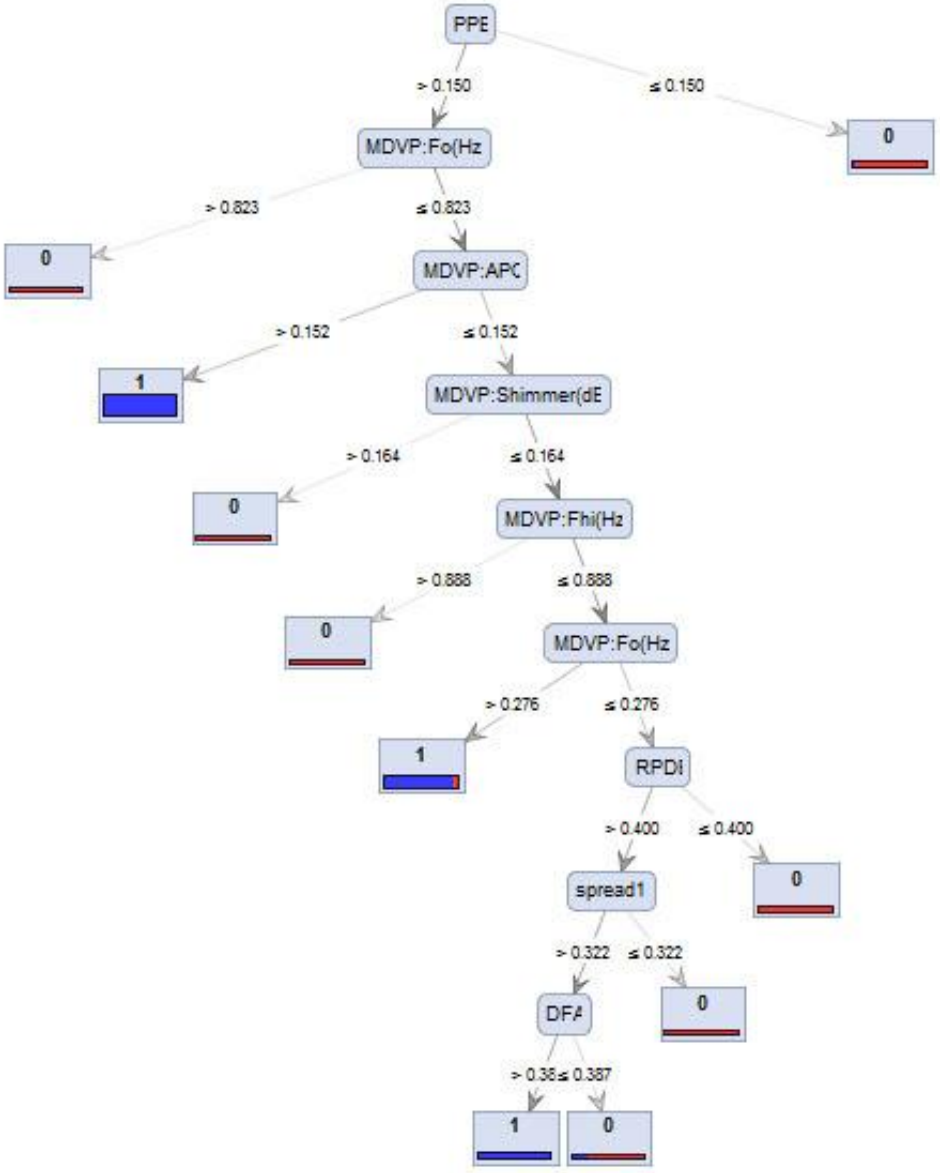
f_measure: 87.45 %

สามารถคำนวณหาค่า Recall ได้จากสมการ

$$F\_measure = \frac{(2\ x\ Precision\ x\ Recall)}{(Precision + \ Recall\ )}$$
$$= \frac{(2\ x\ 0.8372\ \ x\ 0.9153)}{(0.8372 + \ 0.9153\ )}$$
$$= \ 0.8745\ x\ 100$$
$$= \ 87.45\%$$

(4.21)

Mahasarakham University



ภาพประกอบ 4.5 โมเดลต้นไม้การตัดสินใจ

จากภาพประกอบ 4.5 เป็นการสร้างโมเดล decision tree จะทำการคัดเลือกคุณลักษณะที่มีความสัมพันธ์กับคลาสคำตอบมากที่สุดขึ้นมาเป็นโหนดบนสุดของ tree (root node) ในที่นี้คือคุณลักษณะ PPEหลังจากนั้นก็จะหาคุณลักษณะถัดไปเรื่อยๆ ในการหาความสัมพันธ์ของคุณลักษณะนี้จะใช้ตัววัด ที่เรียกว่า Information Gain (IG) จากโมเดลนี้สามารถนำไปสร้างเป็นกฎได้ดังนี้

Mahasarakham University

Tree

PPE > 0.150

| MDVP:Fo(Hz) > 0.823: 0 {1=0, 0=2}

| MDVP:Fo(Hz) ≤ 0.823

| | MDVP:APQ > 0.152: 1 {1=67, 0=0}

| | MDVP:APQ ≤ 0.152

| | | MDVP:Shimmer(dB) > 0.164: 0 {1=0, 0=3}

| | | MDVP:Shimmer(dB) ≤ 0.164

| | | | MDVP:Fhi(Hz) > 0.888: 0 {1=0, 0=2}

| | | | MDVP:Fhi(Hz) ≤ 0.888

| | | | | MDVP:Fo(Hz) > 0.276: 1 {1=33, 0=2}

| | | | | MDVP:Fo(Hz) ≤ 0.276

| | | | | | RPDE > 0.400

| | | | | | | spread1 > 0.322

| | | | | | | | DFA > 0.387: 1 {1=16, 0=0}

| | | | | | | | DFA ≤ 0.387: 0 {1=1, 0=3}

| | | | | | | spread1 ≤ 0.322: 0 {1=0, 0=3}

| | | | | | RPDE ≤ 0.400: 0 {1=0, 0=6}

PPE ≤ 0.150: 0 {1=1, 0=17}

4.2.4 ผลการทดลองขั้นตอนวิธีต้นไม้การตัดสินใจ ในส่วนข้อมูลทดสอบ

Accuracy: 69.23%

Confusion Matrix:

True:	1	0
1:	17	0
0:	12	10

สามารถคำนวณหาค่า accuracy ได้จากสมการ

$$Accuracy = \frac{TP + TN}{(TP + FN + TN + FP)} = \frac{17 + 10}{(17 + 0 + 10 + 0)} = 0.6923\ x\ 100 = 69.23\%$$

(4.22)

Mahasarakham University

classification_error: 30.77%

Confusion Matrix:

True: 1 0

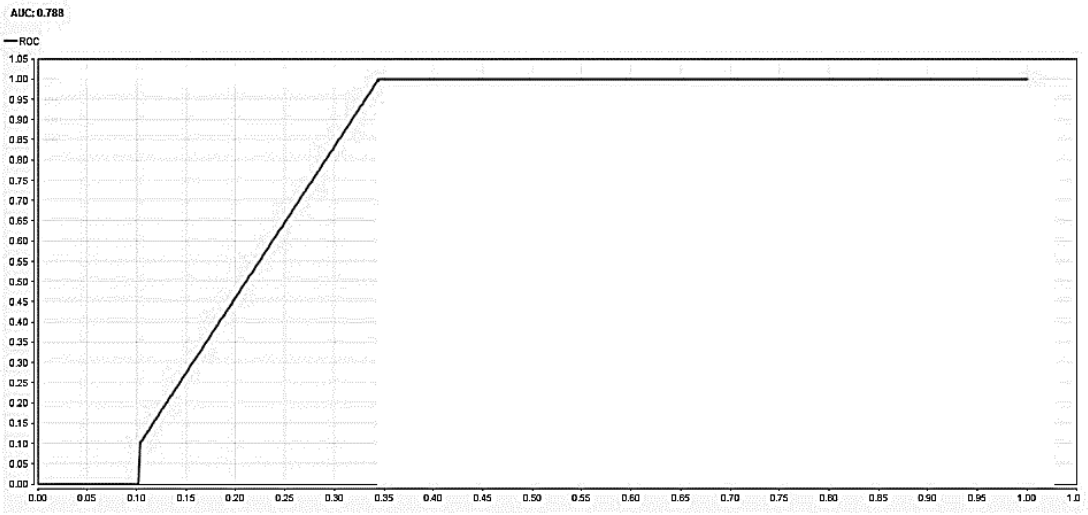
1: 17 0

0: 12 10

สามารถคำนวณหาค่า accuracy ได้จากสมการ

$$\text{classification\_error} = 100 - \text{accuracy}$$
$$= 100 - 69.23$$
$$= 30.77\% \quad (4.23)$$

AUC: 0.788



ภาพประกอบ 4.6 ค่าพื้นที่ใต้กราฟของขั้นตอนวิธีต้นไม้การตัดสินใจส่วนข้อมูลทดสอบ

จากภาพประกอบ 4.6 พบว่าค่าพื้นที่ใต้กราฟเข้าใกล้ 1 แสดงว่าขั้นตอนวิธีนี้นี้เป็นเทคนิคที่มีประสิทธิภาพในการคัดกรองผู้ป่วยโรคพาร์กินสันพอสมควร

Precision: 100%

ConfusionMatrix:

True: 1 0

1: 17 0

0: 12 10

Mahasarakham University

สามารถคำนวณหาค่า precision ได้จากสมการ

$$\begin{aligned} Precision &= \frac{TP}{(TP+FP)} \\ &= \frac{17}{(17+0)} \\ &= 1\ x\ 100 \\ &= 100\ \% \end{aligned} \tag{4.24}$$

Recall: 58.62%

Confusion Matrix:

True: 1 0

1: 17 0

0: 12 10

สามารถคำนวณหาค่า Recall ได้จากสมการ

$$\begin{aligned} Recall &= \frac{TP}{(TP+FN)} \\ &= \frac{17}{(17+12)} \\ &= 0.5862\ x\ 100 \\ &= 58.62\ \% \end{aligned} \tag{4.25}$$

f_measure: 73.91 %

สามารถคำนวณหาค่า Recall ได้จากสมการ

$$\begin{aligned} F\_measure &= \frac{(2\ x\ Precision\ x\ Recall)}{(Precision + Recall\ )} \\ &= \frac{(2\ x\ 1\ x\ 0.5862)}{(1 + 0.5862\ )} \\ &= 0.7391\ x\ 100 \\ &= 73.91\% \end{aligned} \tag{4.26}$$

Mahasarakham University

4.2.5 ผลการทดลองขั้นตอนนาอีพเบย์ในส่วนข้อมูลการสอน

Accuracy: 70.51% +/- 10.55%

Confusion Matrix:

True: 1 0

1: 76 4

0: 42 34

สามารถคำนวณหาค่า accuracy ได้จากสมการ

$$Accuracy = \frac{TP + TN}{(TP + FN + TN + FP)}$$
$$= \frac{76 + 34}{(76 + 4 + 34 + 42)}$$
$$= 0.7051\ x\ 100$$
$$= 70.51\%$$

(4.27)

classification_error: 29.49% +/- 10.55%

Confusion Matrix:

True: 1 0

1: 76 4

0: 42 34

สามารถคำนวณหาค่า accuracy ได้จากสมการ

classification_error = 100 – accuracy

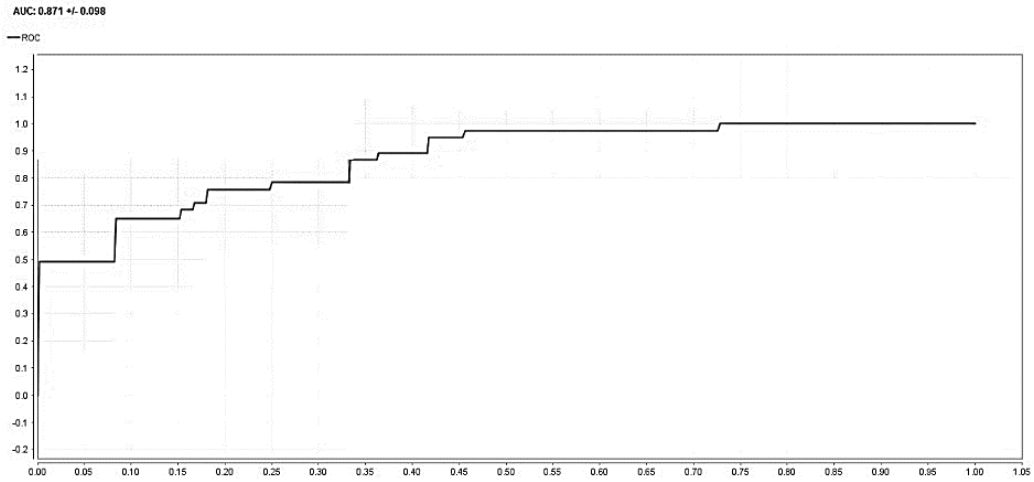
= 100 - 70.51

= 29.49 %

(4.28)

Mahasarakham University

AUC: 0.871 +/- 0.098



ภาพประกอบ 4.7 ค่าพื้นที่ใต้กราฟของขั้นตอนวิธีนาอีพเบย์ส่วนข้อมูลสอน

จากภาพประกอบ 4.7 พบว่าค่าพื้นที่ใต้กราฟเข้าใกล้ 1 แสดงว่าขั้นตอนวิธีนี้นี้เป็นเทคนิคที่มีประสิทธิภาพดีในการคัดกรองผู้ป่วยโรคพาร์กินสัน

Precision: 95.00%

Confusion Matrix:

True:	1	0
1:	76	4
0:	42	34

สามารถคำนวณหาค่า precision ได้จากสมการ

$$Precision = \frac{TP}{(TP+FP)}$$
$$= \frac{76}{(76+4)}$$
$$= 0.95 \ x \ 100$$
$$= 95\ \%$$

(4.29)

Mahasarakham University

Recall: 64.40%

Confusion Matrix:

True: 1 0

1: 76 4

0: 42 34

สามารถคำนวณหาค่า Recall ได้จากสมการ

$$\begin{aligned} Recall &= \frac{TP}{(TP+FN)} \\ &= \frac{76}{(76+42)} \\ &= 0.6440 \ x\ 100 \\ &= 64.41\ \% \end{aligned}$$

(4.30)

f_measure: 76.77%

Confusion Matrix:

True: 1 0

1: 76 4

0: 42 34

สามารถคำนวณหาค่า Recall ได้จากสมการ

$$\begin{aligned} F\_measure &= \frac{(2\ x\ Precision\ x\ Recall)}{(Precision + \ Recall\ )} \\ &= \frac{(2\ x\ 0.9500\ \ x\ 0.6441)}{(0.9500 + \ 0.6441\ )} \\ &= \ 0.7677\ x\ 100 \\ &= \ 76.77\% \end{aligned}$$

(4.31)

Mahasarakham University

4.2.6 ผลการทดลองขั้นตอนวิธีนาอีพเบย์ ในส่วนข้อมูลทดสอบ

Accuracy: 64.10%

Confusion Matrix:

True: 1 0

1: 17 2

0: 12 8

สามารถคำนวณหาค่า accuracy ได้จากสมการ

$$Accuracy = \frac{TP + TN}{(TP + FN + TN + FP)}$$

$$= \frac{17 + 8}{(17 + 2 + 8 + 12)}$$

$$= 0.6410\ x\ 100$$

$$= 64.10\%$$

(4.32)

classification_error: 35.90%

Confusion Matrix:

True: 1 0

1: 17 2

0: 12 8

สามารถคำนวณหาค่า accuracy ได้จากสมการ

classification_error = 100 – accuracy

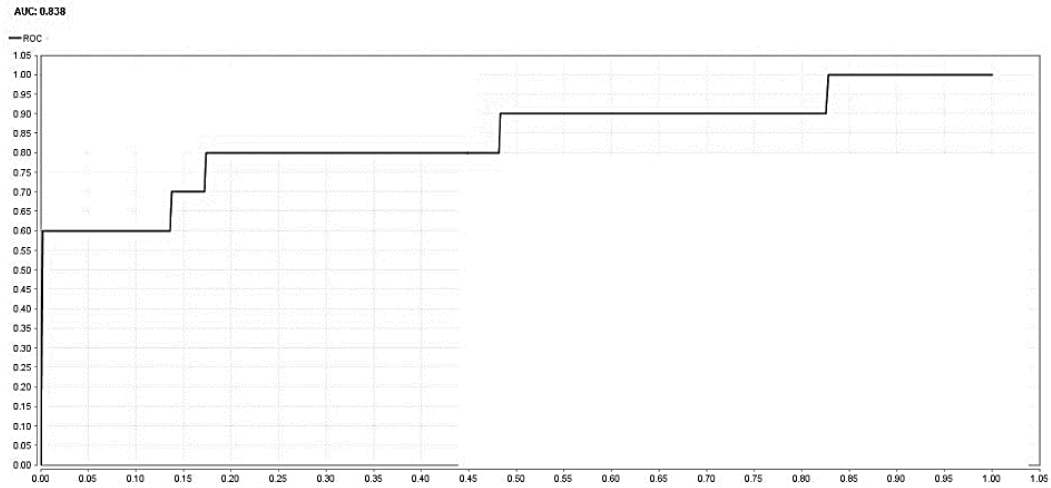
= 100 - 64.10

= 35.90%

(4.33)

Mahasarakham University

AUC: 0.839



ภาพประกอบ 4.8 ค่าพื้นที่ใต้กราฟของขั้นตอนวิธีนาอีพเบย์ส่วนข้อมูลทดสอบ

จากภาพประกอบ 4.8 พบว่าค่าพื้นที่ใต้กราฟเข้าใกล้ 1 แสดงว่าขั้นตอนวิธีนี้นี้เป็นเทคนิคที่มีประสิทธิภาพดีในการคัดกรองผู้ป่วยโรคพาร์กินสัน

Precision: 89.47%

Confusion Matrix:

True:	1	0
1:	17	2
0:	12	8

สามารถคำนวณหาค่า precision ได้จากสมการ

$$Precision = \frac{TP}{(TP+FP)}$$
$$= \frac{17}{(17+2)}$$
$$= 0.8947 \ x\ 100$$
$$= 89.47\ \%$$

(4.34)

Mahasarakham University

Recall: 58.62 %

Confusion Matrix:

True: 1 0

1: 17 2

0: 12 8

สามารถคำนวณหาค่า Recall ได้จากสมการ

$$Recall = \frac{TP}{(TP + FN)} = \frac{17}{(17 + 12)} = 0.5862 \ x\ 100 = 58.62\ \%$$

(4.35)

f_measure: 70.83%

Confusion Matrix:

True: 1 0

1: 17 2

0: 12 8

สามารถคำนวณหาค่า Recall ได้จากสมการ

$$F\_measure = \frac{(2\ x\ Precision\ x\ Recall)}{(Precision + Recall\ )} = \frac{(2\ x\ 0.8947\ x\ 0.5862)}{(0.8947 + 0.5862\ )} = 0.7083\ x\ 100 = 70.83\%$$

(4.36)

4.2.7 ผลการทดลองด้วยขั้นตอนวิธี SVM ในส่วนข้อมูลการสอน

Accuracy: 75.64% +/- 2.62% จากการทดลองพบว่าการใช้ขั้นตอนซัพพอร์ตเวคเตอร์แมชชีนอย่างเดียวในส่วนการใช้ชุดข้อมูลการสอนให้ค่าความถูกต้องเท่ากับร้อยละ 75.64 และค่าส่วนเบี่ยงเบนมาตรฐานร้อยละ 2.62

Confusion Matrix:

True: 1 0

1: 118 38

0: 0 0

สามารถคำนวณหาค่า accuracy ได้จากสมการ

$$Accuracy = \frac{TP + TN}{(TP + FN + TN + FP)}$$
$$= \frac{118 + 0}{(118 + 38 + 0 + 0)}$$
$$= 0.7564\ x\ 100$$
$$= 75.64\%$$

(4.37)

classification_error: 24.36% +/- 2.62%

Confusion Matrix:

True: 1 0

1: 118 38

0: 0 0

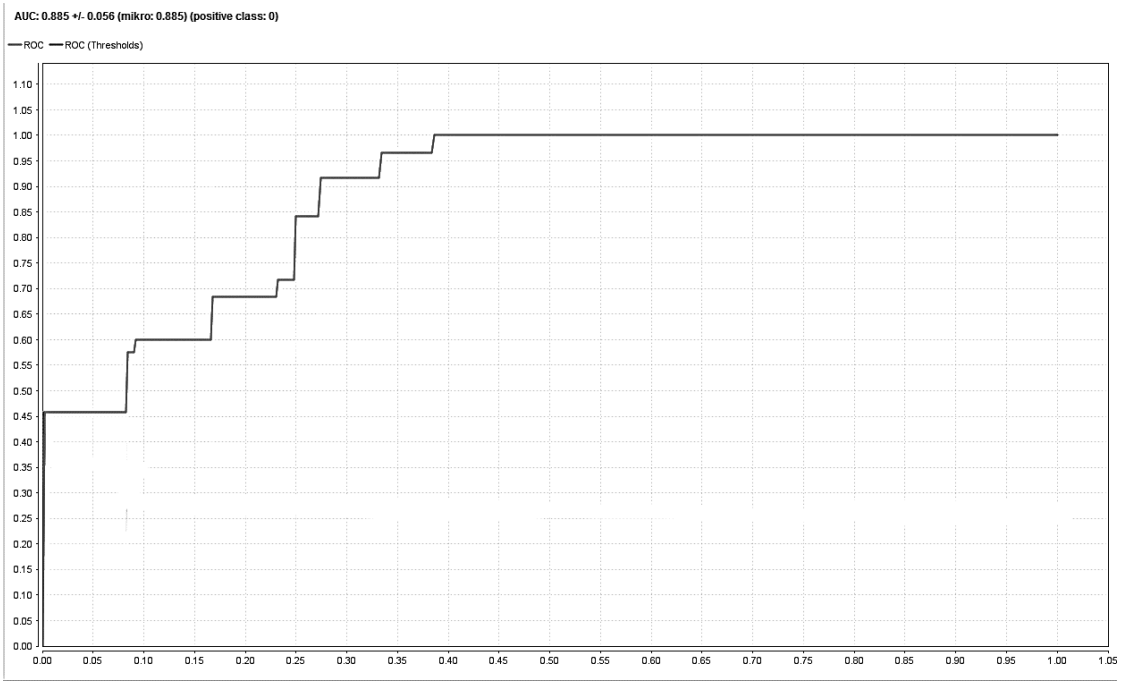
สามารถคำนวณหาค่า accuracy ได้จากสมการ

$$Classification\_error = 100 - accuracy$$
$$= 100 - 75.64$$
$$= 24.36\ \%$$

(4.38)

Mahasarakham University

AUC: 88.5



ภาพประกอบ 4.9 ค่าพื้นที่ใต้กราฟของขั้นตอนวิธี Relief + SVM + Grid ส่วนข้อมูลสอน

จากภาพประกอบ 4.9 พบว่าค่าพื้นที่ใต้กราฟเท่ากับ 1 แสดงให้เห็นถึงประสิทธิภาพในการจำแนกกลุ่มข้อมูลของผู้ที่ป่วยเป็นโรคพาร์กินสันและผู้ที่ไม่ได้เป็นผู้ป่วยเป็นโรคพาร์กินสันออกจากกันได้ และแสดงให้เห็นว่าขั้นตอนวิธีนี้นี้เป็นเทคนิคที่มีประสิทธิภาพสูงในการคัดกรองผู้ป่วยโรคพาร์กินสัน

Precision: 75.64%

Confusion Matrix:

True:	1	0
1:	118	38
0:	0	0

สามารถคำนวณหาค่า precision ได้จากสมการ

$$\begin{aligned} Precision &= \frac{TP}{(TP+FP)} \\ &= \frac{118}{(118+38)} \\ &= 0.7564\ x\ 100 \\ &= 75.64\ \% \end{aligned}$$

(4.39)

Mahasarakham University

Recall: 100.00% +/- 0.00%

Confusion Matrix:

True: 1 0

1: 118 38

0: 0 0

สามารถคำนวณหาค่า Recall ได้จากสมการ

$$Recall = \frac{TP}{(TP + FN)} = \frac{118}{(118+0)} = 1\ x\ 100 = 100\%$$

(4.40)

F_measure: 86.13%

สามารถคำนวณหาค่า Recall ได้จากสมการ

$$F\_measure = \frac{(2\ x\ Precision\ x\ Recall)}{(Precision + Recall\ )} = \frac{(2\ x\ 0.7564\ x\ 1)}{(0.7564 + 1\ )} = 0.8613x\ 100 = 86.13\%$$

4.2.8 ผลการทดลองด้วยขั้นตอนวิธี SVM ในส่วนข้อมูลการทดสอบ

Accuracy: 74.35% จากการทดลองพบว่าการใช้ขั้นตอนซัพพอร์ตเวคเตอร์แมชชีนอย่างเดียวในส่วนการใช้ชุดข้อมูลการทดสอบ ให้ค่าความถูกต้องเท่ากับร้อยละ 74.35

Confusion Matrix:

True: 1 0

1: 29 10

0: 0 0

Mahasarakham University

สามารถคำนวณหาค่า accuracy ได้จากสมการ

$$Accuracy = \frac{TP + TN}{(TP + FN + TN + FP)}$$
$$= \frac{29 + 0}{(29 + 10 + 0 + 0)}$$
$$= 0.7435\ x\ 100$$
$$= 74.35\%$$

(4.41)

classification_error: 25.65%

Confusion Matrix:

True: 1 0

1: 29 10

0: 0 0

สามารถคำนวณหาค่า accuracy ได้จากสมการ

Classification_error = 100 – accuracy

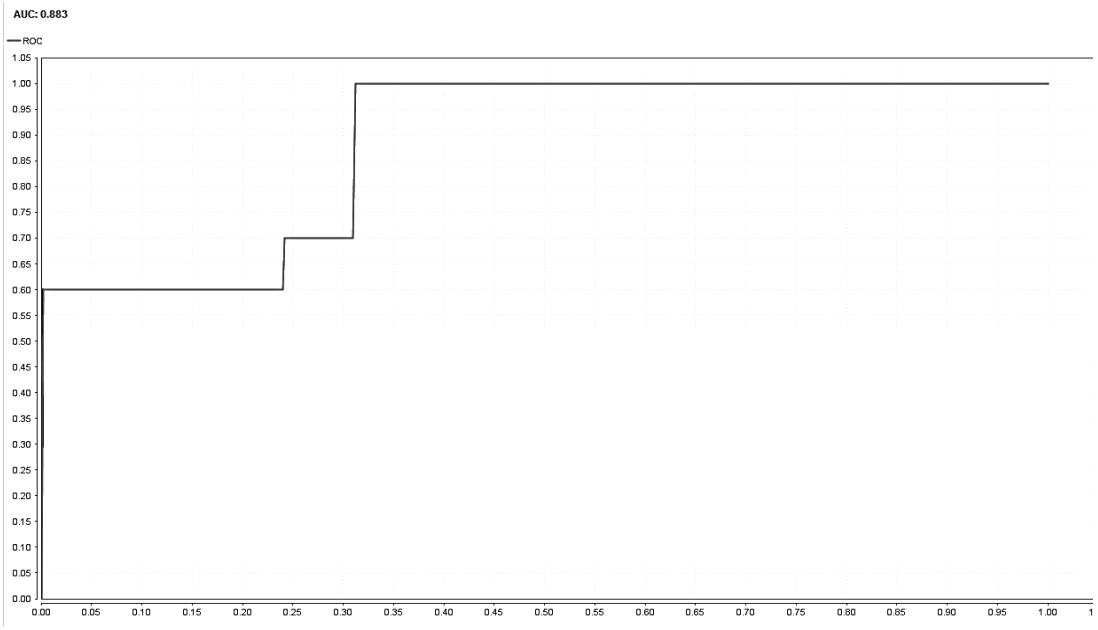
= 100 – 74.35

= 25.65 %

(4.42)

Mahasarakham University

AUC: 88.3



ภาพประกอบ 4.10 ค่าพื้นที่ใต้กราฟของขั้นตอนวิธี Relief + SVM + Grid ส่วนข้อมูลสอน

จากภาพประกอบ 4.10 พบว่าค่าพื้นที่ใต้กราฟเท่ากับ 1 แสดงให้เห็นถึงประสิทธิภาพในการจำแนกกลุ่มข้อมูลของผู้ที่ป่วยเป็นโรคพาร์กินสันและผู้ที่ไม่ได้เป็นผู้ป่วยเป็นโรคพาร์กินสันออกจากกันได้ และแสดงให้เห็นว่าขั้นตอนวิธีนี้นี้เป็นเทคนิคที่มีประสิทธิภาพสูงในการคัดกรองผู้ป่วยโรคพาร์กินสัน

Precision: 74.35%

Confusion Matrix:

True:	1	0
1:	29	10
0:	0	0

สามารถคำนวณหาค่า precision ได้จากสมการ

$$Precision = \frac{TP}{(TP + FP)}$$
$$= \frac{29}{(29 + 10)}$$
$$= 0.7435\ x\ 100$$
$$= 74.35\ \%$$

(4.43)

Mahasarakham University

Recall: 100.00% +/- 0.00%

Confusion Matrix:

True: 1 0

1: 29 10

0: 0 0

สามารถคำนวณหาค่า Recall ได้จากสมการ

$$Recall = \frac{TP}{(TP + FN)} = \frac{29}{(29+0)} = 1\ x\ 100 = 100\%$$

(4.44)

F_measure: 85.29%

สามารถคำนวณหาค่า Recall ได้จากสมการ

$$F\_measure = \frac{(2\ x\ Precision\ x\ Recall)}{(Precision + Recall\ )} = \frac{(2\ x\ 0.7435\ x\ 1)}{(0.7435 + 1\ )} = 0.8529\ x\ 100 = 85.29\%$$

(4.45)

Mahasarakham University

4.2.9 ผลการทดลองด้วยขั้นตอนวิธีโครงข่ายประสาทเทียม (Back Propagation) ในส่วนข้อมูลการสอน

Accuracy: 87.18% +/- 3.01% จากการทดลองพบว่าการใช้ขั้นตอนโครงข่ายประสาทเทียม ในส่วนการใช้ชุดข้อมูลการสอนให้ค่าความถูกต้องเท่ากับร้อยละ 87.18 และค่าส่วนเบี่ยงเบนมาตรฐานร้อยละ 3.01

Confusion Matrix:

True:	1	0
1:	108	10
0:	10	28

สามารถคำนวณหาค่า accuracy ได้จากสมการ

$$Accuracy = \frac{TP + TN}{(TP + FN + TN + FP)}$$

$$= \frac{108 + 28}{(108 + 10 + 28 + 10)}$$

$$= 0.8718\ x\ 100$$

$$= 87.18\%$$

(4.46)

classification_error: 12.82% +/- 3.01%

Confusion Matrix:

True:	1	0
1:	108	10
0:	10	28

สามารถคำนวณหาค่า accuracy ได้จากสมการ

$$Classification\_error = 100 - accuracy$$

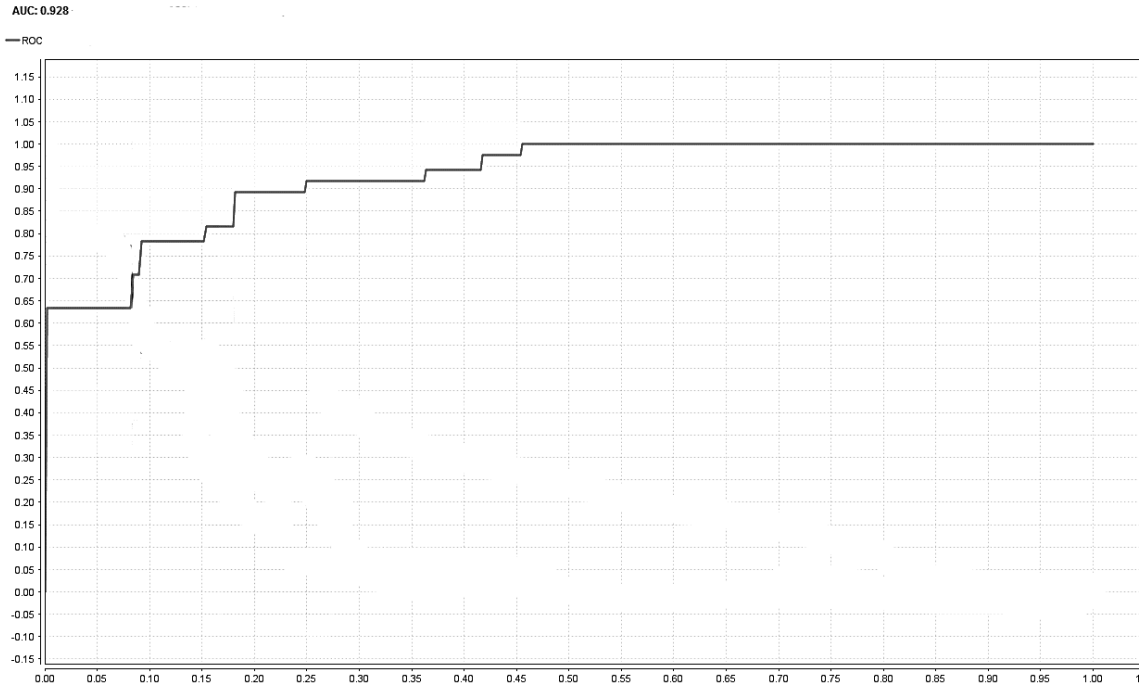
$$= 100 - 87.18$$

$$= 12.82\ \%$$

(4.47)

Mahasarakham University

AUC: 92.8



ภาพประกอบ 4.11 ค่าพื้นที่ใต้กราฟของขั้นตอนวิธี Relief + SVM + Grid ส่วนข้อมูลสอน

จากภาพประกอบ 4.11 พบว่าค่าพื้นที่ใต้กราฟเท่ากับ 1 แสดงให้เห็นถึงประสิทธิภาพในการจำแนกกลุ่มข้อมูลของผู้ที่ป่วยเป็นโรคพาร์กินสันและผู้ที่ไม่ได้เป็นผู้ป่วยเป็นโรคพาร์กินสันออกจากกันได้ และแสดงให้เห็นว่าขั้นตอนวิธีนี้นี้เป็นเทคนิคที่มีประสิทธิภาพสูงในการคัดกรองผู้ป่วยโรคพาร์กินสัน

Precision: 91.53%

Confusion Matrix:

True:	1	0
1:	108	10
0:	10	28

สามารถคำนวณหาค่า precision ได้จากสมการ

$$\begin{aligned} Precision &= \frac{TP}{(TP+FP)} \\ &= \frac{108}{(108+10)} \\ &= 0.9153\ x\ 100 \\ &= 91.53\ \% \end{aligned}$$

(4.48)

Mahasarakham University

Recall: 100.00% +/- 0.00%

Confusion Matrix:

True: 1 0

1: 108 10

0: 10 28

สามารถคำนวณหาค่า Recall ได้จากสมการ

$$\begin{aligned} Recall &= \frac{TP}{(TP + FN)} \\ &= \frac{108}{(108 + 10)} \\ &= 0.9153\ x\ 100 \\ &= 91.53\% \end{aligned}$$

(4.49)

F_measure: 91.53%

สามารถคำนวณหาค่า Recall ได้จากสมการ

$$\begin{aligned} F\_measure &= \frac{(2\ x\ Precision\ x\ Recall)}{(Precision + Recall\ )} \\ &= \frac{(2\ x\ 0.9153\ x\ 0.9153)}{(0.91.53 + 0.9153\ )} \\ &= 0.9153x\ 100 \\ &= 91.53\% \end{aligned}$$

(4.50)

4.2.10 ผลการทดลองด้วยขั้นตอนวิธีโครงข่ายประสาทเทียม (Back Propagation) ในส่วนข้อมูลการทดสอบ

Accuracy: 82.05% จากการทดลองพบว่าการใช้ขั้นตอนโครงข่ายประสาทเทียมในส่วนการใช้ชุดข้อมูลการทดสอบ ให้ค่าความถูกต้องเท่ากับร้อยละ 82.05

Mahasarakham University

Confusion Matrix:

True: 1 0

1: 22 0

0: 7 10

สามารถคำนวณหาค่า accuracy ได้จากสมการ

$$Accuracy = \frac{TP + TN}{(TP + FN + TN + FP)}$$
$$= \frac{22 + 10}{(22 + 0 + 10 + 7)}$$
$$= 0.8205\ x\ 100$$
$$= 82.05\%$$

(4.51)

classification_error: 25.65%

Confusion Matrix:

True: 1 0

1: 22 0

0: 7 10

สามารถคำนวณหาค่า accuracy ได้จากสมการ

Classification_error = 100 – accuracy

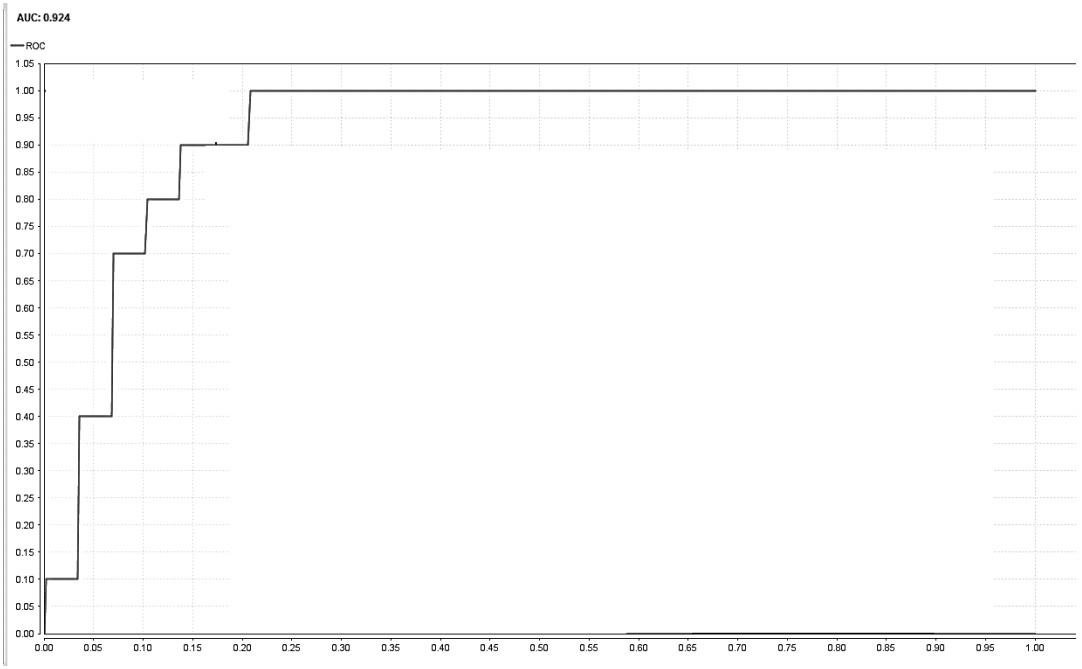
= 100 – 82.05

= 17.94 %

(4.52)

Mahasarakham University

AUC: 92.4



ภาพประกอบ 4.12 ค่าพื้นที่ใต้กราฟของขั้นตอนวิธี Relief + SVM + Grid ส่วนข้อมูลสอน

จากภาพประกอบ 4.12 พบว่าค่าพื้นที่ใต้กราฟเท่ากับ 1 แสดงให้เห็นถึงประสิทธิภาพในการจำแนกกลุ่มข้อมูลของผู้ที่ป่วยเป็นโรคพาร์กินสันและผู้ที่ไม่ได้เป็นผู้ป่วยเป็นโรคพาร์กินสันออกจากกันได้ และแสดงให้เห็นว่าขั้นตอนวิธีนี้นี้เป็นเทคนิคที่มีประสิทธิภาพสูงในการคัดกรองผู้ป่วยโรคพาร์กินสัน

Precision: 100%

Confusion Matrix:

True:	1	0
1:	22	0
0:	7	10

สามารถคำนวณหาค่า precision ได้จากสมการ

$$Precision = \frac{TP}{(TP+FP)}$$
$$= \frac{22}{(22+0)}$$
$$= 1\ x\ 100$$
$$= 100\ \%$$

(4.53)

Mahasarakham University

Recall: 75.86%

Confusion Matrix:

True: 1 0

1: 22 0

0: 7 10

สามารถคำนวณหาค่า Recall ได้จากสมการ

$$Recall = \frac{TP}{(TP + FN)}$$
$$= \frac{22}{(22+7)}$$
$$= 0.7586\ x\ 100$$
$$= 75.86\%$$

(4.54)

F_measure: 86.27%

สามารถคำนวณหาค่า Recall ได้จากสมการ

$$F\_measure = \frac{(2\ x\ Precision\ x\ Recall)}{(Precision + Recall\ )}$$
$$= \frac{(2\ x\ 1\ x\ 0.7586)}{(0.7586 + 1\ )}$$
$$= 0.8627\ x\ 100$$
$$= 86.27\%$$

(4.55)

4.2.11 ผลการหาค่าพารามิเตอร์ที่ความเหมาะสมที่สุดด้วยการค้นหาแบบกริด

ในการศึกษานี้ได้สร้างกริดด้วยค่า C ต่ำสุดที่ 0 และค่า C มากสุดที่ 10 โดยแบ่งช่วงกริดค่า C ออกเป็น100 สเกล และใช้เสกลแบบเชิงเส้น ส่วนค่า Gamma ค่าต่ำสุดคือ 0 ค่ามากที่สุดคือ 10 แบ่งช่วงกริดค่า Gamma ออกเป็น 10 สเกล โดยใช้สเกลแบบเชิงเส้นเช่นกัน ทำให้เกิดเป็นกริดขนาด 1111 ค่าขึ้น ซึ่งถือเป็นขนาดที่ไม่ใหญ่เกินไปสำหรับข้อมูลชุดนี้ ทำให้ไม่ใช้เวลาในการประมวลผล

Mahasarakham University

มากเกินไป โดยใช้เวลาประมวลผลทั้งหมด 19 วินาที จากการทดลองพบว่าค่าพารามิเตอร์ C ที่เหมาะสมที่สุดคือ 3.6 และค่าพารามิเตอร์ที่เหมาะสมที่สุดของ Gamma คือ 9 เนื่องจากให้ค่าประสิทธิภาพ (performance) สูงที่สุดในสำหรับข้อมูลชุดนี้ สามารถดูผลการหาค่าพารามิเตอร์ที่เหมาะสมที่สุดทั้งหมดได้จากภาคผนวก ข

4.3 สรุปการทดลอง

จากผลการทดลองทั้งหมดเมื่อดูค่าพื้นที่ใต้กราฟ AUC จากภาพประกอบ 4.12 และ 4.13 จะพบว่าวิธีที่ให้ค่า AUC เข้าใกล้ 1 ที่สุดคือ SVM รองลงมาคือ Naïve Bayes และ Decision Tree ตามลำดับ แสดงว่าขั้นตอนซัพพอร์ตเวคเตอร์ที่ศึกษามีประสิทธิภาพในการจำแนกข้อมูลผู้ป่วยโรคพาร์กินสันที่สุดเมื่อเปรียบเทียบกับวิธีอื่นๆ เมื่อวัดผลของการประเมินประสิทธิภาพการจำแนกประเภทข้อมูลโดยใช้ค่าอื่นที่นิยมใช้ในการวัดประสิทธิภาพประกอบด้วย accuracy, precision, recall และ f_measure พบว่าขั้นตอนซัพพอร์ตเวคเตอร์ที่ศึกษามีประสิทธิภาพสูงที่สุด ดังแสดงในตาราง 4.6

ตาราง 4.6 เปรียบเทียบผลการทดลองของแต่ละขั้นตอนวิธีในส่วนข้อมูลสอน

ค่าที่ใช้วัด	**ข้อมูลสอน**				
	Relief+SVM +Grid	Decision Tree	Naïve Bayes	SVM	Neural network
accuracy	1.0000	0.8013	0.7051	0.7564	0.8717
precision	1.0000	0.8372	0.9500	0.7564	0.9152
recall	1.0000	0.9153	0.6441	1.0000	0.9152
f_measure	1.0000	0.8745	0.7677	0.8613	0.9152
	ข้อมูลทดสอบ				
	Relief+SVM +Grid	Decision Tree	Naïve Bayes	SVM	Neural network
accuracy	0.8974	0.6923	0.6410	0.7435	0.8205
precision	1.0000	1.0000	0.8947	0.7435	1.0000
recall	0.8620	0.5862	0.5862	1.0000	0.7586
f_measure	0.9259	0.7391	0.7083	0.8529	0.8627

Mahasarakham University

บทที่ 5

สรุปผล และข้อเสนอแนะ

5.1 สรุปผล

งานวิจัยนี้ได้เสนอเทคนิคการคัดกรองผู้ป่วยโรคพาร์กินสันจากข้อมูลการวิเคราะห์เสียง โดยใช้ขั้นตอนวิธีซัพพอร์ตเวคเตอร์แมชชีนและการค้นหาแบบกริด ซึ่งเป็นเทคนิคการเรียนรู้ของเครื่องที่มีประสิทธิภาพในการจำแนกประเภทข้อมูลได้ดี โดยนำเสนอแบบจำลองการจำแนกผู้ป่วยโรคพาร์กินสันจากข้อมูลการวิเคราะห์เสียง ทดสอบประสิทธิภาพแบบจำลองด้วยขั้นตอนวิธีซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ และเนอีฟเบย์ โดยวัดประสิทธิภาพจากค่าความถูกต้อง (Accuracy) ของแบบจำลอง พบว่าขั้นตอนวิธีซัพพอร์ตเวกเตอร์แมชชีนร่วมกับการค้นหาแบบกริดให้ประสิทธิภาพในการจำแนกดีที่สุด ให้ค่าความถูกต้อง (Accuracy) สูงที่สุดเท่ากับ 100% สำหรับข้อมูลที่ใช้สอน เมื่อลดมิติข้อมูลลงเหลือ 13 คุณลักษณะด้วยขั้นตอนรีลีฟ และการปรับสมดุลข้อมูลด้วยวิธีการ cost-sensitive approach โดยการกำหนดค่าน้ำหนัก (weight) ให้แต่ละคลาสคำตอบไม่เท่ากัน และให้ค่าความถูกต้อง (Accuracy) สูงที่สุดเท่ากับ 89.74% สำหรับข้อมูลทดสอบ เมื่อเปรียบเทียบผลการทดลองกับขั้นตอนวิธีต้นไม้ตัดสินใจ เนอีฟเบย์ ซัพพอร์ตเวคเตอร์แมชชีนอย่างเดียว และโครงข่าวประสาทเทียม พบว่าขั้นตอนวิธีต้นไม้ตัดสินใจให้ค่าความถูกต้อง 80.13% สำหรับข้อมูลที่ใช้สอน 69.23% สำหรับข้อมูลทดสอบ ขั้นตอนวิธีเนอีฟเบย์ให้ค่าความถูกต้อง 70.51% สำหรับข้อมูลที่ใช้สอน 64.10% สำหรับข้อมูลทดสอบ ซัพพอร์ตเวคเตอร์แมชชีนอย่างเดียวให้ค่าความถูกต้อง 75.64% สำหรับข้อมูลที่ใช้สอน 74.35 % สำหรับข้อมูลทดสอบ และขั้นตอนวิธี โครงข่าวประสาทเทียมให้ค่าความถูกต้อง 87.17% สำหรับข้อมูลที่ใช้สอน 82.05% สำหรับข้อมูลทดสอบ

ผลการทดลองจากงานวิจัยนี้ พบว่าขั้นตอนวิธีรีลีฟมีประสิทธิภาพดีในการลดคุณลักษณะของข้อมูล จากคุณลักษณะที่สกัดได้ทั้งหมดทำให้พบว่าชุดข้อมูลประเภทนี้สามารถลดมิติของข้อมูลด้วยวิธีรีลีฟลงได้อย่างมาก โดยไม่กระทบต่อประสิทธิภาพในการจำแนกข้อมูลแต่อย่างใดและ ยิ่งเมื่อลดมิติของข้อมูลลง ยิ่งส่งผลให้ค่าความถูกต้องในการจำแนกเพิ่มสูงขึ้น ซึ่งการลดมิติของข้อมูลนี้ สามารถลดทรัพยากรของระบบและลดระยะเวลาในการประมวลผลได้เป็นอย่างมาก และเนื่องจากข้อมูลส่วนใหญ่จะเป็นข้อมูลที่ไม่สมดุลคือคลาสคำตอบมีจำนวนแตกต่างกันมาก การปรับสมดุลข้อมูลด้วยวิธีการ Undersampling โดยการกำหนดค่าน้ำหนัก (weight) จึงช่วยในการลดการโน้มเอียงของข้อมูลได้ ขั้นตอนวิธีพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) มีประสิทธิภาพในการจำแนก โดยเฉพาะเมื่อนำมาใช้ร่วมกับวิธีการหาค่าที่เหมาะสมให้กับพารามิเตอร์ด้วยการค้นหาแบบกริด ยิ่งส่งผลให้ค่าความถูกต้องในการจำแนกเพิ่มสูงขึ้นอย่างมีนัยสำคัญ

Mahasarakham University

ผลจากงานวิจัยนี้สามารถนำแบบจำลองที่นำเสนอไปประยุกต์ใช้กับการจำแนกข้อมูลกับประเภทอื่นๆ เช่น การคัดกรองโรคอื่น ๆ ที่มีลักษณะเดียวกันได้

5.2 ขอเสนอแนะ

สิ่งที่นาสนใจที่จะทำต่อไปคือ ทดลองแก้ปัญหาข้อมูลไม่สมดุลที่มีการใช้พื้นที่ร่วมกันเพื่อให้สามารถจำแนกประเภทข้อมูลไม่สมดุลได้อย่างมีประสิทธิภาพ และการนำขั้นตอนวิธีนี้ไปทดสอบกับข้อมูลชุดอื่น ๆ เพื่อหาประสิทธิภาพว่าสามารถนำขั้นตอนวิธีนี้ไปใช้กับข้อมูลที่มีลักษณะแตกต่างกันไปได้หรือไม่ นอกจากนี้ในเรื่องการสร้างแบบจำลองให้มีเทคนิคการเรียนรู้ของเครื่องหลายๆ วิธีมีการเรียนรู้ร่วมกัน เพื่อปรับปรุงประสิทธิภาพร่วมกับเทคนิคอื่น ๆ

Mahasarakham University

เอกสารอ้างอิง

Mahasarakham University

เอกสารอ้างอิง

[1] Esposito, F., Malerba, D. and Semeraro, G. A comparative analysis of methods for pruning decision trees. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1997; 19(5): 476-491.

[2] Kenneth, WL, Ian, B. and Robin, C. Parkinson's disease. In: Thomas F. editor. Neurology and Neurosurgery Illustrated. 4th ed. Philadelphia : Churchill Livingstone; 2004.

[3] Cao, D.Z., Su-Lin Pang, Yuan-Huai Bai. Forecasting Exchange Rate Using Support Vector Machines. Machine Learning and Cybernetics. 2005; 6: 3448-3452.

[4] Deepa Shenoy, Sandhya Joshi. Classification of Alzheimer's disease and Parkinson's Disease by using Machine Learning and Neural Network Methods. Conference on Machine Learning and Computing; 2010 (IEEE)(8). p. 218-222.

[5] นฤพนต์ ว่องประชานุกูล. วิธีที่เหมาะสมสำหรับการตัดกิ่งต้นไม้ตัดสินใจของการทำเหมืองข้อมูลทางด้านวิทยาศาสตร์ [วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต บัณฑิตวิทยาลัย]. นครราชสีมา: มหาวิทยาลัยเทคโนโลยีสุรนารี; 2548.

[6] Breiman, L., Friedman, J., Stone, C. J., Olshen, R. A. Classification and Regression Trees. n.p.: CRC press; 1984.

[7] Breslow, L. A., & Aha, D. W. Simplifying decision trees: A survey. The Knowledge Engineering Review. 1997; 12(1): 1-40.

[8] ปริญญา สงวนสัตย์. การเรียนรู้ของเครื่อง. กรุงเทพฯ: สถาบันการจัดการปัญญาภิวัฒน์; 2558.

[9] Malerba, F. D., Semeraro, G. and Esposito, Tamma V.. The effects of pruning methods on the predictive accuracy of induced decision trees. Applied Stochastic Models in Business and Industry. 1999; 15: 277-299.

[10] H.X. Zhang, M.C. Liu Z.D., and B.T. Fan Zhao C.Y. Application of Support Vector Machine for Prediction Toxic Activity of Different Data Sets. Toxicology. 2006; 217: 105-119.

[11] ก้องศักดิ์ จงเกษมวงศ์. การตัดเล็มอย่างอ่อนสำหรับต้นไม้ตัดสินใจโดยการใช้แบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก. กรุงเทพฯ: จุฬาลงกรณ์มหาวิทยลัย; 2543.

[12] Ramani, R. Geetha and Sivagami, G. Parkinson Disease Classification using Data Mining Algorithms. International Journal of Computer Applications (IJCA). 2011; 32(7): 46-53.

Mahasarakham University

[13] Das, R. Turkey A comparison of multiple classification methods for diagnosis of Parkinson disease. Expert System with Application. 2010; 37: 1568-1572.

[14] Setiono, R. Extracting rules from pruned neural networks for breast cancer diagnosis. Artificial Intelligence in Medicine. 1996; 1(8): 37–51.

[15] เอกสิทธิ์ พัชรวงศ์ศักดา. การวิเคราะห์ข้อมูลด้วยเทคเนิคดาต้า ไมน์นิ่ง เบื้องต้น. กรุงเทพฯ: เอเชีย ดิจิตอลการพิมพ์; 2557.

[16] Nancy, P. and Ramani, Geetha. A Comparison on Performance of Data Mining Algorithms in Classification of Social Network Data. International Journal of Computer Applications. 2011; 32: 47-54.

[17] Little, M. A., McSharry, P. E., Roberts, S. J., Costello, D. A., abd Moroz, I. M. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. BioMedical Engineering OnLine. 2007; 6(1): 23.

Shomona Gracia Jacob and Geetha Ramani. (2011). Discovery of Knowledge Patterns in Clinical Data through Data Mining Algorithms:Multi-class Categorization of Breast Tissue Data. International Journal of Computer Applications (IJCA)(32(7)), 46-53.

[18] Jacob, S. G. and Ramani, R. G. Discovery of knowledge patterns in clinical data through data mining algorithms: Multi-class categorization of breast tissue data. International Journal of Computer Applications (IJCA). 2011; 32(7): 46-53.

[19] Vapnik, V. The Nature of Statistical Learning Theory. New York: Springer-Verlag; 1995.

[20] David, G.A. and Magnus, J.B. Diagnosing Parkinson by using Artificial Neural Network and Support Vector Machine. *Global* Journal of Computer Science and Technology. 2009; 9: 63-71.

[21] Engin, Mehmet, Serdar Demirag, Erkan Zeki Engin, Gu¨rbu¨z, Fisun Ersan, Erden Asena and Zafer Colakog. The classification of human tremor signal using artificial neural network. Expert Systems with Applications. 2007; 33(3): 754-761.

[22] Ene, M. Neural network-based approach to discriminate healthy people from those with Parkinson's disease. Annals of the University of Craiova, Math. Comp. Sci. Ser. 2008; 35: 112-116.

Mahasarakham University

ภาคผนวก

Mahasarakham University

ภาคผนวก ก

ชุดข้อมูลที่ใช้ในการศึกษาเป็นข้อมูลจาก UCI จำนวน 195 ระเบียน

Mahasarakham University

ชุดข้อมูลที่ใช้ในการศึกษาเป็นข้อมูลจาก UCI จำนวน 195 ระเบียน

name,MDVP:Fo(Hz),MDVP:Fhi(Hz),MDVP:Flo(Hz),MDVP:Jitter(%),MDVP:Jitter(Abs),MDVP:RAP,MDVP:PPQ,Jitter:DDP,MDVP:Shimmer,MDVP:Shimmer(dB),Shimmer:APQ3,Shimmer:APQ5,MDVP:APQ,Shimmer:DDA,NHR,HNR,status,RPDE,DFA,spread1,spread2,D2,PPE

phon_R01_S01_1,119.99200,157.30200,74.99700,0.00784,0.00007,0.00370,0.00554,0.01109,0.04374,0.42600,0.02182,0.03130,0.02971,0.06545,0.02211,21.03300,1,0.414783,0.815285,-4.813031,0.266482,2.301442,0.284654

phon_R01_S01_2,122.40000,148.65000,113.81900,0.00968,0.00008,0.00465,0.00696,0.01394,0.06134,0.62600,0.03134,0.04518,0.04368,0.09403,0.01929,19.08500,1,0.458359,0.819521,-4.075192,0.335590,2.486855,0.368674

phon_R01_S01_3,116.68200,131.11100,111.55500,0.01050,0.00009,0.00544,0.00781,0.01633,0.05233,0.48200,0.02757,0.03858,0.03590,0.08270,0.01309,20.65100,1,0.429895,0.825288,-4.443179,0.311173,2.342259,0.332634

phon_R01_S01_4,116.67600,137.87100,111.36600,0.00997,0.00009,0.00502,0.00698,0.01505,0.05492,0.51700,0.02924,0.04005,0.03772,0.08771,0.01353,20.64400,1,0.434969,0.819235,-4.117501,0.334147,2.405554,0.368975

phon_R01_S01_5,116.01400,141.78100,110.65500,0.01284,0.00011,0.00655,0.00908,0.01966,0.06425,0.58400,0.03490,0.04825,0.04465,0.10470,0.01767,19.64900,1,0.417356,0.823484,-3.747787,0.234513,2.332180,0.410335

phon_R01_S01_6,120.55200,131.16200,113.78700,0.00968,0.00008,0.00463,0.00750,0.01388,0.04701,0.45600,0.02328,0.03526,0.03243,0.06985,0.01222,21.37800,1,0.415564,0.825069,-4.242867,0.299111,2.187560,0.357775

phon_R01_S02_1,120.26700,137.24400,114.82000,0.00333,0.00003,0.00155,0.00202,0.00466,0.01608,0.14000,0.00779,0.00937,0.01351,0.02337,0.00607,24.88600,1,0.596040,0.764112,-5.634322,0.257682,1.854785,0.211756

phon_R01_S02_2,107.33200,113.84000,104.31500,0.00290,0.00003,0.00144,0.00182,0.00431,0.01567,0.13400,0.00829,0.00946,0.01256,0.02487,0.00344,26.89200,1,0.637420,0.763262,-6.167603,0.183721,2.064693,0.163755

phon_R01_S02_3,95.73000,132.06800,91.75400,0.00551,0.00006,0.00293,0.00332,0.00880,0.02093,0.19100,0.01073,0.01277,0.01717,0.03218,0.01070,21.81200,1,0.615551,0.773587,-5.498678,0.327769,2.322511,0.231571

Mahasarakham University

phon_R01_S02_4,95.05600,120.10300,91.22600,0.00532,0.00006,0.00268,0.00332,0.0080
3,0.02838,0.25500,0.01441,0.01725,0.02444,0.04324,0.01022,21.86200,1,0.54703
7,0.798463,-5.011879,0.325996,2.432792,0.271362
phon_R01_S02_5,88.33300,112.24000,84.07200,0.00505,0.00006,0.00254,0.00330,0.0076
3,0.02143,0.19700,0.01079,0.01342,0.01892,0.03237,0.01166,21.11800,1,0.61113
7,0.776156,-5.249770,0.391002,2.407313,0.249740
phon_R01_S02_6,91.90400,115.87100,86.29200,0.00540,0.00006,0.00281,0.00336,0.0084
4,0.02752,0.24900,0.01424,0.01641,0.02214,0.04272,0.01141,21.41400,1,0.58339
0,0.792520,-4.960234,0.363566,2.642476,0.275931
phon_R01_S04_1,136.92600,159.86600,131.27600,0.00293,0.00002,0.00118,0.00153,0.00
355,0.01259,0.11200,0.00656,0.00717,0.01140,0.01968,0.00581,25.70300,1,0.460
600,0.646846,-6.547148,0.152813,2.041277,0.138512
phon_R01_S04_2,139.17300,179.13900,76.55600,0.00390,0.00003,0.00165,0.00208,0.004
96,0.01642,0.15400,0.00728,0.00932,0.01797,0.02184,0.01041,24.88900,1,0.4301
66,0.665833,-5.660217,0.254989,2.519422,0.199889
phon_R01_S04_3,152.84500,163.30500,75.83600,0.00294,0.00002,0.00121,0.00149,0.003
64,0.01828,0.15800,0.01064,0.00972,0.01246,0.03191,0.00609,24.92200,1,0.4747
91,0.654027,-6.105098,0.203653,2.125618,0.170100
phon_R01_S04_4,142.16700,217.45500,83.15900,0.00369,0.00003,0.00157,0.00203,0.004
71,0.01503,0.12600,0.00772,0.00888,0.01359,0.02316,0.00839,25.17500,1,0.5659
24,0.658245,-5.340115,0.210185,2.205546,0.234589
phon_R01_S04_5,144.18800,349.25900,82.76400,0.00544,0.00004,0.00211,0.00292,0.006
32,0.02047,0.19200,0.00969,0.01200,0.02074,0.02908,0.01859,22.33300,1,0.5673
80,0.644692,-5.440040,0.239764,2.264501,0.218164
phon_R01_S04_6,168.77800,232.18100,75.60300,0.00718,0.00004,0.00284,0.00387,0.008
53,0.03327,0.34800,0.01441,0.01893,0.03430,0.04322,0.02919,20.37600,1,0.6310
99,0.605417,-2.931070,0.434326,3.007463,0.430788
phon_R01_S05_1,153.04600,175.82900,68.62300,0.00742,0.00005,0.00364,0.00432,0.010
92,0.05517,0.54200,0.02471,0.03572,0.05767,0.07413,0.03160,17.28000,1,0.6653
18,0.719467,-3.949079,0.357870,3.109010,0.377429

Mahasarakham University

phon_R01_S05_2,156.40500,189.39800,142.82200,0.00768,0.00005,0.00372,0.00399,0.01116,0.03995,0.34800,0.01721,0.02374,0.04310,0.05164,0.03365,17.15300,1,0.649554,0.686080,-4.554466,0.340176,2.856676,0.322111

phon_R01_S05_3,153.84800,165.73800,65.78200,0.00840,0.00005,0.00428,0.00450,0.01285,0.03810,0.32800,0.01667,0.02383,0.04055,0.05000,0.03871,17.53600,1,0.660125,0.704087,-4.095442,0.262564,2.739710,0.365391

phon_R01_S05_4,153.88000,172.86000,78.12800,0.00480,0.00003,0.00232,0.00267,0.00696,0.04137,0.37000,0.02021,0.02591,0.04525,0.06062,0.01849,19.49300,1,0.629017,0.698951,-5.186960,0.237622,2.557536,0.259765

phon_R01_S05_5,167.93000,193.22100,79.06800,0.00442,0.00003,0.00220,0.00247,0.00661,0.04351,0.37700,0.02228,0.02540,0.04246,0.06685,0.01280,22.46800,1,0.619060,0.679834,-4.330956,0.262384,2.916777,0.285695

phon_R01_S05_6,173.91700,192.73500,86.18000,0.00476,0.00003,0.00221,0.00258,0.00663,0.04192,0.36400,0.02187,0.02470,0.03772,0.06562,0.01840,20.42200,1,0.537264,0.686894,-5.248776,0.210279,2.547508,0.253556

phon_R01_S06_1,163.65600,200.84100,76.77900,0.00742,0.00005,0.00380,0.00390,0.01140,0.01659,0.16400,0.00738,0.00948,0.01497,0.02214,0.01778,23.83100,1,0.397937,0.732479,-5.557447,0.220890,2.692176,0.215961

phon_R01_S06_2,104.40000,206.00200,77.96800,0.00633,0.00006,0.00316,0.00375,0.00948,0.03767,0.38100,0.01732,0.02245,0.03780,0.05197,0.02887,22.06600,1,0.522746,0.737948,-5.571843,0.236853,2.846369,0.219514

phon_R01_S06_3,171.04100,208.31300,75.50100,0.00455,0.00003,0.00250,0.00234,0.00750,0.01966,0.18600,0.00889,0.01169,0.01872,0.02666,0.01095,25.90800,1,0.418622,0.720916,-6.183590,0.226278,2.589702,0.147403

phon_R01_S06_4,146.84500,208.70100,81.73700,0.00496,0.00003,0.00250,0.00275,0.00749,0.01919,0.19800,0.00883,0.01144,0.01826,0.02650,0.01328,25.11900,1,0.358773,0.726652,-6.271690,0.196102,2.314209,0.162999

phon_R01_S06_5,155.35800,227.38300,80.05500,0.00310,0.00002,0.00159,0.00176,0.00476,0.01718,0.16100,0.00769,0.01012,0.01661,0.02307,0.00677,25.97000,1,0.470478,0.676258,-7.120925,0.279789,2.241742,0.108514

Mahasarakham University

phon_R01_S06_6,162.56800,198.34600,77.63000,0.00502,0.00003,0.00280,0.00253,0.00841,0.01791,0.16800,0.00793,0.01057,0.01799,0.02380,0.01170,25.67800,1,0.427785,0.723797,-6.635729,0.209866,1.957961,0.135242

phon_R01_S07_1,197.07600,206.89600,192.05500,0.00289,0.00001,0.00166,0.00168,0.00498,0.01098,0.09700,0.00563,0.00680,0.00802,0.01689,0.00339,26.77500,0,0.422229,0.741367,-7.348300,0.177551,1.743867,0.085569

phon_R01_S07_2,199.22800,209.51200,192.09100,0.00241,0.00001,0.00134,0.00138,0.00402,0.01015,0.08900,0.00504,0.00641,0.00762,0.01513,0.00167,30.94000,0,0.432439,0.742055,-7.682587,0.173319,2.103106,0.068501

phon_R01_S07_3,198.38300,215.20300,193.10400,0.00212,0.00001,0.00113,0.00135,0.00339,0.01263,0.11100,0.00640,0.00825,0.00951,0.01919,0.00119,30.77500,0,0.465946,0.738703,-7.067931,0.175181,1.512275,0.096320

phon_R01_S07_4,202.26600,211.60400,197.07900,0.00180,0.000009,0.00093,0.00107,0.00278,0.00954,0.08500,0.00469,0.00606,0.00719,0.01407,0.00072,32.68400,0,0.368535,0.742133,-7.695734,0.178540,1.544609,0.056141

phon_R01_S07_5,203.18400,211.52600,196.16000,0.00178,0.000009,0.00094,0.00106,0.00283,0.00958,0.08500,0.00468,0.00610,0.00726,0.01403,0.00065,33.04700,0,0.340068,0.741899,-7.964984,0.163519,1.423287,0.044539

phon_R01_S07_6,201.46400,210.56500,195.70800,0.00198,0.000010,0.00105,0.00115,0.00314,0.01194,0.10700,0.00586,0.00760,0.00957,0.01758,0.00135,31.73200,0,0.344252,0.742737,-7.777685,0.170183,2.447064,0.057610

phon_R01_S08_1,177.87600,192.92100,168.01300,0.00411,0.00002,0.00233,0.00241,0.00700,0.02126,0.18900,0.01154,0.01347,0.01612,0.03463,0.00586,23.21600,1,0.360148,0.778834,-6.149653,0.218037,2.477082,0.165827

phon_R01_S08_2,176.17000,185.60400,163.56400,0.00369,0.00002,0.00205,0.00218,0.00616,0.01851,0.16800,0.00938,0.01160,0.01491,0.02814,0.00340,24.95100,1,0.341435,0.783626,-6.006414,0.196371,2.536527,0.173218

phon_R01_S08_3,180.19800,201.24900,175.45600,0.00284,0.00002,0.00153,0.00166,0.00459,0.01444,0.13100,0.00726,0.00885,0.01190,0.02177,0.00231,26.73800,1,0.403884,0.766209,-6.452058,0.212294,2.269398,0.141929

Mahasarakham University

phon_R01_S08_4,187.73300,202.32400,173.01500,0.00316,0.00002,0.00168,0.00182,0.00504,0.01663,0.15100,0.00829,0.01003,0.01366,0.02488,0.00265,26.31000,1,0.396793,0.758324,-6.006647,0.266892,2.382544,0.160691

phon_R01_S08_5,186.16300,197.72400,177.58400,0.00298,0.00002,0.00165,0.00175,0.00496,0.01495,0.13500,0.00774,0.00941,0.01233,0.02321,0.00231,26.82200,1,0.326480,0.765623,-6.647379,0.201095,2.374073,0.130554

phon_R01_S08_6,184.05500,196.53700,166.97700,0.00258,0.00001,0.00134,0.00147,0.00403,0.01463,0.13200,0.00742,0.00901,0.01234,0.02226,0.00257,26.45300,1,0.306443,0.759203,-7.044105,0.063412,2.361532,0.115730

phon_R01_S10_1,237.22600,247.32600,225.22700,0.00298,0.00001,0.00169,0.00182,0.00507,0.01752,0.16400,0.01035,0.01024,0.01133,0.03104,0.00740,22.73600,0,0.305062,0.654172,-7.310550,0.098648,2.416838,0.095032

phon_R01_S10_2,241.40400,248.83400,232.48300,0.00281,0.00001,0.00157,0.00173,0.00470,0.01760,0.15400,0.01006,0.01038,0.01251,0.03017,0.00675,23.14500,0,0.457702,0.634267,-6.793547,0.158266,2.256699,0.117399

phon_R01_S10_3,243.43900,250.91200,232.43500,0.00210,0.000009,0.00109,0.00137,0.00327,0.01419,0.12600,0.00777,0.00898,0.01033,0.02330,0.00454,25.36800,0,0.438296,0.635285,-7.057869,0.091608,2.330716,0.091470

phon_R01_S10_4,242.85200,255.03400,227.91100,0.00225,0.000009,0.00117,0.00139,0.00350,0.01494,0.13400,0.00847,0.00879,0.01014,0.02542,0.00476,25.03200,0,0.431285,0.638928,-6.995820,0.102083,2.365800,0.102706

phon_R01_S10_5,245.51000,262.09000,231.84800,0.00235,0.000010,0.00127,0.00148,0.00380,0.01608,0.14100,0.00906,0.00977,0.01149,0.02719,0.00476,24.60200,0,0.467489,0.631653,-7.156076,0.127642,2.392122,0.097336

phon_R01_S10_6,252.45500,261.48700,182.78600,0.00185,0.000007,0.00092,0.00113,0.00276,0.01152,0.10300,0.00614,0.00730,0.00860,0.01841,0.00432,26.80500,0,0.610367,0.635204,-7.319510,0.200873,2.028612,0.086398

phon_R01_S13_1,122.18800,128.61100,115.76500,0.00524,0.00004,0.00169,0.00203,0.00507,0.01613,0.14300,0.00855,0.00776,0.01433,0.02566,0.00839,23.16200,0,0.579597,0.733659,-6.439398,0.266392,2.079922,0.133867

Mahasarakham University

phon_R01_S13_2,122.96400,130.04900,114.67600,0.00428,0.00003,0.00124,0.00155,0.00373,0.01681,0.15400,0.00930,0.00802,0.01400,0.02789,0.00462,24.97100,0,0.538688,0.754073,-6.482096,0.264967,2.054419,0.128872

phon_R01_S13_3,124.44500,135.06900,117.49500,0.00431,0.00003,0.00141,0.00167,0.00422,0.02184,0.19700,0.01241,0.01024,0.01685,0.03724,0.00479,25.13500,0,0.553134,0.775933,-6.650471,0.254498,1.840198,0.103561

phon_R01_S13_4,126.34400,134.23100,112.77300,0.00448,0.00004,0.00131,0.00169,0.00393,0.02033,0.18500,0.01143,0.00959,0.01614,0.03429,0.00474,25.03000,0,0.507504,0.760361,-6.689151,0.291954,2.431854,0.105993

phon_R01_S13_5,128.00100,138.05200,122.08000,0.00436,0.00003,0.00137,0.00166,0.00411,0.02297,0.21000,0.01323,0.01072,0.01677,0.03969,0.00481,24.69200,0,0.459766,0.766204,-7.072419,0.220434,1.972297,0.119308

phon_R01_S13_6,129.33600,139.86700,118.60400,0.00490,0.00004,0.00165,0.00183,0.00495,0.02498,0.22800,0.01396,0.01219,0.01947,0.04188,0.00484,25.42900,0,0.420383,0.785714,-6.836811,0.269866,2.223719,0.147491

phon_R01_S16_1,108.80700,134.65600,102.87400,0.00761,0.00007,0.00349,0.00486,0.01046,0.02719,0.25500,0.01483,0.01609,0.02067,0.04450,0.01036,21.02800,1,0.536009,0.819032,-4.649573,0.205558,1.986899,0.316700

phon_R01_S16_2,109.86000,126.35800,104.43700,0.00874,0.00008,0.00398,0.00539,0.01193,0.03209,0.30700,0.01789,0.01992,0.02454,0.05368,0.01180,20.76700,1,0.558586,0.811843,-4.333543,0.221727,2.014606,0.344834

phon_R01_S16_3,110.41700,131.06700,103.37000,0.00784,0.00007,0.00352,0.00514,0.01056,0.03715,0.33400,0.02032,0.02302,0.02802,0.06097,0.00969,21.42200,1,0.541781,0.821364,-4.438453,0.238298,1.922940,0.335041

phon_R01_S16_4,117.27400,129.91600,110.40200,0.00752,0.00006,0.00299,0.00469,0.00898,0.02293,0.22100,0.01189,0.01459,0.01948,0.03568,0.00681,22.81700,1,0.530529,0.817756,-4.608260,0.290024,2.021591,0.314464

phon_R01_S16_5,116.87900,131.89700,108.15300,0.00788,0.00007,0.00334,0.00493,0.01003,0.02645,0.26500,0.01394,0.01625,0.02137,0.04183,0.00786,22.60300,1,0.540049,0.813432,-4.476755,0.262633,1.827012,0.326197

Mahasarakham University

phon_R01_S16_6,114.84700,271.31400,104.68000,0.00867,0.00008,0.00373,0.00520,0.01120,0.03225,0.35000,0.01805,0.01974,0.02519,0.05414,0.01143,21.66000,1,0.547975,0.817396,-4.609161,0.221711,1.831691,0.316395

phon_R01_S17_1,209.14400,237.49400,109.37900,0.00282,0.00001,0.00147,0.00152,0.00442,0.01861,0.17000,0.00975,0.01258,0.01382,0.02925,0.00871,25.55400,0,0.341788,0.678874,-7.040508,0.066994,2.460791,0.101516

phon_R01_S17_2,223.36500,238.98700,98.66400,0.00264,0.00001,0.00154,0.00151,0.00461,0.01906,0.16500,0.01013,0.01296,0.01340,0.03039,0.00301,26.13800,0,0.447979,0.686264,-7.293801,0.086372,2.321560,0.098555

phon_R01_S17_3,222.23600,231.34500,205.49500,0.00266,0.00001,0.00152,0.00144,0.00457,0.01643,0.14500,0.00867,0.01108,0.01200,0.02602,0.00340,25.85600,0,0.364867,0.694399,-6.966321,0.095882,2.278687,0.103224

phon_R01_S17_4,228.83200,234.61900,223.63400,0.00296,0.00001,0.00175,0.00155,0.00526,0.01644,0.14500,0.00882,0.01075,0.01179,0.02647,0.00351,25.96400,0,0.256570,0.683296,-7.245620,0.018689,2.498224,0.093534

phon_R01_S17_5,229.40100,252.22100,221.15600,0.00205,0.000009,0.00114,0.00113,0.00342,0.01457,0.12900,0.00769,0.00957,0.01016,0.02308,0.00300,26.41500,0,0.276850,0.673636,-7.496264,0.056844,2.003032,0.073581

phon_R01_S17_6,228.96900,239.54100,113.20100,0.00238,0.00001,0.00136,0.00140,0.00408,0.01745,0.15400,0.00942,0.01160,0.01234,0.02827,0.00420,24.54700,0,0.305429,0.681811,-7.314237,0.006274,2.118596,0.091546

phon_R01_S18_1,140.34100,159.77400,67.02100,0.00817,0.00006,0.00430,0.00440,0.01289,0.03198,0.31300,0.01830,0.01810,0.02428,0.05490,0.02183,19.56000,1,0.460139,0.720908,-5.409423,0.226850,2.359973,0.226156

phon_R01_S18_2,136.96900,166.60700,66.00400,0.00923,0.00007,0.00507,0.00463,0.01520,0.03111,0.30800,0.01638,0.01759,0.02603,0.04914,0.02659,19.97900,1,0.498133,0.729067,-5.324574,0.205660,2.291558,0.226247

phon_R01_S18_3,143.53300,162.21500,65.80900,0.01101,0.00008,0.00647,0.00467,0.01941,0.05384,0.47800,0.03152,0.02422,0.03392,0.09455,0.04882,20.33800,1,0.513237,0.731444,-5.869750,0.151814,2.118496,0.185580

Mahasarakham University

phon_R01_S18_4,148.09000,162.82400,67.34300,0.00762,0.00005,0.00467,0.00354,0.014
00,0.05428,0.49700,0.03357,0.02494,0.03635,0.10070,0.02431,21.71800,1,0.4874
07,0.727313,-6.261141,0.120956,2.137075,0.141958
phon_R01_S18_5,142.72900,162.40800,65.47600,0.00831,0.00006,0.00469,0.00419,0.014
07,0.03485,0.36500,0.01868,0.01906,0.02949,0.05605,0.02599,20.26400,1,0.4893
45,0.730387,-5.720868,0.158830,2.277927,0.180828
phon_R01_S18_6,136.35800,176.59500,65.75000,0.00971,0.00007,0.00534,0.00478,0.016
01,0.04978,0.48300,0.02749,0.02466,0.03736,0.08247,0.03361,18.57000,1,0.5432
99,0.733232,-5.207985,0.224852,2.642276,0.242981
phon_R01_S19_1,120.08000,139.71000,111.20800,0.00405,0.00003,0.00180,0.00220,0.00
540,0.01706,0.15200,0.00974,0.00925,0.01345,0.02921,0.00442,25.74200,1,0.495
954,0.762959,-5.791820,0.329066,2.205024,0.188180
phon_R01_S19_2,112.01400,588.51800,107.02400,0.00533,0.00005,0.00268,0.00329,0.00
805,0.02448,0.22600,0.01373,0.01375,0.01956,0.04120,0.00623,24.17800,1,0.509
127,0.789532,-5.389129,0.306636,1.928708,0.225461
phon_R01_S19_3,110.79300,128.10100,107.31600,0.00494,0.00004,0.00260,0.00283,0.00
780,0.02442,0.21600,0.01432,0.01325,0.01831,0.04295,0.00479,25.43800,1,0.437
031,0.815908,-5.313360,0.201861,2.225815,0.244512
phon_R01_S19_4,110.70700,122.61100,105.00700,0.00516,0.00005,0.00277,0.00289,0.00
831,0.02215,0.20600,0.01284,0.01219,0.01715,0.03851,0.00472,25.19700,1,0.463
514,0.807217,-5.477592,0.315074,1.862092,0.228624
phon_R01_S19_5,112.87600,148.82600,106.98100,0.00500,0.00004,0.00270,0.00289,0.00
810,0.03999,0.35000,0.02413,0.02231,0.02704,0.07238,0.00905,23.37000,1,0.489
538,0.789977,-5.775966,0.341169,2.007923,0.193918
phon_R01_S19_6,110.56800,125.39400,106.82100,0.00462,0.00004,0.00226,0.00280,0.00
677,0.02199,0.19700,0.01284,0.01199,0.01636,0.03852,0.00420,25.82000,1,0.429
484,0.816340,-5.391029,0.250572,1.777901,0.232744
phon_R01_S20_1,95.38500,102.14500,90.26400,0.00608,0.00006,0.00331,0.00332,0.0099
4,0.03202,0.26300,0.01803,0.01886,0.02455,0.05408,0.01062,21.87500,1,0.64495
4,0.779612,-5.115212,0.249494,2.017753,0.260015

Mahasarakham University

phon_R01_S20_2,100.77000,115.69700,85.54500,0.01038,0.00010,0.00622,0.00576,0.01865,0.03121,0.36100,0.01773,0.01783,0.02139,0.05320,0.02220,19.20000,1,0.594387,0.790117,-4.913885,0.265699,2.398422,0.277948

phon_R01_S20_3,96.10600,108.66400,84.51000,0.00694,0.00007,0.00389,0.00415,0.01168,0.04024,0.36400,0.02266,0.02451,0.02876,0.06799,0.01823,19.05500,1,0.544805,0.770466,-4.441519,0.155097,2.645959,0.327978

phon_R01_S20_4,95.60500,107.71500,87.54900,0.00702,0.00007,0.00428,0.00371,0.01283,0.03156,0.29600,0.01792,0.01841,0.02190,0.05377,0.01825,19.65900,1,0.576084,0.778747,-5.132032,0.210458,2.232576,0.260633

phon_R01_S20_5,100.96000,110.01900,95.62800,0.00606,0.00006,0.00351,0.00348,0.01053,0.02427,0.21600,0.01371,0.01421,0.01751,0.04114,0.01237,20.53600,1,0.554610,0.787896,-5.022288,0.146948,2.428306,0.264666

phon_R01_S20_6,98.80400,102.30500,87.80400,0.00432,0.00004,0.00247,0.00258,0.00742,0.02223,0.20200,0.01277,0.01343,0.01552,0.03831,0.00882,22.24400,1,0.576644,0.772416,-6.025367,0.078202,2.053601,0.177275

phon_R01_S21_1,176.85800,205.56000,75.34400,0.00747,0.00004,0.00418,0.00420,0.01254,0.04795,0.43500,0.02679,0.03022,0.03510,0.08037,0.05470,13.89300,1,0.556494,0.729586,-5.288912,0.343073,3.099301,0.242119

phon_R01_S21_2,180.97800,200.12500,155.49500,0.00406,0.00002,0.00220,0.00244,0.00659,0.03852,0.33100,0.02107,0.02493,0.02877,0.06321,0.02782,16.17600,1,0.583574,0.727747,-5.657899,0.315903,3.098256,0.200423

phon_R01_S21_3,178.22200,202.45000,141.04700,0.00321,0.00002,0.00163,0.00194,0.00488,0.03759,0.32700,0.02073,0.02415,0.02784,0.06219,0.03151,15.92400,1,0.598714,0.712199,-6.366916,0.335753,2.654271,0.144614

phon_R01_S21_4,176.28100,227.38100,125.61000,0.00520,0.00003,0.00287,0.00312,0.00862,0.06511,0.58000,0.03671,0.04159,0.04683,0.11012,0.04824,13.92200,1,0.602874,0.740837,-5.515071,0.299549,3.136550,0.220968

phon_R01_S21_5,173.89800,211.35000,74.67700,0.00448,0.00003,0.00237,0.00254,0.00710,0.06727,0.65000,0.03788,0.04254,0.04802,0.11363,0.04214,14.73900,1,0.599371,0.743937,-5.783272,0.299793,3.007096,0.194052

Mahasarakham University

phon_R01_S21_6,179.71100,225.93000,144.87800,0.00709,0.00004,0.00391,0.00419,0.01172,0.04313,0.44200,0.02297,0.02768,0.03455,0.06892,0.07223,11.86600,1,0.590951,0.745526,-4.379411,0.375531,3.671155,0.332086

phon_R01_S21_7,166.60500,206.00800,78.03200,0.00742,0.00004,0.00387,0.00453,0.01161,0.06640,0.63400,0.03650,0.04282,0.05114,0.10949,0.08725,11.74400,1,0.653410,0.733165,-4.508984,0.389232,3.317586,0.301952

phon_R01_S22_1,151.95500,163.33500,147.22600,0.00419,0.00003,0.00224,0.00227,0.00672,0.07959,0.77200,0.04421,0.04962,0.05690,0.13262,0.01658,19.66400,1,0.501037,0.714360,-6.411497,0.207156,2.344876,0.134120

phon_R01_S22_2,148.27200,164.98900,142.29900,0.00459,0.00003,0.00250,0.00256,0.00750,0.04190,0.38300,0.02383,0.02521,0.03051,0.07150,0.01914,18.78000,1,0.454444,0.734504,-5.952058,0.087840,2.344336,0.186489

phon_R01_S22_3,152.12500,161.46900,76.59600,0.00382,0.00003,0.00191,0.00226,0.00574,0.05925,0.63700,0.03341,0.03794,0.04398,0.10024,0.01211,20.96900,1,0.447456,0.697790,-6.152551,0.173520,2.080121,0.160809

phon_R01_S22_4,157.82100,172.97500,68.40100,0.00358,0.00002,0.00196,0.00196,0.00587,0.03716,0.30700,0.02062,0.02321,0.02764,0.06185,0.00850,22.21900,1,0.502380,0.712170,-6.251425,0.188056,2.143851,0.160812

phon_R01_S22_5,157.44700,163.26700,149.60500,0.00369,0.00002,0.00201,0.00197,0.00602,0.03272,0.28300,0.01813,0.01909,0.02571,0.05439,0.01018,21.69300,1,0.447285,0.705658,-6.247076,0.180528,2.344348,0.164916

phon_R01_S22_6,159.11600,168.91300,144.81100,0.00342,0.00002,0.00178,0.00184,0.00535,0.03381,0.30700,0.01806,0.02024,0.02809,0.05417,0.00852,22.66300,1,0.366329,0.693429,-6.417440,0.194627,2.473239,0.151709

phon_R01_S24_1,125.03600,143.94600,116.18700,0.01280,0.00010,0.00743,0.00623,0.02228,0.03886,0.34200,0.02135,0.02174,0.03088,0.06406,0.08151,15.33800,1,0.629574,0.714485,-4.020042,0.265315,2.671825,0.340623

phon_R01_S24_2,125.79100,140.55700,96.20600,0.01378,0.00011,0.00826,0.00655,0.02478,0.04689,0.42200,0.02542,0.02630,0.03908,0.07625,0.10323,15.43300,1,0.571010,0.690892,-5.159169,0.202146,2.441612,0.260375

Mahasarakham University

phon_R01_S24_3,126.51200,141.75600,99.77000,0.01936,0.00015,0.01159,0.00990,0.03476,0.06734,0.65900,0.03611,0.03963,0.05783,0.10833,0.16744,12.43500,1,0.638545,0.674953,-3.760348,0.242861,2.634633,0.378483

phon_R01_S24_4,125.64100,141.06800,116.34600,0.03316,0.00026,0.02144,0.01522,0.06433,0.09178,0.89100,0.05358,0.04791,0.06196,0.16074,0.31482,8.86700,1,0.671299,0.656846,-3.700544,0.260481,2.991063,0.370961

phon_R01_S24_5,128.45100,150.44900,75.63200,0.01551,0.00012,0.00905,0.00909,0.02716,0.06170,0.58400,0.03223,0.03672,0.05174,0.09669,0.11843,15.06000,1,0.639808,0.643327,-4.202730,0.310163,2.638279,0.356881

phon_R01_S24_6,139.22400,586.56700,66.15700,0.03011,0.00022,0.01854,0.01628,0.05563,0.09419,0.93000,0.05551,0.05005,0.06023,0.16654,0.25930,10.48900,1,0.596362,0.641418,-3.269487,0.270641,2.690917,0.444774

phon_R01_S25_1,150.25800,154.60900,75.34900,0.00248,0.00002,0.00105,0.00136,0.00315,0.01131,0.10700,0.00522,0.00659,0.01009,0.01567,0.00495,26.75900,1,0.296888,0.722356,-6.878393,0.089267,2.004055,0.113942

phon_R01_S25_2,154.00300,160.26700,128.62100,0.00183,0.00001,0.00076,0.00100,0.00229,0.01030,0.09400,0.00469,0.00582,0.00871,0.01406,0.00243,28.40900,1,0.263654,0.691483,-7.111576,0.144780,2.065477,0.093193

phon_R01_S25_3,149.68900,160.36800,133.60800,0.00257,0.00002,0.00116,0.00134,0.00349,0.01346,0.12600,0.00660,0.00818,0.01059,0.01979,0.00578,27.42100,1,0.365488,0.719974,-6.997403,0.210279,1.994387,0.112878

phon_R01_S25_4,155.07800,163.73600,144.14800,0.00168,0.00001,0.00068,0.00092,0.00204,0.01064,0.09700,0.00522,0.00632,0.00928,0.01567,0.00233,29.74600,1,0.334171,0.677930,-6.981201,0.184550,2.129924,0.106802

phon_R01_S25_5,151.88400,157.76500,133.75100,0.00258,0.00002,0.00115,0.00122,0.00346,0.01450,0.13700,0.00633,0.00788,0.01267,0.01898,0.00659,26.83300,1,0.393563,0.700246,-6.600023,0.249172,2.499148,0.105306

phon_R01_S25_6,151.98900,157.33900,132.85700,0.00174,0.00001,0.00075,0.00096,0.00225,0.01024,0.09300,0.00455,0.00576,0.00993,0.01364,0.00238,29.92800,1,0.311369,0.676066,-6.739151,0.160686,2.296873,0.115130

Mahasarakham University

phon_R01_S26_1,193.03000,208.90000,80.29700,0.00766,0.00004,0.00450,0.00389,0.013
51,0.03044,0.27500,0.01771,0.01815,0.02084,0.05312,0.00947,21.93400,1,0.4975
54,0.740539,-5.845099,0.278679,2.608749,0.185668
phon_R01_S26_2,200.71400,223.98200,89.68600,0.00621,0.00003,0.00371,0.00337,0.011
12,0.02286,0.20700,0.01192,0.01439,0.01852,0.03576,0.00704,23.23900,1,0.4360
84,0.727863,-5.258320,0.256454,2.550961,0.232520
phon_R01_S26_3,208.51900,220.31500,199.02000,0.00609,0.00003,0.00368,0.00339,0.01
105,0.01761,0.15500,0.00952,0.01058,0.01307,0.02855,0.00830,22.40700,1,0.338
097,0.712466,-6.471427,0.184378,2.502336,0.136390
phon_R01_S26_4,204.66400,221.30000,189.62100,0.00841,0.00004,0.00502,0.00485,0.01
506,0.02378,0.21000,0.01277,0.01483,0.01767,0.03831,0.01316,21.30500,1,0.498
877,0.722085,-4.876336,0.212054,2.376749,0.268144
phon_R01_S26_5,210.14100,232.70600,185.25800,0.00534,0.00003,0.00321,0.00280,0.00
964,0.01680,0.14900,0.00861,0.01017,0.01301,0.02583,0.00620,23.67100,1,0.441
097,0.722254,-5.963040,0.250283,2.489191,0.177807
phon_R01_S26_6,206.32700,226.35500,92.02000,0.00495,0.00002,0.00302,0.00246,0.009
05,0.02105,0.20900,0.01107,0.01284,0.01604,0.03320,0.01048,21.86400,1,0.3315
08,0.715121,-6.729713,0.181701,2.938114,0.115515
phon_R01_S27_1,151.87200,492.89200,69.08500,0.00856,0.00006,0.00404,0.00385,0.012
11,0.01843,0.23500,0.00796,0.00832,0.01271,0.02389,0.06051,23.69300,1,0.4077
01,0.662668,-4.673241,0.261549,2.702355,0.274407
phon_R01_S27_2,158.21900,442.55700,71.94800,0.00476,0.00003,0.00214,0.00207,0.006
42,0.01458,0.14800,0.00606,0.00747,0.01312,0.01818,0.01554,26.35600,1,0.4507
98,0.653823,-6.051233,0.273280,2.640798,0.170106
phon_R01_S27_3,170.75600,450.24700,79.03200,0.00555,0.00003,0.00244,0.00261,0.007
31,0.01725,0.17500,0.00757,0.00971,0.01652,0.02270,0.01802,25.69000,1,0.4867
38,0.676023,-4.597834,0.372114,2.975889,0.282780
phon_R01_S27_4,178.28500,442.82400,82.06300,0.00462,0.00003,0.00157,0.00194,0.004
72,0.01279,0.12900,0.00617,0.00744,0.01151,0.01851,0.00856,25.02000,1,0.4704
22,0.655239,-4.913137,0.393056,2.816781,0.251972

Mahasarakham University

phon_R01_S27_5,217.11600,233.48100,93.97800,0.00404,0.00002,0.00127,0.00128,0.00381,0.01299,0.12400,0.00679,0.00631,0.01075,0.02038,0.00681,24.58100,1,0.462516,0.582710,-5.517173,0.389295,2.925862,0.220657

phon_R01_S27_6,128.94000,479.69700,88.25100,0.00581,0.00005,0.00241,0.00314,0.00723,0.02008,0.22100,0.00849,0.01117,0.01734,0.02548,0.02350,24.74300,1,0.487756,0.684130,-6.186128,0.279933,2.686240,0.152428

phon_R01_S27_7,176.82400,215.29300,83.96100,0.00460,0.00003,0.00209,0.00221,0.00628,0.01169,0.11700,0.00534,0.00630,0.01104,0.01603,0.01161,27.16600,1,0.400088,0.656182,-4.711007,0.281618,2.655744,0.234809

phon_R01_S31_1,138.19000,203.52200,83.34000,0.00704,0.00005,0.00406,0.00398,0.01218,0.04479,0.44100,0.02587,0.02567,0.03220,0.07761,0.01968,18.30500,1,0.538016,0.741480,-5.418787,0.160267,2.090438,0.229892

phon_R01_S31_2,182.01800,197.17300,79.18700,0.00842,0.00005,0.00506,0.00449,0.01517,0.02503,0.23100,0.01372,0.01580,0.01931,0.04115,0.01813,18.78400,1,0.589956,0.732903,-5.445140,0.142466,2.174306,0.215558

phon_R01_S31_3,156.23900,195.10700,79.82000,0.00694,0.00004,0.00403,0.00395,0.01209,0.02343,0.22400,0.01289,0.01420,0.01720,0.03867,0.02020,19.19600,1,0.618663,0.728421,-5.944191,0.143359,1.929715,0.181988

phon_R01_S31_4,145.17400,198.10900,80.63700,0.00733,0.00005,0.00414,0.00422,0.01242,0.02362,0.23300,0.01235,0.01495,0.01944,0.03706,0.01874,18.85700,1,0.637518,0.735546,-5.594275,0.127950,1.765957,0.222716

phon_R01_S31_5,138.14500,197.23800,81.11400,0.00544,0.00004,0.00294,0.00327,0.00883,0.02791,0.24600,0.01484,0.01805,0.02259,0.04451,0.01794,18.17800,1,0.623209,0.738245,-5.540351,0.087165,1.821297,0.214075

phon_R01_S31_6,166.88800,198.96600,79.51200,0.00638,0.00004,0.00368,0.00351,0.01104,0.02857,0.25700,0.01547,0.01859,0.02301,0.04641,0.01796,18.33000,1,0.585169,0.736964,-5.825257,0.115697,1.996146,0.196535

phon_R01_S32_1,119.03100,127.53300,109.21600,0.00440,0.00004,0.00214,0.00192,0.00641,0.01033,0.09800,0.00538,0.00570,0.00811,0.01614,0.01724,26.84200,1,0.457541,0.699787,-6.890021,0.152941,2.328513,0.112856

Mahasarakham University

phon_R01_S32_2,120.07800,126.63200,105.66700,0.00270,0.00002,0.00116,0.00135,0.00349,0.01022,0.09000,0.00476,0.00588,0.00903,0.01428,0.00487,26.36900,1,0.491345,0.718839,-5.892061,0.195976,2.108873,0.183572

phon_R01_S32_3,120.28900,128.14300,100.20900,0.00492,0.00004,0.00269,0.00238,0.00808,0.01412,0.12500,0.00703,0.00820,0.01194,0.02110,0.01610,23.94900,1,0.467160,0.724045,-6.135296,0.203630,2.539724,0.169923

phon_R01_S32_4,120.25600,125.30600,104.77300,0.00407,0.00003,0.00224,0.00205,0.00671,0.01516,0.13800,0.00721,0.00815,0.01310,0.02164,0.01015,26.01700,1,0.468621,0.735136,-6.112667,0.217013,2.527742,0.170633

phon_R01_S32_5,119.05600,125.21300,86.79500,0.00346,0.00003,0.00169,0.00170,0.00508,0.01201,0.10600,0.00633,0.00701,0.00915,0.01898,0.00903,23.38900,1,0.470972,0.721308,-5.436135,0.254909,2.516320,0.232209

phon_R01_S32_6,118.74700,123.72300,109.83600,0.00331,0.00003,0.00168,0.00171,0.00504,0.01043,0.09900,0.00490,0.00621,0.00903,0.01471,0.00504,25.61900,1,0.482296,0.723096,-6.448134,0.178713,2.034827,0.141422

phon_R01_S33_1,106.51600,112.77700,93.10500,0.00589,0.00006,0.00291,0.00319,0.00873,0.04932,0.44100,0.02683,0.03112,0.03651,0.08050,0.03031,17.06000,1,0.637814,0.744064,-5.301321,0.320385,2.375138,0.243080

phon_R01_S33_2,110.45300,127.61100,105.55400,0.00494,0.00004,0.00244,0.00315,0.00731,0.04128,0.37900,0.02229,0.02592,0.03316,0.06688,0.02529,17.70700,1,0.653427,0.706687,-5.333619,0.322044,2.631793,0.228319

phon_R01_S33_3,113.40000,133.34400,107.81600,0.00451,0.00004,0.00219,0.00283,0.00658,0.04879,0.43100,0.02385,0.02973,0.04370,0.07154,0.02278,19.01300,1,0.647900,0.708144,-4.378916,0.300067,2.445502,0.259451

phon_R01_S33_4,113.16600,130.27000,100.67300,0.00502,0.00004,0.00257,0.00312,0.00772,0.05279,0.47600,0.02896,0.03347,0.04134,0.08689,0.03690,16.74700,1,0.625362,0.708617,-4.654894,0.304107,2.672362,0.274387

phon_R01_S33_5,112.23900,126.60900,104.09500,0.00472,0.00004,0.00238,0.00290,0.00715,0.05643,0.51700,0.03070,0.03530,0.04451,0.09211,0.02629,17.36600,1,0.640945,0.701404,-5.634576,0.306014,2.419253,0.209191

Mahasarakham University

phon_R01_S33_6,116.15000,131.73100,109.81500,0.00381,0.00003,0.00181,0.00232,0.00542,0.03026,0.26700,0.01514,0.01812,0.02770,0.04543,0.01827,18.80100,1,0.624811,0.696049,-5.866357,0.233070,2.445646,0.184985

phon_R01_S34_1,170.36800,268.79600,79.54300,0.00571,0.00003,0.00232,0.00269,0.00696,0.03273,0.28100,0.01713,0.01964,0.02824,0.05139,0.02485,18.54000,1,0.677131,0.685057,-4.796845,0.397749,2.963799,0.277227

phon_R01_S34_2,208.08300,253.79200,91.80200,0.00757,0.00004,0.00428,0.00428,0.01285,0.06725,0.57100,0.04016,0.04003,0.04464,0.12047,0.04238,15.64800,1,0.606344,0.665945,-5.410336,0.288917,2.665133,0.231723

phon_R01_S34_3,198.45800,219.29000,148.69100,0.00376,0.00002,0.00182,0.00215,0.00546,0.03527,0.29700,0.02055,0.02076,0.02530,0.06165,0.01728,18.70200,1,0.606273,0.661735,-5.585259,0.310746,2.465528,0.209863

phon_R01_S34_4,202.80500,231.50800,86.23200,0.00370,0.00002,0.00189,0.00211,0.00568,0.01997,0.18000,0.01117,0.01177,0.01506,0.03350,0.02010,18.68700,1,0.536102,0.632631,-5.898673,0.213353,2.470746,0.189032

phon_R01_S34_5,202.54400,241.35000,164.16800,0.00254,0.00001,0.00100,0.00133,0.00301,0.02662,0.22800,0.01475,0.01558,0.02006,0.04426,0.01049,20.68000,1,0.497480,0.630409,-6.132663,0.220617,2.576563,0.159777

phon_R01_S34_6,223.36100,263.87200,87.63800,0.00352,0.00002,0.00169,0.00188,0.00506,0.02536,0.22500,0.01379,0.01478,0.01909,0.04137,0.01493,20.36600,1,0.566849,0.574282,-5.456811,0.345238,2.840556,0.232861

phon_R01_S35_1,169.77400,191.75900,151.45100,0.01568,0.00009,0.00863,0.00946,0.02589,0.08143,0.82100,0.03804,0.05426,0.08808,0.11411,0.07530,12.35900,1,0.561610,0.793509,-3.297668,0.414758,3.413649,0.457533

phon_R01_S35_2,183.52000,216.81400,161.34000,0.01466,0.00008,0.00849,0.00819,0.02546,0.06050,0.61800,0.02865,0.04101,0.06359,0.08595,0.06057,14.36700,1,0.478024,0.768974,-4.276605,0.355736,3.142364,0.336085

phon_R01_S35_3,188.62000,216.30200,165.98200,0.01719,0.00009,0.00996,0.01027,0.02987,0.07118,0.72200,0.03474,0.04580,0.06824,0.10422,0.08069,12.29800,1,0.552870,0.764036,-3.377325,0.335357,3.274865,0.418646

Mahasarakham University

phon_R01_S35_4,202.63200,565.74000,177.25800,0.01627,0.00008,0.00919,0.00963,0.02756,0.07170,0.83300,0.03515,0.04265,0.06460,0.10546,0.07889,14.98900,1,0.427627,0.775708,-4.892495,0.262281,2.910213,0.270173

phon_R01_S35_5,186.69500,211.96100,149.44200,0.01872,0.00010,0.01075,0.01154,0.03225,0.05830,0.78400,0.02699,0.03714,0.06259,0.08096,0.10952,12.52900,1,0.507826,0.762726,-4.484303,0.340256,2.958815,0.301487

phon_R01_S35_6,192.81800,224.42900,168.79300,0.03107,0.00016,0.01800,0.01958,0.05401,0.11908,1.30200,0.05647,0.07940,0.13778,0.16942,0.21713,8.44100,1,0.625866,0.768320,-2.434031,0.450493,3.079221,0.527367

phon_R01_S35_7,198.11600,233.09900,174.47800,0.02714,0.00014,0.01568,0.01699,0.04705,0.08684,1.01800,0.04284,0.05556,0.08318,0.12851,0.16265,9.44900,1,0.584164,0.754449,-2.839756,0.356224,3.184027,0.454721

phon_R01_S37_1,121.34500,139.64400,98.25000,0.00684,0.00006,0.00388,0.00332,0.01164,0.02534,0.24100,0.01340,0.01399,0.02056,0.04019,0.04179,21.52000,1,0.566867,0.670475,-4.865194,0.246404,2.013530,0.168581

phon_R01_S37_2,119.10000,128.44200,88.83300,0.00692,0.00006,0.00393,0.00300,0.01179,0.02682,0.23600,0.01484,0.01405,0.02018,0.04451,0.04611,21.82400,1,0.651680,0.659333,-4.239028,0.175691,2.451130,0.247455

phon_R01_S37_3,117.87000,127.34900,95.65400,0.00647,0.00005,0.00356,0.00300,0.01067,0.03087,0.27600,0.01659,0.01804,0.02402,0.04977,0.02631,22.43100,1,0.628300,0.652025,-3.583722,0.207914,2.439597,0.206256

phon_R01_S37_4,122.33600,142.36900,94.79400,0.00727,0.00006,0.00415,0.00339,0.01246,0.02293,0.22300,0.01205,0.01289,0.01771,0.03615,0.03191,22.95300,1,0.611679,0.623731,-5.435100,0.230532,2.699645,0.220546

phon_R01_S37_5,117.96300,134.20900,100.75700,0.01813,0.00015,0.01117,0.00718,0.03351,0.04912,0.43800,0.02610,0.02161,0.02916,0.07830,0.10748,19.07500,1,0.630547,0.646786,-3.444478,0.303214,2.964568,0.261305

phon_R01_S37_6,126.14400,154.28400,97.54300,0.00975,0.00008,0.00593,0.00454,0.01778,0.02852,0.26600,0.01500,0.01581,0.02157,0.04499,0.03828,21.53400,1,0.635015,0.627337,-5.070096,0.280091,2.892300,0.249703

Mahasarakham University

phon_R01_S39_1,127.93000,138.75200,112.17300,0.00605,0.00005,0.00321,0.00318,0.00962,0.03235,0.33900,0.01360,0.01650,0.03105,0.04079,0.02663,19.65100,1,0.654945,0.675865,-5.498456,0.234196,2.103014,0.216638

phon_R01_S39_2,114.23800,124.39300,77.02200,0.00581,0.00005,0.00299,0.00316,0.00896,0.04009,0.40600,0.01579,0.01994,0.04114,0.04736,0.02073,20.43700,1,0.653139,0.694571,-5.185987,0.259229,2.151121,0.244948

phon_R01_S39_3,115.32200,135.73800,107.80200,0.00619,0.00005,0.00352,0.00329,0.01057,0.03273,0.32500,0.01644,0.01722,0.02931,0.04933,0.02810,19.38800,1,0.577802,0.684373,-5.283009,0.226528,2.442906,0.238281

phon_R01_S39_4,114.55400,126.77800,91.12100,0.00651,0.00006,0.00366,0.00340,0.01097,0.03658,0.36900,0.01864,0.01940,0.03091,0.05592,0.02707,18.95400,1,0.685151,0.719576,-5.529833,0.242750,2.408689,0.220520

phon_R01_S39_5,112.15000,131.66900,97.52700,0.00519,0.00005,0.00291,0.00284,0.00873,0.01756,0.15500,0.00967,0.01033,0.01363,0.02902,0.01435,21.21900,1,0.557045,0.673086,-5.617124,0.184896,1.871871,0.212386

phon_R01_S39_6,102.27300,142.83000,85.90200,0.00907,0.00009,0.00493,0.00461,0.01480,0.02814,0.27200,0.01579,0.01553,0.02073,0.04736,0.03882,18.44700,1,0.671378,0.674562,-2.929379,0.396746,2.560422,0.367233

phon_R01_S42_1,236.20000,244.66300,102.13700,0.00277,0.00001,0.00154,0.00153,0.00462,0.02448,0.21700,0.01410,0.01426,0.01621,0.04231,0.00620,24.07800,0,0.469928,0.628232,-6.816086,0.172270,2.235197,0.119652

phon_R01_S42_2,237.32300,243.70900,229.25600,0.00303,0.00001,0.00173,0.00159,0.00519,0.01242,0.11600,0.00696,0.00747,0.00882,0.02089,0.00533,24.67900,0,0.384868,0.626710,-7.018057,0.176316,1.852402,0.091604

phon_R01_S42_3,260.10500,264.91900,237.30300,0.00339,0.00001,0.00205,0.00186,0.00616,0.02030,0.19700,0.01186,0.01230,0.01367,0.03557,0.00910,21.08300,0,0.440988,0.628058,-7.517934,0.160414,1.881767,0.075587

phon_R01_S42_4,197.56900,217.62700,90.79400,0.00803,0.00004,0.00490,0.00448,0.01470,0.02177,0.18900,0.01279,0.01272,0.01439,0.03836,0.01337,19.26900,0,0.372222,0.725216,-5.736781,0.164529,2.882450,0.202879

Mahasarakham University

phon_R01_S42_5,240.30100,245.13500,219.78300,0.00517,0.00002,0.00316,0.00283,0.00949,0.02018,0.21200,0.01176,0.01191,0.01344,0.03529,0.00965,21.02000,0,0.371837,0.646167,-7.169701,0.073298,2.266432,0.100881

phon_R01_S42_6,244.99000,272.21000,239.17000,0.00451,0.00002,0.00279,0.00237,0.00837,0.01897,0.18100,0.01084,0.01121,0.01255,0.03253,0.01049,21.52800,0,0.522812,0.646818,-7.304500,0.171088,2.095237,0.096220

phon_R01_S43_1,112.54700,133.37400,105.71500,0.00355,0.00003,0.00166,0.00190,0.00499,0.01358,0.12900,0.00664,0.00786,0.01140,0.01992,0.00435,26.43600,0,0.413295,0.756700,-6.323531,0.218885,2.193412,0.160376

phon_R01_S43_2,110.73900,113.59700,100.13900,0.00356,0.00003,0.00170,0.00200,0.00510,0.01484,0.13300,0.00754,0.00950,0.01285,0.02261,0.00430,26.55000,0,0.369090,0.776158,-6.085567,0.192375,1.889002,0.174152

phon_R01_S43_3,113.71500,116.44300,96.91300,0.00349,0.00003,0.00171,0.00203,0.00514,0.01472,0.13300,0.00748,0.00905,0.01148,0.02245,0.00478,26.54700,0,0.380253,0.766700,-5.943501,0.192150,1.852542,0.179677

phon_R01_S43_4,117.00400,144.46600,99.92300,0.00353,0.00003,0.00176,0.00218,0.00528,0.01657,0.14500,0.00881,0.01062,0.01318,0.02643,0.00590,25.44500,0,0.387482,0.756482,-6.012559,0.229298,1.872946,0.163118

phon_R01_S43_5,115.38000,123.10900,108.63400,0.00332,0.00003,0.00160,0.00199,0.00480,0.01503,0.13700,0.00812,0.00933,0.01133,0.02436,0.00401,26.00500,0,0.405991,0.761255,-5.966779,0.197938,1.974857,0.184067

phon_R01_S43_6,116.38800,129.03800,108.97000,0.00346,0.00003,0.00169,0.00213,0.00507,0.01725,0.15500,0.00874,0.01021,0.01331,0.02623,0.00415,26.14300,0,0.361232,0.763242,-6.016891,0.109256,2.004719,0.174429

phon_R01_S44_1,151.73700,190.20400,129.85900,0.00314,0.00002,0.00135,0.00162,0.00406,0.01469,0.13200,0.00728,0.00886,0.01230,0.02184,0.00570,24.15100,1,0.396610,0.745957,-6.486822,0.197919,2.449763,0.132703

phon_R01_S44_2,148.79000,158.35900,138.99000,0.00309,0.00002,0.00152,0.00186,0.00456,0.01574,0.14200,0.00839,0.00956,0.01309,0.02518,0.00488,24.41200,1,0.402591,0.762508,-6.311987,0.182459,2.251553,0.160306

Mahasarakham University

phon_R01_S44_3,148.14300,155.98200,135.04100,0.00392,0.00003,0.00204,0.00231,0.00612,0.01450,0.13100,0.00725,0.00876,0.01263,0.02175,0.00540,23.68300,1,0.398499,0.778349,-5.711205,0.240875,2.845109,0.192730

phon_R01_S44_4,150.44000,163.44100,144.73600,0.00396,0.00003,0.00206,0.00233,0.00619,0.02551,0.23700,0.01321,0.01574,0.02148,0.03964,0.00611,23.13300,1,0.352396,0.759320,-6.261446,0.183218,2.264226,0.144105

phon_R01_S44_5,148.46200,161.07800,141.99800,0.00397,0.00003,0.00202,0.00235,0.00605,0.01831,0.16300,0.00950,0.01103,0.01559,0.02849,0.00639,22.86600,1,0.408598,0.768845,-5.704053,0.216204,2.679185,0.197710

phon_R01_S44_6,149.81800,163.41700,144.78600,0.00336,0.00002,0.00174,0.00198,0.00521,0.02145,0.19800,0.01155,0.01341,0.01666,0.03464,0.00595,23.00800,1,0.329577,0.757180,-6.277170,0.109397,2.209021,0.156368

phon_R01_S49_1,117.22600,123.92500,106.65600,0.00417,0.00004,0.00186,0.00270,0.00558,0.01909,0.17100,0.00864,0.01223,0.01949,0.02592,0.00955,23.07900,0,0.603515,0.669565,-5.619070,0.191576,2.027228,0.215724

phon_R01_S49_2,116.84800,217.55200,99.50300,0.00531,0.00005,0.00260,0.00346,0.00780,0.01795,0.16300,0.00810,0.01144,0.01756,0.02429,0.01179,22.08500,0,0.663842,0.656516,-5.198864,0.206768,2.120412,0.252404

phon_R01_S49_3,116.28600,177.29100,96.98300,0.00314,0.00003,0.00134,0.00192,0.00403,0.01564,0.13600,0.00667,0.00990,0.01691,0.02001,0.00737,24.19900,0,0.598515,0.654331,-5.592584,0.133917,2.058658,0.214346

phon_R01_S49_4,116.55600,592.03000,86.22800,0.00496,0.00004,0.00254,0.00263,0.00762,0.01660,0.15400,0.00820,0.00972,0.01491,0.02460,0.01397,23.95800,0,0.566424,0.667654,-6.431119,0.153310,2.161936,0.120605

phon_R01_S49_5,116.34200,581.28900,94.24600,0.00267,0.00002,0.00115,0.00148,0.00345,0.01300,0.11700,0.00631,0.00789,0.01144,0.01892,0.00680,25.02300,0,0.528485,0.663884,-6.359018,0.116636,2.152083,0.138868

phon_R01_S49_6,114.56300,119.16700,86.64700,0.00327,0.00003,0.00146,0.00184,0.00439,0.01185,0.10600,0.00557,0.00721,0.01095,0.01672,0.00703,24.77500,0,0.555303,0.659132,-6.710219,0.149694,1.913990,0.121777

Mahasarakham University

phon_R01_S50_1,201.77400,262.70700,78.22800,0.00694,0.00003,0.00412,0.00396,0.012
35,0.02574,0.25500,0.01454,0.01582,0.01758,0.04363,0.04441,19.36800,0,0.5084
79,0.683761,-6.934474,0.159890,2.316346,0.112838
phon_R01_S50_2,174.18800,230.97800,94.26100,0.00459,0.00003,0.00263,0.00259,0.007
90,0.04087,0.40500,0.02336,0.02498,0.02745,0.07008,0.02764,19.51700,0,0.4484
39,0.657899,-6.538586,0.121952,2.657476,0.133050
phon_R01_S50_3,209.51600,253.01700,89.48800,0.00564,0.00003,0.00331,0.00292,0.009
94,0.02751,0.26300,0.01604,0.01657,0.01879,0.04812,0.01810,19.14700,0,0.4316
74,0.683244,-6.195325,0.129303,2.784312,0.168895
phon_R01_S50_4,174.68800,240.00500,74.28700,0.01360,0.00008,0.00624,0.00564,0.018
73,0.02308,0.25600,0.01268,0.01365,0.01667,0.03804,0.10715,17.88300,0,0.4075
67,0.655683,-6.787197,0.158453,2.679772,0.131728
phon_R01_S50_5,198.76400,396.96100,74.90400,0.00740,0.00004,0.00370,0.00390,0.011
09,0.02296,0.24100,0.01265,0.01321,0.01588,0.03794,0.07223,19.02000,0,0.4512
21,0.643956,-6.744577,0.207454,2.138608,0.123306
phon_R01_S50_6,214.28900,260.27700,77.97300,0.00567,0.00003,0.00295,0.00317,0.008
85,0.01884,0.19000,0.01026,0.01161,0.01373,0.03078,0.04398,21.20900,0,0.4628
03,0.664357,-5.724056,0.190667,2.555477,0.148569

Mahasarakham University

ภาคผนวก ข

ผลการหาค่าพารามิเตอร์ C และ γ(gamma) ที่เหมาะสมที่สุด

Mahasarakham University

ตารางแสดงผลการหาค่าพารามิเตอร์ C และ γ(gamma) ที่เหมาะสมที่สุด

C	γ	Performance		C	γ	Performance		C	γ	Performance		C	γ	Performance		C	γ	Performance
0	0	0.546		2	2	0.895		4	4	0.936		6	6	0.963		8	8	0.948
0.1	0	0.634		2.1	2	0.864		4.1	4	0.961		6.1	6	0.946		8.1	8	0.946
0.2	0	0.534		2.2	2	0.829		4.2	4	0.961		6.2	6	0.934		8.2	8	0.959
0.3	0	0.634		2.3	2	0.780		4.3	4	0.898		6.3	6	0.934		8.3	8	0.963
0.4	0	0.621		2.4	2	0.864		4.4	4	0.938		6.4	6	0.938		8.4	8	0.961
0.5	0	0.584		2.5	2	0.854		4.5	4	0.898		6.5	6	0.921		8.5	8	0.961
0.6	0	0.534		2.6	2	0.868		4.6	4	0.986		6.6	6	0.946		8.6	8	0.936
0.7	0	0.559		2.7	2	0.871		4.7	4	0.911		6.7	6	0.950		8.7	8	0.923
0.8	0	0.609		2.8	2	0.818		4.8	4	0.938		6.8	6	0.932		8.8	8	0.895
0.9	0	0.584		2.9	2	0.855		4.9	4	0.963		6.9	6	0.946		8.9	8	0.946
1	0	0.596		3	2	0.880		5	4	0.934		7	6	0.963		9	8	0.950
1.1	0	0.609		3.1	2	0.932		5.1	4	0.934		7.1	6	0.961		9.1	8	0.948
1.2	0	0.668		3.2	2	0.882		5.2	4	0.932		7.2	6	0.961		9.2	8	0.963
1.3	0	0.643		3.3	2	0.841		5.3	4	0.850		7.3	6	0.920		9.3	8	0.909
1.4	0	0.718		3.4	2	0.907		5.4	4	0.921		7.4	6	0.921		9.4	8	0.932
1.5	0	0.695		3.5	2	0.921		5.5	4	0.945		7.5	6	0.898		9.5	8	0.975

Mahasarakham University

1.6	0	0.680		3.6	2	0.868		5.6	4	0.923		7.6	6	0.963		9.6	8	0.913
1.7	0	0.671		3.7	2	0.909		5.7	4	0.946		7.7	6	0.973		9.7	8	0.923
1.8	0	0.671		3.8	2	0.827		5.8	4	0.907		7.8	6	0.961		9.8	8	0.975
1.9	0	0.761		3.9	2	0.905		5.9	4	0.896		7.9	6	0.932		9.9	8	0.886
2	0	0.857		4	2	0.909		6	4	0.948		8	6	0.948		10	8	0.946
2.1	0	0.695		4.1	2	0.843		6.1	4	0.907		8.1	6	0.986		0	9	0.959
2.2	0	0.746		4.2	2	0.938		6.2	4	0.870		8.2	6	0.936		0.1	9	0.659
2.3	0	0.732		4.3	2	0.923		6.3	4	0.946		8.3	6	0.961		0.2	9	0.684
2.4	0	0.736		4.4	2	0.814		6.4	4	0.909		8.4	6	0.938		0.3	9	0.739
2.5	0	0.684		4.5	2	0.909		6.5	4	0.936		8.5	6	0.930		0.4	9	0.871
2.6	0	0.730		4.6	2	0.880		6.6	4	0.948		8.6	6	0.920		0.5	9	0.946
2.7	0	0.739		4.7	2	0.845		6.7	4	0.945		8.7	6	0.961		0.6	9	0.921
2.8	0	0.750		4.8	2	0.961		6.8	4	0.950		8.8	6	0.930		0.7	9	0.893
2.9	0	0.739		4.9	2	0.854		6.9	4	0.893		8.9	6	0.975		0.8	9	0.923
3	0	0.813		5	2	0.886		7	4	0.923		9	6	0.963		0.9	9	0.921
3.1	0	0.696		5.1	2	0.921		7.1	4	0.909		9.1	6	0.963		1	9	0.920
3.2	0	0.798		5.2	2	0.920		7.2	4	0.959		9.2	6	0.871		1.1	9	0.973
3.3	0	0.736		5.3	2	0.936		7.3	4	0.920		9.3	6	0.948		1.2	9	0.948

Mahasarakham University

3.4	0	0.779		5.4	2	0.843		7.4	4	0.963		9.4	6	0.948		1.3	9	0.895
3.5	0	0.718		5.5	2	0.907		7.5	4	0.934		9.5	6	0.936		1.4	9	0.921
3.6	0	0.707		5.6	2	0.938		7.6	4	0.921		9.6	6	0.959		1.5	9	0.886
3.7	0	0.668		5.7	2	0.936		7.7	4	0.943		9.7	6	0.975		1.6	9	0.934
3.8	0	0.730		5.8	2	0.934		7.8	4	0.950		9.8	6	0.938		1.7	9	0.946
3.9	0	0.770		5.9	2	0.934		7.9	4	0.946		9.9	6	0.950		1.8	9	0.946
4	0	0.709		6	2	0.961		8	4	0.907		10	6	0.923		1.9	9	0.938
4.1	0	0.688		6.1	2	0.911		8.1	4	0.936		0	7	0.870		2	9	0.959
4.2	0	0.754		6.2	2	0.896		8.2	4	0.934		0.1	7	0.671		2.1	9	0.945
4.3	0	0.770		6.3	2	0.932		8.3	4	0.930		0.2	7	0.773		2.2	9	0.971
4.4	0	0.736		6.4	2	0.879		8.4	4	0.893		0.3	7	0.886		2.3	9	0.934
4.5	0	0.748		6.5	2	0.861		8.5	4	0.934		0.4	7	0.920		2.4	9	0.923
4.6	0	0.700		6.6	2	0.925		8.6	4	0.936		0.5	7	0.936		2.5	9	0.973
4.7	0	0.725		6.7	2	0.896		8.7	4	0.948		0.6	7	0.911		2.6	9	0.948
4.8	0	0.679		6.8	2	0.880		8.8	4	0.932		0.7	7	0.909		2.7	9	0.920
4.9	0	0.752		6.9	2	0.893		8.9	4	0.907		0.8	7	0.909		2.8	9	0.921
5	0	0.668		7	2	0.936		9	4	0.946		0.9	7	0.882		2.9	9	0.870
5.1	0	0.750		7.1	2	0.911		9.1	4	0.920		1	7	0.948		3	9	0.932

Mahasarakham University

5.2	0	0.723		7.2	2	0.857		9.2	4	0.884		1.1	7	0.934		3.1	9	0.896
5.3	0	0.729		7.3	2	0.850		9.3	4	0.959		1.2	7	0.921		3.2	9	0.936
5.4	0	0.736		7.4	2	0.923		9.4	4	0.961		1.3	7	0.973		3.3	9	0.961
5.5	0	0.682		7.5	2	0.921		9.5	4	0.934		1.4	7	0.961		3.4	9	0.950
5.6	0	0.825		7.6	2	0.950		9.6	4	0.920		1.5	7	0.946		3.5	9	0.921
5.7	0	0.829		7.7	2	0.895		9.7	4	0.905		1.6	7	0.936		**3.6**	**9**	**1.000**
5.8	0	0.707		7.8	2	0.898		9.8	4	0.923		1.7	7	0.975		3.7	9	0.934
5.9	0	0.741		7.9	2	0.904		9.9	4	0.936		1.8	7	0.938		3.8	9	0.936
6	0	0.759		8	2	0.923		10	4	0.959		1.9	7	0.946		3.9	9	0.946
6.1	0	0.707		8.1	2	0.870		0	5	0.936		2	7	0.948		4	9	0.909
6.2	0	0.680		8.2	2	0.932		0.1	5	0.634		2.1	7	0.986		4.1	9	0.909
6.3	0	0.763		8.3	2	0.925		0.2	5	0.784		2.2	7	0.907		4.2	9	0.973
6.4	0	0.727		8.4	2	0.882		0.3	5	0.845		2.3	7	0.961		4.3	9	0.921
6.5	0	0.789		8.5	2	0.898		0.4	5	0.868		2.4	7	0.913		4.4	9	0.921
6.6	0	0.827		8.6	2	0.920		0.5	5	0.854		2.5	7	0.934		4.5	9	0.925
6.7	0	0.720		8.7	2	0.913		0.6	5	0.868		2.6	7	0.934		4.6	9	0.921
6.8	0	0.764		8.8	2	0.884		0.7	5	0.880		2.7	7	0.893		4.7	9	0.959
6.9	0	0.713		8.9	2	0.911		0.8	5	0.909		2.8	7	0.961		4.8	9	0.884

Mahasarakham University

7	0	0.657		9	2	0.855		0.9	5	0.896		2.9	7	0.936		4.9	9	0.921
7.1	0	0.718		9.1	2	0.921		1	5	0.882		3	7	0.988		5	9	0.961
7.2	0	0.682		9.2	2	0.957		1.1	5	0.829		3.1	7	0.934		5.1	9	0.946
7.3	0	0.687		9.3	2	0.836		1.2	5	0.895		3.2	7	0.948		5.2	9	0.934
7.4	0	0.752		9.4	2	0.855		1.3	5	0.895		3.3	7	0.948		5.3	9	0.921
7.5	0	0.736		9.5	2	0.830		1.4	5	0.909		3.4	7	0.884		5.4	9	0.975
7.6	0	0.764		9.6	2	0.866		1.5	5	0.923		3.5	7	0.920		5.5	9	0.882
7.7	0	0.709		9.7	2	0.936		1.6	5	0.961		3.6	7	0.936		5.6	9	0.975
7.8	0	0.716		9.8	2	0.855		1.7	5	0.905		3.7	7	0.946		5.7	9	0.891
7.9	0	0.737		9.9	2	0.891		1.8	5	0.920		3.8	7	0.973		5.8	9	0.948
8	0	0.761		10	2	0.888		1.9	5	0.909		3.9	7	0.946		5.9	9	0.934
8.1	0	0.682		0	3	0.909		2	5	0.920		4	7	0.896		6	9	0.886
8.2	0	0.723		0.1	3	0.654		2.1	5	0.907		4.1	7	0.916		6.1	9	0.905
8.3	0	0.743		0.2	3	0.793		2.2	5	0.843		4.2	7	0.961		6.2	9	0.946
8.4	0	0.721		0.3	3	0.871		2.3	5	0.946		4.3	7	0.921		6.3	9	0.882
8.5	0	0.677		0.4	3	0.818		2.4	5	0.895		4.4	7	0.946		6.4	9	0.907
8.6	0	0.739		0.5	3	0.859		2.5	5	0.932		4.5	7	0.950		6.5	9	0.959
8.7	0	0.752		0.6	3	0.866		2.6	5	0.963		4.6	7	0.948		6.6	9	0.918

Mahasarakham University

8.8	0	0.780		0.7	3	0.832		2.7	5	0.921		4.7	7	0.882		6.7	9	0.945
8.9	0	0.777		0.8	3	0.845		2.8	5	0.961		4.8	7	0.932		6.8	9	0.975
9	0	0.661		0.9	3	0.871		2.9	5	0.909		4.9	7	0.975		6.9	9	0.921
9.1	0	0.657		1	3	0.852		3	5	0.948		5	7	0.923		7	9	0.946
9.2	0	0.736		1.1	3	0.882		3.1	5	0.945		5.1	7	0.948		7.1	9	0.986
9.3	0	0.834		1.2	3	0.895		3.2	5	0.907		5.2	7	0.911		7.2	9	0.921
9.4	0	0.873		1.3	3	0.843		3.3	5	0.948		5.3	7	0.893		7.3	9	0.938
9.5	0	0.748		1.4	3	0.866		3.4	5	0.934		5.4	7	0.975		7.4	9	0.948
9.6	0	0.684		1.5	3	0.868		3.5	5	0.961		5.5	7	0.936		7.5	9	0.957
9.7	0	0.739		1.6	3	0.921		3.6	5	0.946		5.6	7	0.946		7.6	9	0.961
9.8	0	0.741		1.7	3	0.884		3.7	5	0.948		5.7	7	0.938		7.7	9	0.905
9.9	0	0.698		1.8	3	0.870		3.8	5	0.934		5.8	7	0.938		7.8	9	0.973
10	0	0.727		1.9	3	0.934		3.9	5	0.905		5.9	7	0.870		7.9	9	0.946
0	1	0.870		2	3	0.911		4	5	0.923		6	7	0.907		8	9	0.946
0.1	1	0.675		2.1	3	0.841		4.1	5	0.946		6.1	7	0.946		8.1	9	0.921
0.2	1	0.725		2.2	3	0.907		4.2	5	0.963		6.2	7	0.961		8.2	9	0.961
0.3	1	0.816		2.3	3	0.829		4.3	5	0.936		6.3	7	0.945		8.3	9	0.934
0.4	1	0.818		2.4	3	0.905		4.4	5	0.920		6.4	7	0.921		8.4	9	0.957

Mahasarakham University

0.5	1	0.789		2.5	3	0.920		4.5	5	0.886		6.5	7	0.950		8.5	9	0.905
0.6	1	0.789		2.6	3	0.857		4.6	5	0.943		6.6	7	0.948		8.6	9	0.923
0.7	1	0.805		2.7	3	0.880		4.7	5	0.950		6.7	7	0.938		8.7	9	0.920
0.8	1	0.791		2.8	3	0.921		4.8	5	0.932		6.8	7	0.963		8.8	9	0.959
0.9	1	0.845		2.9	3	0.923		4.9	5	0.945		6.9	7	0.975		8.9	9	0.932
1	1	0.830		3	3	0.909		5	5	0.923		7	7	0.973		9	9	0.905
1.1	1	0.798		3.1	3	0.909		5.1	5	0.946		7.1	7	0.896		9.1	9	0.961
1.2	1	0.779		3.2	3	0.893		5.2	5	0.934		7.2	7	0.921		9.2	9	0.973
1.3	1	0.843		3.3	3	0.946		5.3	5	0.948		7.3	7	0.923		9.3	9	0.948
1.4	1	0.788		3.4	3	0.909		5.4	5	0.921		7.4	7	0.909		9.4	9	0.934
1.5	1	0.745		3.5	3	0.905		5.5	5	0.934		7.5	7	0.907		9.5	9	0.932
1.6	1	0.775		3.6	3	0.946		5.6	5	0.934		7.6	7	0.870		9.6	9	0.905
1.7	1	0.852		3.7	3	0.882		5.7	5	0.920		7.7	7	0.948		9.7	9	0.896
1.8	1	0.779		3.8	3	0.920		5.8	5	0.911		7.8	7	0.961		9.8	9	0.945
1.9	1	0.870		3.9	3	0.896		5.9	5	0.934		7.9	7	0.938		9.9	9	0.936
2	1	0.827		4	3	0.895		6	5	0.945		8	7	0.923		10	9	0.923
2.1	1	0.795		4.1	3	0.911		6.1	5	0.907		8.1	7	0.930		0	10	0.911
2.2	1	0.779		4.2	3	0.916		6.2	5	0.934		8.2	7	0.963		0.1	10	0.646

Mahasarakham University

2.3	1	0.814		4.3	3	0.893		6.3	5	0.920		8.3	7	0.946		0.2	10	0.659
2.4	1	0.782		4.4	3	0.934		6.4	5	0.911		8.4	7	0.934		0.3	10	0.759
2.5	1	0.829		4.5	3	0.895		6.5	5	0.932		8.5	7	0.921		0.4	10	0.907
2.6	1	0.721		4.6	3	0.959		6.6	5	0.843		8.6	7	0.907		0.5	10	0.905
2.7	1	0.838		4.7	3	0.920		6.7	5	0.945		8.7	7	0.918		0.6	10	0.932
2.8	1	0.870		4.8	3	0.938		6.8	5	0.932		8.8	7	0.959		0.7	10	0.938
2.9	1	0.798		4.9	3	0.950		6.9	5	0.973		8.9	7	0.913		0.8	10	0.988
3	1	0.804		5	3	0.891		7	5	0.957		9	7	0.946		0.9	10	0.932
3.1	1	0.868		5.1	3	0.870		7.1	5	0.961		9.1	7	0.921		1	10	0.975
3.2	1	0.907		5.2	3	0.907		7.2	5	0.938		9.2	7	0.946		1.1	10	0.920
3.3	1	0.804		5.3	3	0.961		7.3	5	0.973		9.3	7	0.963		1.2	10	0.936
3.4	1	0.895		5.4	3	0.923		7.4	5	0.920		9.4	7	0.923		1.3	10	0.948
3.5	1	0.791		5.5	3	0.859		7.5	5	0.975		9.5	7	0.961		1.4	10	0.923
3.6	1	0.841		5.6	3	0.945		7.6	5	0.921		9.6	7	0.934		1.5	10	0.963
3.7	1	0.836		5.7	3	0.946		7.7	5	0.945		9.7	7	0.988		1.6	10	0.921
3.8	1	0.857		5.8	3	0.921		7.8	5	0.870		9.8	7	0.961		1.7	10	0.911
3.9	1	0.773		5.9	3	0.884		7.9	5	0.884		9.9	7	0.948		1.8	10	0.948
4	1	0.805		6	3	0.918		8	5	0.950		10	7	0.948		1.9	10	0.963

Mahasarakham University

4.1	1	0.857		6.1	3	0.936		8.1	5	0.938		0	8	0.909		2	10	0.921
4.2	1	0.882		6.2	3	0.945		8.2	5	0.934		0.1	8	0.621		2.1	10	0.909
4.3	1	0.893		6.3	3	0.963		8.3	5	0.907		0.2	8	0.670		2.2	10	0.975
4.4	1	0.823		6.4	3	0.936		8.4	5	0.932		0.3	8	0.884		2.3	10	0.961
4.5	1	0.843		6.5	3	0.920		8.5	5	0.934		0.4	8	0.871		2.4	10	0.904
4.6	1	0.789		6.6	3	0.925		8.6	5	0.946		0.5	8	0.905		2.5	10	0.921
4.7	1	0.807		6.7	3	0.934		8.7	5	0.961		0.6	8	0.918		2.6	10	0.934
4.8	1	0.843		6.8	3	0.934		8.8	5	0.948		0.7	8	0.948		2.7	10	0.946
4.9	1	0.832		6.9	3	0.911		8.9	5	0.911		0.8	8	0.963		2.8	10	0.938
5	1	0.852		7	3	0.961		9	5	0.988		0.9	8	0.893		2.9	10	0.923
5.1	1	0.855		7.1	3	0.918		9.1	5	0.934		1	8	0.948		3	10	0.920
5.2	1	0.921		7.2	3	0.879		9.2	5	0.896		1.1	8	0.921		3.1	10	0.918
5.3	1	0.825		7.3	3	0.938		9.3	5	0.959		1.2	8	0.959		3.2	10	0.957
5.4	1	0.841		7.4	3	0.861		9.4	5	0.857		1.3	8	0.896		3.3	10	0.934
5.5	1	0.907		7.5	3	0.891		9.5	5	0.934		1.4	8	0.907		3.4	10	0.920
5.6	1	0.864		7.6	3	0.959		9.6	5	0.896		1.5	8	0.870		3.5	10	0.963
5.7	1	0.813		7.7	3	0.898		9.7	5	0.896		1.6	8	0.905		3.6	10	0.950
5.8	1	0.884		7.8	3	0.882		9.8	5	0.936		1.7	8	0.934		3.7	10	0.975

Mahasarakham University

5.9	1	0.857		7.9	3	0.882		9.9	5	0.909		1.8	8	0.959		3.8	10	0.946
6	1	0.759		8	3	0.877		10	5	0.938		1.9	8	0.904		3.9	10	0.973
6.1	1	0.814		8.1	3	0.898		0	6	0.934		2	8	0.948		4	10	0.936
6.2	1	0.870		8.2	3	0.905		0.1	6	0.659		2.1	8	0.936		4.1	10	0.963
6.3	1	0.807		8.3	3	0.918		0.2	6	0.720		2.2	8	0.950		4.2	10	0.973
6.4	1	0.845		8.4	3	0.918		0.3	6	0.830		2.3	8	0.961		4.3	10	0.936
6.5	1	0.841		8.5	3	0.950		0.4	6	0.961		2.4	8	0.963		4.4	10	0.963
6.6	1	0.825		8.6	3	0.920		0.5	6	0.882		2.5	8	0.934		4.5	10	0.950
6.7	1	0.918		8.7	3	0.930		0.6	6	0.866		2.6	8	0.948		4.6	10	0.959
6.8	1	0.855		8.8	3	0.961		0.7	6	0.907		2.7	8	0.921		4.7	10	0.975
6.9	1	0.870		8.9	3	0.946		0.8	6	0.934		2.8	8	0.936		4.8	10	0.961
7	1	0.813		9	3	0.932		0.9	6	0.913		2.9	8	0.959		4.9	10	0.946
7.1	1	0.821		9.1	3	0.932		1	6	0.921		3	8	0.921		5	10	0.907
7.2	1	0.879		9.2	3	0.936		1.1	6	0.882		3.1	8	0.934		5.1	10	0.936
7.3	1	0.789		9.3	3	0.936		1.2	6	0.920		3.2	8	0.986		5.2	10	0.905
7.4	1	0.845		9.4	3	0.946		1.3	6	0.936		3.3	8	0.921		5.3	10	0.909
7.5	1	0.855		9.5	3	0.934		1.4	6	0.957		3.4	8	0.936		5.4	10	0.921
7.6	1	0.845		9.6	3	0.896		1.5	6	0.946		3.5	8	0.948		5.5	10	0.907

Mahasarakham University

7.7	1	0.798		9.7	3	0.921		1.6	6	0.896		3.6	8	0.882		5.6	10	0.920
7.8	1	0.814		9.8	3	0.896		1.7	6	0.936		3.7	8	0.946		5.7	10	0.959
7.9	1	0.880		9.9	3	0.896		1.8	6	0.988		3.8	8	0.959		5.8	10	0.936
8	1	0.932		10	3	0.986		1.9	6	0.936		3.9	8	0.923		5.9	10	0.921
8.1	1	0.811		0	4	0.921		2	6	0.893		4	8	0.932		6	10	0.948
8.2	1	0.868		0.1	4	0.648		2.1	6	0.932		4.1	8	0.921		6.1	10	0.918
8.3	1	0.880		0.2	4	0.818		2.2	6	0.950		4.2	8	0.946		6.2	10	0.920
8.4	1	0.866		0.3	4	0.805		2.3	6	0.909		4.3	8	0.948		6.3	10	0.938
8.5	1	0.857		0.4	4	0.859		2.4	6	0.920		4.4	8	0.886		6.4	10	0.973
8.6	1	0.895		0.5	4	0.852		2.5	6	0.948		4.5	8	0.923		6.5	10	0.907
8.7	1	0.854		0.6	4	0.845		2.6	6	0.911		4.6	8	0.934		6.6	10	0.946
8.8	1	0.893		0.7	4	0.893		2.7	6	0.948		4.7	8	0.914		6.7	10	0.975
8.9	1	0.920		0.8	4	0.895		2.8	6	0.936		4.8	8	0.948		6.8	10	0.936
9	1	0.839		0.9	4	0.909		2.9	6	0.946		4.9	8	0.911		6.9	10	0.948
9.1	1	0.846		1	4	0.896		3	6	0.975		5	8	0.911		7	10	0.959
9.2	1	0.804		1.1	4	0.895		3.1	6	0.921		5.1	8	0.925		7.1	10	0.973
9.3	1	0.857		1.2	4	0.948		3.2	6	0.923		5.2	8	0.963		7.2	10	0.930
9.4	1	0.854		1.3	4	0.891		3.3	6	0.932		5.3	8	0.934		7.3	10	0.945

Mahasarakham University

9.5	1	0.893		1.4	4	0.871		3.4	6	0.950		5.4	8	0.973		7.4	10	0.909
9.6	1	0.920		1.5	4	0.920		3.5	6	0.909		5.5	8	0.921		7.5	10	0.975
9.7	1	0.845		1.6	4	0.920		3.6	6	0.936		5.6	8	0.934		7.6	10	0.963
9.8	1	0.855		1.7	4	0.895		3.7	6	0.907		5.7	8	0.959		7.7	10	0.973
9.9	1	0.855		1.8	4	0.959		3.8	6	0.907		5.8	8	0.948		7.8	10	0.870
10	1	0.866		1.9	4	0.934		3.9	6	0.884		5.9	8	0.936		7.9	10	0.911
0	2	0.841		2	4	0.884		4	6	0.934		6	8	0.921		8	10	0.936
0.1	2	0.763		2.1	4	0.923		4.1	6	0.898		6.1	8	0.918		8.1	10	0.963
0.2	2	0.748		2.2	4	0.936		4.2	6	0.950		6.2	8	0.896		8.2	10	0.920
0.3	2	0.857		2.3	4	0.934		4.3	6	0.973		6.3	8	0.946		8.3	10	0.963
0.4	2	0.859		2.4	4	0.921		4.4	6	0.948		6.4	8	0.884		8.4	10	0.973
0.5	2	0.779		2.5	4	0.934		4.5	6	0.921		6.5	8	0.893		8.5	10	0.959
0.6	2	0.802		2.6	4	0.945		4.6	6	0.946		6.6	8	0.920		8.6	10	0.948
0.7	2	0.854		2.7	4	0.884		4.7	6	0.950		6.7	8	0.911		8.7	10	0.959
0.8	2	0.805		2.8	4	0.934		4.8	6	0.932		6.8	8	0.971		8.8	10	0.905
0.9	2	0.829		2.9	4	0.934		4.9	6	0.909		6.9	8	0.946		8.9	10	0.946
1	2	0.852		3	4	0.923		5	6	0.911		7	8	0.961		9	10	0.988
1.1	2	0.779		3.1	4	0.868		5.1	6	0.920		7.1	8	0.959		9.1	10	0.934

Mahasarakham University

1.2	2	0.846		3.2	4	0.946		5.2	6	0.936		7.2	8	0.950		9.2	10	0.946
1.3	2	0.877		3.3	4	0.923		5.3	6	0.961		7.3	8	0.934		9.3	10	0.959
1.4	2	0.839		3.4	4	0.868		5.4	6	0.961		7.4	8	0.920		9.4	10	0.946
1.5	2	0.907		3.5	4	0.938		5.5	6	0.914		7.5	8	0.946		9.5	10	0.921
1.6	2	0.845		3.6	4	0.907		5.6	6	0.959		7.6	8	0.936		9.6	10	0.948
1.7	2	0.888		3.7	4	0.932		5.7	6	0.930		7.7	8	0.882		9.7	10	0.920
1.8	2	0.868		3.8	4	0.884		5.8	6	0.963		7.8	8	0.934		9.8	10	0.938
1.9	2	0.846		3.9	4	0.909		5.9	6	0.945		7.9	8	0.948		9.9	10	0.920
																10	10	0.907

Mahasarakham University

ประวัติย่อผู้วิจัย

Mahasarakham University

ประวัติย่อผู้วิจัย

ชื่อ นามสกุล	นายณัฐพล แสนคำ
วัน เดือน ปีเกิด	วันที่ 27 มิถุนายน พ.ศ. 2516
จังหวัด และประเทศที่เกิด	จังหวัดขอนแก่น ประเทศไทย
ประวัติการศึกษา	พ.ศ. 2532 มัธยมศึกษาตอนต้น โรงเรียนกมลาไสย อำเภอกมลาไสย จังหวัดกาฬสินธุ์
	พ.ศ. 2535 มัธยมศึกษาตอนปลาย โรงเรียนสมเด็จพิทยาคม อำเภอสมเด็จ จังหวัดกาฬสินธุ์
	พ.ศ. 2539 ปริญญาวิทยาศาสตรบัณฑิต (วท.บ.) วิชาเอกวิทยาการคอมพิวเตอร์ สถาบันราชภัฏมหาสารคาม
	พ.ศ. 2546 ปริญญาวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาวิชาเทคโนโลยีสารสนเทศ สถาบันพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
	พ.ศ. 2559 ปริญญาปรัชญาดุษฎีบัณฑิต (ปร.ด.) สาขาวิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ มหาวิทยาลัยมหาสารคาม
ตำแหน่ง สถานที่ทำงาน	อาจารย์ มหาวิทยาลัยราชภัฏบุรีรัมย์ ตำบลในเมือง อำเภอเมืองบุรีรัมย์ จังหวัดบุรีรัมย์ 31000
ที่อยู่ที่สามารถติดต่อได้	บ้านเลขที่ 439/21 ถนนจิระ ตำบลในเมือง อำเภอเมือง จังหวัดบุรีรัมย์ 31000

Mahasarakham University