

บทที่ 2

เอกสารและงานวิจัยที่เกี่ยวข้อง

ในการทำวิจัยเรื่อง ระบบสนับสนุนการตัดสินใจในการเลือกใช้แพคเกจอินเทอร์เน็ตมือถือ โดยใช้ต้นไม้ตัดสินใจ ผู้วิจัยได้ทฤษฎี และเทคโนโลยี รวมถึงงานวิจัยที่เกี่ยวข้อง เพื่อใช้เป็นแนวทางสำหรับการดำเนินงานวิจัย ประกอบไปด้วยหัวข้อหลักๆ ดังนี้

- 2.1 แพคเกจอินเทอร์เน็ตมือถือ (Mobile Internet Package)
- 2.2 ระบบสนับสนุนการตัดสินใจ (Decision Support System)
- 2.3 การตัดสินใจแบบโครงสร้างต้นไม้ (Decision Tree)
- 2.4 ซอฟต์แวร์ Weka
- 2.5 งานวิจัยที่เกี่ยวข้อง

2.1 แพคเกจอินเทอร์เน็ตมือถือ (Mobile Internet Package)

อินเทอร์เน็ตมือถือ (Mobile Internet) เป็นการใช้งานอินเทอร์เน็ตผ่านอุปกรณ์ที่เรียกว่า โทรศัพท์มือถืออาจเพื่อการใช้งานจดหมายอิเล็กทรอนิกส์ หรือ อีเมล (e-Mail) การสนทนา (Chat) การค้นหาข้อมูล การดาวน์โหลด เกม เพลง ไฟล์ข้อมูล เป็นต้น โดยอาศัยเทคโนโลยี ดังนี้

WAP ย่อมาจาก Wireless Application Protocol เป็นมาตรฐานการสื่อสารสากลของการใช้งานโทรศัพท์มือถือ ผ่านทางอินเทอร์เน็ต โดยสามารถเชื่อมต่ออินเทอร์เน็ตผ่านทางโทรศัพท์มือถือได้

GPRS ย่อมาจาก General Packet Radio Service เป็นระบบการสื่อสารไร้สาย (Wireless) ที่สามารถรับ - ส่งข้อมูลด้วยความเร็วสูงสุดถึง 171.2 kbps ลักษณะการส่งข้อมูลจะมีการแบ่งย่อยออกเป็นส่วนๆ ที่เรียกว่า แพคเกจ (Package) และสามารถติดต่อไปยังอินเทอร์เน็ตได้โดยผ่านโทรศัพท์มือถือ หรือคอมพิวเตอร์

EDGE ย่อมาจาก Enhanced Data Rates for Global Evolution เป็นเทคโนโลยีระดับ G3 ที่ถูกกำหนดโดย ITU (International Telecommunication Union) เพื่อตอบสนองความต้องการในการรับส่งข้อมูลความเร็วสูงผ่านทางเครือข่ายไร้สาย โดยเฉพาะสำหรับมือถือในระบบ GSM โดยเครือข่าย EDGE นี้มีการพัฒนามาตามมาตรฐานเดียวกันกับ GPRS แต่มีความเร็วสูงถึง 236 kbps หรือถึง 473.6 kbps เทียบเท่ากับการใช้งาน ADSL

3G หรือ Third Generation เป็นเทคโนโลยีการสื่อสารที่ผสมผสานการนำเสนอข้อมูลและเทคโนโลยีเข้าด้วยกัน เช่น โทรศัพท์มือถือ กล้องถ่ายรูป และอินเทอร์เน็ต เป็นต้น โดยเทคโนโลยี

3G จะมีช่องสัญญาณความถี่ และความจุในการรับส่งข้อมูลที่มากขึ้น รวมถึงสามารถรับส่งข้อมูลด้วยความเร็วสูง ทำให้ประสิทธิภาพในการรับส่งข้อมูล และการใช้บริการมัลติมีเดียทำได้เต็มที่และสมบูรณ์แบบมากขึ้น

ด้วยเทคโนโลยีอินเทอร์เน็ตมือถือในปัจจุบันทำให้การใช้งานอินเทอร์เน็ตผ่านโทรศัพท์มือถือเป็นที่นิยมมากขึ้น ผู้ให้บริการด้านเครือข่ายโทรศัพท์มือถือต่างนำเสนอการให้บริการเกี่ยวกับอินเทอร์เน็ตบนมือถือที่หลากหลายมาก ทำให้เกิดเป็นแพคเกจอินเทอร์เน็ตมือถือที่มีให้เลือกใช้มากมายในปัจจุบัน

แพคเกจอินเทอร์เน็ตมือถือ (Mobile Internet Package) คือ นำเสนอแพคเกจการให้บริการเกี่ยวกับอินเทอร์เน็ตบนมือถือ อาจแบ่งได้ 2 รูปแบบ ดังนี้

2.1.1 แพคเกจอินเทอร์เน็ตแบบคิดตามเวลา คือ แพคเกจอินเทอร์เน็ตที่คิดตามเวลาที่ใช้งาน โดยมีให้เลือกตั้งแต่ แบบ 1 ชั่วโมงจนถึง ไม่จำกัดชั่วโมงการใช้งาน ซึ่งจะคิดการใช้งานอินเทอร์เน็ตบนมือถือแบบไม่จำกัดปริมาณการใช้งาน แต่คิดตามชั่วโมงการใช้งานจริง เช่น การใช้เซิร์ฟเวอร์ หรือเปิดหน้าเว็บไซต์ครั้งละ 5-10 นาที ระบบก็จะทำการหักเวลาการใช้งานไปเรื่อยๆ จนครบแพคเกจที่เลือกไว้

2.1.2 แพคเกจอินเทอร์เน็ตแบบคิดตามปริมาณการใช้งาน คือ คิดค่าบริการตามการใช้งานอินเทอร์เน็ตบนมือถือของผู้ใช้ โดยจำกัดปริมาณการใช้งานตามปริมาณในแพคเกจที่ผู้ใช้เลือก แต่ไม่จำกัดเวลา เช่น แพคเกจปริมาณการใช้งานแบบ 30 Mb ระบบจะทำการหักตามปริมาณที่ใช้ไปเรื่อยๆ จนครบปริมาณการใช้งานที่เลือกไว้

2.2 ระบบสนับสนุนการตัดสินใจ (Decision Support System)

ระบบสนับสนุนการตัดสินใจ คือ ระบบสารสนเทศที่ช่วยในการจัดเตรียมข้อมูล จัดรูปแบบหรือนำแบบจำลองในการแก้ปัญหา และข้อมูลที่มีความจำเป็นมาช่วยในการตัดสินใจ

2.2.1 ลักษณะของปัญหา

2.2.1.1 ปัญหาแบบมีโครงสร้าง (Structured Problem) เป็นปัญหาที่มีวิธีการแก้ไขปัญหาคืออย่างชัดเจนแน่นอน หรือสามารถจำลองปัญหาได้ด้วยสูตรทางคณิตศาสตร์ (แบบจำลองทางคณิตศาสตร์) และแทนค่าในสูตรจนสามารถคำนวณหาคำตอบได้อย่างชัดเจน หรือปัญหาที่ผู้ตัดสินใจมีข้อมูลและสารสนเทศประกอบการตัดสินใจอย่างครบถ้วนและสามารถนำไปใช้แก้ปัญหาได้โดยการเขียนโปรแกรม (กิตติ, 2550)

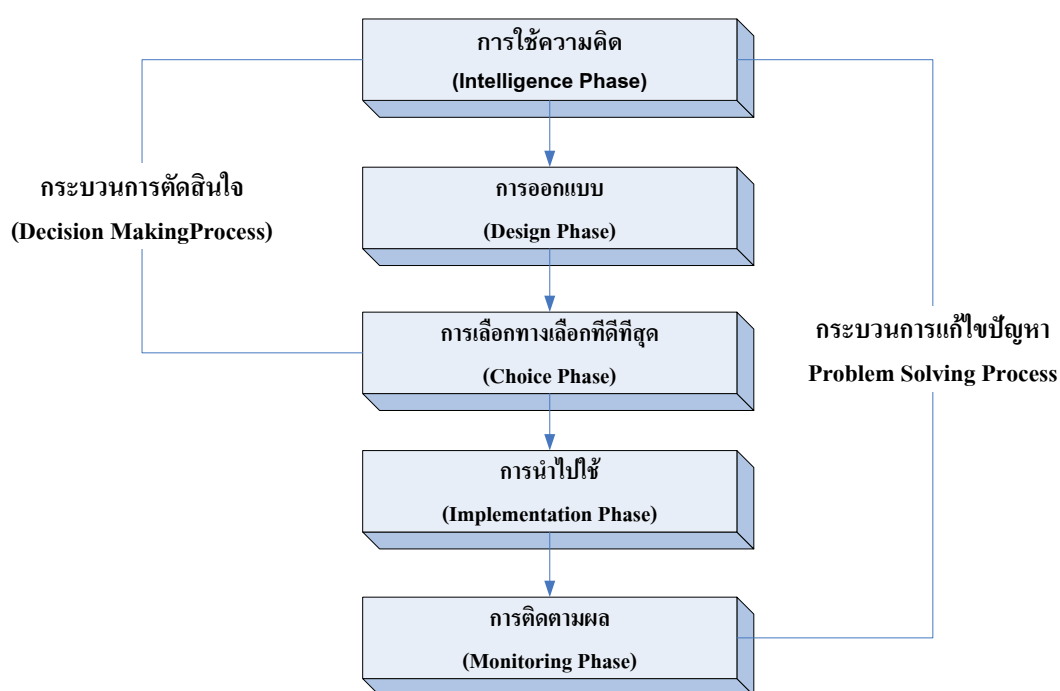
2.2.1.2 ปัญหาที่ไม่มีโครงสร้าง (Unstructured Problem) เป็นปัญหาที่ไม่สามารถหาวิธีในการแก้ไขได้อย่างชัดเจน และแน่นอน ไม่สามารถจำลองได้ด้วยสูตรทางคณิตศาสตร์หรือปัญหา

ที่ผู้ตัดสินใจที่มีข้อมูล และสารสนเทศไม่เพียงพอต่อการแก้ไขปัญหา จึงต้องอาศัยประสบการณ์ของผู้ตัดสินใจในการแก้ไขปัญหา (กิตติ, 2550)

2.2.1.3 ปัญหาแบบกึ่งโครงสร้าง (Semi Structured Problem) เป็นปัญหาแบบที่มีลักษณะเฉพาะ ส่วนมากจะไม่เกิดซ้ำ และไม่มีกระบวนการดำเนินการมาตรฐาน หรือเป็นปัญหาที่มีวิธีการแก้ไขเพียงบางส่วนเท่านั้น ส่วนที่เหลือจะต้องอาศัยประสบการณ์ หรือความชำนาญในการตัดสินใจแก้ไขปัญหา ส่วนเทคโนโลยีสารสนเทศให้ได้แค่การสนับสนุนเท่านั้น (กิตติ, 2550)

2.2.2 การตัดสินใจและการแก้ปัญหา

การตัดสินใจจัดว่าเป็นหนึ่งของกระบวนการแก้ปัญหาของมนุษย์เมื่อพบว่าปัญหาเกิดขึ้นในเรื่องหนึ่งเรื่องใดแล้ว การแก้ปัญหาก็ผ่านขั้นตอนการตัดสินใจเพื่อแก้ปัญหาต่อไป ส่วนกระบวนการตัดสินใจ (Decision Making Process) คือ การกำหนดขั้นตอนในการตัดสินใจแก้ปัญหาที่เกิดขึ้นภายในองค์กรอย่างมีหลักเกณฑ์ ด้วยการกำหนดขั้นตอนต่างๆ โดยผู้ที่กำหนดขั้นตอนคือ George Huber ได้นำมารวมเข้ากับกระบวนการแก้ปัญหา จึงทำให้กระบวนการตัดสินใจมีทั้งหมด 5 ขั้นตอน ดังภาพที่ 2-1



ภาพที่ 2-1 กระบวนการตัดสินใจและแก้ปัญหา (กิตติ, 2550)

กระบวนการตัดสินใจ (Decision Making Process) จะประกอบด้วยการใช้ความคิด (Intelligence Phase) การออกแบบ (Design Phase) การเลือกทางเลือกที่ดีที่สุด (Choice

Phase) โดยในส่วนของ การใช้ความคิดนั้นจะเป็นการระบุถึงปัญหาที่พบรวมทั้งจำแนกปัญหาออกมาเป็นส่วนย่อยๆ และคิดวิธีที่แก้ปัญหานั้นๆ โดยผลลัพธ์ในขั้นตอนนี้จะเรียกว่า “Decision Statement” ขั้นต่อไปเป็นการออกแบบเพื่อวิเคราะห์ทางเลือกที่ใช้ในการตัดสินใจ โดยขั้นตอนนี้อาจมีการสร้างแบบจำลองขึ้นมาเพื่อดูทางเลือกต่างๆ เช่นการสร้างแผนภาพการตัดสินใจด้วยต้นไม้ (Decision Tree) หรือ ตารางการตัดสินใจ (Decision Table) และในขั้นสุดท้ายของกระบวนการตัดสินใจ เป็นการเลือกทางเลือกที่ดีที่สุดเพื่อนำกระบวนการนั้นไปใช้ในการแก้ปัญหาจริง

กระบวนการแก้ไขปัญหา (Problem Solving Process) จะเป็นการดำเนินงานต่อจากกระบวนการตัดสินใจ โดยเพิ่ม การนำไปใช้ (Implementation Phase) และการติดตามผล (Monitoring Phase) ในส่วนของการนำไปใช้นั้นเป็นการลงมือปฏิบัติเพื่อแก้ปัญหามาจริงโดยอาจจะประสบความสำเร็จ หรือไม่ประสบความสำเร็จก็ได้ โดยขั้นตอนที่จะบอกสิ่งนี้ได้คือขั้นตอนการติดตามผล

2.3 การตัดสินใจแบบโครงสร้างต้นไม้ (Decision Tree)

เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เป็นเทคนิคหนึ่งของ Classification ซึ่งเป็นวิธีการแบ่งประเภทหรือแยกหมวดหมู่ข้อมูล โดย Classification นั้นเป็นเทคนิคหนึ่งของเหมืองข้อมูล (Data Mining)

2.3.1 เหมืองข้อมูล (Data Mining) คือ การวิเคราะห์ข้อมูล เพื่อแยกประเภท จำแนกรูปแบบและความสัมพันธ์ของข้อมูลจากฐานข้อมูลที่มีขนาดใหญ่หรือคลังข้อมูล โดยมีเทคนิคต่างๆหลายวิธี (กิตติ, 2550) ซึ่งรูปแบบการทำเหมืองข้อมูลนั้นได้รวมความรู้จากหลายแขนงเข้าไว้ด้วยกันที่ประกอบด้วย ระบบการเรียนรู้ของเครื่องจักร (Machine Learning) รวมกับวิทยาศาสตร์สารสนเทศ (Information Science) สถิติ (Statistics) และ ระบบฐานข้อมูล โดยทั่วไปแล้วเทคนิคที่นำมาใช้ส่วนใหญ่มี 5 ประเภท (ศุภชัย, 2551)

2.3.1.1 เทคนิค Classification เป็นเทคนิคในการจำแนกกลุ่มข้อมูลด้วยคุณลักษณะต่างๆที่ได้มีการกำหนดไว้แล้ว เทคนิคประเภทนี้เหมาะกับการสร้างแบบจำลองเพื่อการพยากรณ์ค่าข้อมูล (Predictive Modeling) ในอนาคตจากการที่ได้จำแนกกลุ่มข้อมูลตัวอย่างไว้แล้ว ซึ่งในลักษณะดังกล่าวนี้เรียกว่า “Supervised Learning” เทคนิคการ Classification มี 2 รูปแบบ ได้แก่ Tree Induction และ Neural Induction และเป็นกระบวนการสร้างแบบจำลองเพื่อจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดตัวอย่างเช่น การแบ่งประเภทลูกค้าว่า เชื่อถือได้หรือไม่ซึ่งเป็นการสร้างแบบจำลองโดยการเรียนรู้จากข้อมูลที่ได้กำหนดกลุ่มไว้เรียบร้อยแล้ว

2.3.1.2 เทคนิค Association Rule Discovery เป็นเทคนิคที่ใช้ในการค้นหาความสัมพันธ์ของฐานข้อมูลที่มีขนาดใหญ่ เพื่อที่จะทำการวิเคราะห์ข้อมูลและหาสิ่งที่ซ่อนอยู่ในข้อมูลนั้น เช่น การวิเคราะห์ข้อมูลการซื้อขายในซูเปอร์-มาร์เก็ต เพื่อทำการจัดทำตารางแผนเพื่อจัดการส่งเสริมการขาย (Promotion) และเตรียมการวางแผนการเรียงชั้นวางสินค้า (Shelf) เช่น การวางน้ำอัดลมกับข้าวโพดคั่วไว้ใกล้กัน

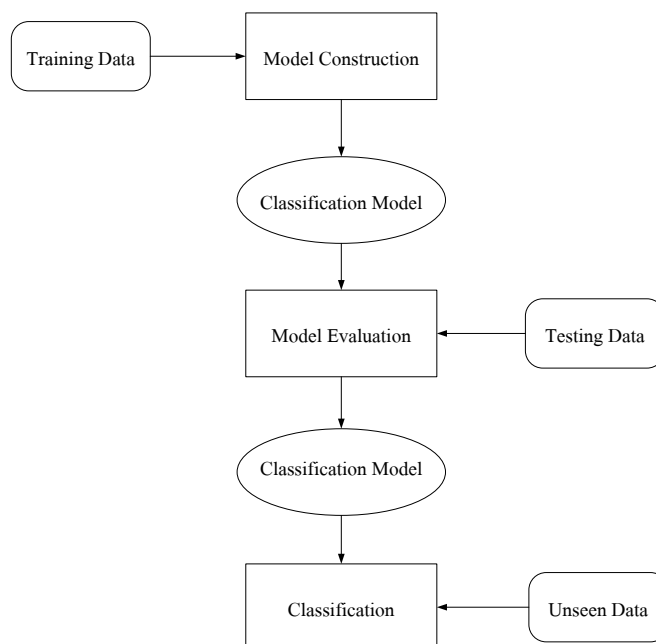
2.3.1.3 เทคนิค Clustering เป็นเทคนิคการลดขนาดของข้อมูลด้วยการรวมกลุ่มตัวแปรที่มีลักษณะเดียวกันไว้ด้วยกัน ทำให้สามารถค้นหาข้อมูลที่ถูกละเลยไปได้ เทคนิคนี้มักถูกใช้เป็นขั้นตอนเบื้องต้นในการทำเหมืองข้อมูล และเหมาะกับข้อมูลที่ยังไม่มีกลุ่มอย่างชัดเจนจึงทำการ Cluster เพื่อหากลุ่มต่างๆของข้อมูล โดยจำนวนกลุ่มของข้อมูลมักจะใช้ตัว k แทน ซึ่งผู้ที่ใช้เทคนิคนี้จะเป็นผู้กำหนดจำนวนกลุ่ม วิธีนี้อาจเรียกอีกอย่างว่า K-mean

2.3.1.4 เทคนิค Deviation Detection เป็นเทคนิควิธีในการหาค่าที่แตกต่างไปจากค่ามาตรฐาน หรือค่าที่คาดคิดไว้ว่าต่างไปมาน้อยเพียงใด โดยทั่วไปมักใช้วิธีการทางสถิติ หรือการแสดงให้เห็นภาพ สำหรับเทคนิคนี้ใช้ในการตรวจสอบ ลายเซ็นปลอม หรือบัตรเครดิตปลอม เป็นต้น

2.3.1.5 เทคนิค Sequential Analysis เป็นเทคนิคในการวิเคราะห์ลำดับเพื่อค้นพบรูปแบบของการปรากฏของข้อมูล ซึ่งปรากฏในรายการที่แยกออกมา เช่นถ้าผู้ซื้อซื้อสินค้า A แล้วเขาจะซื้อสินค้า B ในภายหลัง เทคนิคนี้จะแตกต่างจากเทคนิค Association Rule Discovery เพราะคำนึงถึงลำดับการซื้อด้วย

2.3.2 วิธีการแบ่งประเภทหรือแยกหมวดหมู่ข้อมูล Classification คือ กระบวนการสร้างแบบจำลองเพื่อจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนด เป็นการสร้างแบบจำลองการจัดหมวดหมู่ได้จากกลุ่มตัวอย่างของข้อมูลที่ได้กำหนดไว้ล่วงหน้า และสามารถพยากรณ์กลุ่มของข้อมูลที่ยังไม่เคยนำมาจัดหมวดหมู่ได้ แบบจำลองที่ได้อาจอยู่ในรูปแบบการตัดสินใจแบบต้นไม้ตัดสินใจ (Decision Tree) หรือแบบโครงข่ายประสาทเทียม (Neural Network)

ในการจัดหมวดหมู่จำเป็นต้องมีกลุ่มข้อมูลสำหรับการเรียนรู้ (Training Data) เพื่อให้ข้อมูลเรียนรู้ และสร้างแบบจำลอง (Model Construction) และทดสอบโดยกลุ่มข้อมูลสำหรับการทดสอบ (Testing Data) เพื่อประเมินความถูกต้องของโมเดล (Model Evaluation) อีกทั้งใช้ชุดข้อมูลที่ไม่เคยเห็นมาก่อน (Unseen Data) เพื่อทำการกำหนด Class ให้กับข้อมูลใหม่ที่ได้มา หรือทำนายค่าออกมาตามที่ต้องการ เช่น การจัดหมวดหมู่ของผู้ยื่นขอเครดิต (Credits) เป็นระดับต่ำระดับกลาง และระดับสูง ของความเสี่ยงที่จะได้รับ หรือ การอนุมัติบุคคลเข้ารับทำงานในลักษณะงานต่างๆ (พูน, 2548)



ภาพที่ 2-2 กระบวนการ Classification (พูน, 2548)

2.3.3 เทคนิคต้นไม้ตัดสินใจ (Decision Tree) คือ แบบจำลองทางคณิตศาสตร์ เพื่อการหาทางเลือกที่ดีที่สุด โดยการนำข้อมูลมาสร้างแบบจำลองการพยากรณ์ในรูปแบบของโครงสร้างต้นไม้ ซึ่งมีการเรียนรู้ข้อมูลแบบมีผู้สอน (Supervised Learning) สามารถสร้างแบบจำลองการจัดหมวดหมู่ (Clustering) ได้จากกลุ่มตัวอย่างของข้อมูลที่กำหนดไว้ล่วงหน้า (Training set) ได้โดยอัตโนมัติ และสามารถพยากรณ์กลุ่มของรายการที่ยังไม่เคยนำมาจัดหมวดหมู่ได้อีกด้วย

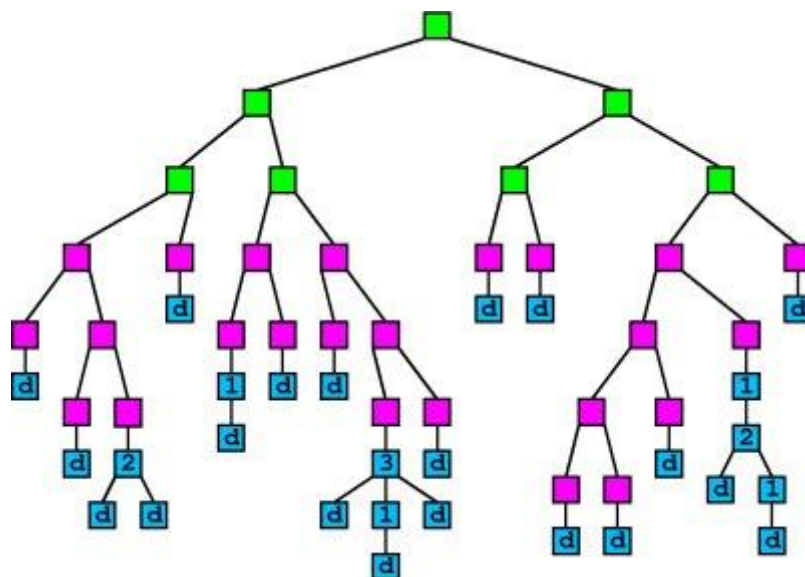
โดยปกติมักประกอบด้วยกฎในรูปแบบ “ถ้า เงื่อนไข แล้ว ผลลัพธ์” เช่น (พูน, 2548)

“If Income = High and Married = No THEN Risk = Poor”

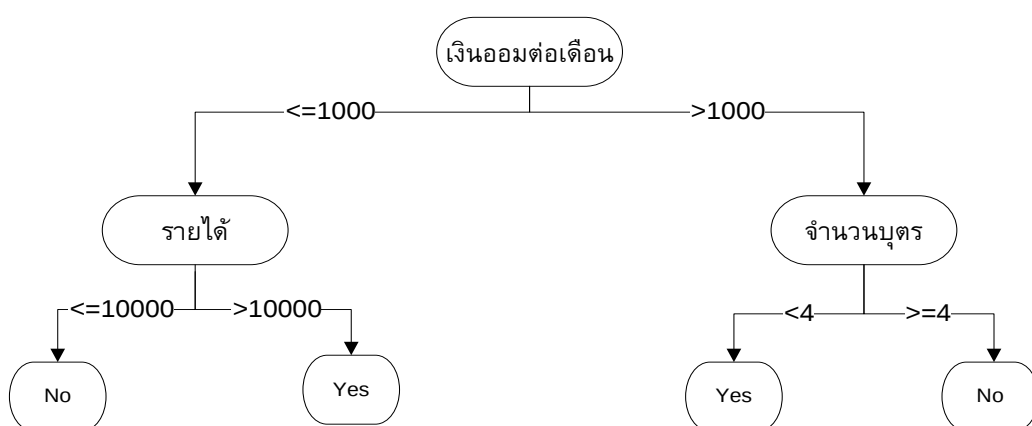
“If Income = High and Married = Yes THEN Risk = Good”

2.3.3.1 ส่วนประกอบของต้นไม้ตัดสินใจ ประกอบด้วย (จิตตภู, 2550)

- 1) โหนด (Node) คือ คุณสมบัติต่างๆ เป็นจุดที่แยกข้อมูลว่าจะให้ไปในทิศทางใด ซึ่งโหนดที่อยู่สูงสุดเรียกว่า โหนดราก (Root Node)
- 2) กิ่ง (Branch) คือ คุณสมบัติของคุณสมบัติในโหนดที่แตกออกมา โดยจำนวนของกิ่งจะเท่ากับคุณสมบัติของโหนด
- 3) ใบ (Leaf) คือ กลุ่มของผลลัพธ์ในการแยกแยะข้อมูล โดยสามารถแสดงส่วนประกอบของต้นไม้ตัดสินใจ ดังภาพที่ 1 และแสดงตัวอย่างของต้นไม้ตัดสินใจได้ดังภาพที่ 2-3



ภาพที่ 2-3 ส่วนประกอบของต้นไม้ตัดสินใจ



ภาพที่ 2-4 ตัวอย่างต้นไม้ตัดสินใจ (ศุภชัย ประครองศิลป์, 2551)

2.3.3.2 การสร้างต้นไม้ตัดสินใจ หลักการพื้นฐานของการสร้างต้นไม้ตัดสินใจ เป็นการสร้างในลักษณะจากบนลงล่าง (Top-Down) คือเริ่มจากการสร้างรากของต้นไม้ก่อนแล้วจึงแตกกิ่งไปจนถึงใบ โดยแสดงขั้นตอนการสร้างต้นไม้ตัดสินใจได้ดังนี้ (Han and Kamber, 2001)

- 1) ต้นไม้เริ่มต้นโดยมีโหนดเพียงโหนดเดียวแสดงถึงชุดข้อมูลฝึก (Training Set)
- 2) ถ้าข้อมูลทั้งหมดอยู่ในกลุ่มเดียวกันแล้ว ให้โหนดนั้นเป็นใบและตั้งชื่อแยกตามกลุ่มของข้อมูลนั้น
- 3) ถ้าในโหนดมีข้อมูลหลายกลุ่มปะปนอยู่ จะต้องวัดค่าเกน (Gain) ของแต่ละแอททริบิวต์เพื่อที่จะใช้เป็นเกณฑ์ในการคัดเลือกแอททริบิวต์ ที่มีความสามารถในการแบ่งแยก

ข้อมูลออกเป็นกลุ่มต่างๆ ได้ดีที่สุด โดยแอททริบิวต์ที่มีค่าเกินมากที่สุดจะถูกเลือกให้เป็นตัวทดสอบหรือแอททริบิวต์ใช้ในการตัดสินใจ โดยแสดงในรูปของโหนดบนต้นไม้

4) กิ่งของต้นไม้ ถูกสร้างขึ้นจากค่าต่างๆ ที่เป็นไปได้ของโหนดทดสอบ และข้อมูลจะถูกแบ่งออกตามกิ่งต่างๆ ที่สร้างขึ้น

5) ทำการวนซ้ำเพื่อหาแอททริบิวต์ที่มีค่าเกินมากที่สุด สำหรับข้อมูลที่ถูกแบ่งแยกออกมาในแต่ละกิ่งเพื่อนำแอททริบิวต์นี้มาสร้างเป็นโหนดตัดสินใจต่อไป โดยที่แอททริบิวต์ที่ถูกเลือกมาเป็นโหนดแล้วจะไม่ถูกเลือกมาอีก สำหรับโหนดในระดับต่อไป

6) ทำการวนซ้ำเพื่อแบ่งข้อมูลและแตกกิ่งของต้นไม้ไปเรื่อยๆ โดยการวนซ้ำจะสิ้นสุดก็ต่อเมื่อเงื่อนไขข้อใดข้อหนึ่งต่อไปนี้จริง

2.3.3.3 การคำนวณค่า Information Gain ต้นไม้ตัดสินใจ (Decision Tree) (ขจรศักดิ์ ศรีอ่อน, 2552) เป็นโครงสร้างที่ใช้แสดงกฎที่ได้จากเทคนิคการจำแนกประเภทข้อมูล โดยต้นไม้ตัดสินใจจะมีลักษณะคล้ายโครงสร้างต้นไม้ ที่แต่ละโหนดแสดงคุณลักษณะ (Attribute) ในการสร้างต้นไม้ตัดสินใจ ปัญหาสำคัญที่ต้องพิจารณาคือ ควรจะตัดสินใจเลือกแอททริบิวต์ใดมาทำหน้าที่เป็นโหนดราก ในแต่ละขั้นตอนของการสร้างต้นไม้และต้นไม้ย่อย (Subtree) ของต้นไม้ตัดสินใจ เกณฑ์ที่ใช้ช่วยประกอบการเลือก แอททริบิวต์คือการคำนวณค่ามาตรฐานเกน (Gain Criterion) ซึ่ง เป็นค่าที่บ่งบอกว่าแอททริบิวต์นั้น สามารถจำแนกกลุ่มของข้อมูลได้ดีเพียงใด โดยทดลองเลือก แต่ละแอททริบิวต์ที่เป็นไปได้จากชุดข้อมูลมาทำหน้าที่เป็นโหนดราก ถ้าแอททริบิวต์ใดให้ค่าเกินสูงสุด แสดงว่าแอททริบิวต์นั้นสามารถจำแนกกลุ่มของข้อมูลได้ดีที่สุด การใช้ค่า Information Gain จะช่วยลดจำนวนครั้งของการทดสอบ ในการแยกแยะข้อมูลอีกทั้งยังรับประกันว่าต้นไม้ตัดสินใจที่ได้ไม่มีความซับซ้อนมากเกินไป

ซึ่งค่า Information Gain นั้นสามารถคำนวณได้จากสมการดังสมการที่ 2-1

$$I(s_1, s_2, \dots, s_n) = - \sum_{i=1}^n \frac{s_i}{s} \log_2 \frac{s_i}{s} \quad (2-1)$$

เมื่อ

S เป็นเซตของข้อมูลซึ่งประกอบด้วยข้อมูล s เรคคอร์ด

n เป็นจำนวนกลุ่มทั้งหมดที่ต่างกันของข้อมูลชุดนั้น

C_i แทนกลุ่มในลำดับ ที่ i โดย ที่ i มีค่าระหว่าง 1 ถึง n

s_i แทนจำนวนข้อมูลที่เป็นสมาชิกของ S และอยู่ในกลุ่ม C_i

s_{ij} แทนจำนวนข้อมูลที่เป็นสมาชิกของ S ในกลุ่ม C_i จากการแบ่งข้อมูลด้วยค่าที่เป็นไปได้ของแอตทริบิวต์ A

j ค่าระหว่าง 1 ถึง v

ค่าเอ็นโทรปีของแอตทริบิวต์ A ซึ่งมีค่าของแอตทริบิวต์เป็น $(a_1, a_2, a_3, \dots, a_v)$ หาได้ในสมการที่ 2-2

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

ดังนั้นจะสามารถพิจารณาค่ามาตรฐานเกณฑ์ดังสมการที่ 2-3

$$Gain(A) = I(s_{1j}, s_{2j}, \dots, s_{nj}) - E(A) \quad (2-3)$$

2.3.3.4 อัลกอริธึม ID3 (ID3 Algorithm) (สุภชัย ประคองศิลป์, 2551)
ID3 (Iterative Dichotomiser 3) เป็นอัลกอริธึมพื้นฐานที่ใช้ในการสร้างการตัดสินใจแบบโครงสร้างต้นไม้ที่ใช้หลักการของการใช้ทฤษฎีข่าวสาร (Information Theory) และค่าที่วัดได้จะนำมาใช้ตัดสินใจว่าจะใช้ตัวแปรใดใช้ในการทำนาย หรือแบ่งประเภทของข้อมูล โดยที่ชุดตัวอย่าง (Sample) คือชุดของข้อมูลที่ใช้ในการเรียนรู้ (Training Sample) ตัวแปรเป้าหมาย (Target Attribute) คือตัวแปรที่นำค่าไปใช้ในการทำนายผลในโครงสร้างต้นไม้และแอตทริบิว (Attributes) คือตัวแปรอื่นๆที่ใช้ในการสร้างโหนดในต้นไม้ และไม่ใช้ตัวแปรเป้าหมาย (Target Attribute) (Tom, 1997) ซึ่งมีลักษณะของอัลกอริธึมดังภาพที่ 2-5

ID3(Examples, Target_Attribute, Attributes)

- Create a root node for the tree
- If all examples are positive, Return the single-node tree Root label = +.
- If all examples are negative, Return the single-node tree, with = -.
- If number of predicting attributes is empty, then Return the single node tree Root, with label = most common value of the target attribute in the examples.
- Otherwise Begin
- A = The Attribute that best classifies examples.
- Decision Tree attribute for Root = A.
- For each possible value v_i of A,
- Add a new tree branch below Root, corresponding to the test $A = v_i$
- Let Examples (v_i)₃ be the subset of examples that have the value v_i for A
- If Examples (v_i) is empty
- Then below this new branch add a leaf node with label = most common target value in the examples
- Else below this new branch add the subtree ID3 (Examples(v_i)₃ Target_Attribute, Attributes – {A})
- End
- Return Root

ภาพที่ 2-5 กระบวนการทำงานของอัลกอริทึม ID3 (Tom, 1997)

2.3.3.4 อัลกอริทึม C4.5 (J48) (พวงทิพย์ แท่นแสง, 2550) เป็นอัลกอริทึมการสร้างกฎจากต้นไม้ตัดสินใจ (Decision Tree) มีพื้นฐานเดียวกับอัลกอริทึม ID3 ถูกออกแบบโดย Quinlan (1992) ซึ่งได้พัฒนาเพิ่มเติมจาก ID3 ดังนี้

- 1) สามารถหลีกเลี่ยงการสร้างโครงสร้างต้นไม้ที่ใหญ่เกินไป เนื่องจากมีข้อมูลจำนวนมาก อย่างไรก็ตามขึ้นอยู่กับข้อกำหนดความลึก เมื่อมีการเจริญเติบโตของ Decision Tree
- 2) ความผิดพลาดลดลง เพราะมีการตัดทอนความผิดพลาดออกไป (Pruning node)
- 3) มีการสร้างกฎหลังการตัดทอนข้อมูลที่ผิดพลาดออก
- 4) สามารถใช้กับข้อมูลที่มีความต่อเนื่อง (Continuous Attributes) ที่เป็นตัวเลข

ได้ เช่น ค่าอนุกรมวิธาน จำนวนเงิน เป็นต้น

- 5) การเลือก Attribute ที่วัดการเลือกให้เหมาะสม
- 6) สามารถใช้กับชุดข้อมูล (Training Data) ที่มีค่าผิดพลาด (Missing Attribute)
- 7) สามารถใช้สำหรับ Attribute กับ Costs ที่แตกต่างกันได้มีการยอมรับประสิทธิภาพกระบวนการคำนวณที่ปรับปรุง

2.4 ซอฟต์แวร์ Weka

Weka ย่อมาจาก Waikato Environment for Knowledge Analysis (ขจรศักดิ์ ศรีอ่อน, 2552) ซึ่งเป็นซอฟต์แวร์สำเร็จภาพประกอบประเภทซอฟต์แวร์ให้เปล่า (Free Ware) ที่สามารถดาวน์โหลดได้จากเว็บไซต์ ซึ่งอยู่ภายใต้การควบคุมของ GPL License ซึ่งโปรแกรม Weka ได้ถูกพัฒนามาจากภาษาจาวาทั้งหมด ซึ่งเขียนมาโดยเน้นกับงานทางด้านการเรียนรู้ด้วยเครื่อง (Machine Learning) และ การทำเหมืองข้อมูลโปรแกรมประกอบไปด้วยโมดูลย่อยๆ สำหรับใช้ในการจัดการข้อมูล และเป็นโปรแกรมที่สามารถใช้ Graphic User Interface (GUI) และ ใช้คำสั่งในการให้ซอฟต์แวร์ประมวลผลดังภาพที่ 2-6



ภาพที่ 2-6 หน้าจอหลักของโปรแกรม Weka Explore

2.4.1 โปรแกรมหลักของซอฟต์แวร์ Weka

2.4.1.1 Simple CLI (Command Line Interface) เป็นโปรแกรมรับคำสั่งการทำงานผ่านการพิมพ์

- 2.4.1.2 Explorer เป็นโปรแกรมที่ออกแบบในลักษณะ GUI
- 2.4.1.3 Experimenter เป็นโปรแกรมที่ออกแบบการทดลองและการทดสอบผล
- 2.4.1.4 KnowledgeFlow เป็นโปรแกรมออกแบบผังการไหลของความรู้
- 2.4.1.5 ArffViewer เป็นโปรแกรมที่ใช้สำหรับแก้ไขเพิ่มประเภท Arff
- 2.4.1.6 Log เป็นโปรแกรมที่ใช้อ่านข้อความบันทึกเก็บระหว่างการทำงาน

2.4.2 เมนูหลักของ Explorer

- 2.4.2.1 Preprocess การเตรียมข้อมูล
- 2.4.2.2 Classify รวมโมดูลการทำเหมืองข้อมูลแบบจัดจำแนกประเภท
- 2.4.2.3 Cluster รวมโมดูลการทำเหมืองข้อมูลแบบการเกาะกลุ่ม
- 2.4.2.4 Associate รวมโมดูลการทำเหมืองข้อมูลแบบกฎเชื่อมโยง
- 2.4.2.5 Select attributes รวมโมดูลสำหรับการวิเคราะห์ความเกี่ยวข้องของลักษณะ

ประจำ

- 2.4.2.6 Visualize นำเสนอข้อมูลด้วยภาพนามธรรมสองมิติ

2.4.3 ประเภทของแฟ้มข้อมูลที่ได้รับได้

- 2.4.3.1 แฟ้มข้อมูลที่ได้รับต้องอยู่ในรูปแบบ ASCII อาจเป็น arff, csv, C45
- 2.4.3.2 ในกรณีแฟ้มข้อมูลอยู่ในเครือข่ายผู้ใช้สามารถเรียกใช้โดยอาศัย URL
- 2.4.3.3 หรืออาจใช้ข้อมูลที่อยู่ในฐานข้อมูลที่เชื่อมโยงผ่าน JDBC

2.5 งานวิจัยที่เกี่ยวข้อง

วาสนา วงษ์กำภู (2552) ได้พัฒนาระบบสืบค้นงานวิจัยด้านจิตเวชศาสตร์ สำหรับจัดเก็บข้อมูลงานวิจัย และช่วยในการสืบค้นงานวิจัยได้อย่างเป็นระบบ โดยนำวิธีการของต้นไม้ตัดสินใจ (Decision Tree) เข้ามาช่วยในการจัดแบ่งหมวดหมู่งานวิจัย ทำให้สามารถแบ่งแยกข้อมูลงานวิจัยออกเป็นประเภทต่างๆ และสามารถแสดงผลลัพธ์ออกเป็นเงื่อนไขที่ผู้ใช้ต้องการ ทำให้การสืบค้นข้อมูลวิจัยมีประสิทธิภาพมากขึ้น

ศุภชัย ประครองศิลป์ (2551) ได้นำเสนอการพัฒนาระบบสนับสนุนการตัดสินใจในการอนุมัติลูกบ้านเข้าโครงการ โดยใช้เทคนิคต้นไม้ตัดสินใจและในอัลกอริทึม ID3 ในการเรียนรู้จากข้อมูลลูกบ้านในอดีตเพื่อสร้างโมเดลสำหรับตัดสินใจอนุมัติลูกบ้านเข้าโครงการ โดยทำการทดสอบความแม่นยำของอัลกอริทึม ID3 เปรียบเทียบกับอัลกอริทึมอื่นที่เป็นเทคนิคต้นไม้ตัดสินใจเหมือนกัน ด้วยการนำข้อมูลในอดีตของบุคคลที่เคยเข้าร่วมโครงการทั้งหมด 963 รายการและวัดความแม่นยำของอัลกอริทึม ID3 โดยใช้ Confusion Matrix ได้ผลความแม่นยำคือ 93.67 % ซึ่งมากกว่าอัลกอริทึม

ตัวอื่นๆ และผู้วิจัยได้วัดประสิทธิภาพโมเดลของระบบที่พัฒนาขึ้น เปรียบเทียบกับโปรแกรม Weka 3.5.8 โดยใช้ข้อมูลสำหรับการสอน (Training Data) ที่เหมือนกัน ผลที่ได้คือ สามารถสร้างโมเดลได้เหมือนกับระบบที่ผู้วิจัยพัฒนา จึงสรุปได้ว่าโมเดลที่พัฒนาขึ้นมีความถูกต้องและแม่นยำอยู่ในระดับดีมาก สามารถนำไปสนับสนุนงานได้จริง

มลธิดา ฤทธิสมบุรณ์ (2550) ได้นำเสนอการพัฒนากระบวนการสนับสนุนการพิจารณาอนุมัติให้สินเชื่อเพื่อการเช่าซื้อสินค้า โดยใช้เทคนิคการแบ่งกลุ่มแบบโครงสร้างต้นไม้ และใช้อัลกอริทึมในการเรียนรู้แบบ ID3 โดยใช้โปรแกรม Weka 3.4 ในการทดลองเพื่อสร้างโมเดล อัลกอริทึมแบบ ID3 จะนำค่าชี้วัดมาใช้ในการแบ่งข้อมูล โดยกลุ่มตัวอย่าง คือชุดข้อมูลที่ใช้ในการเรียนรู้ เป็นตัวแปรเป้าหมาย และแอททริบิว คือตัวแปรอื่นๆ ที่ใช้ในการสร้างโหนดในต้นไม้ และไม่ใช่ตัวแปรเป้าหมาย ส่วนต้นไม้ตัดสินใจจะนำมาใช้ในการเลือกรากของต้นไม้โดยดูจากแอททริบิวที่มีความสำคัญมากที่สุดก่อนแล้วลดลงตามลำดับเพื่อช่วยในการพิจารณาการอนุมัติ การให้สินเชื่อโดยวิธีการดังกล่าวเหมาะกับการจำแนกข้อมูลออกเป็นกลุ่มอย่างชัดเจน โดยระบบที่ได้พัฒนาขึ้นนั้นอยู่ในระดับดี

พยุชน พานิชย์กุล (2548) ได้พัฒนาระบบค้ำไม้หนึ่งโดยใช้ ต้นไม้ตัดสินใจ (Decision Tree) โดยนำเทคนิคของโครงสร้างต้นไม้การตัดสินใจมาพัฒนาระบบงาน มีวิธีการคือ ระบบจะรับข้อมูล (Data) มาประมวลผล เพื่อสร้างโครงสร้างต้นไม้ ซึ่งโครงสร้างของต้นไม้ที่ได้จะแสดงเป็นโครงสร้างที่มีกฎต่างๆ ตามเป้าหมายของการใช้งาน เพื่อนำมาแบ่งกลุ่มข้อมูลตามที่กำหนดไว้ โดยผลการพัฒนาระบบสามารถแบ่งประเภทได้ในระดับดี และสามารถหาความน่าเชื่อถือของการแบ่งประเภทได้ในระดับดี และระบบที่พัฒนาสามารถนำไปใช้ได้กับข้อมูลอื่นๆ เพราะไม่ได้กำหนดขอบเขตไว้กับงานใดงานหนึ่ง