บทที่ 2

ทฤษฎีที่เกี่ยวข้องกับการสกัดกฎ

วิทยานิพนธ์นี้ได้นำทฤษฎีต่างๆ มาช่วยในการสกัดกฎเพื่อให้สามารถสกัดกฎได้มีประสิทธิภาพ ซึ่งประกอบด้วย การเตรียมข้อมูล (Data Preprocessing) โครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น (Multilayer Perceptron Neural Network: MLP) และการวัด ประสิทธิภาพ (Evaluation) โดยรายละเอียดมีดังนี้

2.1 การเตรียมข้อมูล (Data Preprocessing)

การทำเตรียมข้อมูล (Data Preprocessing) เป็นขั้นตอนในการเตรียมข้อมูลก่อนการสกัดกฎ เนื่องจากฐานข้อมูลที่นำมาสกัดกฎอาจมีข้อมูลที่ไม่สมบูรณ์ เช่น ในบางแถวข้อมูลอาจมีค่าตัวแปรข้อมูลเข้าหายไป (Missing Values) ซึ่งสามารถจัดการกับค่าที่หายไปได้ด้วย 3 วิธี [5] ดังนี้

2.1.1 ลบแถวข้อมูลที่มีค่าตัวแปรข้อมูลเข้าหายไปทิ้ง วิธีนี้เหมาะกับข้อมูลที่มีจำนวนแถวข้อมูลที่หายไปไม่มากนัก เมื่อเปรียบเทียบกับจำนวนแถวข้อมูลทั้งหมดที่มีอยู่ เพื่อให้ได้ข้อมูลที่ถูกต้องทุกแถวในการทำงาน [5] ตัวอย่างเช่น Setiono [28] ได้ลบแถวข้อมูลที่มี ค่าตัวแปรข้อมูลเข้าหายไปจำนวน 16 แถวจากข้อมูลทั้งหมด 699 แถวของฐานข้อมูลโรคมะเร็งเต้านมก่อนการสกัดกฎ เป็นต้น

2.1.2 แทนค่าตัวแปรข้อมูลเข้าที่หายไปด้วยค่าเฉลี่ยของกลุ่ม (Class Mean) วิธีนี้เหมาะกับค่าตัวแปรข้อมูลเข้าที่เป็นเลขจำนวนจริง [5] ตัวอย่าง เช่น Duch [31] และ Wettayaprasit [27] ได้แทนค่าตัวแปรข้อมูลเข้าที่หายไปของฐานข้อมูลโรคมะเร็งเต้านมด้วย ค่าเฉลี่ยของกลุ่ม เป็นต้น

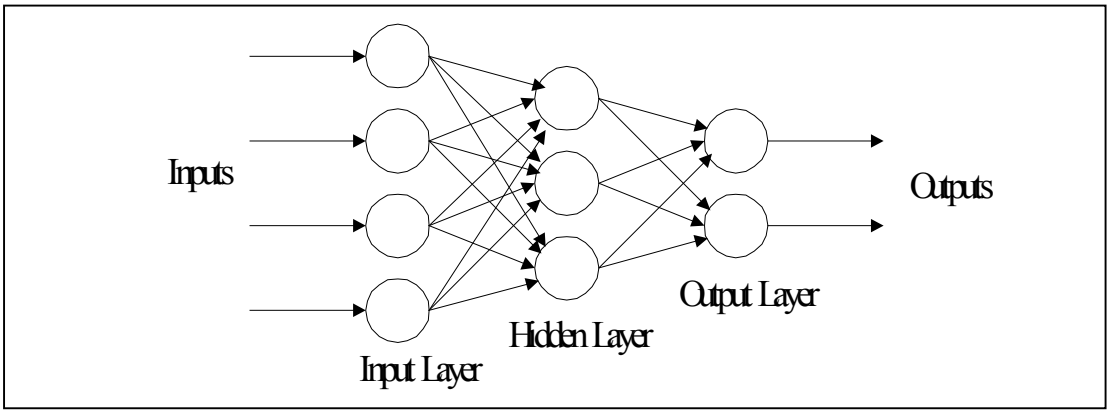
2.1.3 แทนค่าตัวแปรข้อมูลเข้าที่หายไปด้วยไปด้วยค่ามัธยฐานของกลุ่ม (Class Median) วิธีนี้เหมาะกับค่าตัวแปรข้อมูลเข้าที่เป็นเลขจำนวนนับ [5]

2.2 โครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น (Multilayer Perceptron Neural Network: MLP)

ในวิทยานิพนธ์นี้ใช้โครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น (Multilayer Perceptron Neural Network: MLP) มาสกัดกฎภาษาธรรมชาติ ซึ่งมีสถาปัตยกรรม และวิธีการเรียนรู้แบบแพร่ย้อนกลับ (Backpropagation Learning) โดยรายละเอียดมีดังนี้

2.2.1 สถาปัตยกรรมของโครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น (Multilayer Perceptron Neural Network: MLP)

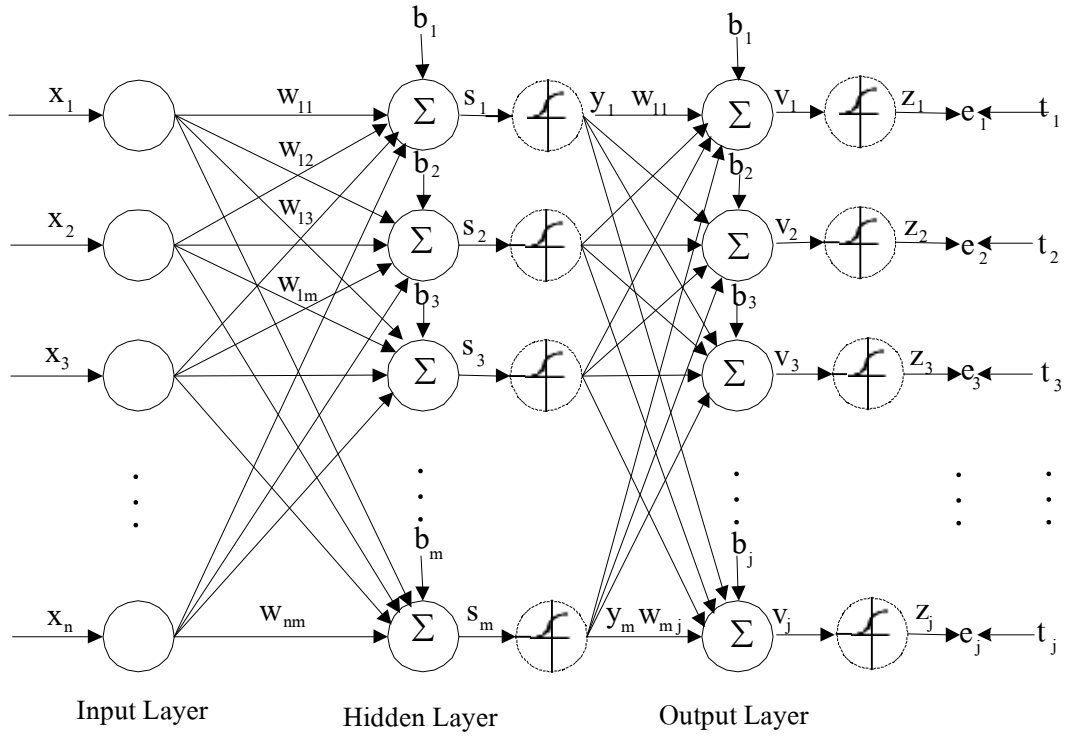
โครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้นประกอบด้วยชั้นต่างๆ คือ ชั้นข้อมูลเข้า (Input Layer) 1 ชั้น ชั้นซ่อน (Hidden Layer) กี่ชั้นก็ได้ และชั้นผลลัพธ์ (Output Layer) 1 ชั้น โดยที่ชั้นซ่อนจะอยู่ระหว่างชั้นข้อมูลเข้าและชั้นผลลัพธ์ ภาพประกอบ 2.1 แสดง ตัวอย่างที่มีชั้นซ่อน 1 ชั้น [32] ในการเชื่อมต่อระหว่างชั้นต่างๆ ทุกๆ โหนดในชั้นข้อมูลเข้าจะมีเส้นเชื่อมของค่าน้ำหนักเพื่อส่งสัญญาณไปยังทุกๆ โหนดในชั้นซ่อน จากนั้นทุกๆ โหนดใน ชั้นซ่อนจะส่งสัญญาณไปยังทุกๆโหนดในชั้นผลลัพธ์ ตัวอย่างดังภาพประกอบ 2.1 [32]



ภาพประกอบ 2.1 สถาปัตยกรรมของโครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น

2.3.2 วิธีการเรียนรู้แบบแพร่ย้อนกลับ (Backpropagation Learning)

ในการเรียนรู้แบบแพร่ย้อนกลับของโครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้นที่มีชั้นซ่อน 1 ชั้น แสดงดังภาพประกอบ 2.2 ซึ่งมีขั้นตอนวิธีในการเรียนรู้ดังนี้ [5, 32, 33]



โดยที่ x_n = โหนดข้อมูลเข้าที่ n มีทั้งหมด N โหนด

s_m = ค่าผลรวมของโหนดซ่อนที่ m มีทั้งหมด M โหนด

y_m = ค่าของโหนดซ่อนที่ m หลังทำการปรับค่าด้วยฟังก์ชันกระตุ้น

v_j = ค่าผลรวมของโหนดผลลัพธ์ที่ j มีทั้งหมด J โหนด

z_j = ค่าของโหนดผลลัพธ์ที่ j หลังทำการปรับค่าด้วยฟังก์ชันกระตุ้น

t_j = ค่าผลลัพธ์ที่ต้องการ (Target Value) จากโหนดผลลัพธ์ที่ j

e_j = ค่าความผิดพลาดของค่าของโหนดผล

ภาพประกอบ 2.2 ตัวแปรของวิธีการเรียนรู้แบบแพร่ย้อนกลับ

1. กำหนดจำนวนโหนดข้อมูลเข้า (N) จำนวนโหนดซ่อน (M) จำนวนโหนดผลลัพธ์ (J) และจำนวนรอบสูงสุดที่จะทำการเรียนรู้ (R)

2. สุ่มค่าน้ำหนักเริ่มต้นให้กับทุกๆเส้นเชื่อมภายในโครงข่ายประสาทเทียมมีค่าอยู่ในช่วง [-1,1]

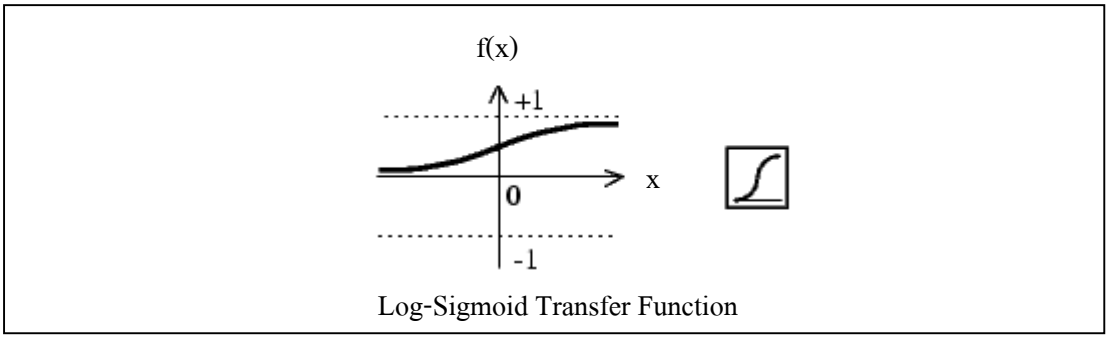
3. รับค่าข้อมูลเข้าของข้อมูลแถวแรก เพื่อใช้ในการคำนวณหาค่าผลลัพธ์ของโครงข่ายประสาทเทียม

4. คำนวณค่าผลรวมของโหนดซ่อน (s_m) ด้วยสมการที่ (2.1) จากนั้นทำการปรับค่ารวมด้วยฟังก์ชันกระตุ้น (Activation Function) ฟังก์ชันซิกมอยด์ (Sigmoid Function) ดังสมการที่ (2.2) ซึ่งจะได้ค่าของโหนดซ่อน (y_m) หลังการปรับค่าดังสมการที่ (2.3) โดยที่ค่าที่ได้จะอยู่ในช่วง [0,1] ดังภาพประกอบ 2.3 [34]

$$s_m = \sum_{n=1}^{N} x_n w_{nm} + b_m \qquad (2.1)$$

$$f(x) = \frac{1}{1+e^{-x}} \qquad (2.2)$$

$$y_m = f(s_m) \qquad (2.3)$$



ภาพประกอบ 2.3 แสดงค่าที่ได้จากฟังก์ชันซิกมอยด์

5. คำนวณค่าผลรวมของโหนดผลลัพธ์ (v_j) ด้วยสมการที่ (2.4) จากนั้นทำการปรับค่าผลรวมด้วยฟังก์ชันกระตุ้น (Activation Function) ฟังก์ชันซิกมอยด์ (Sigmoid Function) ดังสมการที่ (2.2) ซึ่งจะได้ค่าของโหนดผลลัพธ์ (z_j) หลังการปรับค่าดังสมการที่ (2.5)

$$v_j = \sum_{m=1}^{M} y_m w_{mj} + b_j \tag{2.4}$$

$$z_j = f(v_j) \tag{2.5}$$

6. คำนวณค่าความผิดพลาดของค่าของโหนดผลลัพธ์ดังสมการที่ (2.6)

$$e_j = z_j(1 - z_j)(t_j - z_j) \tag{2.6}$$

7. คำนวณค่าความผิดพลาดของค่าของโหนดซ่อนดังสมการที่ (2.7)

$$e_m = y_m(1 - y_m)\sum_{j=1}^{J} e_j w_{mj} \tag{2.7}$$

8. ปรับค่าน้ำหนักของเส้นเชื่อมระหว่างโหนดซ่อนและโหนดผลลัพธ์จากค่าความผิดพลาดของโหนดผลลัพธ์ดังสมการที่ (2.8) โดยที่ η คืออัตราการเรียนรู้มีค่าอยู่ในช่วง [0,1]

$$w_{mj} = w_{mj} + \eta e_j z_j \tag{2.8}$$

9. ปรับค่าน้ำหนักของเส้นเชื่อมระหว่างโหนดข้อมูลเข้าและโหนดซ่อนจากค่าความผิดพลาดของโหนดซ่อนดังสมการที่ (2.9)

$$w_{nm} = w_{nm} + \eta e_m z_m \tag{2.9}$$

10. รับค่าข้อมูลเข้าของข้อมูลแถวถัดไปและกลับไปทำข้อ 4 แต่ถ้าเป็น ข้อมูลแถวสุดท้ายให้ไปทำข้อ 11

11. คำนวณค่าความผิดพลาดเฉลี่ย (Mean Squared Error: MSE) ในทุกแถวข้อมูลดังสมการที่ (2.10) ถ้าค่าความผิดพลาดเฉลี่ยมีค่าน้อยกว่าค่าที่ยอมรับได้ ให้จบการเรียนรู้ แต่ถ้าค่าความผิดพลาดเฉลี่ยมากกว่าค่าที่ยอมรับได้ให้ตรวจสอบว่าได้ทำการเรียนรู้ครบตามจำนวนรอบ (R) ที่กำหนดไว้หรือไม่ ถ้าครบแล้วให้จบการเรียนรู้ แต่ถ้ายังไม่ครบให้กลับไปทำข้อ 3 ซึ่งก็คือเริ่มต้นเรียนรู้รอบใหม่

$$MSE = \frac{\sum_{=1}^{I}\sum_{j=1}^{J}(t_{ij} - z_{ij})^2}{ij} \qquad (2.10)$$

โดยที่ t_{ji} = ค่าผลลัพธ์ที่ต้องการจากโหนดผลลัพธ์ที่ j ในแถวข้อมูลเข้าที่ i ซึ่งมีทั้งหมด I แถว

z_{ji} = ค่าของโหนดผลลัพธ์ที่ j ในแถวข้อมูลเข้าที่ i ซึ่งมีทั้งหมด I แถว

i = ลำดับของแถวข้อมูล

j = ลำดับของโหนดผลลัพธ์

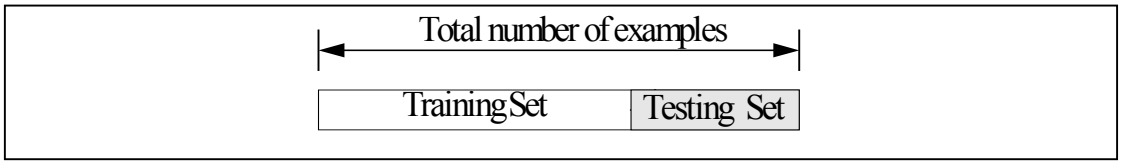
2.3 การวัดประสิทธิภาพ

ในการทดลองจะต้องมีการคำนวณหาค่าความถูกต้องของข้อมูล (Accuracy) จากโครงข่ายประสาทเทียม และจากกฎที่สกัดได้ เพื่อวัดว่าโครงข่ายประสาทเทียม และกฎสามารถทำนายค่าข้อมูลได้ถูกต้องมากน้อยเพียงใดซึ่งคิดเป็นร้อยละ [6] ดังสมการที่ (2.11)

$$\text{ค่าความถูกต้องของข้อมูล (\%)} = \frac{\text{จำนวนแถวข้อมูลที่ทำนายถูกต้อง}}{\text{จำนวนแถวข้อมูลทั้งหมด}} \times 100 \qquad (2.11)$$

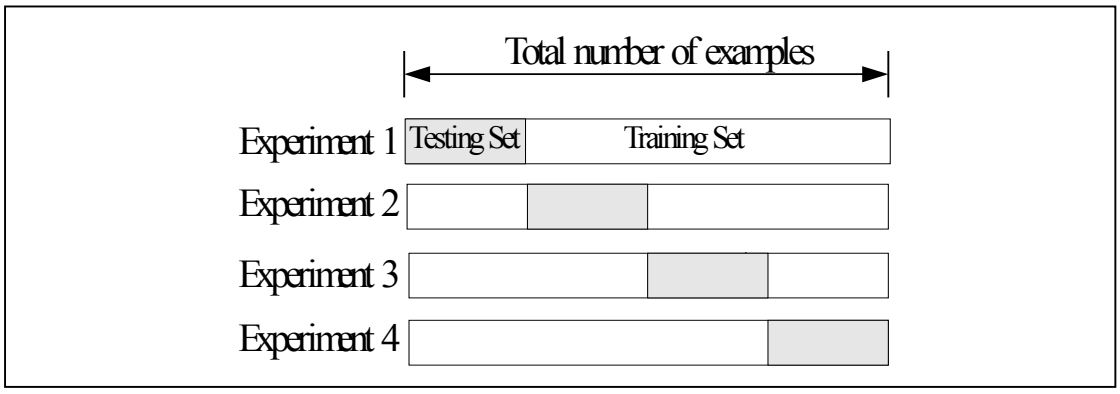
ในการวัดประสิทธิภาพของวิธีการสกัดกฎ โดยใช้ค่าความถูกต้องของข้อมูล (Accuracy) จะต้องทำการเลือกข้อมูลชุดสอน (Training Set) และข้อมูลชุดทดสอบ (Testing Set) ซึ่งมีหลักการเลือก 3 วิธีคือเลือกสุ่มข้อมูลแบบร้อยละ เลือกสุ่มข้อมูลแบบความเที่ยงตรง K กลุ่ม (K-Fold Cross Validation) และเลือกสุ่มข้อมูลแบบ Leave-one-out Cross Validation ซึ่งมีรายละเอียดดังนี้ [35]

1. การเลือกสุ่มข้อมูลแบบร้อยละ (Percentage) จะเลือกสุ่มข้อมูลชุดสอนตามร้อยละที่กำหนด สำหรับข้อมูลที่เหลือจะเป็นข้อมูลชุดทดสอบดังภาพประกอบ 2.4 ข้อดีของการเลือกสุ่มข้อมูลแบบร้อยละคือเป็นวิธีการเลือกสุ่มข้อมูลที่ง่าย แต่ข้อเสียคือข้อมูลทุกตัวไม่ได้ถูกนำมาเป็นข้อมูลชุดสอนและชุดทดสอบ



ภาพประกอบ 2.4 การเลือกสุ่มข้อมูลแบบร้อยละ

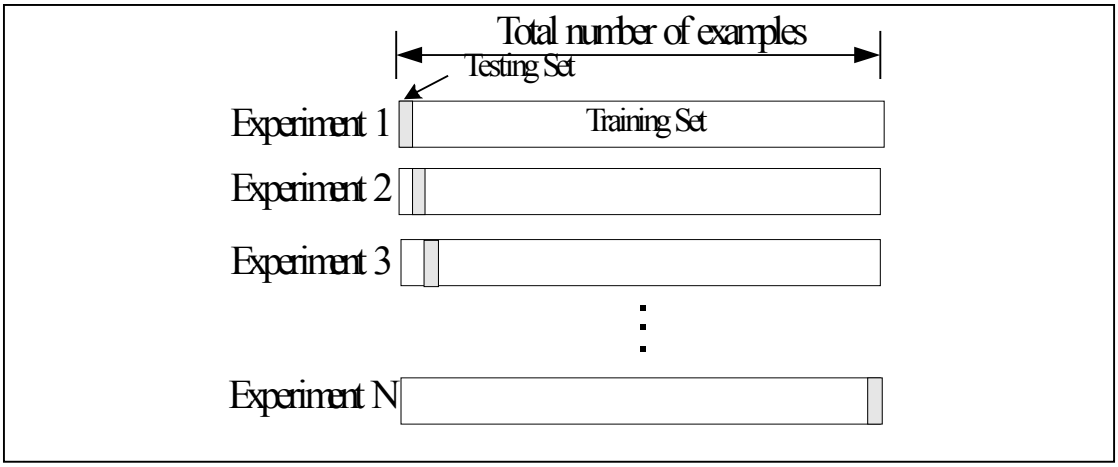
2. การเลือกสุ่มข้อมูลแบบความเที่ยงตรง K กลุ่ม (K-Fold Cross Validation) จะเลือกสุ่มข้อมูลออกเป็น K ชุดเท่ากัน ในการทดลองครั้งแรกข้อมูลชุดที่ 1 เป็น ข้อมูลชุดทดสอบ และข้อมูลชุดที่เหลือเป็นข้อมูลชุดสอน ในการทดลองครั้งที่สองข้อมูลชุดที่ 2 เป็นข้อมูลชุดทดสอบ และข้อมูลชุดที่เหลือเป็นข้อมูลชุดสอน ทำจนกระทั่งข้อมูลทุกชุดได้ถูกนำมาเป็นข้อมูลชุดทดสอบ ซึ่งมีการทดลองทั้งหมด K ครั้ง [6] ตัวอย่างการเลือกสุ่มข้อมูลแบบ ความเที่ยงตรง K กลุ่ม เมื่อ K = 4 แสดงดังภาพประกอบ 2.5 ข้อดีของการเลือกสุ่มข้อมูลแบบความเที่ยงตรง K กลุ่มคือข้อมูลทุกตัวจะถูกนำมาเป็นข้อมูลชุดสอนและข้อมูลชุดทดสอบ แต่ ข้อเสียคือใช้เวลานานในการทดลอง เนื่องจากต้องทดลองข้อมูลทั้งหมด K ครั้ง [36] ทั้งนี้นิยมกำหนดให้ค่า K มีค่าเท่ากับ 10



ภาพประกอบ 2.5 การเลือกสุ่มข้อมูลแบบความเที่ยงตรง K กลุ่ม เมื่อ K = 4

3. เลือกสุ่มข้อมูลแบบ Leave-one-out Cross Validation คือการเลือกสุ่มข้อมูลแบบความเที่ยงตรง K กลุ่ม เมื่อกำหนดให้ K มีค่าเท่ากับจำนวนแถวข้อมูลทั้งหมด (N) ดังภาพประกอบ 2.6 ข้อดีของการเลือกสุ่มข้อมูลแบบ Leave-one-out Cross Validation คือเหมาะสำหรับ

ข้อมูลขนาดเล็ก แต่ข้อเสียคือไม่เหมาะสำหรับข้อมูลขนาดใหญ่เนื่องจากต้องทดลองหลายครั้งทำให้ใช้เวลาในการทดลองนาน [6]



ภาพประกอบ 2.6 การเลือกสุ่มข้อมูลแบบ Leave-one-out Cross Validation

วิทยานิพนธ์นี้ได้นำทฤษฎีต่างๆ มาช่วยในการสกัดกฎประกอบด้วย การเตรียมข้อมูล (Data Preprocessing) เพื่อจัดการค่าตัวแปรข้อมูลเข้าที่หายไป (Missing Values) โครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น (Multilayer Perceptron Neural Network: MLP) ซึ่งมีการเรียนรู้แบบแพร่ย้อนกลับ และการวัดประสิทธิภาพ (Evaluation) ด้วย ค่าความถูกต้องของข้อมูล โดยสามารถเลือกสุ่มข้อมูลได้ 3 แบบคือแบบร้อยละ แบบความเที่ยงตรง K กลุ่ม และแบบ Leave-one-out Cross Validation