บทที่ 2
ทฤษฎีและรายงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 วิธีความใกล้เคียงกันมากที่สุด (K-Nearest Neighbor (KNN) Method)

ไม่มีการสร้างตัวแบบจากข้อมูลฝึกหัดเก็บไว้ ทำนายข้อมูลใหม่โดยอาศัยการเปรียบเทียบกับข้อมูลฝึกหัดจำนวน k ตัว ที่อยู่ใกล้เคียงกันมากที่สุด ใช้คำตอบของข้อมูลฝึกหัดที่อยู่ใกล้เคียงกันมากที่สุด k ตัว ที่พบมากที่สุดเป็นคำตอบ วิธีนี้ทำนายได้เฉพาะข้อมูลเชิงกลุ่ม (nominal data) เท่านั้น

วิธีความใกล้เคียงกันมากที่สุดเป็นวิธีการที่ได้รับความนิยมในการใช้งานอย่างมาก เนื่องจากเป็นวิธีการที่ง่ายและมีประสิทธิภาพซึ่งสามารถนำไปประยุกต์ใช้กับงานได้หลายอย่าง เช่น งานทางด้านการจำแนกกลุ่ม รวมถึงงานทางด้านการแทนที่ข้อมูลสูญหาย ซึ่งมีขั้นตอนการดำเนินการดังนี้ (Tan, P-N. and et al. : 2006)

2.1.1.1 กำหนดค่า k เพื่อใช้พิจารณาสมาชิกที่อยู่ใกล้เคียงกันมากที่สุด เช่น k = 3 คือ จะพิจารณาเฉพาะข้อมูล 3 ตัวแรกที่อยู่ใกล้กับจุดที่ต้องการจะทำนาย

2.1.1.2 คำนวณหาระยะห่างระหว่างข้อมูลตัวอย่างที่สนใจกับข้อมูลอื่น ๆ ทุกตัวด้วยวิธีระยะห่างยูคลิเดียน (euclidean distance) จากสมการดังนี้

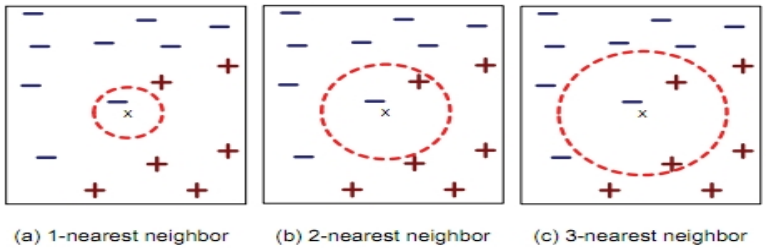
$$\mathrm{dist}(x_i, x_j) = \sqrt{\sum_{k=1}^{n}(x_{i,k} - x_{j,k})^2} \qquad (2.1)$$

โดยที่ $\mathrm{dist}(x_i, x_j)$ คือ ระยะห่างระหว่างตัวอย่าง x_i กับตัวอย่าง x_j

n คือ จำนวนคุณสมบัติทั้งหมดของตัวอย่าง

$x_{i,k}$ คือ คุณสมบัติที่ k ของตัวอย่าง x_i

2.1.1.3 เลือกค่าข้อมูลที่มีค่าระยะห่างน้อยที่สุด k ตัว เพื่อนำมาพิจารณาหาคำตอบ ดังรูปที่ 2.1



(a) ความใกล้เคียงกันมากที่สุดโดยพิจารณาจากข้อมูล 1 ตัว

(b) ความใกล้เคียงกันมากที่สุดโดยพิจารณาจากข้อมูล 2 ตัว

(c) ความใกล้เคียงกันมากที่สุดโดยพิจารณาจากข้อมูล 3 ตัว

รูปที่ 2.1 ตัวอย่างของวิธีความใกล้เคียงกันมากที่สุด

2.1.2 วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ (Decision Tree Method)

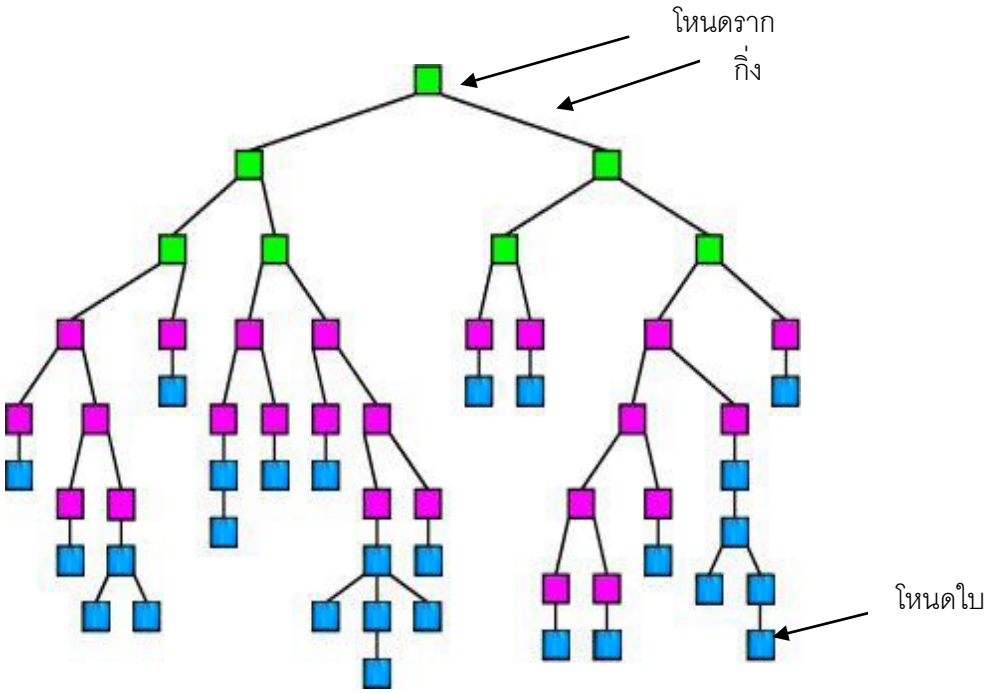
วิธีแผนภาพต้นไม้เพื่อการตัดสินใจเป็นตัวแบบทางคณิตศาสตร์เพื่อหาทางเลือกที่ดีที่สุดโดยการนำข้อมูลมาสร้างตัวแบบการพยากรณ์ในรูปแบบของโครงสร้างต้นไม้ซึ่งมีการเรียนรู้ข้อมูลแบบมีผู้สอน (supervised learning) สามารถสร้างตัวแบบการจำแนกกลุ่มได้จากกลุ่มตัวอย่างของข้อมูลฝึกหัดโดยอัตโนมัติและสามารถพยากรณ์กลุ่มของรายการที่ยังไม่เคยนำมาจำแนกกลุ่มได้อีกด้วย

2.1.2.1 ส่วนประกอบของแผนภาพต้นไม้เพื่อการตัดสินใจ

1) โหนด (Node) คือคุณสมบัติต่าง ๆ เป็นจุดที่แยกข้อมูลว่าจะให้ไปในทิศทางใด ซึ่งโหนดที่อยู่สูงสุดเรียกว่า โหนดราก (root node)

2) กิ่ง (Branch) คือคุณสมบัติของโหนดที่แตกออกมาโดยจำนวนของกิ่งจะเท่ากับคุณสมบัติของโหนด

3) ใบ (Leaf) คือกลุ่มของผลลัพธ์ในการแยกแยะข้อมูล ซึ่งโหนดที่อยู่ล่างสุดเรียกว่า โหนดใบ (leaf node) โดยสามารถแสดงส่วนประกอบของแผนภาพต้นไม้เพื่อการตัดสินใจดังรูปที่ 2.2



รูปที่ 2.2 ส่วนประกอบของแผนภาพต้นไม้เพื่อการตัดสินใจ

2.1.2.2 การสร้างแผนภาพต้นไม้เพื่อการตัดสินใจ

หลักการพื้นฐานของการสร้างแผนภาพต้นไม้เพื่อการตัดสินใจเป็นการสร้างจากบนลงล่าง คือ เริ่มจากการสร้างรากของต้นไม้ก่อน แล้วจึงแตกกิ่งไปจนถึงใบ โดยแสดงขั้นตอนการสร้างแผนภาพต้นไม้เพื่อการตัดสินใจได้ดังนี้

1) ต้นไม้เริ่มต้นโดยมีโหนดเพียงโหนดเดียวแสดงถึงชุดข้อมูลฝึกหัด (training data set)

2) ถ้าข้อมูลทั้งหมดอยู่ในกลุ่มเดียวกันแล้ว ให้โหนดนั้นเป็นใบและตั้งชื่อแยกตามกลุ่มของข้อมูลนั้น

3) ถ้าในโหนดมีข้อมูลหลายกลุ่มปะปนอยู่ จะต้องวัดค่าผลกำไร (gain) ของแต่ละคุณลักษณะ (attribute) เพื่อที่จะใช้เป็นเกณฑ์ (criterion) ในการคัดเลือกคุณลักษณะที่มีความสามารถในการแบ่งข้อมูลออกเป็นกลุ่มต่าง ๆ ได้ดีที่สุด โดยคุณลักษณะที่มีผลกำไรมากที่สุดจะถูกเลือกให้เป็นตัวทดสอบหรือคุณลักษณะที่ใช้ในการตัดสินใจโดยแสดงในรูปของโหนดบนต้นไม้

4) กิ่งของต้นไม้ถูกสร้างขึ้นจากค่าต่าง ๆ ที่เป็นไปได้ของโหนดทดสอบและข้อมูลจะถูกแบ่งออกตามกิ่งต่าง ๆ ที่สร้างขึ้น

5) ทำการวนซ้ำเพื่อหาคุณลักษณะที่มีผลกำไรมากที่สุด สำหรับข้อมูลที่ถูกแบ่งแยกออกมาในแต่ละกิ่งเพื่อนำคุณลักษณะนี้มาสร้างเป็นโหนดตัดสินใจต่อไป โดยที่คุณลักษณะที่ถูกเลือกมาเป็นโหนดแล้วจะไม่ถูกเลือกมาอีกสำหรับโหนดในระดับต่อ ๆ ไป

6) ทำการวนซ้ำเพื่อแบ่งข้อมูลและแตกกิ่งของต้นไม้ไปเรื่อย ๆ โดยการวนซ้ำจะสิ้นสุดก็ต่อเมื่อเงื่อนไขข้อใดข้อหนึ่งข้างบนนี้เป็นจริง

2.1.2.3 การคำนวณค่าผลกำไรสารสนเทศ (Information Gain)

แผนภาพต้นไม้เพื่อการตัดสินใจเป็นโครงสร้างที่ใช้แสดงกฎที่ได้จากเทคนิคการจำแนกกลุ่มของข้อมูล โดยแผนภาพต้นไม้เพื่อการตัดสินใจจะมีลักษณะคล้ายโครงสร้างต้นไม้ โดยที่แต่ละโหนดแสดงคุณลักษณะ ในการสร้างแผนภาพต้นไม้เพื่อการตัดสินใจ ปัญหาที่สำคัญที่ต้องพิจารณาคือควรจะตัดสินใจเลือกคุณลักษณะใดมาทำหน้าที่เป็นโหนดราก ในแต่ละขั้นตอนของการสร้างต้นไม้และต้นไม้ย่อย (subtree) ของแผนภาพต้นไม้ เพื่อการตัดสินใจ เกณฑ์ที่ใช้ช่วยประกอบการเลือกคุณลักษณะเพื่อการคำนวณเกณฑ์ผลกำไร (gain criterion) ซึ่งเป็นค่าที่บ่งบอกว่าคุณลักษณะที่เป็นไปได้จากชุดข้อมูลมาทำหน้าที่เป็นโหนดราก ถ้าคุณลักษณะใดให้ผลกำไรสูงที่สุดแสดงว่าคุณลักษณะนั้นสามารถจำแนกกลุ่มของข้อมูลได้ดีที่สุด การใช้ผลกำไรสารสนเทศจะช่วยลดจำนวนครั้งของการทดสอบในการแยกแยะข้อมูล อีกทั้งยังรับประกันว่าแผนภาพต้นไม้เพื่อการตัดสินใจที่ได้จะไม่มีความซับซ้อนมากเกินไป (overfitting) ซึ่งผลกำไรสารสนเทศนั้นสามารถคำนวณได้ดังนี้

$$I(S_1, S_2, ..., S_n) = -\sum_{i=1}^{n} \frac{S_i}{S} \log_2 \frac{S_i}{S} \qquad (2.2)$$

เมื่อ S คือ เซตของข้อมูลซึ่งประกอบด้วยข้อมูล S ระเบียน (record)

n คือ จำนวนกลุ่มทั้งหมดที่ต่างกันของข้อมูลชุดนั้น

S_i คือ จำนวนข้อมูลที่เป็นสมาชิกของ S และอยู่ในกลุ่ม C_i

C_i คือ กลุ่มในลำดับที่ i โดยที่ i มีค่าระหว่าง 1 ถึง n

ค่าเอ็นโทรปี (Entropy) ของคุณลักษณะ A ซึ่งมีค่าของคุณลักษณะเป็น $(a_1, a_2, ..., a_v)$ หาได้ดังนี้

$$E(A) = \sum_{j=1}^{V} \frac{S_{1j} + ... + S_{nj}}{S} I(S_{1j} + ... + S_{nj}) \qquad (2.3)$$

S_{ij} คือ จำนวนข้อมูลที่เป็นสมาชิกของ S และอยู่ในกลุ่ม C_i จากการแบ่งข้อมูลด้วยค่าที่เป็นไปได้ของคุณลักษณะ A

ดังนั้นจะสามารถพิจารณาเกณฑ์ผลกำไรได้ดังนี้

$$\text{Gain}(A) = I(S_{1j}, S_{2j}, ..., S_{nj}) - E(A) \quad (2.4)$$

ข้อดี

1) เข้าใจได้ง่าย
2) สร้างกฎได้จากต้นไม้
3) เลือกเฉพาะคุณลักษณะ (attribute) ที่สำคัญในการสร้างตัวแบบ

ข้อเสีย

1) ใช้ได้กับคำตอบ (class) ที่เป็นข้อมูลเชิงกลุ่ม (nominal data) เท่านั้น
2) ความถูกต้องในการทำนายไม่สูง

การประยุกต์ใช้แผนภาพต้นไม้เพื่อการตัดสินใจ (Decision tree application)

แผนภาพต้นไม้เพื่อการตัดสินใจใช้ตอบคำถามที่ต้องการจำแนกประเภทข้อมูลที่ต้องการความเข้าใจประกอบ

ใช้ในการพิจารณาให้สินเชื่อแก่บุคลต่าง ๆ

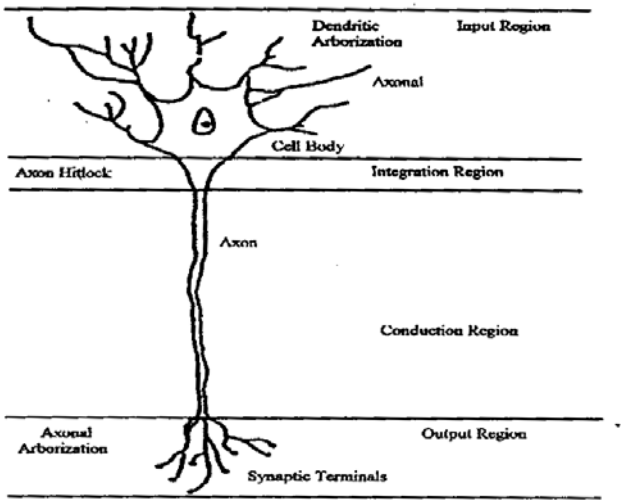
ใช้ในการทำนายว่าลูกค้าคนไหนที่มีโอกาสจะยกเลิกการใช้บริการและเหตุผลเพราะอะไร

2.1.3 วิธีโครงข่ายประสาทเทียม (Artificial Neural Network Method)

วิธีโครงข่ายประสาทเทียมใช้หลักการเลียนแบบการทำงานของสมองมนุษย์ เส้นเชื่อมแต่ละเส้นจะมีน้ำหนักถ่วง (weight) เพื่อใช้กำหนดน้ำหนักถ่วงหรือความสำคัญของข้อมูลเข้า (input data) กำหนดค่าเริ่มต้นโดยการสุ่ม ในแต่ละโหนดทำการคำนวณค่าผลรวมเชิงเส้นแบบถ่วงน้ำหนักและผ่านฟังก์ชันกระตุ้น (activation function) คำนวณค่าความคลาดเคลื่อน (error) ระหว่างคำตอบที่ทำนายได้กับเฉลย ถ้ามีความคลาดเคลื่อนเกิดขึ้น ระบบจะทำการปรับปรุงค่าน้ำหนักถ่วงของแต่ละการเชื่อมต่อ (connection) ทำนายข้อมูลได้ทั้งข้อมูลเชิงกลุ่ม (nominal data) และข้อมูลเชิงตัวเลข (numeric data) วิธีโครงข่ายประสาทเทียมอยู่ในหมวดวิธีที่เป็น Functions ใช้สมการทางคณิตศาสตร์ในการสร้างตัวแบบ

โครงข่ายประสาทเทียมเป็นศาสตร์ที่จำลองแบบความสามารถของมนุษย์ด้านการเรียนรู้ จดจำและจำแนกสิ่งต่าง ๆ ซึ่งใช้สมองเป็นส่วนสำคัญ ในการประมวลระบบของโครงข่ายประสาทเทียมนั้นจะเลียนแบบการทำงานของระบบสมองคือมีการส่งผ่านข้อมูลระหว่างกันโดยมีการเชื่อมต่อของเซลล์ประสาท (neuron) กันเป็นโครงข่ายร่างแหจำนวนมากและมีการประมวลผลในลักษณะขนาน (parallel processing) สาเหตุหลักที่โครงข่ายประสาทเทียมเป็นที่นิยมมากขึ้นเนื่องจากมีความยืดหยุ่นในการทำงานสูงและสามารถปรับตัวเองให้ทำงานในสภาพที่เปลี่ยนแปลงได้ดี อีกทั้งยังไม่จำเป็นต้องทราบตัวแบบทางคณิตศาสตร์ที่แน่นอนของกระบวนการ เพียงแต่ใช้ชุดข้อมูลที่ประกอบด้วยข้อมูลเข้า (input data) และข้อมูลเป้าหมาย (target data) ของกระบวนการในจำนวนที่มากพอมาใช้ในการสอนโครงข่ายประสาทเทียม

2.1.3.1 ความรู้พื้นฐานของระบบประสาท (Neural system knowledge)



รูปที่ 2.3 โครงข่ายของเซลล์ประสาท

ภายในสมองมนุษย์ประกอบด้วยหน่วยประมวลผลขนาดเล็ก เรียกว่า เซลล์ประสาท (neuron) ซึ่งจะมีประมาณ 10 หน่วย ในเซลล์ประสาทแต่ละหน่วยดังแสดงในรูปที่ 2.3 ประกอบด้วยใยประสาท (dendrites) ตัวเซลล์ (cell body) และเส้นใยประสาท (axon) ซึ่งแบ่งออกเป็น 4 บริเวณคือ

1) บริเวณนำกระแสประสาทเข้า (Input region) เป็นบริเวณที่จะมีการนำกระแสประสาท (nerve impulse) จากเซลล์ประสาทอื่นเข้ามาภายในตัวเซลล์โดยผ่านทางใยประสาท ซึ่งมีลักษณะแตกเป็นกิ่งก้านคล้ายต้นไม้และมีจำนวนตั้งแต่ 1 ใยขึ้นไป

2) บริเวณการรวมกระแสประสาทเข้า (Integration region) เป็นบริเวณที่มีการรวมกระแสประสาทก่อนที่จะเข้าสู่บริเวณการนำแสประสาทรวมออกจากเซลล์

3) บริเวณการนำกระแสประสาทรวมออกจากเซลล์ (Conduction region) เป็นบริเวณที่จะนำกระแสประสาทรวมออกจากเซลล์โดยใช้เส้นใยประสาทเป็นทางผ่านซึ่งมีเพียง 1 เส้นใยต่อเซลล์เท่านั้น

4) บริเวณการนำกระแสประสาทรวมออก (Output region) เป็นบริเวณส่วนปลายของเส้นใยประสาทที่มีการแตกแขนงใช้ในการถ่ายทอดกระแสประสาทข้ามเซลล์ไปยังเซลล์ประสาทอื่นโดยผ่านทางใยประสาทของเซลล์ประสาทนั้น

2.1.3.2 การเรียนรู้ของโครงข่ายประสาทเทียม (Artificial neural network learning)

การเรียนรู้ของโครงข่ายประสาทเทียมจะมีประสิทธิภาพเพียงใดนั้นขึ้นอยู่กับค่าถ่วงน้ำหนัก (weight) ของโครงข่ายที่ทำการออกแบบ ซึ่งการฝึกหัดโครงข่ายประสาทเทียมคือการหาค่าถ่วงน้ำหนักที่เหมาะสมให้กับโครงข่ายประสาทเทียมนั้น ๆ โดยทั่วไปสามารถจำแนกวิธีการเรียนรู้ของโครงข่ายประสาทเทียมได้เป็น 2 ประเภท คือ การเรียนรู้แบบมีผู้สอนและการเรียนรู้แบบไม่มีผู้สอน

1) การเรียนรู้แบบมีผู้สอน (Supervised learning)

การเรียนรู้แบบมีผู้สอนจะกำหนดข้อมูลฝึกหัด (training data set) ให้กับโครงข่ายประสาท ซึ่งกลุ่มนี้ประกอบด้วยข้อมูลเข้า (Input data) และข้อมูลเป้าหมาย (target data) ที่ต้องการ จากนั้นโครงข่ายประสาทเทียมจะทำการคำนวณค่าถ่วงน้ำหนักที่เหมาะสมให้กับข้อมูลฝึกหัด โดยคำตอบที่ได้จากโครงข่ายประสาทเทียมจะถูกคำนวณค่าความผิดพลาด (error value) ว่ามีความห่างจากคำตอบที่ต้องการของข้อมูลนำเข้าในชุดเดียวกันมากน้อยเพียงใด ถ้ายังมีความผิดพลาดสูงอยู่ การฝึกหัดจะดำเนินต่อไปจนกว่าค่าความผิดพลาดจะลดลงต่ำกว่าค่าที่ยอมรับได้ (accept level) จึงจะหยุดฝึกหัด สุดท้ายค่าถ่วงน้ำหนักที่ได้จะเป็นเหมือนฟังก์ชันที่ใช้ในการแปลงข้อมูล

2) การเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning)

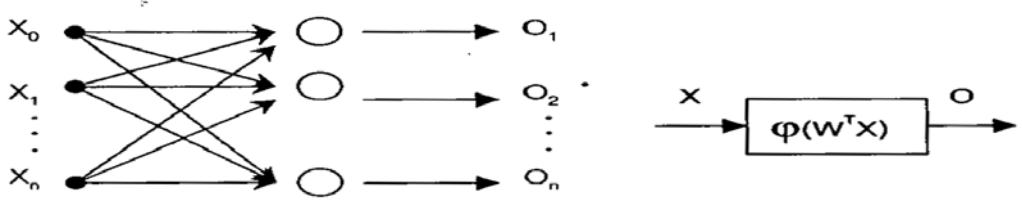
การเรียนรู้แบบไม่มีผู้สอนจะอาศัยชุดข้อมูลเข้าเพียงอย่างเดียวในการฝึกหัดโครงข่ายประสาทเทียมโดยไม่มีข้อมูลเป้าหมาย แต่จะใช้ข้อมูลออก (output data) จากโครงข่ายประสาทเทียมแทน เมื่อป้อนข้อมูลเข้าสู่โครงข่ายประสาทเทียม โครงข่ายประสาทเทียมจะคำนวณค่าความสัมพันธ์ที่มีอยู่ภายในกลุ่มข้อมูลเข้า โดยอาศัยค่าถ่วงน้ำหนักเป็นตัวแยกความแตกต่างของข้อมูลเข้าและนำไปเก็บไว้ในโหนดข้อมูลออกของโครงข่ายประสาทเทียม ซึ่งมีวัตถุประสงค์เพื่อใช้ในการจำแนกชุดข้อมูล (classification)

2.1.3.3 การเชื่อมโยงของโครงข่ายประสาทเทียม (Artificial neural network linking)

เพื่อให้โครงข่ายประสาทเทียมสามารถเรียนรู้ได้อย่างมีประสิทธิภาพ จำเป็นต้องมีการเชื่อมโยงกันระหว่างเซลล์ประสาท โดยทั่วไปสามารถแบ่งการเชื่อมโยงของโครงข่ายได้ 2 ลักษณะคือ

1) โครงข่ายแบบส่งสัญญาณไปข้างหน้า (Feedforward network)

เป็นโครงข่ายที่การประมวลผลจะอาศัยชุดข้อมูลปัจจุบันและส่งค่าที่ประมวลผลได้ไปยังชั้นถัด ๆ ไป กล่าวคือ โครงข่ายชนิดนี้จะประกอบด้วยชั้นต่าง ๆ โดยชั้นแรกจะเป็นชั้นข้อมูลเข้า (input layer) และชั้นสุดท้ายเป็นชั้นข้อมูลออก (output layer) ส่วนระหว่างชั้นข้อมูลเข้ากับชั้นข้อมูลออกอาจจะมีหรือ ไม่มีชั้นซ่อน (hidden layer) อยู่ภายในก็ได้ ซึ่งขึ้นกับกฎการเรียนรู้ (learning rule) ที่ใช้ในการสอนโครงข่าย เช่น ถ้าเป็นโครงข่ายเพอร์เซปตรอนแบบหลายชั้น (multi-layer perceptron) จะมีชั้นซ่อนอยู่ระหว่างชั้นข้อมูลเข้ากับชั้นข้อมูลออก ซึ่งอาจมีมากกว่าหนึ่งชั้นได้ การเชื่อมต่อระหว่างชั้นของโครงข่ายแบบส่งสัญญาณไปข้างหน้าจะมีค่าถ่วงน้ำหนัก (weight) เป็นตัวเชื่อมและสัญญาณนำเข้าที่เข้ามาจะถูกส่งไปตามทิศทางของลูกศรจนถึงชั้นข้อมูลออกโดยไม่มีการย้อนกลับ สามารถแสดงตัวแบบโครงข่ายแบบส่งสัญญาณไปข้างหน้าได้ดังรูปที่ 2.4

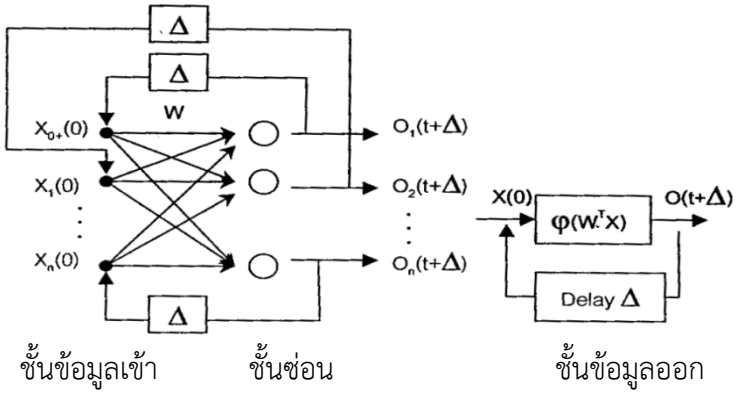


ชั้นข้อมูลเข้า ชั้นซ่อน ชั้นข้อมูลออก

รูปที่ 2.4 ลักษณะโครงข่ายประสาทเทียมแบบส่งสัญญาณไปข้างหน้า

2) โครงข่ายแบบมีการย้อนกลับ (Feedback network)

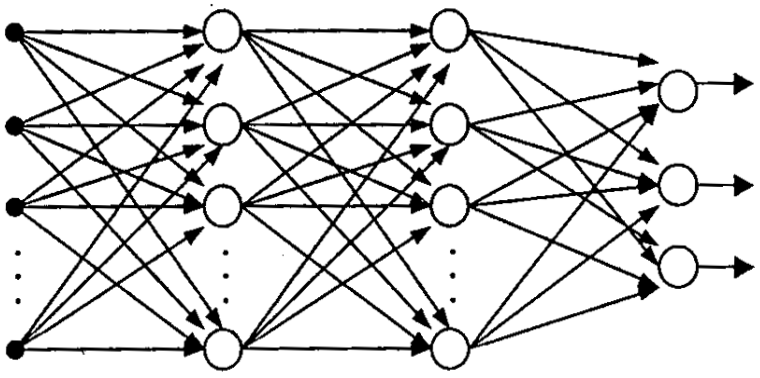
โครงข่ายชนิดนี้มีชื่อเรียกอีกชื่อหนึ่งว่า โครงข่ายหันกลับ (recurrent network) เป็นโครงข่ายที่จะอาศัยทั้งข้อมูลในปัจจุบันและข้อมูลที่มีการประวิงเวลามาใช้ในการประมวลผลของโครงข่ายประสาทเทียม สามารถแสดงตัวแบบโครงข่ายแบบมีการย้อนกลับได้ดังรูปที่ 2.5



รูปที่ 2.5 ลักษณะโครงข่ายประสาทเทียมแบบมีการย้อนกลับ

2.1.3.4 การแพร่แบบย้อนกลับ (Back-propagation)

การแพร่แบบย้อนกลับเป็นขั้นตอนที่ใช้สอนโครงข่ายประสาทเทียมแบบเพอร์เซปตรอนหลายชั้น (multi-layer perceptron) ซึ่งตัวแบบโครงข่ายประสาทเทียมมีการเชื่อมโยงกันเป็นโครงข่ายแบบเป็นชั้น ๆ โครงข่ายชนิดนี้มีการเชื่อมโยงกัน 3 ชั้น ประกอบด้วยชั้นข้อมูลเข้า (input layer) ถัดมาเป็นชั้นซ่อน (hidden layer) และชั้นสุดท้ายคือชั้นข้อมูลออก (output layer) สามารถแสดงโครงข่ายประสาทเทียมแบบเพอร์เซปตรอนหลายชั้นที่มีชั้นซ่อน 2 ชั้น ได้ดังรูปที่ 2.6



รูปที่ 2.6 ลักษณะโครงข่ายประสาทเทียมแบบเพอร์เซปตรอนหลายชั้น

ที่มาของชื่อการแพร่แบบย้อนกลับนั้นมาจากจุดที่ว่า วิธีการปรับค่าถ่วงน้ำหนักเพื่อให้ได้ค่าที่เหมาะสมนั้นจะใช้วิธีสอนว่าค่าเป้าหมาย (target) ของแต่ละข้อมูลเข้านั้นคืออะไร และใช้ค่าความผิดพลาด (error) ของข้อมูลออกมาใช้เป็นตัวชี้นำในการปรับค่าถ่วงน้ำหนัก ดังนั้นการแพร่แบบย้อนกลับ จึงเป็นกระบวนการเรียนรู้แบบมีผู้สอน แต่ปัญหาที่เกิดขึ้นคือไม่มีค่าเป้าหมายของสัญญาณที่ออกมาจาก แต่ละเซลล์ประสาทในชั้นซ่อน ดังนั้นจึงต้องอาศัยการแพร่ความผิดพลาดจากชั้นข้อมูลออกกลับมายังชั้นซ่อนนั่นเอง

2.1.3.5 ปัจจัยที่ส่งผลต่อการเรียนรู้การแพร่แบบย้อนกลับ

1) การกำหนดค่าเริ่มต้นของค่าถ่วงน้ำหนัก

ก่อนที่จะทำการสอนโครงข่ายประสาทเทียมแบบหลายชั้น จำเป็นต้องกำหนดค่าเริ่มต้นให้กับค่าถ่วงน้ำหนักที่เชื่อมโยงระหว่างชั้นทุกชั้น โดยค่านี้จะเป็นเลขจำนวนจริงที่มีค่าน้อย ๆ ที่ได้มาจากการสุ่มค่าเริ่มต้น (randomness)

2) การกำหนดเกณฑ์การหยุดฝึกหัด

เกณฑ์ในการหยุดฝึกหัดนั้นขึ้นกับผู้ที่ทำการออกแบบโครงข่ายประสาทเทียมว่าต้องการที่จะให้โครงข่ายประสาทเทียมมีความแม่นยำเพียงใด โดยทั่วไปนิยมใช้ค่าดัชนีที่ชี้ถึงค่าความผิดพลาดของระบบได้ ในงานวิจัยส่วนใหญ่ใช้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error, MSE)

3) อัตราการเรียนรู้ (Learning rate, η)

อัตราการเรียนรู้เป็นค่าสัมประสิทธิ์ที่แสดงถึงการเรียนรู้ของโครงข่าย โดยทั่วไปค่าที่เหมาะสมจะอยู่ในช่วง 0.05 ถึง 0.5 ถ้าอัตราการเรียนรู้มีค่าสูง แสดงว่ากำหนดให้โครงข่ายมีการเปลี่ยนแปลงค่าถ่วงน้ำหนักที่มาก ในทางตรงกันข้ามถ้ามีอัตราการเรียนรู้ต่ำ แสดงว่ากำหนดให้โครงข่ายมีการเปลี่ยนแปลงค่าถ่วงน้ำหนักที่น้อย ซึ่งจำเป็นต้องใช้เวลาในการเรียนรู้ที่มากขึ้น แต่จะมีข้อดีคือโครงข่ายจะมีเสถียรภาพและไม่เกิดการแกว่ง (oscillation) ขณะที่ทำการเรียนรู้

4) ค่าคงที่โมเมนตัม (Momentum constant, α)

ค่าคงที่โมเมนตัมเป็นค่าสัมประสิทธิ์ที่ช่วยหน่วงไม่ให้การเปลี่ยนแปลงค่าถ่วงน้ำหนักนั้นมีค่ามากเกินไป เป็นการเพิ่มเสถียรภาพให้กับโครงข่ายประสาทเทียมได้อีกทางหนึ่ง ซึ่งค่าโมเมนตัมที่เหมาะสมจะมีค่าเข้าใกล้ 1.0 และควรจะกำหนดให้สอดคล้องกับอัตราการเรียนรู้ด้วย เช่น ถ้าอัตราการเรียนรู้สูงก็ควรที่จะมีค่าโมเมนตัมที่ต่ำ ทำให้การเปลี่ยนแปลงค่าถ่วงน้ำหนักนั้นไม่มากจนเกินไป แต่ถ้าอัตราการเรียนรู้ต่ำก็ควรจะมีค่าโมเมนตัมที่สูง

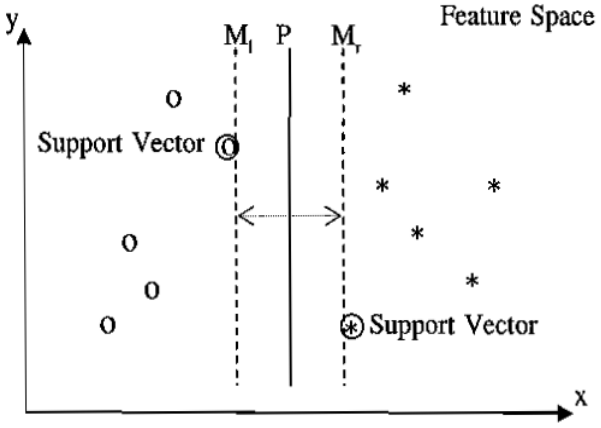
2.1.4 วิธีซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine Method)

เป้าหมายของวิธีซัพพอร์ทเวกเตอร์แมชชีนคือกระบวนการสอนเครื่องแบบมีผู้สอน (supervised learning) เพื่อให้สามารถสร้างตัวจำแนกข้อมูล (classifier) ที่มีความทั่วไป (generalize) สูง นั่นคือสามารถทำงานได้ดีกับตัวอย่างที่ไม่รู้จัก (unknown database) ด้วยกระบวนการปรับรูปแบบข้อมูลจากข้อมูลที่มีมิติต่ำ (low dimension dataset) บนพื้นที่ข้อมูลนำเข้า (input space) ให้อยู่ในรูปแบบของข้อมูลที่มีมิติสูง (high dimension dataset) บนพื้นที่ข้อมูลคุณลักษณะ (feature space) โดยใช้ฟังก์ชันในการปรับรูปแบบข้อมูลเรียกว่า **ฟังก์ชันเคอร์เนล (kernel function)** ซึ่งความสามารถดังกล่าวช่วยให้การสร้างตัวจำแนกข้อมูลด้วยสมการกำลังสอง (quadratic equation) บนพื้นที่ข้อมูลคุณลักษณะเป็นไปได้ง่ายขึ้นและมี

ความชัดเจนในการจำแนกกลุ่มมากยิ่งขึ้นด้วย นอกจากนี้ตัวจำแนกข้อมูลที่ดีควรมีโครงสร้างแบบเส้นตรง (linear classifier) และสามารถสร้างพื้นที่ระยะห่างระหว่างตัวจำแนกข้อมูลกับค่าที่ใกล้ที่สุดของแต่ละกลุ่มข้อมูลได้มากที่สุดเพื่อประสิทธิภาพในการแยกแยะประเภทของชุดข้อมูลแต่ละประเภทออกจากกันอย่างชัดเจน ซึ่งเส้นที่เหมาะสมดังกล่าวเรียกว่า ระนาบแบ่งเขตข้อมูลที่เหมาะสม (optimal separating hyperplane)

2.1.4.1 แนวความคิดของซัพพอร์ทเวกเตอร์แมชชีน

ซัพพอร์ทเวกเตอร์แมชชีนเป็นสมการที่ใช้ในการจำแนกค่าคุณลักษณะของ 2 กลุ่ม ที่วางตัวอยู่ในพื้นที่คุณลักษณะ (feature space) ออกจากกันโดยจะสร้างเส้นแบ่ง (plane) ที่เป็นเส้นตรงขึ้นมาและเพื่อให้ทราบว่าเส้นตรงที่แบ่ง 2 กลุ่มออกจากกันนั้น เส้นตรงใดที่เป็นเส้นที่ดีที่สุด โดยเส้นตรงนั้นจะเพิ่มเส้นขอบ (margin) ออกไปทั้งสองข้าง โดยเส้นขอบที่เพิ่มนั้นจะขนานกับเส้นเดิมเสมอ เส้นขอบที่เพิ่มขึ้นมานี้จะขยายออกไปจนกว่าจะสัมผัสกับค่าของกลุ่มตัวอย่างที่ใกล้ที่สุด ดังรูปที่ 2.7



รูปที่ 2.7 การขยายตัวของเส้นขอบ

จากรูปที่ 2.7 เส้น $\mathbf{M}_1$ และ $\mathbf{M}_r$ คือเส้นขอบที่ขยายออกไปด้านซ้ายและขวาตามลำดับ และ P คือเส้นแบ่งข้อมูลทั้ง 2 กลุ่ม เมื่อเส้น $\mathbf{M}_1$ และ $\mathbf{M}_r$ ขยายออกจนไปสัมผัสค่าข้อมูลที่ใกล้ที่สุด ซึ่งข้อมูลที่อยู่บนเส้นขอบของทั้งสองฝั่งนั้นเรียกว่า **ซัพพอร์ทเวกเตอร์ (support vector)** จะวัดค่าระยะความห่างของเส้นขอบ โดยเส้น P จะเปลี่ยนความชันไปเรื่อย ๆ เพื่อที่จะหาความกว้างสูงสุดของเส้นขอบ

กระบวนการโดยรวมของซัพพอร์ทเวกเตอร์แมชชีนนั้นเป็นการหาค่าความชันของเส้น P ที่เส้นขอบมีความกว้างสูงสุด

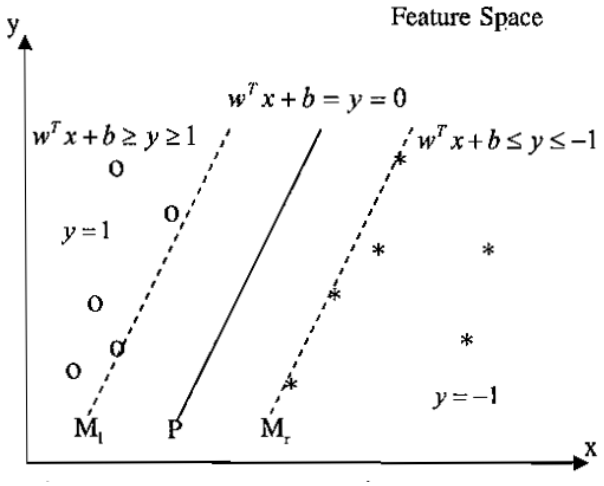
2.1.4.2 สมการพื้นฐานของซัพพอร์ทเวกเตอร์แมชชีน

ถ้านำแนวคิดของซัพพอร์ทเวกเตอร์แมชชีนที่กล่าวไปแล้วในข้อ 2.1.4.1 มาเขียนเป็นสมการเพื่อใช้ในการแก้ปัญหา โดยข้อมูลที่นำมาวางลงในพื้นที่คุณลักษณะนั้นเป็นกลุ่มข้อมูลที่อยู่ในรูปของเวกเตอร์

$$x = ((x_1, y_1),...,(x_i, y_i)) \quad (2.5)$$

เมื่อ x คือ ชุดค่าคุณลักษณะ

ค่าคุณลักษณะที่วางตัวอยู่ในพื้นที่คุณลักษณะจะถูกแบ่งด้วยเส้นตรงดังรูปที่ 2.8 และเมื่อนำเส้นตรงมาแทนค่าด้วยสมการเส้นตรง $y = mx + b$ โดยมีการกำหนดกลุ่มของข้อมูลทั้งสองฝั่งเป็นเพียง 2 ค่า ที่ซึ่งแทนด้วยค่า y เพื่อให้ข้อมูลที่อยู่ในกลุ่มเดียวกันที่มาจากหลายค่ากลายเป็นค่าเดียว ดังสมการในรูปที่ 2.8



รูปที่ 2.8 เส้นขอบและเส้นแบ่งเมื่อแทนด้วยสมการเส้นตรง

จากรูปที่ 2.8 เส้นตรง M_1 แทนด้วยสมการ $w^T x + b \geq y \geq 1$ ซึ่งข้อมูล y ที่มากกว่า 1 ก็จะถูกกำหนดค่าใหม่โดยให้ y เท่ากับ 1 และพจน์ w ก็คือค่าความชัน เช่นเดียวกันกับเส้นตรง M_r ที่ค่าของ y จะถูกกำหนดค่าใหม่เมื่อ y ที่น้อยกว่า –1 ให้เท่ากับ –1 ดังนั้นสมการที่เกิดขึ้นใหม่ จากสมการเส้นขอบ 2.6 และ 2.7 สามารถกำหนดได้ดังสมการที่ 2.8

เมื่อ $w^T x + b \geq y$ กำหนด y = 1 (2.6)

$w^T x + b \leq y$ กำหนด y = –1 (2.7)

$y(w^T x + b) - 1 \geq 0$ (2.8)

โดย y คือ ค่ากลุ่มข้อมูล (1, –1)
w คือ ค่าความชัน
x คือ ค่าคุณลักษณะ
b คือ ค่าคงที่ (ค่าตัดแกน y)

2.1.4.3 ค่าความกว้างเส้นขอบ (Margin)

การคำนวณความกว้างของเส้นขอบต้องทำการคำนวณพจน์ w ให้อยู่ในรูปปกติมาตรฐาน (normalization) โดยคำนวณจากสมการที่ 2.6 และ 2.7 เมื่อแทนค่า y ลงไปแล้ว

$w^T x^- + b = 1$

$w^T x^- + b = -1$

$w^T (x^+ - x^-) = 2$

$$M = \left(\frac{w}{\|w\|}\right)^{T}\left(x^{+}-x^{-}\right)$$

$$= \frac{2}{\|w\|} \quad (2.9)$$

$$w = \sum_{i=1}^{N}\alpha_i y_i x_i \quad (2.10)$$

โดยที่ M คือ ความกว้างของเส้นขอบ

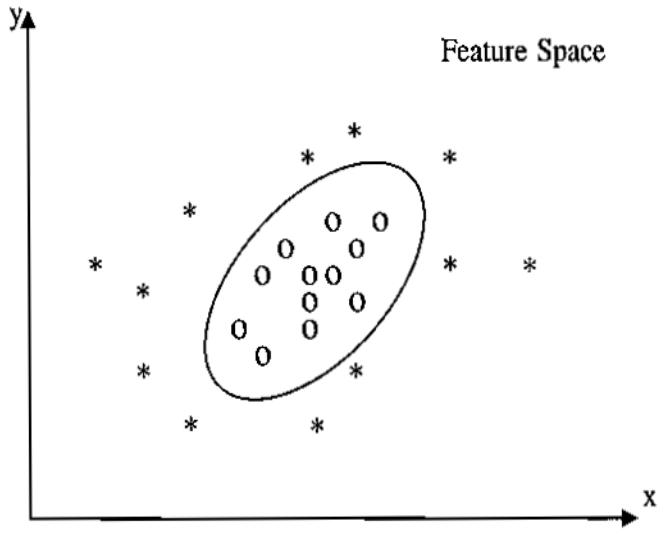
α คือ สัมประสิทธิ์คงที่

เมื่อนำค่า w ไปใส่ในสมการที่ 2.8 ซึ่งเป็นสมการในการหาเส้นแบ่ง จะได้

$$y_i\left(\sum_{i=1}^{N}\alpha_i y_i\left(x_i^{T}, x_j\right)+b\right)-1 \geq 0 \quad (2.11)$$

2.1.4.4 เคอร์เนล (Kernel)

ในความเป็นจริงนั้นข้อมูล 2 กลุ่ม ไม่ได้วางตัวในพื้นที่คุณลักษณะ และไม่สามารถแบ่งได้โดยเส้นตรง แต่ข้อมูลอาจจะจับกลุ่มกันในตำแหน่งต่าง ๆ ดังนั้นจึงเป็นปัญหาทำให้ไม่สามารถที่จะใช้สมการซัพพอร์ทเวกเตอร์แมชชีนแบบเชิงเส้นได้ ดังนั้นจะต้องมีเครื่องมือมาช่วยให้ข้อมูลเหล่านั้นเรียงตัวใหม่ในพื้นที่ เรียกว่า **พื้นที่หลายมิติ (higher dimensional space)**



รูปที่ 2.9 รูปแบบการวางตัวที่ไม่สามารถแบ่งด้วยเส้นตรงได้

ในเคอร์เนลนั้นคือการคูณกันของชุดเวกเตอร์ของ x ใด ๆ

$$K\left(x_i, x_j\right) = x_i^{T}x_j \quad (2.12)$$

เคอร์เนลที่นิยมใช้มีอยู่ 3 ชนิด คือ

1) โพลิโนเมียล (Polynomial)

$$K(x_i, x_j) = (< x_i, x_j > +1)^d \quad (2.13)$$

เมื่อ d คือ ค่าเลขยกกำลัง

2) ฟังก์ชันเบสิสเรเดียล (Radial Basis Function : RBF)

$$K(x_i, x_j) = \exp\left(-\frac{x_i - x_j^2}{2\sigma^2}\right) \quad (2.14)$$

เมื่อ σ คือ ค่าพารามิเตอร์

3) ซิกมอยด์ (Sigmoid)

$$K(x_i, x_j) = \tanh(k < x_i, x_j > +\mu) \quad (2.15)$$

เมื่อ k, μ คือค่าพารามิเตอร์

ดังนั้นจากสมการของเคอร์เนลนั้นสามารถที่จะแทนลงไปในตำแหน่งของ x_i^T, x_j ในสมการที่ 2.11 จึงเขียนเป็นสมการใหม่ดังนี้

$$y_i \sum_{i=1}^{N} (\alpha_i y_i K(x_i, x_j) + b) - 1 \geq 0 \quad (2.16)$$

สมการที่ 2.16 เป็นสมการที่ใช้ในขั้นตอนที่จะเรียนรู้ว่าจะวางตำแหน่งเส้นแบ่งไว้ที่ตำแหน่งใด โดยทำงานร่วมกับเคอร์เนล เพื่อแปลงให้ข้อมูลที่ยากต่อการแบ่งแบบเชิงเส้นสามารถแบ่งได้เมื่อทำให้เป็นข้อมูลแบบหลายมิติ (higher dimension) ดังนั้นจึงมีอีกสมการหนึ่งที่ใช้ค่า w และ b เดิมมาจัดตำแหน่งของข้อมูลเพื่อที่ให้ทราบว่าข้อมูลนั้นเป็นกลุ่มใด กำหนดได้ดังสมการที่ 2.17

$$f(x) = \text{sgn}\left(\sum_{i=1}^{\infty} (\alpha_i y_i K(x_i, x_j) + b)\right) \quad (2.17)$$

เมื่อ $f(x)$ คือค่า y หาในรูปของ x

2.1.5 วิธีฐานกฎ (Rule Based Method) ใช้ชุดลำดับของกฎมาสร้างรูปแบบการแยกประเภทข้อมูล โดยส่วนใหญ่แล้วจะใช้กฎที่เป็น If ... then ซึ่งเป็นกฎอย่างง่าย (Murti, S. and Mahantappa, M., 2012) ใช้อัลกอริทึม Decision Table เป็นเครื่องมือที่ใช้แสดงเงื่อนไขการตัดสินใจและเลือกการทำงานหรือกระทำกิจกรรมภายใต้เหตุการณ์ของเงื่อนไขที่ระบุ วิธีการตัดสินใจแบบ Decision Table จะเป็นตาราง 2 มิติ

2.1.6 วิธีการถดถอยโลจิสติกแบบ 2 กลุ่ม (Binary Logistic Regression Method) เป็นการวิเคราะห์การถดถอยแบบหนึ่งโดยที่ตัวแปรตามเป็นตัวแปรเชิงคุณภาพที่มีค่าได้เพียง 2 ค่า (dichotomous or binary variable) ส่วนตัวแปรอิสระอาจจะเป็นตัวแปรเชิง

ปริมาณหรือเชิงคุณภาพ หรืออาจจะมีทั้งตัวแปรเชิงปริมาณและตัวแปรเชิงคุณภาพก็ได้ (กัลยา วานิชย์บัญชา, 2552)

2.1.7 วิธีนาอีฟ เบย์ (Naïve Bayes Method) คืออัลกอริทึมที่ใช้หลักการของความน่าจะเป็นในการคัดกรองแต่ละคำตอบ (Class) โดยมีคำตอบ 2 คำตอบ (สายชล สินสมบูรณ์ทอง, 2558)

2.2 รายงานวิจัยที่เกี่ยวข้อง

พิจิตรา จอมศรี (2549) ได้ทำการวิจัยเรื่องการทำนายเนื้อหาของเว็บโดยใช้เทคนิคเหมืองข้อมูล กรณีศึกษามหาวิทยาลัยศิลปากร งานวิจัยนี้ได้นำเทคนิคกฎการค้นหาความสัมพันธ์ซึ่งเป็นเทคนิคหนึ่งในเทคนิคเหมืองข้อมูลมาประยุกต์ใช้สร้างแบบเพื่อทำนายข้อมูลการเรียกใช้เว็บในอนาคต โมเดลที่สร้างขึ้นสามารถทำนายเนื้อหาเว็บที่จะถูกเรียกใช้ในวันถัดมาได้ ผลของการใช้เทคนิคเหมืองข้อมูลพบว่าตัวแบบที่สร้างขึ้นสามารถทำนายเนื้อหาเว็บที่จะถูกเรียกใช้ได้โดยมีความถูกต้องร้อยละ 66.67

กมลวรรณ คงทรัพย์ (2551) ทำการวิจัยเรื่องความคิดเห็นของผู้ปกครองต่อการเล่นเกมออนไลน์ของนักเรียน ซึ่งเป็นการวิจัยเชิงสำรวจ (Survey Research) พบว่าผลกระทบต่อการศึกษาที่เกิดจากการเล่นออนไลน์ของนักเรียน ผู้ปกครอง ส่วนใหญ่มีความเห็นว่าการเล่นเกมออนไลน์มีผลทำให้นักเรียนไปโรงเรียนสายและนักเรียนไม่อ่านหนังสือทบทวนบทเรียนเลย ทำให้ผลการเรียนวิชาคณิตศาสตร์ตกต่ำที่สุด

สุรพล สีห์สุรางค์ (2551) ได้ทำการวิจัยเรื่องพฤติกรรมการเล่นเกมคอมพิวเตอร์ของเด็กนักเรียนในเขตอำเภอเมืองเชียงใหม่ โดยการสัมภาษณ์และใช้แบบสอบถามจากกลุ่มตัวอย่าง ซึ่งเป็นเด็กนักเรียนที่เล่นเกมคอมพิวเตอร์จากร้านเกมคอมพิวเตอร์ในเขตอำเภอเมืองเชียงใหม่ พบว่าการเล่นเกมคอมพิวเตอร์ทำให้ผลการเรียนลดลง มีผลทำให้ตาพร่า/ปวดศีรษะ/สุขภาพอ่อนแอ และทำให้ไม่รู้จักแบ่งเวลา

กาญจนา หฤหรรษพงศ์ (2552) ทำการวิจัยเรื่องการค้นหาความรู้จากฐานข้อมูลนักศึกษาโดยใช้เทคนิกการทำเหมืองข้อมูล : กรณีศึกษามหาวิทยาลัยวลัยลักษณ์ งานวิจัยนี้นำเสนอการทำเหมืองข้อมูลโดยใช้เทคนิกการค้นหากฎความสัมพันธ์ การจำแนกประเภทข้อมูลและได้นำเสนออัลกอริทึมใหม่ในการค้นหารูปแบบลำดับ ความรู้ที่ได้สามารถนำไปใช้เป็นแนวทางในการเสนอแนะแนวการเรียนแก่นักเรียน

ธานินทร์ เสวกจันทร์ (2552) ทำการวิจัยเรื่องการศึกษาพฤติกรรมและผลกระทบจากการเปิดรับสื่อเกมออนไลน์ของนักเรียนมัธยมศึกษาในเขตอำเภอเมือง จังหวัดสุราษฎร์ธานี โดยใช้แบบสอบถามในการเก็บรวบรวมข้อมูล ผลการศึกษาพบว่ากลุ่มตัวอย่างมีผลกระทบด้านสุขภาพมากที่สุด (=3.02) รองลงมาคือผลกระทบด้านการศึกษา (=2.87) และผลกระทบด้านการเงิน (=2.53) ตามลำดับ

ภัทรพงศ์ พงศ์ภัทรกานต์ (2552) ได้นำเสนอการเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลของแบบจำลอง C5.0, CART, SVM และ SVM ร่วมกับ C5.0 ภายใต้หลักการทำงานของ

เหมืองข้อมูลโดยใช้ชุดข้อมูลจำนวน 9 ชุด ทำการเลือกชุดข้อมูลที่เป็นไบนารีคลาสและมีคุณลักษณะ (attribute) กับจำนวนตัวอย่างที่หลากหลายรูปแบบ SVM จะทำการคัดแยกชุดข้อมูลออกเป็นไบนารีคลาสได้ดีมาเปรียบเทียบประสิทธิภาพในการจำแนกกลุ่มข้อมูลโดยใช้ C5.0, CART, SVM และ SVM ผสมผสานกับ C5.0 โดยทำการทดลองวัดประสิทธิภาพความถูกต้องเปรียบเทียบกัน ซึ่งผลการทดลองพบว่าแบบจำลอง SVM ผสมผสานกับ C5.0 ที่ผู้วิจัยนำเสนอมีประสิทธิภาพสูงที่สุดทุกชุดข้อมูลจากจำนวน 9 ชุด ที่ได้ทำการทดลอง ซึ่งสรุปผลได้ว่าการใช้ตัวแบบผสมผสานกันสามารถจำแนกประเภทข้อมูลประเภทไบนารีคลาสได้อย่างมีประสิทธิภาพสูงกว่าการใช้ตัวแบบ C5.0, CART และ SVM เพียงอย่างเดียว

อุมาพร กีรติบัญชร (2552) ทำการวิจัยเรื่องการศึกษาการทำเหมืองข้อมูลผู้ใช้บริการเว็ปไซต์ งานวิจัยนี้มีวัตถุประสงค์เพื่อวิเคราะห์ถึงการศึกษาการทำเหมืองข้อมูลผู้ใช้บริการเว็ปไซต์และการสร้างโปรแกรมสำหรับทำเหมืองข้อมูล จากข้อมูลพฤติกรรมของผู้เข้าใช้เว็บบอร์ดของมหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี โดยใช้เทคนิคการหากฎความสัมพันธ์ ผลการวิจัยทำให้สามารถวิเคราะห์ข้อมูลจากการทำเหมืองข้อมูลเพื่อทำนายปัจจัยที่เกี่ยวข้องกับการเข้าใช้เว็บบอร์ดของนักศึกษา และสามารถจัดการบริหารเว็บบอร์ดให้มีประสิทธิภาพโดยโปรแกรมสามารถกำหนดเพิ่มตัวแปรที่สนใจได้

วาทินี นุ้ยเพียร และคณะ (2553) ได้ทำการศึกษาเพื่อเปรียบเทียบวิธีการจำแนกข้อมูลโดยเลือกใช้อัลกอริทึมโครงข่ายประสาทแบบมัลติเลเยอร์เปอร์เซ็ปตรอน ซัพพอร์ตเวกเตอร์แมชชีน นาอีฟเบย์และความใกล้เคียงกันมากที่สุดเพื่อประเมินประสิทธิภาพค่าความถูกต้อง (accuracy) ค่าความแม่นยำ (precision) ค่าความระลึก (recall) และค่าความถ่วงดุล (F-measure) ใช้ข้อมูลจาก UCI ประกอบด้วย Ozone Days และ Adult เลือกกลุ่มข้อมูลโดยมีจำนวนคำตอบ (class) เท่ากันในข้อมูลแต่ละชุด เป็นการทดลองแบบมีการเรียนรู้ จากผลการวิจัยอัลกอริทึมที่ดีที่สุดของข้อมูล Ozone Days คือซัพพอร์ตเวกเตอร์แมชชีน ฟังก์ชันเคอร์เนลแบบ Rbf มีค่าความถูกต้อง 94.83% ค่าความแม่นยำ 96% ค่าความระลึก 96% และค่าความถ่วงดุล 96% ส่วนข้อมูล Adult คือซัพพอร์ตเวกเตอร์แมชชีน ฟังก์ชันเคอร์เนลแบบโพลิโนเมียล มีค่าความถูกต้อง 79.66% ค่าความแม่นยำ 80% ค่าความระลึก 80% และค่าความถ่วงดุล 80% อัลกอริทึมที่สามารถเลือกใช้ได้ดีคือ ซัพพอร์ตเวกเตอร์แมชชีน สามารถใช้กับลักษณะข้อมูลที่เป็นตัวเลขหรือข้อมูลเชิงกลุ่มแบบข้อความ ซึ่งเทคนิคเหล่านี้สามารถประยุกต์ใช้กับการสร้างเทคโนโลยีการจัดเก็บและนำเสนอเนื้อหาแบบมีโครงสร้าง โดยสามารถวิเคราะห์จำแนกหรือจัดแบ่งข้อมูลที่มีความสัมพันธ์กับข้อมูลอื่น ๆ แบบเชิงความหมายได้ต่อไป

ชุติมา อุตมะมุณีย์ และประสงค์ ปราณีตพลกรัง (2554) ทำการวิจัยการพัฒนาตัวแบบระบบสนับสนุนการตัดสินใจแบบอัตโนมัติออนไลน์สำหรับการเลือกสาขาวิชาเรียนของนักศึกษาระดับอุดมศึกษา โดยใช้เทคนิคการจัดทำเหมืองข้อมูลและใช้โปรแกรม WEKA ในการสร้างแบบจำลอง ผลการวิจัยพบว่าเทคนิคข่ายงานเบย์ได้ผลลัพธ์ดีที่สุด สามารถบ่งบอกตัวแปรสำคัญที่มีผลต่อการตัดสินใจในการเลือกสาขาวิชาเรียนได้

รุจิรา ธรรมสมบัติ (2555) ได้ทำวิจัยเรื่องระบบสนับสนุนการตัดสินใจในการเลือกใช้แพคเกจอินเทอร์เน็ตมือถือ โดยใช้ต้นไม้ตัดสินใจ งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาระบบสนับสนุนการตัดสินใจในการเลือกใช้แพคเกจอินเทอร์เน็ตมือถือ โดยใช้ต้นไม้ตัดสินใจ และเป็น

แนวทางในการตัดสินใจเลือกใช้แพคเกจอินเทอร์เน็ตมือถือจากพฤติกรรมของผู้ใช้บริการแต่ละคน โดยการเปรียบเทียบโมเดลที่ถูกสร้างขึ้นจากอัลกอริทึมต้นไม้ตัดสินใจคือ ID3 และ C4.5 (J48) เพื่อหาโมเดลที่มีค่าความถูกต้องมากที่สุดก่อนนำมาพัฒนาระบบ ในการพัฒนาระบบในภาษา ASP.NET ด้วย C# และฐานข้อมูล SQL Server 2008 โดยพัฒนาขึ้นในลักษณะของเว็บแอพพลิเคชั่น (Web-Based Application) และใช้โปรแกรม Weka 3.6.2 เพื่อตรวจสอบความถูกต้องของโมเดลที่ระบบได้สร้างขึ้นมา ผลที่ได้คือโมเดลที่ถูกสร้างจากอัลกอริทึมต้นไม้ตัดสินใจ ID3 มีค่าความถูกต้องมากกว่า C4.5 (J48) โดยมีค่าความถูกต้อง (Correctly Classified Instances) เมื่อทดสอบกับกลุ่มข้อมูลสำหรับการเรียนรู้ (Training Data) จำนวน 1,000 ชุด เท่ากับ 92.3% และเมื่อนำอัลกอริทึมต้นไม้ตัดสินใจ ID3 ไปทดสอบกับชุดข้อมูลทดสอบ (Testing Data) จำนวน 500 ชุด ให้ผลการทดสอบโดยมีค่าความถูกต้องเท่ากับ 92.2% และเมื่อพิจารณาค่า Confusion Matrix พบว่าผลของการทำนายจากโมเดลมีจำนวนข้อมูลค่าจริงกับจำนวนข้อมูลจากการทำนายของโมเดลมีผลลัพธ์ตรงกัน ได้ค่าเฉลี่ยรวมเท่ากับ 83.06% ซึ่งเป็นค่าเฉลี่ยที่อยู่ในระดับค่อนข้างสูง สามารถนำโมเดลที่ได้ไปพัฒนาระบบต่อไป

ศิรมณ์ เสถียรรัมย์ และกฤษฎา ศรีแผ้ว (2557) ได้ทำวิจัยเรื่องความสัมพันธ์ระหว่างพฤติกรรมการเล่นเกมออนไลน์จากข้อมูลจราจรคอมพิวเตอร์และผลสัมฤทธิ์ทางการเรียนของนักเรียน กรณีศึกษาโรงเรียนธัญรัตน์ โดยมีวัตถุประสงค์เพื่อนำเสนอการศึกษาเชิงประจักษ์ในการทำนายผลสัมฤทธิ์ทางการเรียนโดยใช้ข้อมูลพื้นฐานของนักเรียนและพฤติกรรมการเล่นเกมออนไลน์ เป้าหมายเพื่อศึกษาความสัมพันธ์ของพฤติกรรมการเล่นเกมออนไลน์และผลสัมฤทธิ์ทางการเรียนว่ามีความสัมพันธ์กันหรือไม่ ข้อมูลที่ใช้ในการศึกษาครั้งนี้ประกอบด้วย ข้อมูลการเล่นเกมออนไลน์ของนักเรียนจากข้อมูลบันทึกจราจรทางคอมพิวเตอร์ที่เก็บที่ Proxy เครื่องแม่ข่าย โดยใช้ข้อมูลตลอดระยะเวลา 1 ภาคเรียน จำนวน 700,000 ระเบียน จากนักเรียน 2,398 คน ผลการทดลองเมื่อทำการวิเคราะห์ข้อมูลด้วยวิธีหาปัจจัยอิทธิพลพบว่าพฤติกรรมการเล่นเกมออนไลน์ไม่ได้สัมพันธ์กับผลสัมฤทธิ์ทางการเรียน แต่ผลสัมฤทธิ์ทางการเรียนโดยส่วนใหญ่จะขึ้นอยู่กับปัจจัยเรื่องเพศ ชั้นปี มากกว่าปัจจัยอื่น ๆ

Yiming Ma, Bing Liu, ChingKian Wong, Philip, S. Yu and Shuik Ming Lee (2000) ทำการวิจัยเกี่ยวกับการกำหนดเป้าหมายการเรียนที่ถูกต้องโดยใช้เทคนิคการขุดค้นเหมืองข้อมูล ซึ่งการศึกษานี้มีความน่าสนใจและมีความท้าทายในเทคนิคการทำเหมืองข้อมูลมาประยุกต์ใช้โปรแกรมเหล่านี้สามารถช่วยได้ทั้งนักการศึกษาและนักเรียนและปรับปรุงคุณภาพของการศึกษา

BehrouzMinaei-Bidgoli, Deborah, A. Kashy, GerdKortemeyer and William F. Punch (2003) ทำการวิจัยเรื่องการทำนายสมรรถภาพของนักเรียนที่มีการเรียนตามระบบ LON-CAPA โดยใช้เทคนิคเหมืองข้อมูลเพื่อจำแนกกลุ่มนักศึกษา เพื่อการทำนายผลการเรียนเมื่อสิ้นสุดการเรียนแต่ละวิชา โดยดูจากประวัติการเข้าเรียน (Logged Data) ผ่านทางเว็บโดยวิเคราะห์ความสามารถของผู้เรียนและจัดกลุ่มนักศึกษาที่มีความคล้ายคลึงกัน

Xhemali, D. et. al. (2009) ได้ทำวิจัยเรื่องวิธีนาอีฟเบย์ แผนภาพต้นไม้เพื่อการตัดสินใจ และโครงข่ายประสาทเทียมในการจำแนกกลุ่มของการฝึกหัดหน้าเวป (web page) โดยการจำแนกกลุ่มเวปใช้เทคโนโลยีที่แตกต่างกันจำนวนมาก ในการศึกษานี้เน้นการเปรียบเทียบวิธีนาอีฟเบย์ วิธีแผนภาพต้นไม้เพื่อการตัดสินใจ และวิธีโครงข่ายประสาทเทียมสำหรับการวิเคราะห์

อัตโนมัติและการจำแนกกลุ่มของข้อมูลคุณลักษณะจากหลักสูตรฝึกหัดหน้าเวป เขาได้แนะนำวิธีนาอีฟเบย์และประมวลผลตัวอย่างข้อมูลชุดเดียวกันโดยใช้วิธีแผนภาพต้นไม้เพื่อการตัดสินใจและวิธีโครงข่ายประสาทเทียม เพื่อหาอัตราความสำเร็จของวิธีการเหล่านี้ในหลักสูตรฝึกหัดหน้าเวป ผลการวิจัยนี้พบว่าโดยทั้งหมด วิธีนาอีฟเบย์ดีที่สุดสำหรับหลักสูตร ฝึกหัดหน้าเวป มีค่าความถ่วงดุล (F-measure) มากกว่า 97% ทั้ง ๆ ที่ผึกหัดด้วยขนาดตัวอย่างที่น้อยกว่าการจำแนกกลุ่มที่เคยใช้

Tuisima, S. et. al. (2012) ศึกษาเรื่องการจัดจำแนกกลุ่มของระดับการติดเกมคอมพิวเตอร์ในนักศึกษาชั้นมัธยมศึกษาปีที่ 1-3 โดยใช้โครงข่ายประสาทเทียม โดยใช้กลุ่มตัวอย่างนักศึกษาจำนวน 32 ราย ซึ่งเกมมีบทบาทต่อการอยู่อาศัยในแต่ละวัน เก็บรวบรวมข้อมูลระหว่างวันที่ 18 พฤษภาคม ถึงวันที่ 26 กรกฎาคม ปี ค.ศ. 2011 โดยจำแนกระดับการติดเกมโดยใช้โครงข่ายประสาทเทียมด้วยอัลกอริทึมการแพร่แบบย้อนกลับและเปรียบเทียบผลกับแผนภาพต้นไม้ตัดสินใจ ทำการวัดความถูกต้องของตัวแบบโดยใช้วิธี 10-fold Cross Validation งานวิจัยนี้จำแนกเกมคอมพิวเตอร์พิจารณาจากคุณลักษณะ 4 ประเภท คือ เกมระยะยาว เกมที่ไม่มีกฎเกณฑ์แน่นอน เกมระยะเวลาจริง และเกมพื้นฐาน ผลการทดลองพบว่าความถูกต้องของโครงข่ายประสาทเทียมด้วยอัลกอริทึมการแพร่แบบย้อนกลับสำหรับเกมระยะยาว เกมที่ไม่มีกฎเกณฑ์แน่นอน เกมระยะเวลาจริง และเกมพื้นฐานคือ 97.75, 91.35, 90.00 และ 97.73 ตามลำดับ ส่วนความถูกต้องของอัลกอริทึมแผนภาพต้นไม้ตัดสินใจคือ 88.76, 92.63, 87.50 และ 90.91 ตามลำดับ

Chou, C. H. (2013) ศึกษาการใช้ Tic-Tac-Toe สำหรับเรียนรู้การจำแนกกลุ่มการทำเหมืองข้อมูลและการประเมินผล โดยเกม Tic-Tac-Toe เป็นเกมที่ได้รับความนิยม ใช้ผู้เล่น 2 คน วัตถุประสงค์ของการศึกษานี้เพื่อตรวจสอบว่าการเรียนรู้กลไกลวิธีไหนที่ประสบความสำเร็จในการจำแนกกลุ่มของเกมนี้ ผลการทดลองพบว่าวิธี 3-ความใกล้เคียงกันมากที่สุดมีอัตราความถูกต้องในการจำแนกกลุ่มของเกม Tic-Tac-Toe มากที่สุดคือ 99% วิธีการจำแนกกลุ่มทั้ง 4 วิธี คือ วิธีโครงข่ายประสาทเทียม วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีโลจิสติก และวิธี 3-ความใกล้เคียงกันมากที่สุด มีอัตราความถูกต้องมากกว่า 98% ซึ่งวิธีการทั้ง 4 วิธี ไม่มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ