บทที่ 2

เอกสารและงานวิจัยที่เกี่ยวข้อง

ในงานวิจัยนี้ผู้วิจัยได้ศึกษาแนวคิด ทฤษฎี ตลอดจนงานวิจัยที่เกี่ยวข้อง เพื่อเป็นประโยชน์ในการทำวิจัย ประกอบด้วยเนื้อหาต่าง ๆ ดังนี้

การพยากรณ์

การพยากรณ์ หมายถึง การคาดคะเน หรือการทำนายเหตุการณ์ที่จะเกิดขึ้นในอนาคต โดยศึกษารูปแบบการเกิดของเหตุการณ์จากข้อมูลที่เก็บรวบรวมไว้อย่างมีระบบ (ทรงศิริ แต้สมบัติ, 2549) การพยากรณ์มีความสำคัญอย่างยิ่งต่อบุคคล และองค์กรในงานสาขาต่าง ๆ เพราะจะทำให้ทราบว่าจะเกิดเหตุการณ์ใดในอนาคตด้วยความเชื่อมั่นระดับหนึ่ง จะทำให้เกิดประโยชน์สูงสุดแก่องค์กร

โดยปกติจะจำแนกการพยากรณ์ออกเป็น 2 กลุ่ม ใหญ่ ๆ คือ

การพยากรณ์เชิงปริมาณ (Quantitative Forecasting Method) (สุรเชษฐ์ เถื่อนแก้วสิงห์, 2552) เป็นการพยากรณ์ที่ใช้รูปแบบทางคณิตศาสตร์หนึ่งอย่าง หรือมากกว่า โดยอาศัยข้อมูล หรือตัวเลขจากอดีตมาสร้างเป็นตัวแบบในการพยากรณ์ ซึ่งใช้เพื่อคาดคะเนเหตุการณ์ในอนาคต ส่วนมากนำไปใช้ในทางธุรกิจ เทคนิคที่นิยมใช้ในการพยากรณ์ คือ การพยากรณ์อนุกรมเวลา (Time Series Forecasting Method) และการพยากรณ์เชิงสาเหตุ หรือการพยากรณ์ความสัมพันธ์ (Causal Forecasting Method)

การพยากรณ์เชิงคุณภาพ (Qualitative Forecasting Method) (สุรเชษฐ์ เถื่อนแก้วสิงห์, 2552) เป็นการพยากรณ์ที่อาศัยความรู้ ความสามารถ ประสบการณ์ ความชำนาญ รวมทั้งวิจารณญาณของผู้ทำการพยากรณ์โดยตรง โดยไม่ใช้รูปแบบทางคณิตศาสตร์ จึงทำการตรวจสอบความแม่นยำของการพยากรณ์ได้ยากกว่าการพยากรณ์เชิงปริมาณ ซึ่งในการเลือกใช้การพยากรณ์เชิงคุณภาพอาจเกิดจาก ผู้ทำการพยากรณ์ไม่ได้เก็บรวบรวมข้อมูลไว้ ไม่ชำนาญในการพยากรณ์เชิงปริมาณ หรือไม่มีความเชื่อถือในการพยากรณ์เชิงปริมาณ เป็นต้น การพยากรณ์เชิงคุณภาพมีหลายวิธี เช่น การคาดคะเน หรือประมาณการ การระดมความคิด การพยากรณ์โดยการสำรวจตลาด การพยากรณ์โดยยึดอดีตเป็นหลัก การพยากรณ์ด้วยเทคนิคเดลไฟ (Delphi Method) เป็นต้น

1. องค์ประกอบของการพยากรณ์ที่ดี

วิธีการที่จะทำให้การพยากรณ์สามารถประเมินผลได้อย่างแม่นยำ และมีความถูกต้องใกล้เคียงกับความจริง (สุรเชษฐ์ เถื่อนแก้วสิงห์, 2552) คือ

1. ระบุวัตถุประสงค์ในการนำเอาผลการพยากรณ์ไปใช้ และช่วงเวลาที่การพยากรณ์จะครอบคลุมถึง เพื่อจะเลือกใช้วิธีการพยากรณ์ได้อย่างถูกต้อง และเหมาะสม

2. การรวบรวมข้อมูลต้องทำอย่างเป็นระบบ ถูกต้องตามความเป็นจริง เพราะคุณภาพของข้อมูลมีผลกระทบโดยตรงต่อการพยากรณ์

3. ระบุถึงข้อข้อจำกัด ขอบเขต และสมมติฐานที่ตั้งไว้ในการพยากรณ์ เพื่อทำให้ทราบถึงข้อจำกัด หรือเงื่อนไขต่าง ๆ ที่มีผลต่อการพยากรณ์

4. ตรวจสอบความถูกต้องแม่นยำของค่าพยากรณ์ที่ได้กับค่าที่เกิดขึ้นจริงเป็นระยะเพื่อจะได้นำมาปรับวิธีการ ค่าต่าง ๆ หรือสมการที่ใช้ในการคำนวณให้มีความเหมาะสมเมื่อเวลาเปลี่ยนแปลงไป

2. ระยะเวลาที่ใช้ในการพยากรณ์ (Forecasting Time Horizons)

การพยากรณ์ถูกจำแนกประเภทด้วยการพิจารณาจากระยะเวลาที่ครอบคลุมในอนาคต โดยแบ่งออกเป็น 4 ช่วงเวลา (สุรเชษฐ์ เถื่อนแก้วสิงห์, 2552) คือ

1. การพยากรณ์หนึ่งหน่วยเวลาล่วงหน้า (Immediate-Range Forecast)

เป็นการพยากรณ์ที่มีช่วงเวลาน้อยกว่า 1 เดือน โดยเป้าหมายของการพยากรณ์จะมุ่งเพื่อปรับปรุงวิธีการทำงานให้ดีขึ้นมากกว่าการเปลี่ยนแปลงวิธีการ

2. การพยากรณ์ระยะสั้น (Short-Range Forecast)

การพยากรณ์ชนิดนี้มีช่วงเวลาภายใน 1 ปี แต่โดยทั่วไปแล้วจะไม่น้อยกว่า 3 เดือน มักใช้สำหรับการวางแผนการซื้อ การจัดตารางการทำงาน การวางแผนระดับของกำลังแรงงาน การมอบหมายงาน และระดับการผลิต เป็นต้น

3. การพยากรณ์ระยะปานกลาง (Medium-Range Forecast)

การพยากรณ์ระยะปานกลาง มักเป็นช่วงเวลาตั้งแต่ 3 เดือน ถึง 3 ปี โดยส่วนมาก**ระยะเวลาที่นิยมพยากรณ์ คือ 1 ปี เนื่องจากเท่ากับหนึ่งรอบของระยะเวลาบัญชีพอดี มีประโยชน์ในด้านการวางแผนการขาย การวางแผนการผลิต การวางแผนงบประมาณ การวางแผนด้านบุคลากร และการวิเคราะห์แผนการปฏิบัติการต่าง ๆ เป็นต้น**

4. การพยากรณ์ระยะยาว (Long-Range Forecast)

โดยทั่วไปจะมีระยะเวลา 3 ปี หรือมากกว่า การพยากรณ์ระยะยาวใช้ในการวางแผนสำหรับผลิตภัณฑ์ใหม่ ๆ การใช้จ่ายในการลงทุน การให้ความสะดวกเกี่ยวกับทำเลที่ตั้งหรือการขยายทำเลที่ตั้ง และงานวิจัยและพัฒนา เป็นต้น

3. ขั้นตอนการพยากรณ์

ขั้นตอนของการพยากรณ์ แบ่งออกเป็น 6 ขั้นตอน (สุรเชษฐ์ เถื่อนแก้วสิงห์, 2552) คือ

1. กำหนดเป้าหมาย

ทำการกำหนดเป้าหมายของการพยากรณ์ว่าจะนำผลที่ได้ไปใช้ประโยชน์อย่างไร นำไปช่วยในการตัดสินใจเรื่องอะไร หรือนำไปแก้ปัญหาได้หรือไม่ เพราะหากกำหนดเป้าหมายไม่ชัดเจน การพยากรณ์ก็จะไม่สามารถนำไปใช้ได้ตามวัตถุประสงค์ที่ตั้งไว้

2. กำหนดตัวแปร

ต้องวิเคราะห์ถึงตัวแปรต่าง ๆ ที่ส่งผลกระทบต่อเป้าหมายที่ต้องการพยากรณ์ และทำการกำหนดว่าตัวแปรอะไรบ้างที่จะใช้สำหรับการพยากรณ์

3. กำหนดระยะเวลา

ระยะเวลาของการพยากรณ์แบ่งออกเป็น 2 ระยะ คือ จำนวนช่วงระยะเวลาที่ต้องการพยากรณ์ และเวลาที่ต้องการนำค่าพยากรณ์ไปใช้

4. พิจารณาข้อมูล

ทำการพิจารณาว่า จะนำข้อมูลมาจากแหล่งไหน เพื่อนำมาใช้ในการพยากรณ์ ซึ่งอาจนำข้อมูลมาจากทั้งภายในองค์การ และภายนอกองค์กร

5. เลือกวิธีการพยากรณ์

เลือกวิธีที่จะใช้สำหรับการพยากรณ์ โดยพิจารณาให้เหมาะสมต่อหลักเกณฑ์ต่าง ๆ เช่น ประเภท และจำนวนข้อมูล ลักษณะข้อมูลที่เกิดขึ้นในอดีต ช่วงระยะเวลาของการพยากรณ์ ความซับซ้อนของวิธีการพยากรณ์ และความเร่งด่วนที่จะใช้ค่าพยากรณ์

6. ตรวจสอบความแม่นยำของการพยากรณ์

ทำการตรวจสอบผลที่ได้จากการพยากรณ์กับค่าที่เกิดขึ้นจริง ว่าใกล้เคียงกันมากน้อยเท่าไร ถ้าค่าความคลาดเคลื่อนในการพยากรณ์ยิ่งต่ำ แสดงว่าวิธีการพยากรณ์นั้นมีความแม่นยำสูง

สถิติ (Statistics)

สถิติ หมายถึง ตัวเลขหรือข้อมูลที่ได้จากการเก็บรวบรวม เช่น จำนวนผู้ประสบอุบัติเหตุบนท้องถนน อัตราการเกิดของเด็กทารก ปริมาณน้ำฝนในแต่ละปี เป็นต้น สถิติในความหมายที่กล่าวมานี้เรียกอีกอย่างหนึ่งว่า ข้อมูลทางสถิติ (Statistical data) และจะรวมถึงวิธีการที่ว่าด้วยการเก็บรวบรวมข้อมูล การนำเสนอข้อมูล การวิเคราะห์ข้อมูล และการตีความหมายข้อมูล (ฉัตรศิริ ปิยะพิมลสิทธิ์, 2544)

1. มาตราการวัด (Scale of Measurement)

สามารถจำแนกตามมาตราการวัด ได้ 4 ประเภท (ธีระดา ภิญโญ และคณะ, 2552; นุชรินทร์ ทิพยวรรณากร, 2547; ชูศรี วงศ์รัตนะ, 2553) คือ

1. มาตรานามบัญญัติ (Nominal Scale) เป็นมาตราวัดค่าที่ง่ายที่สุด เพราะแบ่งข้อมูลออกเป็นกลุ่ม (Category) และให้ชื่อที่แตกต่างกัน เช่น เพศ มี 2 เพศ คือ เพศชาย และเพศหญิง เป็นต้น ไม่สามารถเปรียบเทียบความแตกต่างระหว่างข้อมูลที่อยู่ต่างกลุ่มกันได้ บางครั้งเรียกข้อมูลชนิดนี้ว่า ข้อมูลเชิงกลุ่ม (Categorical data)

2. มาตราเรียงอันดับ (Ordinal Scale) เป็นมาตรการวัดที่อาศัยการให้ลำดับที่ (Ranked) มาตราวัดแบบนี้ไม่สามารถบอกค่าแตกต่างที่แท้จริงได้ว่ามากกว่า หรือน้อยกว่าเป็นจำนวนเท่าใด ตัวเลขที่ผู้วิจัยกำหนดขึ้นเพียงแต่ให้ลำดับที่ของความสำคัญเท่านั้น เช่น รหัส 1 แทน ต่ำกว่าปริญญาตรี รหัส 2 แทน ปริญญาตรี รหัส 3 แทน สูงกว่าปริญญาตรี เป็นต้น ไม่สามารถบอกได้ว่าผู้ที่มีการศึกษาต่ำกว่าปริญญาตรีมีความรู้น้อยกว่าผู้ที่มีการศึกษาในระดับปริญญาตรีเป็นปริมาณเท่าใด แต่สามารถบอกได้ว่าการศึกษาที่ต่ำกว่าปริญญาตรีมีการศึกษาต่ำกว่าการศึกษาปริญญาตรี

3. มาตราอันตรภาค (Interval Scale) เป็นมาตรการวัดที่ให้ความสำคัญในเรื่องความมากกว่า หรือน้อยกว่าของตัวเลข ผลต่างของตัวเลขมีความหมายแน่นอน สามารถบอกได้ว่าข้อมูลที่อยู่ต่างกลุ่มกันนั้นมีปริมาณที่แตกต่างกันเท่าใด เช่น 27 และ 37 องศาเซลเซียส สามารถบอกได้ว่าอุณหภูมิของทั้ง 2 ค่านั้นแตกต่างกัน โดยวันที่มีอุณหภูมิ 37 องศาเซลเซียส มีอุณหภูมิสูงกว่าวันที่มีอุณหภูมิ 27 องศาเซลเซียส อยู่ 10 องศาเซลเซียส แต่ถ้าวันใดวัดได้ 0 องศาเซลเซียสไม่ได้หมายความว่าวันนั้นไม่มีอุณหภูมิเลย นั่นคือ มาตราอันตรภาคไม่มีศูนย์แท้

4. มาตราอัตราส่วน (Ratio Scale) เป็นมาตรการวัดที่มีความสมบูรณ์มากที่สุด เช่น ความสูง 0 เซนติเมตร หมายถึง ไม่มีความสูงเลย จำนวนลูกค้าที่เข้ามาซื้อสินค้าในร้านค้าเป็น 0 หมายถึง ไม่มีลูกค้าที่เข้ามาซื้อสินค้าเลย นั่นคือมาตราอัตราส่วนจะมีคุณลักษณะคล้ายมาตราอันตรภาค แต่มาตราอัตราส่วนมีศูนย์แท้

2. ลักษณะของข้อมูล

สามารถจำแนกตามลักษณะของข้อมูล ได้ 2 ประเภท (ธีระดา ภิญโญ และคณะ, 2552; นุชรินทร์ ทิพยวรรณากร, 2547) คือ

1. ข้อมูลเชิงปริมาณ (Quantitative Data) เป็นข้อมูลที่วัดค่าได้ว่ามีค่ามากหรือมีค่าน้อย แสดงเป็นตัวเลขจำนวนเต็ม ทศนิยมหรือเศษส่วน สามารถนำมาคิดคำนวณได้ เช่น รายได้ น้ำหนัก เป็นต้น ข้อมูลเชิงปริมาณเป็นข้อมูลที่อยู่ในมาตราวัดแบบอันตรภาค หรือมาตราวัดแบบอัตราส่วน ซึ่งแบ่งได้เป็น 2 ประเภท คือ

1.1 ข้อมูลเชิงปริมาณที่ไม่ต่อเนื่อง (Discrete Data) หมายถึง ข้อมูลที่มีค่าเป็นตัวเลขที่ไม่ต่อเนื่องกัน ข้อมูลที่มีค่าเป็นจำนวนนับ เช่น จำนวนสินค้าที่ขายได้ในแต่ละวัน จำนวนลูกค้าที่เข้ามาซื้อของในร้านค้าในแต่ละวัน เป็นต้น

1.2 ข้อมูลเชิงปริมาณที่ต่อเนื่อง (Continuous Data) หมายถึง ข้อมูลที่มีค่าเป็นตัวเลขจำนวนใด ๆ ที่ต่อเนื่องกันเป็นเลขทศนิยมได้ เช่น ส่วนสูง น้ำหนัก เป็นต้น

2. ข้อมูลเชิงคุณภาพ (Qualitative Data) เป็นข้อมูลที่อยู่ในรูปของข้อความ จึงไม่สามารถระบุได้ว่ามากหรือน้อย ไม่สามารถนำมาคิดคำนวณได้ ในการจำแนกข้อมูลเชิงคุณภาพจะใช้คุณสมบัติ หรือชนิดของข้อมูลเป็นหลักในการจำแนก เช่น เพศ สีผิว ระดับการศึกษา เป็นต้น ข้อมูลเชิงคุณภาพเป็นข้อมูลที่อยู่ในมาตราวัดแบบนามบัญญัติ หรือมาตราวัดแบบเรียงลำดับ

การทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูล คือ กระบวนการสกัดความรู้ที่น่าสนใจจากข้อมูลที่มีอยู่จำนวนมาก เพื่อค้นหารูปแบบ และความสัมพันธ์ทั้งหมดซึ่งมีอยู่จริงแต่อาจถูกซ่อนอยู่ในข้อมูลจำนวนมากนั้น ปัจจุบันการทำเหมืองข้อมูลได้ถูกนำไปประยุกต์ใช้ในงานหลายประเภทที่จะสามารถช่วยในการตัดสินใจของผู้บริหาร เช่น ปัญหาที่ต้องใช้การวิเคราะห์ถึงสาเหตุที่ส่งผลเชิงประจักษ์ และปัญหาที่เกี่ยวข้องกับเศรษฐศาสตร์ ด้านการเงินและธุรกิจ เป็นต้น (ชนวัฒน์ ศรีสอ้าน, 2551)

การทำเหมืองข้อมูลเป็นขั้นตอนที่สำคัญที่เป็นส่วนหนึ่งของกระบวนการค้นหาความรู้ (Knowledge Discovery in Database Process) (Han, et al., 2006) ดังภาพ 1 ซึ่งประกอบไปด้วย 5 ขั้นตอนหลัก คือ

1. การรวบรวมข้อมูล (Data Integration) เป็นการนำข้อมูลจากหลายแหล่ง อาจจะหลายรูปแบบมารวมเข้าไว้ด้วยกัน

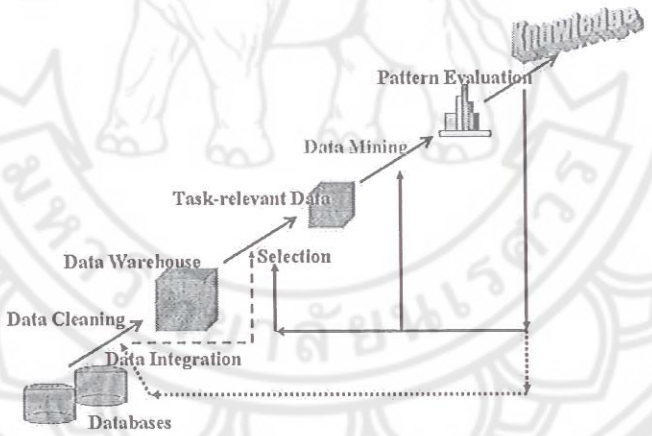
2. การทำความสะอาดข้อมูล (Data Cleaning) เป็นการตัดข้อมูลที่ไม่มีประโยชน์ออกไป เติมเต็มข้อมูลที่มีความสำคัญ และไม่สมบูรณ์ รวมทั้งการปรับเปลี่ยนรูปแบบของข้อมูลที่มาจากหลายแหล่งข้อมูลให้อยู่ในรูปแบบเดียวกัน

3. การคัดเลือกข้อมูล (Data Selection) เป็นการคัดเลือกเฉพาะข้อมูลที่มีความจำเป็นที่เกี่ยวข้อง และสามารถใช้ในการวิเคราะห์ของเทคนิคการทำเหมืองข้อมูลต่าง ๆ

4. การทำเหมืองข้อมูล (Data Mining) เป็นการใช้เทคนิควิเคราะห์ข้อมูล ซึ่งมีมากมายหลายวิธี รวมทั้งการสร้างตัวแบบที่ใช้ในการพยากรณ์

5. การประเมินผล เป็นขั้นตอนการประเมินประสิทธิภาพของเทคนิคเหมืองข้อมูลที่เลือกใช้ว่ามีความเหมาะสม ตอบโจทย์ปัญหาที่ต้องการหรือไม่

โดยทั่วไปขั้นตอนหลักของกระบวนการค้นหาความรู้ ทั้ง 5 ขั้นตอน อาจจะไม่สามารถดำเนินการสำเร็จในครั้งเดียว อาจจะมีการทำซ้ำเพื่อปรับปรุงผลการทำงาน เช่น อาจจะต้องกลับมาคัดเลือกข้อมูลชุดใหม่ อาจจะต้องกลับมาศึกษาวิเคราะห์ข้อมูลด้วยเทคนิคการทำเหมืองข้อมูลเทคนิคอื่น เป็นต้น



ภาพ 1 แสดงกระบวนการค้นหาความรู้

ประเภทของการทำเหมืองข้อมูล

การทำเหมืองข้อมูล สามารถจำแนกได้ 2 ประเภท (Tan, et al., 2006) คือ

1. การสร้างตัวแบบในการทำนาย (Predictive Modeling) หรือเรียกว่า เป็นการเรียนรู้แบบมีผู้สอน (Supervised Learning) คือ การนำข้อมูลที่มีในอดีต และนำข้อมูลเหล่านั้นมาสร้างตัวแบบ เพื่อการทำนายอนาคต โดยมีการใช้ข้อมูลในการสอน (Train) ให้กับระบบก่อน เทคนิคใน

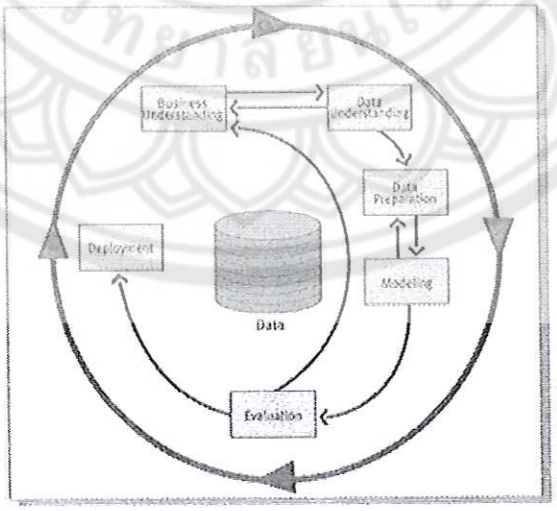
ลักษณะนี้ ได้แก่ เทคนิคการจัดหมวดหมู่ (Classification) เทคนิคการประมาณค่า (Estimation) และเทคนิคการพยากรณ์ (Prediction) เป็นต้น

2. การสร้างตัวแบบในการบรรยาย (Descriptive Modeling) หรือเรียกว่า เป็นการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) คือ การนำข้อมูลที่มีอยู่มาศึกษา เป็นการเรียนรู้จากข้อมูลที่มีอยู่ เทคนิคในลักษณะนี้ ได้แก่ เทคนิคการจัดกลุ่มข้อมูล (Clustering) การค้นพบกฎความสัมพันธ์ (Association Rule Discovery) และการค้นพบรูปแบบเชิงลำดับ (Sequential Pattern Discovery)

การทำเหมืองข้อมูลแต่ละเทคนิคจะประสบความสำเร็จกับการประยุกต์ใช้งานบางกลุ่มเท่านั้น ไม่มีเทคนิคใดที่สามารถแก้ปัญหาได้ครบทุกประเด็น การทดลองเลือกเทคนิคที่เหมาะสมจึงเป็นสิ่งที่จำเป็นที่จะนำไปสู่วิธีแก้ปัญหาที่ดีที่สุด

กระบวนการสำหรับการทำเหมืองข้อมูล (CRISP-DM: CRoss Industry Standard Process for Data Mining)

กระบวนการสำหรับการทำเหมืองข้อมูล (Roiger, et al., 2003; Chapman, et al., 2000) เป็นกระบวนการมาตรฐานอุตสาหกรรมที่เป็นขั้นตอนการแก้ปัญหาที่มักใช้ในทางธุรกิจ เป็นการสร้างตัวแบบด้วยวงจรชีวิต 6 ขั้นตอน แสดงดังภาพ 2 แต่ละขั้นตอนอาจทำตามลำดับและสามารถกลับมาทำซ้ำ ขึ้นอยู่กับความสัมพันธ์ระหว่างงานย่อย เป้าหมายของงาน และผลลัพธ์ความสำเร็จของแต่ละขั้นตอน



ภาพ 2 แสดงขั้นตอนของกระบวนการสำหรับการทำเหมืองข้อมูล

วงจรชีวิตของกระบวนการสำหรับการทำเหมืองข้อมูลทั้ง 6 ขั้นตอน มีรายละเอียด ดังนี้

1. การทำความเข้าใจโจทย์ที่ต้องการทำ (Business Understanding) ขั้นตอนแรกนี้มุ่งเน้นทำความเข้าใจวัตถุประสงค์ของโครงการ และความต้องการจากมุมมองของธุรกิจ เป็นการนิยามปัญหา วางแผน และออกแบบเบื้องต้นเพื่อให้บรรลุวัตถุประสงค์

2. การทำความเข้าใจข้อมูล (Data Understanding) เป็นขั้นตอนการทำความเข้าใจข้อมูล เริ่มต้นด้วยการเก็บรวบรวมข้อมูล ทำความคุ้นเคยกับข้อมูล ระบุปัญหาคุณภาพของข้อมูล เชิงลึก รวมถึงการตั้งสมมติฐานสำหรับการค้นหาความรู้ที่ซ่อนอยู่

3. การเตรียมข้อมูล (Data Preparation) เป็นขั้นตอนที่ครอบคลุมกิจกรรมทั้งหมดในการสร้างชุดข้อมูลขั้นสุดท้ายจากข้อมูลดิบที่มีเริ่มต้น ข้อมูลการเตรียมงานอาจดำเนินการหลายครั้ง รวมถึงการบันทึกข้อมูล การเลือกตัวแปรที่สำคัญ การเปลี่ยนแปลงข้อมูล ตลอดจนการทำความสะอาดข้อมูล

4. การสร้างตัวแบบ (Modeling) ในขั้นตอนนี้เป็นขั้นตอนสร้างตัวแบบด้วยเทคนิคเหมืองข้อมูล การเลือกข้อมูลที่ใช้และการปรับพารามิเตอร์ของเทคนิคเหมืองข้อมูล เพื่อเปรียบเทียบประสิทธิภาพของหลายเทคนิคเหมืองข้อมูลที่ใช้แก้ปัญหาที่นิยามไว้ในขั้นตอนแรก เทคนิคบางอย่างอาจต้องการคุณลักษณะเฉพาะเจาะจงในรูปแบบของข้อมูล ดังนั้นอาจจำเป็นต้องย้อนกลับไปที่ขั้นตอนการเตรียมข้อมูล

5. การประเมินผล (Evaluation) ก่อนที่จะนำตัวแบบไปใช้งานขั้นสุดท้ายจะต้องประเมินผลตัวแบบอย่างละเอียด ทบทวนขั้นตอนในการดำเนินการสร้างรูปแบบ ตรวจสอบความถูกต้องของการบรรลุวัตถุประสงค์ ในตอนท้ายของขั้นตอนนี้จะต้องมีการตัดสินใจเลือกรูปแบบที่เหมาะสมที่สุด

6. การใช้งาน (Deployment) เมื่อสร้างแบบจำลองได้แล้ว ก็จะถึงการประยุกต์ใช้ตัวแบบ ซึ่งส่วนใหญ่จะถูกนำไปใช้ประโยชน์ในการตัดสินใจขององค์กรตามสภาพปัญหา และวัตถุประสงค์ของโครงการ

เทคนิคเคมีน (K-means)

เทคนิคเคมีนเป็นเทคนิคการทำเหมืองข้อมูล ที่นิยมใช้ในการจัดกลุ่มข้อมูล เป็นการเรียนรู้แบบไม่มีผู้สอน มีประสิทธิภาพในด้านของความเข้าใจง่าย สามารถใช้ได้กับข้อมูลหลายประเภท เทคนิคเคมีนจะรวมสิ่งที่มีลักษณะคล้ายกันไว้ในกลุ่มเดียวกัน โดยจะทำการแบ่งกลุ่มข้อมูลออกเป็น K กลุ่ม ตามที่ต้องการ (Han, et al., 2006; Tan, et al., 2006) **ถูกนำเสนอโดย**

MacQueen (T. Katanyukul, 2000 อ้างอิงใน สมชาย จำปาทอง, 2549) ซึ่งสามารถแสดงขั้นตอนวิธีการทำงานของเทคนิคเคมีนได้ดังนี้

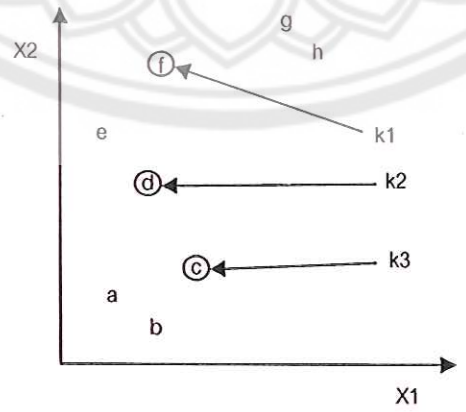
ตาราง 1 แสดงตัวอย่างข้อมูลสำหรับการแบ่งกลุ่มเคมีน

	Observation	
	X1	X2
A	1	2
B	2	1
C	3	3
D	2.5	5
E	1.5	6
F	3	7
G	5	8
H	6	7

ที่มา: T.Katanyukul, 2000 อ้างอิงใน สมชาย จำปาทอง, 2549

ขั้นตอนวิธีของเคมีน

1. ทำการสุ่มส่วนของข้อมูลเป็นกลุ่มเซตย่อย K เซต โดยผู้ที่ทำการแบ่งกลุ่มข้อมูล (Cluster) จะต้องทราบจำนวนกลุ่ม (K) ที่ต้องการแบ่งซึ่งสามารถแสดงได้ ดังภาพ 2



ภาพ 3 แสดงการสุ่มข้อมูลเพื่อทำการแบ่งกลุ่มเคมีน

จากภาพ 3 มีการแบ่งกลุ่มข้อมูลออกเป็น 3 กลุ่ม โดยทำการสุ่มจุดเริ่มต้นของแต่ละกลุ่ม นำข้อมูลทุกตัวมาคำนวณระยะทางกับ k การคำนวณระยะทางทำได้หลายวิธี เช่น ระยะทางยูคลิเดียน (Euclidean distance) ดังสมการ (1) โดยถ้า d คือระยะทางยูคลิเดียนแล้ว จะได้ระยะทาง คือ

$$d(i,j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + ... + |x_{ip} - x_{jp}|^2)} \quad (1)$$

โดยที่

$$d(i,j) \geq 0,\ d(i,i) = 0,\ d(i,j) = d(j,i),\ d(i,j) \leq d(i,k) + d(k,j)$$

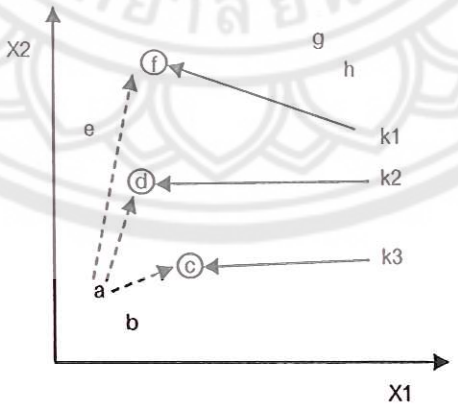
ซึ่งสามารถคำนวณหาระยะทางยูคลิเดียน โดยใช้ข้อมูลจากตาราง 1 และภาพ 3 ซึ่งได้มีการกำหนดจุดเริ่มต้นของแต่ละกลุ่มข้อมูล คือ จุด c, จุด d และจุด f ดังนั้นเมื่อคำนวณระยะทางจากจุด a ไปยังจุดเริ่มต้นของแต่ละกลุ่มข้อมูลจะได้ระยะทาง ดังนี้

ระยะทางจาก a ไป c = $\sqrt{(|1-3|^2 + |2-3|^2)} = 2.24$

ระยะทางจาก a ไป d = $\sqrt{(|1-2.5|^2 + |2-5|^2)} = 3.12$

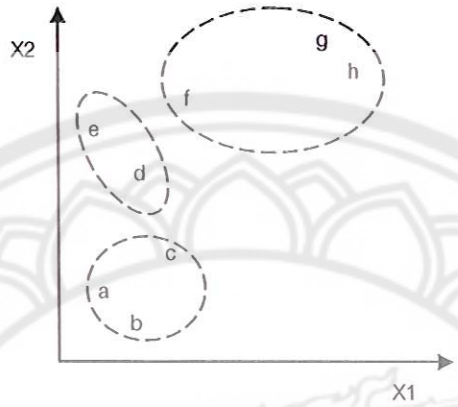
ระยะทางจาก a ไป f = $\sqrt{(|1-3|^2 + |2-7|^2)} = 5.39$

จากการคำนวณระยะทางระหว่างข้อมูลกับกลุ่มข้อมูลนั้นสามารถ ดังภาพ 4



ภาพ 4 แสดงระยะทางของข้อมูลกับข้อมูลเริ่มต้นของกลุ่ม

หลังจากนั้นทำการหาระยะทางของข้อมูลทุกจุดกับกลุ่มข้อมูลทุกกลุ่ม และตรวจดูว่าข้อมูลห่างจากกลุ่มข้อมูลชุดไหนน้อยที่สุดก็นำข้อมูลไปสังกัดอยู่ในกลุ่มข้อมูลนั้น และได้กลุ่มข้อมูลขึ้นมา ดังภาพ 5



ภาพ 5 แสดงข้อมูลที่ได้รับการแบ่งโดยอาศัย K

2. เมื่อได้กลุ่มข้อมูลแต่ละกลุ่มแล้ว แล้วทำการคำนวณหาจุดศูนย์กลางของกลุ่มที่เรียกว่า เซ็นทรอย (Centroid) โดยจุดศูนย์กลางสามารถคำนวณ ดังสมการ (2)

$$M_k = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ik} \qquad (2)$$

โดยที่

$x = \{x \in k \mid k \text{ คือ กลุ่มข้อมูล}\}$

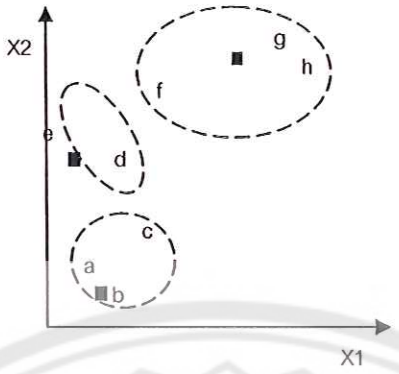
$x = \{x_1, x_2, \ldots, x_i\}$

M_k คือ ค่ากลางของข้อมูลในกลุ่มที่ k

n_k คือ จำนวนข้อมูลของกลุ่มที่ k

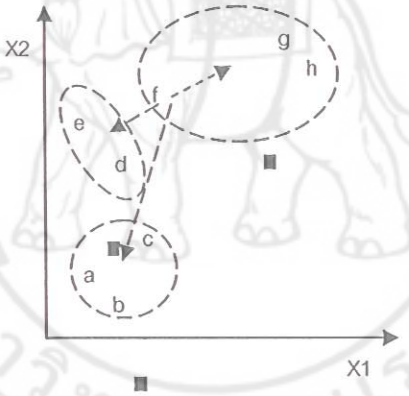
x_{ik} คือ ข้อมูล x ตัวที่ i ของกลุ่มข้อมูลที่ k

เมื่อได้จุดศูนย์กลางของข้อมูลแล้วนำข้อมูลทุกตัวในแต่ละกลุ่มข้อมูลไปคำนวณกับจุดศูนย์กลางของทุกกลุ่มข้อมูล ดังภาพ 6

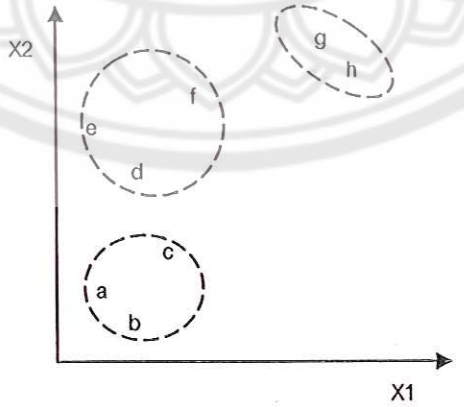


ภาพ 6 แสดงจุดศูนย์กลางของข้อมูลในแต่ละกลุ่มข้อมูล

3. ตรวจดูระยะทางระหว่างข้อมูลกับจุดศูนย์กลางข้อมูล โดยข้อมูลที่ห่างจากจุดศูนย์กลางข้อมูลไหนสั้นที่สุดก็นำข้อมูลไปสังกัดกับกลุ่มข้อมูลดังกล่าว จากจุดนี้จะทำให้เกิดการย้ายกลุ่มข้อมูล ดังภาพ 7-8



ภาพ 7 แสดงการวัดระยะทางของข้อมูลกับจุดศูนย์กลาง



ภาพ 8 แสดงกลุ่มข้อมูลใหม่

4. ทำการคำนวณจุดศูนย์กลางของข้อมูล และกำหนดข้อมูลให้กลุ่มข้อมูลใหม่ไปเรื่อย ๆ จนกว่าจะถึงเงื่อนไขให้หยุด (stopping criterion) โดยเงื่อนไขให้หยุด คือ 1) ข้อมูลในแต่ละกลุ่มถึงสภาวะเสถียรไม่สามารถเปลี่ยนกลุ่มที่ดีกว่านี้ได้ 2) ค่าผิดพลาดที่ต่ำสุด (Minimum Square Error) สำหรับทุกกลุ่มข้อมูลโดยหาค่าผิดพลาดที่ต่ำสุดได้ ดังสมการ (3)

$$E_k^2 = \sum_{k=1}^{K} e_k^2 \text{ โดยที่ } e_k^2 = \sum_{i=1}^{n_k} (x_{ik} - M_k)^2 \qquad (3)$$

โดยที่

$x = \{x \in k \mid k \text{ คือ กลุ่มข้อมูล}\}$

$x = \{x_1, x_2, \ldots, x_i\}$

x_{ik} คือ ข้อมูลที่ i ของกลุ่ม k

M_k คือ ค่ากลางของข้อมูลในกลุ่มที่ k

n_k คือ จำนวนข้อมูลของกลุ่มที่ k

e_k^2 คือ ค่าผิดพลาดในกลุ่มที่ k

E_k^2 คือ ค่าผิดพลาดรวมของทุกกลุ่มข้อมูล

เทคนิคการคัดเลือกตัวแปร (Attribute Selection)

เทคนิคการคัดเลือกตัวแปรเป็นเทคนิคที่ช่วยลดจำนวนตัวแปรที่จะใช้ในการพยากรณ์ อาจกระทำเพื่อเลือกตัวแปรที่ดีที่สุดเพียงตัวเดียว หรือเลือกกลุ่มของตัวแปรที่มีความสำคัญ การเลือกชุดตัวแปรสามารถทำได้หลายวิธี เทคนิคการคัดเลือกชุดตัวแปรแบบอัตราส่วนเกนร่วมกับวิธีการค้นหาแบบจัดลำดับ เป็นวิธีคัดเลือกตัวแปรโดยมีหลักการเช่นเดียวกับการเลือกตัวแปรของการสร้างต้นไม้ตัดสินใจ เพื่อให้ได้ตัวแปรที่เป็นตัวแบ่งข้อมูลออกเป็นกลุ่มย่อยที่มีสมาชิกภายในกลุ่มเป็นชนิดเดียวกันมากที่สุด (Homogeneous) ด้วยมาตรวัดการได้ประโยชน์จากการแบ่งกลุ่มย่อย เรียกว่า อัตราส่วนเกน (Gain Ratio) ดังสมการ (4) ซึ่งเป็นอัตราส่วนของค่าเกน (Gain หรือ Information Gain) กับ ค่าสารสนเทศการแบ่งกลุ่ม (SplitInfo) อันเป็นการลดอิทธิพลของตัวแปรที่มีค่าหลายค่า ผลที่ได้รับจากการใช้เทคนิคนี้จะได้ลำดับของตัวแปร ซึ่งตัวแปรที่อยู่ลำดับแรก ๆ จะถือว่ามีอิทธิพลในการพยากรณ์ต่อตัวแปรเป้าหมายมากกว่าตัวแปรในลำดับถัดไป ทำให้เรา

สามารถพิจารณาเลือกจำนวนตัวแปรที่เหมาะสมได้อย่างมีประสิทธิภาพ (Tan, et al., 2006; Witten, et al., 2011)

$$GainRatio(S,A) = \frac{GAIN(S,A)}{SplitINFO(A)} \quad (4)$$

โดยที่

$$GAIN(S,A) = Entropy(S) - \left(\sum_{v \in Values(A)} \frac{|Sv|}{|S|} Entropy(Sv) \right)$$

$$Entropy(S) = -\sum_{i=1}^{m} P_i \log_2 P_i$$

$$SplitINFO(A) = -\sum_{v \in Values(A)} \frac{P_i}{N} \log_2 \frac{P_i}{N}$$

P_i หมายถึง สัดส่วนของจำนวนสมาชิกของกลุ่ม i กับจำนวนสมาชิกทั้งหมดของกลุ่มตัวอย่าง

S หมายถึง ตัวแปรที่นำมาคำนวณหาค่า Entropy

A หมายถึง ตัวแปร A

$|Sv|$ หมายถึง สมาชิกของตัวแปร A ที่มีค่าเป็น v

$|S|$ หมายถึง จำนวนสมาชิกของกลุ่มตัวอย่าง

$SplitINFO\ (A)$ หมายถึง สัดส่วนของข้อมูลของตัวแปร A ที่มีค่า P_i ต่อจำนวนข้อมูลทั้งหมดของตัวแปร A

สัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน (Person Correlation Coefficient)

ค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน จะมีค่าอยู่ระหว่าง -1 ถึง 1 ค่าสูงสุดมีค่าเป็น 1 ซึ่งตีความหมายได้ว่าตัวแปรนั้นมีความสัมพันธ์เชิงเส้นต่อกันมากที่สุด และถ้าค่าสัมพันธ์มีค่าเป็น 0 แสดงว่าได้ว่าตัวแปรนั้นไม่มีความสัมพันธ์เชิงเส้นต่อกัน

การวิเคราะห์สัมประสิทธิ์สหสัมพันธ์แบบเพียร์สันเป็นการวัดความสัมพันธ์ของตัวแปร 2 ชุด เมื่อตัวแปรทั้งสองชุดนั้นเป็นตัวแปรต่อเนื่อง หรือเป็นข้อมูลอันตรภาคชั้นหรืออัตราส่วนสูตรในการคำนวณหาสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน สามารถคำนวณหาได้ ดังสมการ (5)

$$r_{XY} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{\left[N\sum X^2 - (\sum X)^2\right]\left[N\sum Y^2 - (\sum Y)^2\right]}} \quad (5)$$

โดยที่

r_{XY} **คือ ค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สันระหว่างตัวแปร X , Y**

$\sum X$ คือ ผลรวมของข้อมูลที่วัดได้จากตัวแปรตัวที่ 1 (X)

$\sum Y$ คือ ผลรวมของข้อมูลที่วัดได้จากตัวแปรตัวที่ 2 (Y)

$\sum XY$ คือ ผลรวมของผลคูณระหว่างข้อมูลตัวแปรที่ 1 และ 2

$\sum X^2$ คือ ผลรวมของกำลังสองของข้อมูลที่วัดได้จากตัวแปรตัวที่ 1

$\sum Y^2$ คือ ผลรวมของกำลังสองของข้อมูลที่วัดได้จากตัวแปรตัวที่ 2

N คือ ขนาดของกลุ่มตัวอย่าง

สัมประสิทธิ์อีต้า (Eta Coefficient)

สัมประสิทธิ์อีต้า บางครั้งจะ เรียกว่า อัตราส่วนความสัมพันธ์ (Correlation Ratio) ค่าสัมประสิทธิ์อีต้า ควรจะใช้เมื่อจุดบนแผนภาพกระจัดกระจายที่เกิดจากตัวแปร 2 ตัว มีการเรียงตัวในลักษณะไม่เป็นเส้นตรง (ฉัตรศิริ ปิยะพิมลสิทธิ์, 2544) สัมประสิทธิ์อีต้าสามารถคำนวณหาได้ดังสมการ (6)

$$\eta^2 = 1 - \frac{\sum\left(Y - \bar{Y}_k\right)^2}{\sum\left(Y - \bar{Y}_T\right)^2} \quad (6)$$

โดยที่

η^2 คือ กำลังสองสัมประสิทธิ์อีต้า

$\sum\left(Y - \bar{Y}_k\right)^2$ คือ ผลรวมของกำลังสองของค่าเบี่ยงเบนจากค่าเฉลี่ยในแต่ละกลุ่ม

$\sum\left(Y - \bar{Y}_T\right)^2$ คือ ผลรวมของกำลังสองของค่าเบี่ยงเบนจากค่าเฉลี่ยรวม

การแปลผลค่าสัมประสิทธิ์อีต้าก็เหมือนกับสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน เมื่อยกกำลังสองแล้ว ก็คือสัดส่วนของความผันแปรในตัวแปร Y ที่สามารถอธิบายได้ด้วยตัวแปร X

ความแตกต่างที่สำคัญระหว่างสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน กับ สัมประสิทธิ์อีต้า คือ พิสัยของค่าสัมประสิทธิ์อีต้า จะมีค่าตั้งแต่ 0 ถึง 1 ค่าสัมประสิทธิ์อีต้าจะไม่มีค่าติดลบ

มีค่าระดับความสัมพันธ์ค่าสัมประสิทธิ์อีต้า จะอยู่ระหว่าง 0-1 ถ้าค่าความสัมพันธ์มีค่าเข้าใกล้ 1 หรือเท่ากับ 1 หมายความว่า ตัวแปรทั้งสองมีความสัมพันธ์กันมาก แต่ถ้าค่าความสัมพันธ์มีค่าเท่ากับ 0 แสดงว่า ตัวแปรทั้งสองไม่มีความสัมพันธ์กัน

การแปลงค่าข้อมูล (Data Transformation)

การแปลงค่าข้อมูล เป็นกระบวนการปรับขอบเขตของข้อมูลให้อยู่ในช่วงที่เหมาะสมต่อการนำไปใช้งาน จะต้องแปลงค่าข้อมูลด้วยวิธีการนอมัลไลเซชั่น (Normalization) ซึ่งเป็นการลดค่าของข้อมูลให้อยู่ในขอบเขตที่น้อยลง เพื่อให้เหมาะสมกับการใช้งาน โดยการแปลงค่าข้อมูลในลักษณะเป็นเชิงเส้น (Min-Max Normalization) สามารถคำนวณหาได้ (Larose, 2005) ดังสมการ (7)

$$X^* = \frac{X - \min(X)}{\max(X) - \min(X)} \qquad (7)$$

โดยที่

X^* คือ ค่าของข้อมูลที่ได้หลังจากผ่านการคำนวณหาจากสมการ

X คือ ค่าของข้อมูลก่อนผ่านการคำนวณหาจากสมการ

$\min(X)$ คือ ค่าของข้อมูลที่มีค่าต่ำสุดก่อนผ่านการคำนวณหาจากสมการ

$\max(X)$ คือ ค่าของข้อมูลที่มีค่าสูงสุดก่อนผ่านการคำนวณหาจากสมการ

เทคนิคเคเนียเรสเนเบอร์ (K-Nearest Neighbor: K-NN)

หลักการของเทคนิคเคเนียเรสเนเบอร์เป็นเทคนิคการจัดหมวดหมู่แบบมีผู้สอน มีวิธีการทำงานที่เรียบง่ายไม่ซับซ้อน โดยจะพิจารณาจากชุดของข้อมูลที่มีลักษณะคล้ายคลึงกัน ซึ่งความคล้ายคลึงกันนั้น สามารถใช้ค่ามาตรวัดที่แตกต่างกันตามความเหมาะสมของข้อมูล อาจวัดเป็นระยะทาง (Distance) ถ้าระยะทางระหว่างข้อมูลที่ต้องการพิจารณากับข้อมูลชุดใดมีค่าน้อยที่สุด (เพื่อนบ้านที่ใกล้ที่สุด) ก็จะพิจารณาว่าข้อมูลนั้นควรจะจัดหมวดหมู่ให้เป็นชนิดเดียวกัน (Han, et al., 2006; Tan, et al., 2006)

เทคนิคเคเนียเรสเนเบอร์ อาจพิจารณาจากเพื่อนบ้านที่ใกล้ที่สุดหรือข้อมูลที่คล้ายคลึงที่สุดเพียงชุดเดียว (k = 1) หรือหลายชุด (k > 1) ก็ได้ โดยนิยมทำแบบเพิ่มจำนวน k ไปครั้งละ 1 ทำไปเรื่อย ๆ จนกว่าจะได้ค่า k ที่เหมาะสม ดังสมการ (8)

$$dist(p_i, q_i) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \quad (8)$$

โดยที่

$dist(p_i, q_i)$ คือ ระยะห่างระหว่างตัวอย่าง p_i กับ ตัวอย่าง q_i

n คือ จำนวนตัวแปร

p_i คือ สมาชิกที่กำลังสนใจในการจัดหมวดหมู่

q_i คือ สมาชิกอื่น ๆ ที่รู้อยู่แล้วว่าอยู่ในหมวดหมู่ใด

เมื่อได้สมาชิก q_i จำนวน k ตัว ก็จะนำมาใช้พิจารณาตัดสินว่า p_i ควรจะจัดอยู่ในหมวดหมู่ใด ตามหลักการเสียงข้างมาก (Majority Vote)

เทคนิคเนอีฟเบย์ (Naïve Bayes)

เทคนิคเนอีฟเบย์เป็นเทคนิคการจัดหมวดหมู่ ที่มีประสิทธิภาพอีกวิธีหนึ่ง มีการทำงานที่ไม่ซับซ้อน เหมาะสมกับกรณีของเซตตัวอย่างมีจำนวนมาก แต่มีข้อจำกัดที่ตัวแปรข้อมูลต้องเป็นอิสระไม่ขึ้นต่อกัน (Conditional Independence) สามารถคาดการณ์ผลลัพธ์ด้วยการวิเคราะห์ความน่าจะเป็น ด้วยทฤษฎีเบย์ (Bayes' Theorem) (Han, et al., 2006; Tan, et al., 2006) เป็นวิธีการเรียนรู้หรือวิธีการจัดหมวดหมู่ตามหลักความน่าจะเป็นของเหตุการณ์ต่าง ๆ โดยดูที่ความน่าจะเป็นของเหตุการณ์ที่เคยเกิดขึ้นก่อนหน้านี้ (Prior Probability) และความน่าจะเป็นที่จะเกิดเหตุการณ์ใดเหตุการณ์หนึ่ง (Posterior Probability) ความน่าจะเป็นที่เกิดขึ้นก่อนหน้านี้ ถือว่าเป็นความจริงซึ่งอาจจะเก็บข้อมูลจากการสังเกตการณ์ หรือเก็บสถิติ หรือการสะสมประสบการณ์ จะไม่ขึ้นกับอะไรทั้งสิ้นเนื่องจากเป็นข้อมูลจริงหรือสิ่งที่เกิดขึ้นจริง เช่น พรุ่งนี้ฝนจะตกหรือไม่ Prior Probability คือ ช่วงนี้เป็นฤดูฝนหรือไม่ เพราะจากข้อมูลในอดีตที่ เราทราบกันดี คือ ถ้าช่วงนี้เป็นฤดูฝน จะมีความน่าจะเป็นค่อนข้างสูงที่พรุ่งนี้ฝนจะตก เป็นต้น (ปิยะรัตน์ แสงมหะหมัด, 2551)

ความน่าจะเป็นจะมีค่าอยู่ระหว่าง 0 กับ 1 ถ้าข้อมูลใกล้เคียงความจริง ค่าความน่าจะเป็นจะมีค่าใกล้เคียง 1 มีสูตรการคำนวณหาความน่าจะเป็น (Posterior Probability) (ปิยะรัตน์ แสงมหะหมัด, 2551) ดังสมการ (9)

$$P(A \mid B) = P(A \wedge B) / P(B) \quad (9)$$

ความน่าจะเป็นของเหตุการณ์ A โดยขึ้นอยู่กับเหตุการณ์ B แต่ถ้าเหตุการณ์ A ไม่ขึ้นกับเหตุการณ์ B หรือทั้งสองเหตุการณ์เป็นอิสระไม่ขึ้นต่อกัน คือ การเกิดเหตุการณ์ A ไม่มีผลต่อการเกิดเหตุการณ์ B (ปิยะรัตน์ แสงมหะหมัด, 2551) ดังสมการ (10)

$$P(A \mid B) = P(A) \qquad (10)$$
$$P(B \mid A) = P(B)$$

ให้ B แทน เหตุการณ์ที่จะเกิดขึ้นในเหตุการณ์ทั้งหมด และถ้า A เป็นเหตุการณ์ใด ๆ ที่ $P(A) > 0$ (หมายความว่าเหตุการณ์ A จะต้องเกิดขึ้นเสมอ) ดังสมการ (11) (Han, et al., 2006; Tan, et al., 2006)

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \qquad (11)$$

โดยที่

$P(A \mid B)$ คือ ความน่าจะเป็นที่จะเกิดเหตุการณ์ A เมื่อเกิดเหตุการณ์ B

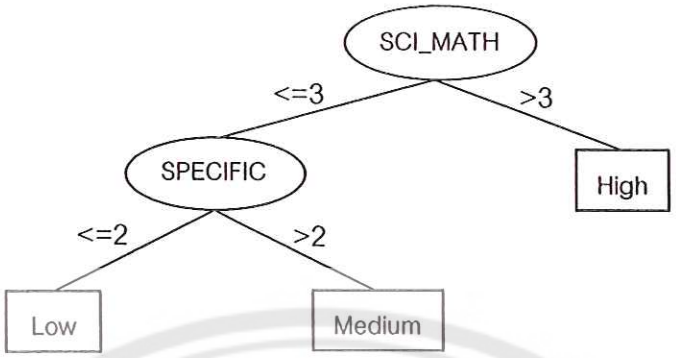
$P(B \mid A)$ คือ ความน่าจะเป็นที่จะเกิดเหตุการณ์ B เมื่อเกิดเหตุการณ์ A

$P(A)$ คือ ความน่าจะเป็นที่จะเกิดเหตุการณ์ A

$P(B)$ คือ ความน่าจะเป็นที่จะเกิดเหตุการณ์ B

เทคนิคต้นไม้ตัดสินใจ (Decision Tree)

ต้นไม้ตัดสินใจเป็นการเรียนรู้แบบมีผู้สอน ต้นไม้ตัดสินใจประกอบด้วยโหนด (Node) แรกสุด เรียกว่า โหนดราก (Root Node) จากโหนดรากจะแตกออกเป็นโหนดลูก (Child Node) จากโหนดลูกก็จะมีลูกของตนเอง ซึ่งโหนดในระดับสุดท้าย เรียกว่า โหนดใบ (Leaf Node) ดังภาพ 9 ต้นไม้ตัดสินใจมีการทำงานเหมือนโครงสร้างต้นไม้ คือ แต่ละโหนดจะแสดงตัวแปรที่ใช้ทดสอบข้อมูล แต่ละกิ่ง (Branch, Link) แสดงผลในการทดสอบ และโหนดใบแสดงกลุ่มหรือคลาส (Class) ที่ได้กำหนดไว้ (Han, et al., 2006; Tan, et al., 2006)



ภาพ 9 แสดงโครงสร้างต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจ C4.5

หลักการสร้างต้นไม้ตัดสินใจ C4.5 (Tan, et al., 2006) คือ ใช้ค่ามาตรฐานอัตราส่วนเกน (Gain Ratio) เพื่อเลือกตัวแปรที่จะใช้เป็นโหนดราก (ตัวแปรเริ่มต้น) สามารถคำนวณหาค่าเกนสารสนเทศ (Information Gain) โดยเริ่มจากการคำนวณค่า *Entropy* ดังสมการ (12)

$$Entropy(S) = -\sum_{i=1}^{m} P_i \log_2 P_i \qquad (12)$$

โดยที่

P_i หมายถึง สัดส่วนของจำนวนสมาชิกของกลุ่ม i กับจำนวนสมาชิกทั้งหมดของกลุ่มตัวอย่าง

S หมายถึง ตัวแปรที่นำมาคำนวณหาค่า Entropy

เมื่อคำนวณค่า *Entropy* เสร็จแล้วต่อไปก็นำค่า Entropy ที่คำนวณได้ทาแทนลงในสมการการหาค่าเกนสารสนเทศ ดังสมการ (13)

$$GAIN(S, A) = Entropy(S) - \left(\sum_{v \in Values(A)} \frac{|Sv|}{|S|} Entropy(Sv) \right) \qquad (13)$$

โดยที่

A หมายถึง ตัวแปร A

$|Sv|$ หมายถึง สมาชิกของตัวแปร A ที่มีค่าเป็น v

$|S|$ หมายถึง จำนวนสมาชิกของกลุ่มตัวอย่าง

คำนวณค่ามาตรฐานอัตราส่วนเกน ดังสมการ (14)

$$GainRatio(S,A) = \frac{GAIN(S,A)}{SplitINFO(A)} \quad (14)$$

โดยที่

$SplitINFO\ (A)$ หมายถึง สัดส่วนของข้อมูลของตัวแปร A ที่มีค่า P_i ต่อจำนวนข้อมูลทั้งหมดของตัวแปร A

คำนวณค่า $SplitINFO$ ดังสมการ (15)

$$SplitINFO(A) = -\sum_{v \in Values(A)} \frac{P_i}{N} \log_2 \frac{P_i}{N} \quad (15)$$

ท้ายสุดจึงเลือกค่ามาตรฐานอัตราส่วนเกนสูงสุดเป็นโหนดราก และเลือกคุณสมบัติถัดไปตามค่ามาตรฐานอัตราส่วนเกนน้อยลงตามลำดับ

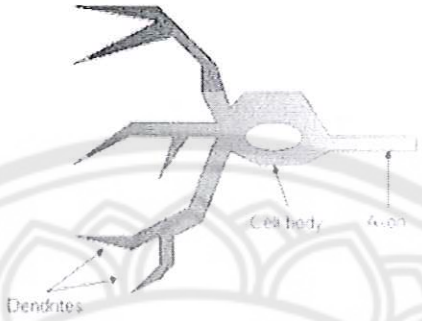
โครงข่ายประสาทเทียม (ANN: Artificial Neural Network)

โครงข่ายประสาทเทียม (ดาริกา เรือนคำ, 2552) คือ โมเดลทางคณิตศาสตร์ เพื่อจำลองการทำงานของเครือข่ายประสาทในสมองมนุษย์ ด้วยวัตถุประสงค์ที่จะสร้างเครื่องมือซึ่งมีความสามารถในการเรียนรู้การจดจำรูปแบบ (Pattern Recognition) และการอุปมานความรู้ (Knowledge Deduction)เช่นเดียวกับความสามารถที่มีในสมองมนุษย์ แนวคิดเริ่มต้นของเทคนิคนี้ได้มาจากการศึกษาข่ายงานไฟฟ้าชีวภาพ (Bioelectric Network) ในสมอง ซึ่งประกอบด้วย เซลล์ประสาท (Neurons) และจุดประสานประสาท (Synapses) แต่ละเซลล์ประสาทประกอบด้วยปลายในการรับกระแสประสาท เรียกว่า เดนไดรท์ (Dendrite) ซึ่งเป็นข้อมูลนำเข้า (Input) และปลายในการส่งกระแสประสาทเรียกว่า แอคซอน (Axon) ซึ่งเป็นเหมือนข้อมูลนำออก (Output) ของเซลล์ เซลล์เหล่านี้ทำงานด้วยปฏิกิริยาไฟฟ้าเคมี เมื่อมีการกระตุ้นด้วยสิ่งเร้าภายนอกหรือกระตุ้นด้วยเซลล์ด้วยกัน กระแสประสาทจะวิ่งผ่านเดนไดรท์เข้าสู่นิวเคลียสซึ่งจะเป็นตัวตัดสินว่าต้องกระตุ้นเซลล์อื่น ๆ ต่อหรือไม่ ถ้ากระแสประสาทแรงพอ นิวเคลียสก็จะกระตุ้นเซลล์อื่น ๆ ต่อไปผ่านทาง

GA
วท
ว
ปวทร
ทนวก
2555

1.5942561
สำนักหอสมุด
27 ก.ค. 2555

แอคซอน ตามแบบจำลองนี้ข่ายงานประสาทเกิดจากการเชื่อมต่อระหว่างเซลล์ประสาท จนเป็นเครือข่ายที่ทำงานร่วมกัน ดังภาพ 10

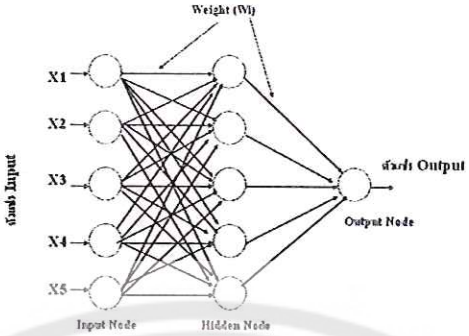


ภาพ 10 แสดงแบบจำลองข่ายประสาทในสมองมนุษย์

ที่มา: ดาริกา เรือนคำ, 2552

โครงสร้างของโครงข่ายประสาทเทียม

นักวิจัยส่วนใหญ่ในปัจจุบันเห็นตรงกันว่าโครงข่ายประสาทเทียมมีโครงสร้างแตกต่างจากโครงข่ายในสมอง ซึ่งเหมือนกับสมอง ในแง่ที่ว่าโครงข่ายประสาทเทียม คือ การรวมกลุ่มแบบขนานของหน่วยประมวลผลย่อย ๆ และการเชื่อมต่อนี้เป็นส่วนสำคัญที่ทำให้เกิดสติปัญญาของโครงข่ายเมื่อพิจารณาขนาดแล้วสมองมีขนาดใหญ่กว่าโครงข่ายประสาทเทียมอย่างมาก รวมทั้งเซลล์ประสาทยังมีความซับซ้อนกว่าหน่วยย่อยของโครงข่าย ดังภาพ 11 อย่างไรก็ดีหน้าที่สำคัญของสมอง เช่นการเรียนรู้ยังคงสามารถถูกจำลองขึ้นอย่างง่ายด้วยโครงข่ายประสาทนี้ (ดาริกา เรือนคำ, 2552)



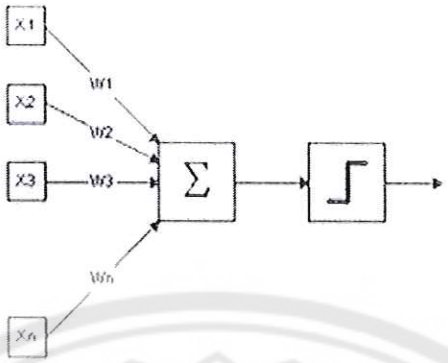
ภาพ 11 แสดงโครงสร้างของโครงข่ายประสาทเทียม

ที่มา: http://www.no-poor.com/article/neural_network.htm

หลักการทำงานของโครงข่ายประสาทเทียม

สำหรับในคอมพิวเตอร์โครงข่าย ประกอบด้วยข้อมูลนำเข้า และข้อมูลนำออก เหมือนกัน โดยจำลองให้ปัจจัยนำเข้าแต่ละอันมีน้ำหนัก (Weight) เพื่อกำหนดน้ำหนักของข้อมูลนำเข้าโดย โครงข่ายประสาทแต่ละหน่วยจะมีค่า Threshold เป็นตัวกำหนดว่าน้ำหนักรวมของข้อมูลนำเข้า ต้องมากขนาดไหนจึงจะสามารถส่งข้อมูลนำออกไปยังเซลล์ประสาทตัวอื่นได้ เมื่อนำเซลล์ประสาท แต่ละหน่วยมาต่อกันให้ทำงานร่วมกันการทำงานนี้ในทางตรรกะแล้วก็จะเหมือนกับปฏิกิริยาเคมีที่ เกิดในสมองเพียงแต่ในคอมพิวเตอร์ทุกอย่างเป็นตัวเลขเท่านั้นเอง

กระบวนการทำงานของโครงข่ายประสาทเทียม (ดาริกา เรือนคำ, 2552) คือ เมื่อมีข้อมูล นำเข้ามายังโครงข่ายก็เอาข้อมูลนำเข้ามาคูณกับน้ำหนักของแต่ละขาผลที่ได้จากข้อมูลนำเข้าทุก ๆ ขาของเซลล์ประสาทจะเอามารวมกันแล้วก็เอามาเทียบกับค่าที่กำหนดไว้ถ้าผลรวมมีค่ามากกว่า ค่าที่กำหนดไว้แล้วเซลล์ประสาทก็จะส่งข้อมูลนำออก ออกไปยังข้อมูลนำเข้าของเซลล์ประสาท อื่น ๆ ที่เชื่อมกันในโครงข่ายถ้าค่าน้อยกว่าค่าที่กำหนดไว้ ก็จะไม่เกิดข้อมูลนำออก ดังภาพ 12



ภาพ 12 แสดงแบบจำลองโครงข่ายประสาทในคอมพิวเตอร์

ที่มา: ดาริกา เรือนคำ, 2552

กระบวนการทำงานของเซลล์ประสาทเทียม สามารถอธิบายได้ (ดาริกา เรือนคำ, 2552) ดังสมการ (16)

$$I = w_1x_1 + w_2x_2 + \ldots + w_nx_n$$
$$O = f(I) \qquad (16)$$

โดยที่

I คือ ค่าผลรวมของสัญญาณขาเข้า
w คือ ค่าน้ำหนักของเส้นเชื่อมต่อ
x คือ ค่าสัญญาณขาเข้า
O คือ ค่าสัญญาณขาออก
$f(I)$ คือ ฟังก์ชั่นถ่ายโอน (Transfer Function)

ข้อมูลนำออกของแต่ละโหนด สามารถอธิบายได้ (ดาริกา เรือนคำ, 2552) ดังสมการ (17)

$$y_i = f(w_i^1x_1 + w_i^2x_2 + \ldots + w_i^nx_n)$$
$$= f(\sum_j w_i^jx_j) \qquad (17)$$

โดยที่

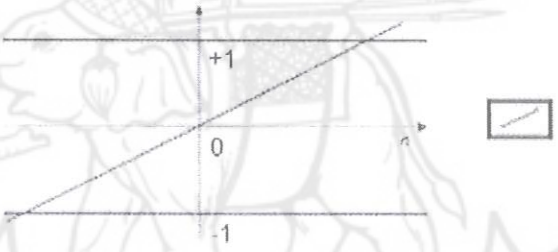
x_j คือ ตัวแปรนำเข้าจากโหนดอื่นๆ

w_i^j คือ น้ำหนัก ของแต่ละเส้น (Connection)

ฟังก์ชั่นถ่ายโอนของส่วนประมวลผลหรือเซลล์ประสาทเทียมนั้นสามารถกำหนดได้โดยฟังก์ชันหลายรูปแบบ แต่โดยทั่วไปจะนิยมใช้ฟังก์ชัน (ดาริกา เรือนคำ, 2552) ต่อไปนี้

1. ฟังก์ชันแบบเชิงเส้น (Linear Transfer Function) สามารถอธิบาย ดังสมการ (18) และดังภาพ 13

$$f(I) = K.I \qquad \text{โดย } K \text{ เป็นค่าคงที่} \tag{18}$$

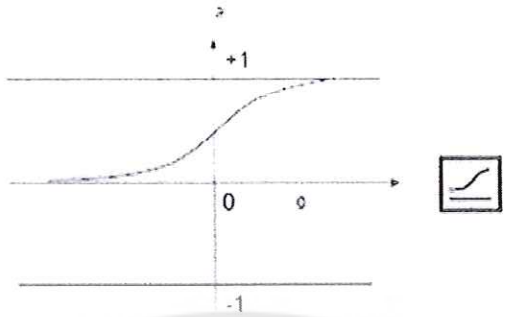


ภาพ 13 แสดงฟังก์ชันแบบเชิงเส้น

ที่มา: ดาริกา เรือนคำ, 2552

2. ซิกมอยด์ฟังก์ชั่น (Sigmoid Transfer Function) สามารถอธิบาย ดังสมการ (19) และดังภาพ 14

$$f(I) = \frac{1}{1+\exp^{-i}} \tag{19}$$

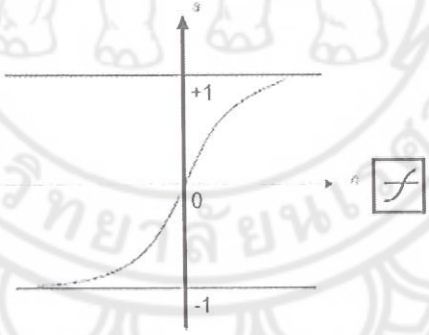


ภาพ 14 แสดงซิกมอยด์ฟังก์ชั่น

ที่มา: ดาริกา เรือนคำ, 2552

3. ไฮเปอร์โบลิคแทนเจ้นท์ฟังก์ชั่น (Hyperbolic Tangent Transfer Function) สามารถอธิบาย ดังสมการ (20) และดังภาพ 15

$$f(I) = \tanh(I) = \frac{\exp^{(I)} - \exp^{(-I)}}{\exp^{(I)} + \exp^{(-I)}} \qquad (20)$$



ภาพ 15 แสดงไฮเปอร์โบลิคแทนเจ้นท์ฟังก์ชั่น

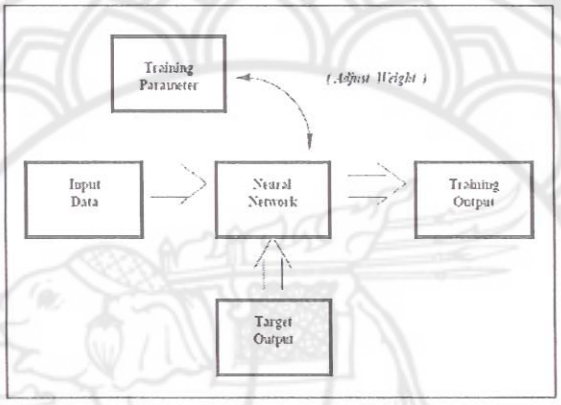
ที่มา: ดาริกา เรือนคำ, 2552

ในจำนวนฟังก์ชันเหล่านี้ซิกมอยด์ฟังก์ชั่นมักจะถูกเลือกใช้เป็นส่วนใหญ่ โดยเฉพาะอย่างยิ่งในโครงข่ายประสาทเทียมแบบหลายชั้น เนื่องจากข้อดีในแง่ของการมีรูปทรงแบบตัว s ซึ่งทำให้มีความต่อเนื่องของความเปลี่ยนแปลงดีกว่าฟังก์ชันแบบขั้น (Step Transfer Function)

การเรียนรู้สำหรับโครงข่ายประสาทเทียม

1. การเรียนแบบมีผู้สอน

การเรียนแบบมีผู้สอน (ดาริกา เรือนคำ, 2552) เป็นการเรียนแบบที่มีการตรวจคำตอบเพื่อให้วงจรข่ายปรับตัว ชุดข้อมูลที่ใช้สอนวงจรข่ายจะมีคำตอบไว้คอยตรวจดูว่าวงจรข่ายให้คำตอบที่ถูกหรือไม่ ถ้าตอบไม่ถูก วงจรข่ายก็จะปรับตัวเองเพื่อให้ได้คำตอบที่ดีขึ้น (เปรียบเทียบกับคนเหมือนกับการสอนนักเรียนโดยมีครูผู้สอนคอยแนะนำ) ดังภาพ 16

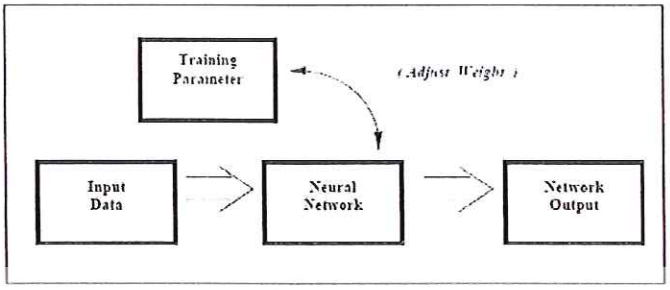


ภาพ 16 แสดงการเรียนรู้แบบมีผู้สอน

ที่มา: ดาริกา เรือนคำ, 2552

2. การเรียนแบบไม่มีผู้สอน

การเรียนแบบไม่มีผู้สอน (ดาริกา เรือนคำ, 2552) เป็นการเรียนแบบไม่มีผู้แนะนำ ไม่มีการตรวจคำตอบว่าถูกหรือผิด วงจรข่ายจะจัดเรียงโครงสร้างด้วยตัวเองตามลักษณะของข้อมูล ผลลัพธ์ที่ได้ วงจรข่ายจะสามารถจัดหมวดหมู่ของข้อมูลได้ (เปรียบเทียบกับคน เช่น การที่เราสามารถแยกแยะพันธุ์พืช พันธุ์สัตว์ตามลักษณะรูปร่างของมันได้เองโดยไม่มีใครสอน) ดังภาพ 17



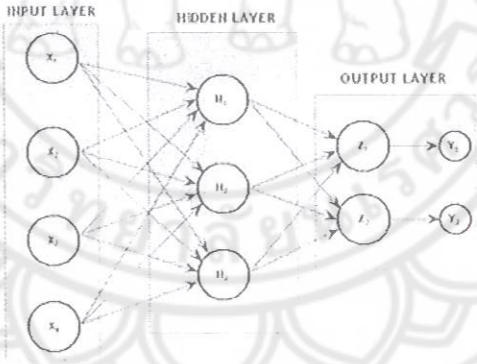
ภาพ 17 แสดงการเรียนรู้แบบไม่มีผู้สอน

ที่มา: ดาริกา เรือนคำ, 2552

สถาปัตยกรรมโครงข่าย (Network Architecture)

1. Feedforward Network

ข้อมูลที่ประมวลผลในวงจรข่ายจะถูกส่งไปในทิศทางเดียวจากโหนดของข้อมูลนำเข้า ส่งต่อมาเรื่อย ๆ จนถึงโหนดของข้อมูลนำออกโดยไม่มีการย้อนกลับของข้อมูล หรือแม้แต่โหนดในชั้นเดียวกันก็ไม่มีการเชื่อมต่อกัน (ดาริกา เรือนคำ, 2552) ดังภาพ 18

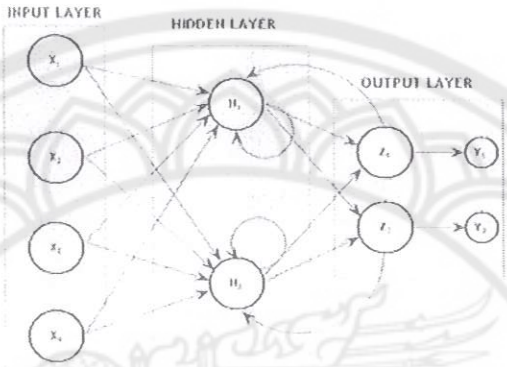


ภาพ 18 แสดงสถาปัตยกรรมของ Feedforward Network

ที่มา: Edward R. Jones, 2008 อ้างอิงใน ดาริกา เรือนคำ, 2552

2. Feedback Network

ข้อมูลที่ประมวลผลในวงจรข่ายจะมีการป้อนกลับเข้าไปยังวงจรโครงข่ายหลาย ๆ ครั้ง จนกระทั่งได้คำตอบออกมา บางทีเรียกว่า Recurrent Network (ดาริกา เรือนคำ, 2552) ดังภาพ 19



ภาพ 19 แสดงสถาปัตยกรรมของ Feedback Network

ที่มา: Edward R. Jones, 2008 อ้างอิงใน ดาริกา เรือนคำ, 2552

โครงข่ายประสาทเทียมแบบหลายชั้น (Multi-Layered Artificial Neural Network)

ในปี ค.ศ.1969 Minskey และ Papert ได้เขียนหนังสืออธิบายเกี่ยวกับขอบเขตของเพอร์เซ็ปตรอนชั้นเดียว (Single-Layer Perceptron) ผลกระทบที่ได้รับจากหนังสือเล่มนั้นร้ายแรง เป็นเหตุให้นักวิจัยสาขาโครงข่ายประสาทเทียม สูญเสียผลประโยชน์ เนื่องจากหนังสือสามารถถ่ายทอดออกมาได้ดี และแสดงข้อมูลในเชิงคำนวณว่าเพอร์เซ็ปตรอนชั้นเดียวไม่สามารถที่จะสร้างรูปแบบการจดจำพื้นฐาน (Basic Pattern Recognition Operation) ได้เช่นการกำหนดความคล้ายคลึงของรูปร่าง หรือกำหนดว่ารูปร่างใดสัมพันธ์กันหรือไม่ แต่สิ่งที่นักวิจัยไม่รู้จนกระทั่งยุค 80s คือ การได้รับการฝึกฝนที่ถูกต้อง ซึ่งเพอร์เซ็ปตรอนหลายชั้น (Multi-layer Perceptron) สามารถดำเนินการแก้ไขสิ่งเหล่านี้ได้ (ดาริกา เรือนคำ, 2552)

องค์ประกอบที่สำคัญของโครงข่ายประสาทเทียมแบบหลายชั้น ประกอบไปด้วย 3 ส่วน หรือ 3 ชั้น ได้แก่ ชั้นของข้อมูลนำเข้า ที่ถูกเชื่อมต่อกับชั้นของชั้นซ่อน ซึ่งเชื่อมต่อกับชั้นของข้อมูลนำออก

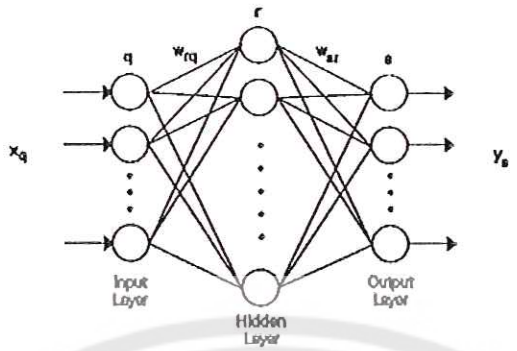
โครงข่ายประสาทเทียมแบบหลายชั้นเป็นโครงข่ายประสาทเทียมรูปแบบหนึ่งที่ได้รับการพัฒนาและนำมาประยุกต์ใช้ในการวิจัยได้อย่างมีประสิทธิภาพ โครงสร้างของโครงข่ายประสาทเทียมแบบหลายชั้นประกอบไปด้วยชั้นของหน่วยประสาทเทียม (Layer) ซึ่งในแต่ละชั้นจะมีหน่วยประสาทเทียมจำนวนหลาย ๆ หน่วยต่ออยู่ด้วยกัน โดยทั่วไปชั้นของหน่วยประสาทเทียมในโครงข่ายประสาทเทียมชนิดนี้แบ่งออกเป็น 3 ชนิด (ดาริกา เรือนคำ, 2552) คือ

1. ชั้นนำเข้า (Input Layer) เป็นชั้นของหน่วยประสาทเทียมที่รับข้อมูลขาเข้า

2. ชั้นซ่อน (Hidden Layer) เป็นชั้นของหน่วยประสาทที่อยู่ภายในโครงข่ายประสาทเทียมระหว่างชั้นข้อมูลนำเข้า และชั้นข้อมูลนำออก ชั้นของหน่วยประสาทนี้ไม่มีการเชื่อมต่อกับข้อมูลภายนอกโดยตรง

3. ชั้นนำออก (Output Layer) เป็นชั้นของหน่วยประสาทเทียมชั้นสุดท้ายที่ส่งข้อมูลออกไปยังภายนอก

โครงข่ายประสาทเทียม (ดาริกา เรือนคำ, 2552) มีชั้นนำเข้า และชั้นนำออก อย่างละหนึ่งชั้น ในขณะที่ชั้นซ่อนอาจจะมีจำนวนมากกว่าหนึ่งชั้นก็ได้ สำหรับจำนวนของหน่วยประสาทเทียมในชั้นนำเข้า และชั้นนำออกก็ขึ้นอยู่กับข้อมูลที่ใช้ในการทำงานของโครงข่ายประสาทเทียมและผลลัพธ์ที่ต้องการได้รับจากโครงข่ายประสาทเทียม ส่วนจำนวนชั้นของชั้นซ่อน และจำนวนของหน่วยประสาทเทียมในชั้นซ่อน นั้นขึ้นอยู่กับความซับซ้อนของปัญหา ปริมาณข้อมูลที่ใช้ ซึ่งการกำหนดจำนวนชั้นและจำนวนหน่วยประสาทเทียมในชั้น ซ่อน แต่ละชั้น ก็ยังไม่มีกฎเกณฑ์ที่ตายตัว แต่จากการวิจัยพบว่าหากจำนวนชั้น และจำนวนหน่วยประสาทเทียมมีน้อยเกินไปโครงข่ายประสาทเทียมก็อาจไม่สามารถตอบปัญหาได้อย่างถูกต้องเท่าที่ควร หรืออาจไม่สามารถเรียนรู้ได้ตามที่ต้องการ ในทางกลับกัน หากจำนวนชั้นและจำนวนหน่วยประสาทเทียมมีมากเกินไปโครงข่ายประสาทเทียมมีแนวโน้มที่จะ "จำ" ตัวอย่างที่ใช้ในกระบวนการเรียนรู้ มากกว่าที่จะเรียนรู้ลักษณะรูปแบบของกลุ่มตัวอย่าง ซึ่งทำให้ไม่สามารถตอบปัญหาได้ดีเท่าที่ควรเช่นกัน นอกจากนั้นยังจะมีผลให้กระบวนการเรียนรู้จะต้องใช้เวลานานขึ้นอีกด้วย ในการใช้งานโครงข่ายประสาทเทียมโดยทั่วไปมักจะใช้จำนวนชั้นซ่อน ประมาณ 1 หรือ 2 ชั้น ก็พอเพียงในการแก้ปัญหาที่ซับซ้อนมาก ๆ ได้ ส วนจำนวนหน่วยประสาทเทียมในชั้นซ่อนเป็นการประมาณและทดลองในรูปแบบลองผิดลองถูกเป็นส่วนใหญ่



ภาพ 20 แสดงโครงข่ายประสาทเทียมแบบหลายชั้น

ที่มา: ดาริกา เรือนคำ, 2552

ภาพ 20 แสดงโครงสร้างของโครงข่ายประสาทเทียมแบบหลายชั้นที่ประกอบไปด้วยชั้นประสาทเทียมจำนวน 3 ชั้นด้วยกัน กล่าวคือ 1) ชั้นนำเข้าหรือชั้น q 2) ชั้นซ่อนหรือชั้น r 3) ชั้นนำออก หรือชั้น s ชั้นประสาทเทียมแต่ละชั้นมีการเชื่อมต่อถึงกันอย่างเต็มรูปแบบ (Fully Interconnected) หมายถึงว่า หน่วยประสาทเทียมแต่ละหน่วยในชั้นหนึ่ง ๆ เชื่อมต่อกับหน่วยประสาทเทียมอื่น ๆ ทุกหน่วยในชั้นก่อนหน้าและชั้นถัดไป แต่ไม่ต่อกับหน่วยประสาทเทียมในชั้นเดียวกัน ในขณะเดียวกันเส้นเชื่อมต่อก็มีค่าน้ำหนักของแต่ละเส้น ที่ทำหน้าที่ปรับระดับของข้อมูลที่ผ่านมาด้วยการคูณ ค่าน้ำหนักของโครงข่ายประสาทเทียมที่แสดงในที่นี้จะเรียกว่า w_{rq} และ w_{sr} ซึ่งค่าน้ำหนักของเส้นเชื่อมต่อเหล่านี้จะถูกปรับเปลี่ยนไปในระหว่างกระบวนการเรียนรู้ของโครงข่ายประสาทเทียม (ดาริกา เรือนคำ, 2552)

ข้อมูลขาเข้าจะถูกรับเข้ามายังโครงข่ายประสาทเทียมผ่านทาง ข้อมูลนำเข้า โดยที่หน่วยประสาทเทียมในชั้น นำเข้า นี้ทำหน้าที่เป็นตัวกระจายข้อมูลไปสู่หน่วยประสาทเทียมในชั้นซ่อนส่วนหน่วยประสาทเทียมแต่ละหน่วยในชั้นซ่อน และชั้นนำออก จะรวบรวมข้อมูลที่ผ่านการปรับระดับโดยเส้นเชื่อมต่อแต่ละเส้น และนำมาตรวจสอบกับฟังก์ชั่นถ่ายโอน เพื่อคำนวณหาผลลัพธ์ในที่นี้ผลลัพธ์ของหน่วยประสาทเทียมในชั้นซ่อนจะถูกส่งต่อไปยัง ชั้นนำออก ในขณะที่ผลลัพธ์ของชั้นนำออก ก็เป็นผลลัพธ์หรือค่าคำตอบของโครงข่ายประสาทเทียมนั่นเอง สำหรับชั้นซ่อนหรือชั้น r ในที่นี้ สามารถคำนวณค่าข้อมูลขาเข้าของแต่ละหน่วยประสาทเทียมได้ (ดาริกา เรือนคำ, 2552) ดังสมการ (21)

$$I_r = \sum_q w_{rq} x_q \quad r = 1,2,...,N_r \tag{21}$$

และสำหรับชั้น นำออก หรือชั้น s ในที่นี้ สามารถคำนวณค่าข้อมูลขาออกของแต่ละหน่วยประสาทเทียมได้ (ดาริกา เรือนคำ, 2552) ดังสมการ (22)

$$o_r = \sum_r w_{sr} o_r \quad s = 1,2,...,N_s \tag{22}$$

โดยที่

x_q คือ ค่าข้อมูลขาเข้าของหน่วยประสาทเทียมในชั้น q

w_{rq} คือ ค่าน้ำหนักเส้นเชื่อมต่อระหว่างชั้น q และ r

w_{sr} คือ ค่าน้ำหนักเส้นเชื่อมต่อระหว่างชั้น r และ s

N_r คือ จำนวนหน่วยประสาทเทียมในชั้น r

N_s คือ จำนวนหน่วยประสาทเทียมในชั้น s

o_r คือ ค่าข้อมูลขาออกของหน่วยประสาทเทียมในชั้น r

ซึ่งผลลัพธ์หรือข้อมูลขาออกของหน่วยประสาทเทียมในชั้น r หรือ o_r สามารถคำนวณหาได้ (ดาริกา เรือนคำ, 2552) ดังสมการ (23)

$$o_r = f(I_r) \quad r = 1,2,...,N \tag{23}$$

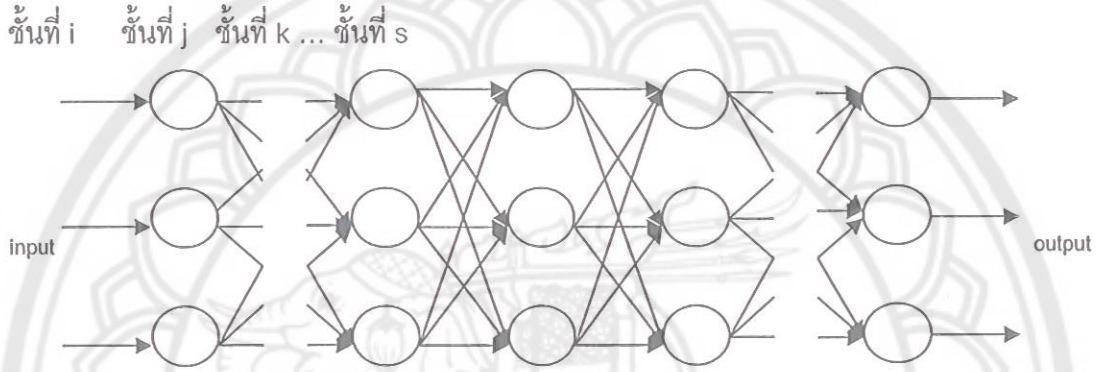
และผลลัพธ์หรือข้อมูลขาออกของหน่วยประสาทเทียมในชั้น s หรือ y_s สามารถคำนวณหาได้ (ดาริกา เรือนคำ, 2552) ดังสมการ (24)

$$y_s = f(I_s) \; s = 1,2,...,N \tag{24}$$

การเรียนรู้โดยวิธีแบบแพร่กลับ (Back-propagation Neural Network Algorithm)

กระบวนการเรียนรู้โดยใช้วิธีแบบแพร่กลับ (ดาริกา เรือนคำ, 2552) เป็นการเรียนรู้แบบหนึ่งที่จัดอยู่ในแบบต้องมีการสอน ซึ่งมีการแสดงทั้งข้อมูลขาเข้า และผลลัพธ์ที่เป็นเป้าหมายให้โครงข่ายประสาทเทียมได้ทราบ จากข้อมูลขาเข้าที่ได้รับ โครงข่ายประสาทเทียมจะทำการประมวล

หาค่าผลลัพธ์ออกมา หลังจากนั้นค่าความผิดพลาดระหว่างผลลัพธ์ที่ได้กับผลลัพธ์ที่เป็นเป้าหมาย จะถูกแพร่ย้อนกลับเข้ามาในโครงข่ายประสาทเทียมเพื่อใช้ในการปรับค่าน้ำหนักเส้นเชื่อมต่อ จนกระทั่งค่าความผิดพลาดลดน้อยลงต่ำกว่าค่าที่กำหนดในช่วงเริ่มต้นของกระบวนการเรียนรู้ ค่าน้ำหนักเส้นเชื่อมต่อถูกกำหนดโดยการสุ่มโดยใช้ค่าน้อย ๆ เมื่อได้รับข้อมูลคู่หนึ่งจากชุดข้อมูลสำหรับการเรียนรู้ โครงข่ายประสาทเทียมก็จะทำการประมวลผลจนได้ค่าผลลัพธ์ y_s ออกมา ซึ่งค่าผลลัพธ์ดังกล่าวเป็นผลจากค่าน้ำหนักเส้นเชื่อมต่อในขณะนั้น ดังภาพ 21



ภาพ 21 แสดงโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับ

ที่มา: ดาริกา เรือนคำ, 2552

ดังนั้นค่าความผิดพลาดที่เกิดขึ้นในชั้นนำออก หรือ δ_s ถูกคำนวณได้ (ดาริกา เรือนคำ, 2552) ดังสมการ (25)

$$\delta_s = f'(I_s)(t_s - y_s) \quad s = 1,2,...,N_s \tag{25}$$

ส่วนค่าความผิดพลาดที่เกิดขึ้นในชั้นซ่อนหรือ δ_r จะถูกคำนวณได้ (ดาริกา เรือนคำ, 2552) ดังสมการ (26)

$$\delta_r = f'(I_r)\sum_r \delta_s w_{sr} \quad r = 1,2,...,N_r \tag{26}$$

โดยที่

$f'(*)$ คือ อนุพันธ์ชั้นที่หนึ่งของฟังก์ชั่นถ่ายโอนเมื่อเทียบกับ I

t_s คือ ค่าผลลัพธ์ที่เป็นเป้าหมาย

สมมุติให้เลือกฟังก์ชั่นซิกมอยด์เป็นฟังก์ชั่นถ่ายโอน ที่ได้รับการใช้อย่างค่อนข้างแพร่หลายในโครงข่ายประสาทเทียม (ดาริกา เรือนคำ, 2552) ดังสมการ (27) และ (28)

$$f'(I_s) = \theta f(I_s)(1 - f(I_s)) \\ = \theta y_s(1 - y_s) \quad (27)$$

และ

$$f'(I_r) = \theta f(I_r)(1 - f(I_r)) \quad = \theta O_r(1 - O_r) \quad (28)$$

ดังนั้น ค่าความผิดพลาดที่เกิดขึ้นในชั้นนำออกและชั้นซ่อนก็สามารถเขียนได้ (ดาริกา เรือนคำ, 2552) ดังสมการ (29)

$$\delta_s = \theta y_s(1 - y_s)(t_s - y_s) \quad s = 1,2,...,N_s \\ \delta_s = \theta O_r(1 - O_r)\sum_s \delta_s w_{sr} \quad r = 1,2,...,N_r \quad (29)$$

ด้วยค่าความผิดพลาดที่คำนวณได้ข้างต้น ค่าน้ำหนักเส้นเชื่อมต่อในชั้นนำออกก็จะถูกปรับเปลี่ยนไป (ดาริกา เรือนคำ, 2552) สามารถคำนวณดังสมการ (30) และ (31)

โดยที่

$$w_{sr}^{new} = w_{sr}^{old} + \Delta w_{sr}^{new} \quad (30)$$

$$\Delta w_{sr}^{new} = (1 - \beta)\eta\delta_s O_r + \beta\Delta w_{sr}^{old} \quad (31)$$

ในขณะเดียวกัน ค่าน้ำหนักเส้นเชื่อมต่อในชั้นซ่อน ก็สามารถปรับเปลี่ยนไป (ดาริกา เรือนคำ, 2552) สามารถคำนวณดังสมการ (32) และ (33)

โดยที่

$$w_{rq}^{new} = w_{rq}^{old} + \Delta w_{rq}^{new} \quad (32)$$

$$\Delta w_{rq}^{new} = (1-\beta)\eta\delta_r O_q + \beta\Delta w_{rq}^{old} \quad (33)$$

ทั้งนี้ค่า η คือ อัตราการเรียนรู้ (Learning Rate) (ดาริกา เรือนคำ, 2552) ซึ่งเป็นตัวกำหนดขนาดขั้นของการปรับเปลี่ยนค่าน้ำหนักเส้นเชื่อมต่อของโครงข่ายประสาทเทียมในแต่ละรอบของการเรียนรู้ โดยทั่วไปมีค่าอยู่ระหว่าง 0 ถึง 1 หาก η มีค่ามาก ๆ การเรียนรู้จะดำเนินไปอย่างรวดเร็ว แต่ก็จะทำให้มีโอกาสข้ามผ่านจุดต่ำสุดของค่าผิดพลาดได้ง่าย ในขณะที่หาก η มีค่าน้อย ๆ การเรียนรู้จะดำเนินไปอย่างละเอียด แต่จะใช้เวลานานขึ้น ส่วน β คือ โมเมนตัม (Momentum Rate) หรือตัวปรับเรียบ มีค่าอยู่ระหว่าง 0 ถึง 1 เช่นกัน ซึ่งเป็นค่าที่ช่วยกำหนดระดับความสำคัญของการปรับเปลี่ยนค่าน้ำหนักเส้นเชื่อมต่อในรอบที่ผ่านมาก่อนหน้า เปรียบเทียบได้กับโมเมนตัมของการปรับเปลี่ยนนั่นคือเมื่อโครงข่ายประสาทเทียมได้ค่าการปรับเปลี่ยนจากการเรียนรู้ในรอบปัจจุบัน ก็ยังคงไม่ละทิ้งการเรียนรู้ในรอบที่ผ่านมาก่อนหน้าเสียทั้งหมด ทั้งหมดนี้เป็นแนวคิดที่จำลองมาจากกระบวนการเรียนรู้ของมนุษย์นั่นเองเมื่อการเรียนรู้รอบหนึ่งผ่านไป ข้อมูลคู่ถัดไปจากชุดข้อมูลสำหรับการเรียนรู้ก็จะถูกแสดงให้กับโครงข่ายประสาทเทียมอีก การคำนวณและปรับเปลี่ยนค่าน้ำหนักเส้นเชื่อมต่อก็จะเกิดขึ้นอีกกระบวนการเรียนรู้นี้จะดำเนินไปซ้ำเป็นรอบ ๆ จนถึงจำนวนรอบการเรียนรู้ที่กำหนดไว้หรือเมื่อค่าความผิดพลาดที่ใช้เป็นเกณฑ์ในการเรียนรู้ลดลงต่ำกว่าค่าที่กำหนดความผิดพลาดที่ใช้เป็นเกณฑ์ในการเรียนรู้มักคำนวณได้ ดังสมการ (34)

$$MSE = \frac{1}{N_T}\sum_{N_T}\sum_{s}(t_s - y_s)^2 \quad (34)$$

โดยที่

N_T คือ จำนวนข้อมูลในชุดข้อมูลสำหรับการเรียนรู้

k-fold Cross Validation

วิธีการ Cross Validation (Tan, et al., 2006; Witten, et al., 2011) คือ การสุ่มตัวอย่างโดยเริ่มจากแบ่งชุดข้อมูลออกเป็นส่วน ๆ และนำบางส่วนจากชุดข้อมูลนั้นมาตรวจสอบ ในการทำ k-fold Cross Validation จะแบ่งข้อมูลออกเป็น k ส่วนเท่า ๆ กัน และทำการคำนวณค่าความผิดพลาด k รอบ โดยแต่ละรอบการคำนวณข้อมูลหนึ่งส่วนจากข้อมูล k ส่วนจะถูกเลือกออกมาเพื่อเป็นส่วนของการทดสอบ และข้อมูลอีก k - 1 ส่วนจะถูกใช้เป็นส่วนของการฝึกสอน

งานวิจัยนี้ได้นำ 10-fold Cross Validation ซึ่งเป็นการแบ่งข้อมูลออกเป็น 10 ส่วนเท่า ๆ กัน ทำการสร้างตัวแบบ 10 ครั้ง โดยใช้ข้อมูลครั้งละเพียง 9 ส่วน เป็นส่วนของการฝึกสอน และข้อมูลอีก 1 ส่วน เป็นส่วนของการทดสอบ ดังภาพ 22 และจะได้ตัวแบบของการเรียนรู้ครบ 10 ตัวแบบ แล้วจะนำค่าความถูกต้องทั้งหมดมาเฉลี่ย เพื่อใช้เป็นตัวแทนของค่าความถูกต้องของข้อมูลทั้งหมด

T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
T1	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**	**T10**

เป็นส่วนของการทดสอบ เป็นส่วนของการฝึกสอน

ภาพ 22 แสดงวิธีการแบบ 10-fold Cross Validation

การวัดประสิทธิภาพของการพยากรณ์

ตาราง Confusion Matrix (Tan, et al., 2006) เป็นตารางในการวัดประสิทธิภาพของตัวแบบเพื่อจัดหมวดหมู่ของข้อมูล โดยอยู่บนพื้นฐานของจำนวนตัวอย่างที่จัดหมวดหมู่ผิดไว้ในเมทริกซ์ ซึ่งมีแถวแทนคลาสจริงของข้อมูล และมีคอลัมน์แทนคลาสที่ตัวแบบทำนายได้ ตาราง Confusion Matrix แสดงดังตาราง 2

ตาราง 2 แสดง Confusion Matrix

ค่าที่แท้จริง (ACTUAL CLASS)	ค่าการทำนาย (PREDICTED CLASS)	
	Class YES	Class NO
Class YES	TP (True Positive)	FN (False Negative)
Class NO	FP (False Positive)	TN (True Negative)

โดยที่

TP คือ ค่าที่บอกความถูกต้องในการจัดหมวดหมู่ของข้อมูล ซึ่งมีค่าที่แท้จริงอยู่ใน Class YES และมีการทำนายว่าอยู่ใน Class YES (ทำนายถูกต้อง)

FP คือ ค่าที่บอกความถูกต้องในการจัดหมวดหมู่ของข้อมูล ซึ่งมีค่าที่แท้จริงอยู่ใน Class NO และมีการทำนายว่าอยู่ใน Class YES (ทำนายผิด)

FN คือ ค่าที่บอกความถูกต้องในการจัดหมวดหมู่ของข้อมูล ซึ่งมีค่าที่แท้จริงอยู่ใน Class YES และมีการทำนายว่าอยู่ใน Class NO (ทำนายผิด)

TN คือ ค่าที่บอกความถูกต้องในการจัดหมวดหมู่ของข้อมูล ซึ่งมีค่าที่แท้จริงอยู่ใน Class NO และมีการทำนายว่าอยู่ใน Class NO (ทำนายถูกต้อง)

วิธีการวัดประสิทธิภาพจะใช้ค่าที่ได้จาก ตาราง Confusion Matrix มาคำนวณค่าวัดประสิทธิภาพต่าง ๆ ดังนี้

1. ค่าความถูกต้อง (Accuracy) คือ ค่าความถูกต้องของตัวแบบที่บอกว่าโปรแกรมสามารถทำนายได้แม่นยำขนาดไหน พิจารณาจากการทายถูกทั้งหมด สามารถคำนวณหาได้ดังสมการ (35)

$$Accuracy = \frac{(TP + TN)}{TP + TN + FP + FN} \quad (35)$$

2. ค่าความแม่นยำ (Precision) หมายถึง ค่าของการทำนายค่าที่จริงในกลุ่มที่พบและถูกต้อง ซึ่งพิจารณาจากจำนวนกลุ่มข้อมูลทั้งหมด สามารถคำนวณหาได้ ดังสมการ (36) ดังนี้

$$\Pr ecision = \frac{TP}{(TP + FP)} \quad (36)$$

3. ค่าความระลึก (Recall) หมายถึง ค่าจำนวนกลุ่มข้อมูลที่พบ ซึ่งพิจารณาจากจำนวนกลุ่มข้อมูลทั้งหมด สามารถคำนวณหาได้ ดังสมการ (37) ดังนี้

$$\mathrm{Re}call = \frac{TP}{(TP + FN)} \quad (37)$$

ค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Root Mean Square Error: RMSE)

ค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง เป็นค่าวัดความแตกต่างระหว่างค่าจริง และค่าที่ประมาณได้จากตัวแบบ หากค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองมีค่าน้อย แสดงว่าตัวแบบสามารถประมาณค่าได้ใกล้เคียงกับค่าจริง ดังนั้นหากค่านี้มีค่าเท่ากับศูนย์แล้วจะหมายความว่าไม่เกิดความคลาดเคลื่อนในตัวแบบนี้เลย ค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง สามารถคำนวณหาได้ ดังสมการ (38)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(\hat{Y}_i - Y_i\right)^2}{N}} \quad (38)$$

โดยที่

$RMSE$ คือ ค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง

N คือ จำนวนขนาดตัวอย่างที่ใช้ในการประมาณตัวแบบ

$\hat{Y}_i$ คือ ค่าที่แท้จริงค่าข้อมูลจริงที่ได้จากการคำนวณ

Y_i คือ ค่าประมาณจากตัวแบบค่าข้อมูลที่ได้จากการพยากรณ์

การพัฒนาหลักสูตร

การพัฒนาหลักสูตร (สุธี วรประดิษฐ, 2551 อ้างอิงใน วรรณนภา โพธิ์ผลิ, 2552) หลักสูตรเป็นสิ่งสำคัญของการจัดการศึกษา เพราะเป็นสิ่งที่กำหนดแนวทางการปฏิบัติในการจัดการเรียนการสอนให้บรรลุจุดมุ่งหมายที่กำหนดไว้ หลักสูตรที่ดีต้องมีการพัฒนาอยู่เสมอ เพื่อให้มีเนื้อหาสาระทันกับการเปลี่ยนแปลงทางสังคม เศรษฐกิจ เทคโนโลยี และการเมือง

ความหมายของหลักสูตร

มีนักการศึกษาได้ให้ความหมายของหลักสูตรไว้หลายท่าน ดังนี้

โอลิวา (1992, หน้า 8-9 อ้างอิงใน วรรณนภา โพธิ์ผลิ, 2552) ซึ่งได้ให้นิยามความหมายของหลักสูตร โดยแบ่งออกได้เป็น

1. การให้นิยามโดยยึดจุดประสงค์ (Purpose) หลักสูตร จึงมีภาระหน้าที่ที่จะทำให้ผู้เรียนควรจะเป็นอย่างไรหรือมีลักษณะอย่างไร หลักสูตรแนวคิดนี้จึงมีความหมายในลักษณะที่เป็นวิธีการ ที่นำไปสู่ความสำเร็จตามจุดประสงค์หรือจุดมุ่งหมาย เช่น หลักสูตร คือ การถ่ายทอดมรดกทางวัฒนธรรม และ การพัฒนาทักษะการคิดของผู้เรียน เป็นต้น

2. การให้นิยามโดยยึดบริบทหรือสภาพแวดล้อม (Contexts) นิยามหลักสูตรในลักษณะนี้ เป็นการอธิบายถึงลักษณะทั่วไปของหลักสูตร ซึ่งแล้วแต่ว่าเนื้อหาสาระของหลักสูตรมีลักษณะเป็นอย่างไร เช่น หลักสูตรที่ยึดเนื้อหาวิชา หลักสูตรที่ยึดผู้เรียนเป็นศูนย์กลางหลักสูตรเพื่อการปฏิรูปสังคม เป็นต้น

3. การให้นิยามโดยยึดวิธีดำเนินการหรือยุทธศาสตร์ (Strategies) เป็นการให้นิยามหลักสูตร ในเชิงวิธีดำเนินการที่เป็นกระบวนการ ยุทธศาสตร์หรือเทคนิควิธีการที่ใช้ในการจัดการเรียนการสอน เช่น หลักสูตร คือ กระบวนการแก้ปัญหา หลักสูตร คือ การทำงานกลุ่ม หลักสูตร คือ การเรียนรู้รายบุคคล หลักสูตร คือ โครงการหรือแผนการจัดการเรียนการสอน เป็นต้น

โซเวลล์ (1996, หน้า 5 อ้างอิงใน วรรณนภา โพธิ์ผลิ, 2552) ได้กล่าวว่า มีผู้อธิบายความหมายของหลักสูตรไว้อย่างมากมาย เช่น หลักสูตรเป็นการสะสมความรู้ดั้งเดิม เป็นวิธีการคิด เป็นประสบการณ์ที่ถูกกำหนดไว้ เป็นแผนการจัดสภาพการเรียนรู้ เป็นความรู้ และคุณลักษณะของผู้เรียน เป็นเนื้อหาและกระบวนการเป็นแผนการเรียนการสอน เป็นจุดหมายปลายทาง และผลลัพธ์ของการจัดการเรียนการสอน และเป็นผลผลิตของระบบเทคโนโลยี เป็นต้น โซเวลล์ ได้อธิบายว่าเป็นเรื่องปกติที่นิยามความหมายของหลักสูตรมีความแตกต่างกันไปเพราะบางคนให้ความหมายของหลักสูตรในระดับที่แตกต่างกัน หรือไม่ได้แยกหลักสูตรกับการจัดการเรียนการสอน โซเวลล์ ได้สรุปว่า หลักสูตร คือ การสอนอะไรให้กับผู้เรียน ซึ่งมีความหมายที่กว้างขวาง ที่รวมทั้งข้อมูล ข่าวสาร ทักษะ และทัศนคติ ทั้งที่ได้กำหนดไว้ และไม่ได้กำหนดไว้ให้แก่ผู้เรียนในสถานศึกษา

ชมพันธุ์ กุญชร ณ อยุธยา (2540, หน้า 3-5 อ้างอิงใน วรรณนภา โพธิ์ผลิ, 2552) ได้อธิบายความหมายของ หลักสูตรว่า มีความแตกต่างกันไปตั้งแต่ความหมายที่แคบสุดจนถึงกว้างสุด ซึ่งสามารถจำแนกความคิดเห็นของนักการศึกษาที่ได้ให้นิยามความหมายของหลักสูตรแบ่งออกเป็น 2 กลุ่มใหญ่ ๆ ได้ดังนี้

1. หลักสูตร หมายถึง แผนประสบการณ์การเรียน นักการศึกษาที่มีความคิดเห็นว่าหลักสูตร หมายถึง แผนประสบการณ์การเรียนนั้น มองหลักสูตรที่เป็นเอกสาร หรือโครงการของการศึกษาที่สถาบันการศึกษาไว้วางแผนไว้ เพื่อให้ผู้เรียนได้ศึกษาตามแผน หรือโครงการที่กำหนดไว้ หลักสูตรตามความหมายนี้ หมายรวมถึง แผนการเรียน หรือรายวิชาต่าง ๆ ที่กำหนดให้เรียน รวมทั้งเนื้อหาวิชาของรายวิชาต่าง ๆ กิจกรรมการเรียนการสอน และการประเมินผล ซึ่งได้กำหนดไว้ในแผนความคิดเห็นของนักการศึกษากลุ่มนี้ ไม่รวมถึงการนำหลักสูตรไปใช้ หรือการเรียนการสอนที่ปฏิบัติจริง

2. หลักสูตร หมายถึง ประสบการณ์การเรียนของผู้เรียน ที่สถาบันการศึกษาจัดให้แก่ผู้เรียนประกอบด้วย จุดมุ่งหมาย เนื้อหา การจัดกิจกรรมการเรียนการสอน การประเมินผล

รุจิร์ ภู่สาระ (2545, หน้า 1 อ้างอิงใน วรรณนภา โพธิ์ผลิ, 2552) ได้อธิบายความหมายของหลักสูตรว่า หมายถึง แผนการเรียน ประกอบด้วยเป้าหมาย และจุดประสงค์เฉพาะที่จะนำเสนอและจัดการเนื้อหา รวมถึงแบบของการเรียนการสอนตามจุดประสงค์ และท้ายที่สุดจะต้องมีการประเมินผลของการเรียน

จากความหมายข้างต้นสรุปได้ว่า หลักสูตร หมายถึง แนวการจัดประสบการณ์ และ/หรือเอกสาร ที่มีการจัดทำเป็นแผนการจัดสภาพการเรียนรู้หรือโครงการจัดการศึกษา โดยมีการกำหนดวิธีการจัดการเรียนรู้ เพื่อให้ผู้เรียนเกิดผลการเรียนรู้ตามจุดประสงค์ หรือจุดมุ่งหมายตามที่หลักสูตรกำหนดไว้ แผนการจัดสภาพการเรียนรู้ดังกล่าว อาจจัดขึ้นได้ทั้งใน และนอกชั้นเรียนภายใต้การบริหาร และดำเนินงานของสถานศึกษา

ความหมายของการพัฒนาหลักสูตร

การพัฒนาหลักสูตรเป็นภารกิจของสถานศึกษา ที่จะต้องดำเนินการอย่างต่อเนื่อง ดังนั้น สถานศึกษาจึงมีความจำเป็นต้องมีความเข้าใจในความหมายของการพัฒนาหลักสูตร ซึ่งมีผู้ให้ความหมายการพัฒนาหลักสูตรไว้หลายกรณี เช่น

กู๊ด (Carter V. Good, 1973, หน้า 157 – 158 อ้างอิงใน สุนีย์ ภู่พันธ์, 2546, หน้า 158-159) ได้ให้ความเห็นว่า การพัฒนาหลักสูตรเกิดขึ้นได้ 2 ลักษณะ คือ การปรับปรุงและเปลี่ยนแปลงหลักสูตร การปรับปรุงหลักสูตรเป็นวิธีการพัฒนาหลักสูตรอย่างหนึ่ง เพื่อให้เหมาะสมกับโรงเรียนหรือระบบโรงเรียน จุดหมายของการสอน วัสดุอุปกรณ์ วิธีสอนรวมทั้งการประเมินผล ส่วนคำว่า

การเปลี่ยนแปลงหลักสูตร หมายถึงการแก้ไขหลักสูตรให้แตกต่างไปจากเดิม เป็นการสร้างโอกาสทางการเรียนขึ้นไป

ทาบา (Hilda Taba, 1962, หน้า 454 อ้างอิงใน สุนีย์ ภู่พันธ์, 2546, หน้า 158-159) **ได้กล่าวว่า การพัฒนาหลักสูตรหมายถึง การเปลี่ยนแปลงและปรับปรุงหลักสูตรเดิมให้ได้ผลดียิ่งขึ้นทั้งในด้านการวางจุดมุ่งหมาย การจัดเนื้อหาวิชาการเรียนการสอน การวัดผลและการ**ประเมินผลอื่น ๆ เพื่อให้บรรลุถึงจุดมุ่งหมายอันใหม่ที่วางไว้ การเปลี่ยนแปลงหลักสูตรเป็นการเปลี่ยนแปลงทั้งระบบ หรือเปลี่ยนแปลงทั้งหมดตั้งแต่จุดมุ่งหมายและวิธีการ และการเปลี่ยนแปลงหลักสูตรนี้จะมีผลกระทบทางด้านความคิดและความรู้สึกผู้ที่เกี่ยวข้องทุกฝ่าย ส่วนการปรับปรุงหลักสูตร หมายถึง การเปลี่ยนแปลงหลักสูตรเพียงบางส่วนโดยไม่เปลี่ยนแปลงแนวคิดพื้นฐานหรือรูปแบบของหลักสูตร

วิชัย วงษ์ใหญ่ (2525, หน้า 10 อ้างอิงใน สุนีย์ ภู่พันธ์, 2546, หน้า 58-159) กล่าวว่า การพัฒนาหลักสูตร คือ การพยายามวางโครงการที่จะช่วยให้นักเรียนได้เรียนรู้ตรงตามจุดหมายที่กำหนดไว้ หรือการพัฒนาหลักสูตร และการสอนคือระบบโครงสร้างของการจัดโปรแกรมการสอน กำหนดจุดมุ่งหมาย เนื้อหาสาระ การปรับปรุงตำราแบบเรียน คู่มือครู และสื่อการเรียนต่าง ๆ การวัด และประเมินผลการใช้หลักสูตรการปรับปรุงแก้ไข และการให้ การอบรมครูผู้ใช้หลักสูตรให้เป็นไปตามวัตถุประสงค์ของการพัฒนาหลักสูตรและการสอน รวมทั้งการบริหารและบริการหลักสูตร

จากความหมายดังกล่าวพอสรุปได้ว่า การพัฒนาหลักสูตร เป็นการปรับปรุงหรือเปลี่ยนแปลงหลักสูตรที่ใช้อยู่เดิมให้มีความเหมาะสมสอดคล้องกับสภาวะการเปลี่ยนแปลงทางสังคมวัฒนธรรม เศรษฐกิจ การเมืองและการปกครองภายในท้องถิ่น ภายในประเทศ และต่างประเทศ โดยมุ่งเน้นในการกำหนดจุดหมาย เป้าหมาย จุดประสงค์ เนื้อหาสาระ การจัดการเรียนการสอน การวัด และประเมินผล

การพัฒนาหลักสูตรในแต่ละระดับ

ในการพัฒนาหลักสูตรนั้น สุนีย์ ภู่พันธ์ (2546, หน้า 160-161) ได้กล่าวไว้ว่า การพัฒนา**หลักสูตรสามารถพัฒนาได้หลายระดับ ขึ้นอยู่กับจุดมุ่งหมายการพัฒนาหลักสูตรนั้น ๆ ว่าจะพัฒนาส่วนใด ส่วนหนึ่งของหลักสูตร เราสามารถแบ่งการพัฒนาหลักสูตรออกเป็น 3 ระดับ ดังนี้คือ**

1. หลักสูตรระดับชาติ เป็นหลักสูตรแกน หรือหลักสูตรแม่บทในระดับการศึกษาต่าง ๆ เช่น หลักสูตรประถมศึกษา หลักสูตรมัธยมศึกษา หลักสูตรอาชีวะศึกษา หลักสูตรการฝึกหัดครูเป็น

ต้นซึ่งมีหน่วยงานที่ทำหน้าที่รับผิดชอบ เช่น ศูนย์พัฒนาหลักสูตร กรมวิชาการ กระทรวงศึกษาและสถาบัน ส่งเสริมการสอนวิทยาศาสตร์และเทคโนโลยี

2. หลักสูตรระดับท้องถิ่น เป็นการนำหลักสูตรระดับชาติมาพัฒนาปรับปรุงให้สอดคล้องกับสภาพท้องถิ่น ตามลักษณะทางธรรมชาติสิ่งแวดล้อมและลักษณะพิเศษของท้องถิ่นเพื่อให้เหมาะสมกับผู้เรียน ผู้ที่มีหน้าที่ในการพัฒนาหลักสูตรระดับท้องถิ่น ได้แก่ เขตการศึกษาต่างๆ หน่วยศึกษานิเทศก์ ผู้บริหารสถานศึกษา ผู้สอน หรือสถานประกอบการในท้องถิ่น

3. หลักสูตรระดับห้องเรียน เป็นการนำหลักสูตรมาพัฒนาสู่การจัดการเรียนการสอนที่มีประสิทธิภาพเพื่อให้บรรลุตามจุดมุ่งหมายที่วางไว้ ผู้ที่มีหน้าที่พัฒนาหลักสูตในระดับนี้ คือ ครูผู้สอน โดยการสนับสนุนของผู้บริหารสถานศึกษา

จากแนวคิดดังกล่าวซึ่งมีความสอดคล้องกับการใช้หลักสูตรการศึกษาขั้นพื้นฐาน พุทธศักราช 2544 โดยกระทรวงศึกษาธิการได้จัดทำหลักสูตรการศึกษาขั้นพื้นฐาน พุทธศักราช 2544 เป็นหลักสูตรแกนกลาง ตลอดจนจัดทำเอกสาร คู่มือที่เกี่ยวข้องกับหลักสูตร ให้โรงเรียนนำหลักสูตรแกนกลางและเอกสารที่เกี่ยวข้องไปจัดทำหลักสูตรสถานศึกษา โดยมีผู้บริหาร คณะครู และคณะกรรมการการศึกษาขั้นพื้นฐาน เป็นผู้จัดทำซึ่งมีคณะศึกษานิเทศก์ เป็นผู้ให้ความช่วยเหลือ แนะนำ และเอื้ออำนวยความสะดวกเอกสารที่เกี่ยวข้อง ต่อจากนั้นคณะครูก็นำหลักสูตรสถานศึกษาในแต่ละกลุ่มสาระการเรียนรู้ ไปใช้จัดการเรียนการสอน และในแต่ละปีสถานศึกษาจะต้องประเมินผลและรายงานผลการใช้หลักสูตรสถานศึกษาให้ต้นสังกัด และคณะกรรมการการศึกษาขั้นพื้นฐานทราบ

ความสำคัญของหลักสูตร

ความสำคัญของหลักสูตรที่มีต่อการจัดการศึกษานั้น มีนักการศึกษาหลายท่านได้กล่าวตรงกันว่า หลักสูตรมีความสำคัญต่อการจัดการศึกษา ดังนี้

ธำรง บัวศรี (2532, หน้า 6-7 อ้างอิงใน วรรณนภา โพธิ์ผลิ, 2552) ได้กล่าวว่า หลักสูตรมีความสำคัญ เพราะหลักสูตรเป็นส่วนกำหนดมาตรฐานการเรียนรู้ เพื่อให้แน่ใจว่าผู้เรียนได้รับการศึกษาที่มีคุณภาพ โดยมีรายละเอียดที่บ่งชี้ว่า ผู้เรียนควรเรียนรู้อะไร มีเนื้อหาสาระมากน้อยเพียงไร ควรได้รับการฝึกฝนให้มีทักษะในด้านใด และควรมีพัฒนาการทั้งในส่วนของร่างกาย จิตใจ สังคม และสติปัญญาอย่างไร

สุมิตร คุณานุกร (2536, หน้า 199-200 อ้างอิงใน วรรณนภา โพธิ์ผลิ, 2552) กล่าวถึงความสำคัญของหลักสูตรว่าหลักสูตรมีความสำคัญ เพราะเป็นเครื่องชี้นำทางหรือเป็นบทบัญญัติของรัฐในการจัดการศึกษาเพื่อให้ผู้ที่มีส่วนเกี่ยวข้องกับการจัดการศึกษานำไปปฏิบัติ อีกทั้งยังเป็นเกณฑ์มาตรฐานทางการศึกษา และควบคุมการจัดการเรียนการสอนของสถานศึกษา

ปฏล นันทวงศ์ และไพโรจน์ ด้วงวิเศษ (2543, หน้า 9 อ้างอิงใน วรรณนภา โพธิ์ผลิ, 2552) สรุปความสำคัญของหลักสูตรว่า หลักสูตรมีความสำคัญยิ่งในฐานะที่เป็นเอกสารที่กำหนดแนวทางในการจัดการเรียนการสอนของโรงเรียน ซึ่งผู้ที่เกี่ยวข้องในการจัดการศึกษาทุกฝ่ายต้องยึดถือเป็นแนวปฏิบัติ เพื่อพัฒนาบุคคลให้มีประสิทธิภาพตามที่พึงประสงค์ให้แก่สังคม และประเทศชาติ

จากความหมายข้างต้นสรุปได้ว่า หลักสูตรมีความสำคัญ เพราะหลักสูตรเป็นเอกสารซึ่งเป็นแผนการ หรือโครงการจัดการศึกษาที่ระบุแนวทางการจัดมวลประสบการณ์เป็นส่วนกำหนดแนวทางการจัดการศึกษาให้กับบุคลากรที่เกี่ยวข้องกับการจัดการศึกษานำไปปฏิบัติ เพื่อให้ผู้เรียนมีคุณภาพทางการศึกษาตามเกณฑ์มาตรฐานการศึกษาที่หลักสูตรกำหนดไว้

องค์ประกอบของหลักสูตร

ในการพัฒนาหลักสูตรต้องพิจารณาถึงองค์ประกอบของหลักสูตร ซึ่งประกอบด้วยส่วนต่าง ๆ ที่มีความสอดคล้อง และมีความสัมพันธ์กัน โดย เคอร์ (1976, หน้า 16-17 อ้างอิงใน วรรณนภา โพธิ์ผลิ, 2552) ได้นำเสนอองค์ประกอบของหลักสูตรไว้ 4 ส่วน ได้แก่ 1) วัตถุประสงค์ของหลักสูตร2) เนื้อหาสาระ 3) ประสบการณ์การเรียน และ 4) การประเมินผล

ซึ่งสอดคล้องกับ ทาบา (1962, หน้า 10 อ้างอิงใน วรรณนภา โพธิ์ผลิ, 2552) ที่ได้เสนอองค์ประกอบของหลักสูตรไว้ว่า หลักสูตรไม่ว่าจะสร้างขึ้นในลักษณะใด ย่อมประกอบด้วยส่วนสำคัญ 4 ส่วน คือ 1) จุดมุ่งหมาย 2) เนื้อหาสาระ 3) กิจกรรมและรูปแบบการเรียนการสอน และ 4) การประเมินผล

จากความหมายข้างต้นสรุปได้ว่า องค์ประกอบของหลักสูตรที่สำคัญมี 4 ส่วน คือ 1) จุดมุ่งหมายของหลักสูตร 2) เนื้อหาสาระ 3) กระบวนการจัดการเรียนรู้ และ 4) การประเมินผล

ผลสัมฤทธิ์ทางการเรียน

มีหลายท่านได้ให้ความหมายของผลสัมฤทธิ์ทางการเรียน ดังนี้

Good (1973 อ้างอิงใน ปริญญา คล้ายเจริญ, 2545) ให้ความหมายของผลสัมฤทธิ์ทางการเรียน หมายถึง ความรู้หรือทักษะอันเกิดจากการเรียนรู้ในวิชาต่าง ๆ ที่ได้เรียนมาแล้ว ซึ่งได้จากผลการทดสอบของครูผู้สอน หรือผู้รับผิดชอบในการสอน หรือทั้งสองอย่างรวมกัน

Eysenck (1972 อ้างอิงใน ปริญญา คล้ายเจริญ, 2545) ให้ความหมายของผลสัมฤทธิ์ทางการเรียน หมายถึง เป็นความสำเร็จที่ได้จากการเรียนที่อาศัยความสามารถเฉพาะตัวบุคคล โดยตัวชี้วัดผลสัมฤทธิ์ทางการเรียนอาจได้มาจากกระบวนการที่ไม่ต้องอาศัยการทดสอบ เช่น การสังเกต การตรวจการบ้าน หรืออาจได้ในรูปของระดับคะแนนที่ได้จากโรงเรียน ซึ่งต้องอาศัย

กรรมวิธีที่ซับซ้อน และระยะเวลาที่นานพอสมควร หรืออาจวัดด้วยแบบวัดผลสัมฤทธิ์ทางการเรียนทั่วไป

Milliman และ Greene (1989 อ้างอิงใน ปริญญา คล้ายเจริญ, 2545) กล่าวว่า ผลสัมฤทธิ์ทางการเรียน หมายถึง ปริมาณและลักษณะของความรู้ในสาขาวิชาที่บุคคลได้รับ ลักษณะการจัดองค์ประกอบและโครงสร้างของความรู้ และการใช้ประโยชน์โครงสร้างของความรู้ในการแก้ปัญหาในการคิดสร้างสรรค์ ในการประเมินความน่าเชื่อถือของข้ออ้าง และในการศึกษาค้นคว้าต่อไป

จากความหมายข้างต้นสรุปได้ว่า ผลสัมฤทธิ์ทางการเรียน หมายถึง ความรู้ ความสามารถ ทักษะที่ได้รับจากการเรียนที่ทำให้ผู้เรียนมีความสามารถในการประมวลความรู้ ในการคิดวิเคราะห์ ทักษะทางด้านวิชาการ รวมทั้งสมรรถภาพของสมองด้านต่าง ๆ ที่เกิดขึ้นจากการค้นคว้า การอบรม การสั่งสอน หรือประสบการณ์ต่าง ๆ โดยผลของความสำเร็จวัดได้จากคะแนนที่ผู้เรียนได้รับ

งานวิจัยที่เกี่ยวข้อง

มีงานวิจัยทางด้านการทำเหมืองข้อมูลที่เกี่ยวข้องกับการศึกษา ดังนี้

กฤษณะ ไวยมัย และคณะ (2544) ได้นำเสนอเรื่อง การใช้เทคนิคดาต้าไมน์นิงเพื่อพัฒนาคุณภาพการศึกษาคณะวิศวกรรมศาสตร์ ซึ่งเป็นการศึกษา และวิเคราะห์ระบบฐานข้อมูลนิสิต โดยนำความรู้ทางด้านเหมืองข้อมูลมาประยุกต์ใช้กับข้อมูลนิสิตคณะวิศวกรรมศาสตร์ เพื่อเป็นแนวทางการแก้ไขปัญหาต่าง ๆ เช่น ปัญหาการเลือกสาขาวิชาไม่ตรงกับความสามารถที่แท้จริงปัญหาผลการเรียนของนิสิตตกต่ำจนต้องออกจากสถาบันการศึกษาอันเป็นผลทำให้ไม่สามารถสำเร็จการศึกษาได้ตามวัตถุประสงค์ ซึ่งจะทำให้ไม่อาจผลิตบุคลากรที่มีความสามารถมารับใช้สังคมได้

บุษรา ลิ้มพิพัฒนางกูร (2549) ได้นำเสนอเรื่อง การพยากรณ์โอกาสการสำเร็จการศึกษาของนักศึกษา โดยใช้จีนีติกอัลกอริทึมแบบหลายวัตถุประสงค์ เปรียบเทียบกับวิธีทางระบบโครงข่ายประสาทเทียม : กรณีศึกษานักศึกษาภาควิชาวิทยาการคอมพิวเตอร์ และสารสนเทศ งานวิจัยนี้ประยุกต์ใช้วิธีการพยากรณ์โอกาสในการสำเร็จการศึกษาโดยใช้เทคนิคจีเนติกอัลกอริทึมเพื่อเลือกตัวแปรที่เหมาะสมก่อนนำไปใช้ในการพยากรณ์ด้วยเครือข่ายประสาท (GA/ANN) โดยทดสอบทั้งจีเนติกอัลกอริทึมแบบวัตถุประสงค์เดียว (SGA) และจีเนติกอัลกอริทึมแบบหลายวัตถุประสงค์ (MGA) เปรียบเทียบกับเครือข่ายประสาทแบบทั่วไป (ANN) จากการทดสอบพบว่า GA/ANN นั้นสามารถลดจำนวนตัวแปรที่ใช้ในโมเดลได้อย่างชัดเจน โดย SGA/ANN สามารถลด

จำนวนตัวแปรลง 42% และ MGA/ANN สามารถลดจำนวนตัวแปรลง 45% และเทคนิคที่เสนอนั้นยังให้ประสิทธิภาพการทำนายที่ใกล้เคียงกับการใช้เครือข่ายประสาทเพียงอย่างเดียว

ณัฐริน เจริญเกียรติบวร (2549) ได้นำเสนอเรื่อง การใช้เทคนิคเหมืองข้อมูลเพื่อช่วยในการแนะแนวการศึกษาต่อระดับอุดมศึกษา โดยนำผลการเรียนรายวิชาในระดับชั้นมัธยมศึกษาตอนปลาย มาประยุกต์ใช้ในการทำเหมืองข้อมูล เพื่อค้นหารูปแบบความสัมพันธ์หรือกฎที่ซ่อนอยู่และนำความสัมพันธ์เหล่านั้นมาแนะแนวการศึกษาต่อในระดับชั้นอุดมศึกษา การเลือกตัวอย่างใช้การแบ่งตามสัดส่วนข้อมูลชุดฝึกสอน:ข้อมูลชุดทดสอบ 60:40, 65:35, 70:30 และ 75:25 ตามลำดับ ความถูกต้อง ตัวแบบที่พัฒนาขึ้นวัดโดยใช้ค่าเฉลี่ยร้อยละความถูกต้องของตัวแบบที่สร้างขึ้นเมื่อใช้กับข้อมูลชุดทดสอบ จากการทำซ้ำ 10 ครั้ง สำหรับการแบ่งสัดส่วนข้อมูลที่ศึกษาพบว่ามีค่าเฉลี่ยความถูกต้องเป็น 82.32%, 81.87%, 91.78% และ 92.37% ตามลำดับ โดยมีค่าเฉลี่ยความถูกต้องรวมเป็น 82.08% เมื่อสัดส่วนข้อมูลชุดฝึกสอนมากขึ้นทำให้มีความถูกต้องของผลลัพธ์มากขึ้น

จิราพร ยิ่งกว่าชาติ (2549) ได้นำเสนอเรื่อง การประยุกต์ใช้การเรียนรู้แบบเบย์กับการสร้างแบบจำลองสำหรับทำนายผลสำเร็จการศึกษาของนักศึกษา ทำการวิเคราะห์หาตัวแปรที่มีผลต่อการสำเร็จการศึกษาของนักศึกษาระดับอาชีวะศึกษา ระดับปริญญาตรี และระดับปริญญาโท โดยอาศัยเทคนิคทางการทำเหมืองข้อมูล คือ เทคนิคข่ายงานเบย์ นำมาใช้สร้างตัวแบบ ทำการทดสอบผลบนพื้นฐานของวิธี k - fold Cross Validation ผลการทดลองพบว่า เทคนิคข่ายงานเบย์สามารถค้นพบตัวแปรที่มีความสำคัญสำหรับการทำนายผลการสำเร็จการศึกษาได้ และให้ค่าความแม่นยำในการทำนายสูง จากตัวแบบที่ได้ทำให้ทราบตัวแปรสำคัญที่มีผลต่อการสำเร็จการศึกษาของนักศึกษาในระดับอาชีวะศึกษา คือ อาชีพของมารดา และเกรดเฉลี่ยเดิม ตัวแปรที่มีผลต่อการสำเร็จการศึกษาของนักศึกษาในระดับปริญญาตรี คือ เกรดเฉลี่ยในชั้นปีแรก เกรดเฉลี่ยเดิม วุฒิการศึกษา ความสามารถทางด้านภาษา อาชีพของบิดามารดา รายได้และรายจ่ายของนักศึกษา และตัวแปรที่มีผลต่อการสำเร็จการศึกษาของนักศึกษาในระดับปริญญาโท คือ อาชีพของบิดามารดา และอาชีพของนักเรียน โดยมีค่าระดับความแม่นยำในการทำนายผลของแต่ละแบบจำลองในแต่ละระดับการศึกษา คือ 93.25%, 87.25% และ 98.53% ตามลำดับ

สุวนีย์ กุลกรินีธรรม (2549) ได้นำเสนอเรื่อง การใช้เทคนิคเหมืองข้อมูลเพื่อการจัดกลุ่มหลักสูตรตามกลุ่มสาขาวิชา ISCED กรณีศึกษากลุ่มสาขาวิชาวิทยาศาสตร์ มีวัตถุประสงค์เพื่อพัฒนาตัวแบบสำหรับจัดกลุ่มหลักสูตรตามกลุ่มสาขาวิชา ISCED ในกลุ่มสาขาวิชาวิทยาศาสตร์ ด้วยเทคนิคเหมืองข้อมูลโดยใช้เทคนิคการจำแนกประเภทข้อมูลแบบเบย์ การวิจัยได้

แบ่งขั้นตอนการทำงานเป็น 2 ส่วน คือ ขั้นตอนการสร้างการเรียนรู้ข้อมูล และขั้นตอนการทดสอบข้อมูลโดยใช้เทคนิคการจำแนกประเภทข้อมูลแบบเบย์มาใช้ในการหาความน่าจะเป็นของแต่ละกลุ่มสาขาวิชา ผลการวิจัยภายใต้ข้อมูลระหว่างข้อมูลชุดฝึกสอน และข้อมูลชุดทดสอบพบว่าถ้าแบ่งสัดส่วนข้อมูลชุดฝึกสอนและข้อมูลชุดทดสอบเป็น 60:40 ตัวแบบที่พัฒนาขึ้นมีค่าความถูกต้องเฉลี่ยเป็น 78.57% ถ้าแบ่งสัดส่วนข้อมูลเป็น 70:30 ตัวแบบมีค่าความถูกต้องเฉลี่ยเป็น 85.71% และถ้าแบ่งสัดส่วนข้อมูลเป็นร้อยละ 80:20 ตัวแบบมีค่าความถูกต้องเฉลี่ยเป็น 90.48%

สุคนธ์ทิพย์ วงศ์พันธ์ (2551) ได้นำเสนอเรื่อง การเปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะที่เหมาะสมและอัลกอริทึมเพื่อจำแนกพฤติกรรมการกระทำความผิดของนักเรียนระดับอาชีวศึกษา งานวิจัยนี้ได้ทำการเปรียบเทียบการทำเหมืองข้อมูล 2 วิธี ได้แก่ 1) วิธีการ Simple Classification เพื่อจำแนกข้อมูลโดยไม่มีการคัดเลือกตัวแปรที่เหมาะสม เปรียบเทียบประสิทธิภาพการจำแนกของ 4 อัลกอริทึม ได้แก่ 1) ตัวจำแนกประเภทเบย์อย่างง่าย 2) ข่ายงานความเชื่อแบบเบย์ 3) ต้นไม้ตัดสินใจ : C4.5 และ 4) อัลกอริทึม RIPPER 2) วิธีการ Hybrid Classification ประกอบไปด้วย 2 ขั้นตอน คือ ขั้นตอนที่ 1 เป็นการคัดเลือกตัวแปรที่เหมาะสม โดยทำการเปรียบเทียบอัลกอริทึมในการค้นหาชุดของตัวแปรระหว่าง Genetic Search กับ Forward Selection ร่วมกับวิธีการประเมินค่าชุดของตัวแปร 3 วิธีการ ได้แก่ 1) CFS 2) Consistency และ 3) Wrapper และขั้นตอนที่ 2 นำไปประมวลผลต่อในอัลกอริทึมเพื่อจำแนกประเภทข้อมูลซึ่งเป็นวิธีเดียวกับวิธี Simple Classification พบว่า Hybrid Classification ที่ใช้ Genetic Search ร่วมกับ Wrapper โดยใช้อัลกอริทึม C4.5 ให้ค่าความถูกต้องสูงที่สุดและสามารถลดตัวแปรที่ต้องนำมาใช้ในการจำแนกจาก 43 ตัวแปรเหลือเพียง 15 ตัวแปร และนำชุดตัวแปรที่คัดเลือกได้ไปประมวลต่อในอัลกอริทึมการจำแนกประเภท พบว่าอัลกอริทึมที่ให้ค่าความถูกต้องที่สูงที่สุด คือ อัลกอริทึม C4.5 และคัดเลือกตัวแปรจาก 15 เหลือเพียง 8 ตัวแปร คือ คะแนนเฉลี่ยเดิมก่อนเข้าศึกษา เพศ อายุ รายได้บิดา เกรดวิชาพิมพ์ดีดไทย 1 ส่วนสูง กรุ๊ปเลือด และเกรดวิชาเอกสารธุรกิจ

ประภาพรรณ ขันรินทร์ และคณะ (2552) ได้นำเสนอเรื่อง การใช้เหมืองข้อมูลช่วยแนะนำการเลือกรายวิชาเรียนของนิสิต กรณีศึกษา หลักสูตรวิทยาศาสตรบัณฑิต สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวร โดยแบ่งข้อมูลชุดฝึกสอน และข้อมูลชุดทดสอบ ออกเป็น 80:20 โดยค่าความแม่นยำของตัวแบบที่สร้างจากเทคนิคการค้นหากฎความสัมพันธ์ เท่ากับ 65.19% และค่าความแม่นยำของตัวแบบที่สร้างจากเทคนิคการจำแนกประเภทเบย์อย่างง่าย เท่ากับ 62.76%

Guo (2010) ได้นำเสนอเรื่อง การรวมวิธีการทางสถิติและโครงข่ายข่ายประสาทสำหรับการวิเคราะห์ความพึงพอใจ และการทำนายหลักสูตรของนักศึกษา เป็นการศึกษาวิเคราะห์ และพยากรณ์ปัจจัยที่สำคัญในการพยากรณ์ความพึงพอใจทางการเรียนเพื่อส่งเสริมการเลือกสถาบันศึกษาของนิสิตใหม่ที่อยู่ในท้องถิ่น และเป็นการดึงดูดนิสิตปัจจุบันไม่ให้ย้ายไปไหน ผู้วิจัยได้ทำการเก็บรวบรวมข้อมูลจาก 43 หลักสูตร ใน 11 ภาคการศึกษา ระหว่างปี ค.ศ. 2002 - 2007 โดยการใช้หลักการทางสถิติสมการถดถอยเชิงเส้นร่วมกับโครงข่ายประสาทเทียม สร้างตัวแบบสำหรับวิเคราะห์ความพึงพอใจ และทำนายหลักสูตรของนักศึกษา ผลของการศึกษาพบว่า เครือข่ายประสาทเทียมให้ผลลัพธ์การพยากรณ์ดีกว่าสมการถดถอยเชิงเส้น

จากงานวิจัยข้างต้น เป็นการศึกษาที่เกี่ยวข้องกับการศึกษา และมีการนำเทคนิคเหมืองข้อมูลมาประยุกต์ใช้ เช่น เทคนิคเนอีฟเบย์ เทคนิคเคเนียเรสเนเบอร์ เทคนิคต้นไม้ตัดสินใจ และเทคนิคโครงข่ายประสาทเทียม ผลลัพธ์ที่ได้เป็นที่น่าพอใจ แต่ในงานวิจัยเหล่านี้ ไม่ได้ทำการพยากรณ์ผลสัมฤทธิ์ทางการเรียนของนิสิต ผู้วิจัยจึงมีความคิดที่จะนำเทคนิค 4 เทคนิค ได้แก่ เทคนิคเนอีฟเบย์ เทคนิคเคเนียเรสเนเบอร์ เทคนิคต้นไม้ตัดสินใจ และเทคนิคโครงข่ายประสาทเทียมแบบแพร่กลับ มาพยากรณ์ผลสัมฤทธิ์ทางการเรียนด้วยเทคนิคเหมืองข้อมูล โดยใช้ข้อมูลภูมิหลัง และเน้นศึกษาจากข้อมูลผลการเรียนในช่วงชั้นปีที่ 1 และ 2 เป็นหลัก เพื่อประโยชน์ในการปรับแก้ไขการจัดการศึกษา และทำให้นิสิตสามารถปรับปรุงผลการเรียนของตนเองให้ดีขึ้น