บทที่ 3
แบบจำลองการคำนวณ

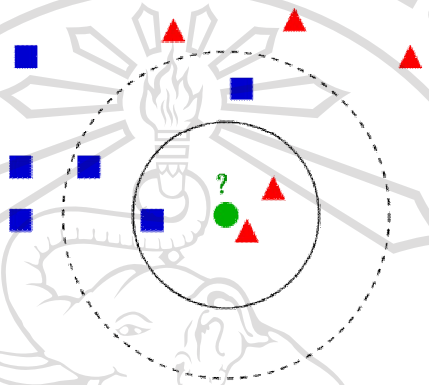
ในบทนี้จะเป็นการกล่าวถึงหลักการและสาระสำคัญในกระบวนการรวมถึงการเรียนรู้ด้วยเครื่องโดยวิธีซัพพอร์ตเวกเตอร์แมชชีนเพื่อเป็นเครื่องมือในการจำแนกข้อมูลสายลำดับ ซึ่งจะเป็นการเปรียบเทียบความคล้ายหรือความเหมือนของข้อมูลในรูปแบบรหัสพันธุกรรมในรูปดีเอ็นเอ โดยในบทที่ 2 ได้วิเคราะห์ลักษณะและโครงสร้างที่สำคัญของข้อมูลสายลำดับ เพื่อนำไปสู่การกำหนดคุณลักษณะให้แก่วิธีซัพพอร์ตเวกเตอร์แมชชีน ซึ่งจะถูกใช้เป็นฐานความรู้ในการรู้จำด้วยเครื่อง และนำไปสู่กระบวนการและวิธีดำเนินการในการคำนวณของแบบจำลองต่างๆ จากนั้นจะเป็นขั้นตอนของการกำหนดรูปแบบการจำแนกชนิดข้อมูลสาบลำดับของวิธีซัพพอร์ตเวกเตอร์-แมชชีนที่จะใช้ในงานวิจัยนี้ ต่อจากนั้นจะอธิบายถึงวิธีการของสเปกตรัมเคอร์เนล เพื่อใช้ในการคำนวณการจำแนกชนิดข้อมูลของโมเลกุลเอชแอลเอ จากนั้นจะเป็นการนิยามหลักการในการจำแนกชนิดข้อมูลของปัญหานี้ เพื่อให้สามารถระบุถึงปัญหาได้อย่างชัดเจน สุดท้ายจะเป็นการนำเสนอขั้นตอนการใช้สเปกตรัมเคอร์เนลในการวัดความคล้ายหรือความเหมือนของข้อมูลสายลำดับ และจะนำเสนอขั้นตอนวิธีการแปลงตัวข้อมูลนำเข้า ซึ่งทำให้เครื่องสามารถเรียนรู้ได้ในลำดับต่อไป ดังนั้นในบทนี้จึงได้แบ่งเนื้อหาออกเป็นหลายส่วนดังนี้

- เค-เนียร์เรสต์เนเบอร์ (K-Nearest Neighbor)
- โครงข่ายประสาทเทียม (Artificial Neural Network)
- ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machines)
    - เคอร์เนลฟังก์ชัน (Kernel function)
    - สตริงเคอร์เนล (String kernel)
    - สเปกตรัมเคอร์เนล (Spectrum kernel)

3.1 เค-เนียร์เรสต์เนเบอร์ (เค-เอ็นเอ็น)

ในรูปแบบของตัวจำแนกชนิดข้อมูลด้วยวิธีการเค-เนียร์เรสต์เนเบอร์ (k -Nearest neighbor pattern classifier) เป็นวิธีการหนึ่งสำหรับแก้ปัญหาแบบประมาณค่าไม่อิงพารามิเตอร์ (Non Parametric estimation) สำหรับการจำแนกชนิดข้อมูลต่างๆ ซึ่งข้อมูลส่วนใหญ่จะมีรูปร่างรูปแบบหรือการจัดเรียงที่ไม่เหมือนกันและกระจัดกระจายตามลักษณะข้อมูลแต่ละชนิดโดย

หลักการของ เค-เอ็นเอ็น (*k*-NN) นั้นจะทำการแปลงคุณลักษณะของข้อมูลเข้าให้อยู่ในรูปของเวกเตอร์ (Vector) และพิจารณาด้วยค่าระยะห่างของข้อมูลที่เราจะพิจารณา ซึ่งสามารถอธิบาย ได้ดังนี้ โดย เค (*k*) หมายถึงจำนวนของชุดข้อมูลที่เราจะพิจารณา เช่น เค=3 หมายถึง เราจะพิจารณาข้อมูลจำนวน 3 ตัวที่อยู่ใกล้เคียงจุดที่เราต้องการพิจารณา โดยส่วนใหญ่จะกำหนดใช้ค่า เค เป็นเลขคี่ ดังรูป 3.15



รูป 3.1 เค-เอ็นเอ็นที่ เค เท่ากับ 3

(ที่มา: http://en.wikipedia.org/wiki/File:KnnClassification.svg)

ซึ่งสมการของเค-เอ็นเอ็นสามารถอธิบายได้ดังสมการนี้

$$F_n\left(\frac{x}{w_i}\right) = \frac{P(x\ Fall\ in\ R(x) with\ volume\ V_n)}{V_n} \tag{3.1}$$

$$F_n\left(\frac{x}{w_i}\right) = \frac{k_n/n}{V_n} \tag{3.2}$$

$$k_n = \sqrt{n} \tag{3.3}$$

$$F_n\left(\frac{x}{w_i}\right) = \frac{1/\sqrt{n}}{V_n} \approx f(x) \tag{3.4}$$

$$V_n = \frac{1}{\sqrt{n} f(x)} \tag{3.5}$$

โดย k_n คือจำนวนของชุดข้อมูลที่ต้องการพิจารณาในแต่ละหน้าต่าง

n คือจำนวนข้อมูลทั้งหมด

V_n คือขนาดของหน้าต่างที่ทำการพิจาณา

ถ้าแปลงสมการข้างต้นให้อยู่ในรูปของเบย์ (Bay's) จะได้สมการดังต่อไปนี้

$$P\left(\frac{w_n}{x}\right) = \frac{F(x/w_i)}{F(x)} = \frac{F(x/w_i)}{\sum_{j=1}^{c} F(x/w_i)} \tag{3.6}$$

$$= \frac{(k_i/n)/V}{\sum_{j=1}^{c}(k_i/n)/V} \tag{3.7}$$

$$= \frac{(k_i/n)/V}{\sum_{j=1}^{c}(k_i/n)/V} \tag{3.8}$$

$$= \frac{k_i}{\sum_{j=1}^{c} k_i} \tag{3.9}$$

$$P\left(\frac{w_n}{x}\right) = \frac{k_i}{k} \tag{3.10}$$

โดย k_i คือจำนวนของชุดข้อมูลทั้งหมดของกลุ่มที่ *i*

k คือจำนวนของชุดข้อมูลทั้งหมดของทุกๆ กลุ่ม

(Cover TM Hart PE, 1967)

3.1.1 ตัวจำแนกชนิดข้อมูลด้วยวิธีการเค-เนียร์เรสต์เนเบอร์ (K-Nearest neighbor classifier)

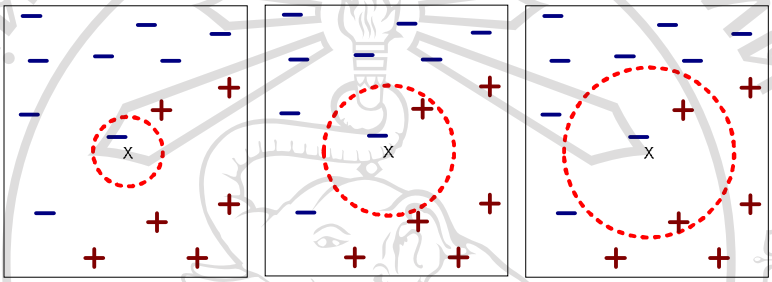
การจำแนกชนิดข้อมูลด้วยวิธีเค-เนียร์เรสต์เนเบอร์นั้น เป็นอัลกอริทึมที่ใช้ในการจัดกลุ่มข้อมูล ซึ่งเป็นอัลกอริทึมที่อยู่ในกลุ่มของการเรียนรู้แบบมีผู้สอน โดยข้อมูลที่นำมาเรียนรู้จะต้องมีการระบุหน้าที่หรือชนิดของข้อมูลแต่ละตัว (Label) ดังนั้นจึงเป็นการจัดกลุ่มข้อมูลที่อยู่ใกล้กันเข้าไว้ด้วยกัน โดยทั่วไปวิธีเค-เนียร์เรสต์เนเบอร์นั้นจะมีการคำนวณอิงกับระยะห่าง (Distance) ที่มีชื่อว่า ยูคลิเดียน (Euclidian) ซึ่งตัวระยะห่างนี้เป็นตัวหลักที่มีประสิทธิภาพในการจำแนกชนิดข้อมูล โดยขั้นตอนการจำแนกนั้นจะมีอยู่ 2 ขั้นตอนได้แก่

ขั้นตอนที่ 1

การใช้กฎของเนียร์เลส เนชเบอร์ (Nearest Neighbor rule) คือการหาจุดทดสอบข้อมูล (Test data point) ที่ใกล้ที่สุดของข้อมูลทั้งหมด ซึ่งจะมีการทำการฝึกสอนเครื่อง โดยใช้ข้อมูลฝึกสอนนั้นเป็นตัวฝึกสอนให้เครื่องรู้จำรูปแบบหรือกลุ่มของข้อมูลแต่ละชนิด ที่เราจะทำการทำนายหรือคาดการณ์ได้ว่าอยู่ในกลุ่มข้อมูลชนิดใด

ขั้นตอนที่ 2

เค-เนียร์เลส เนชเบอร์ (*k*-Nearest Neighbor) คือการกำหนดขนาดของเค โดยพิจารณาจากการกำหนดจุดทดสอบข้อมูลจากขั้นตอนที่ 1 เพื่อหาข้อมูลที่ใกล้มากที่สุด จากนั้นจึง เอาจุดทดสอบข้อมูลมาทดสอบในกลุ่มข้อมูลที่เราจะทำนาย ถ้าค่าระยะห่าง (Sum distance) ของจุดทดสอบข้อมูลที่เรากำหนดไว้มีค่าความห่างน้อย ก็แสดงได้ว่าข้อมูลที่เราสนใจนั้น ควรจะอยู่ในกลุ่มที่เรากำลังพิจารณา ดังรูป 3.2



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

รูป 3.2 ขนาดของเค-เนียร์เลส เนชเบอร์

(ที่มา: http://en.wikipedia.org/wiki/File:KnnClassification.svg)

จากรูปข้างต้นจะแสดงให้เห็นถึงขนาดเค เท่ากับ 1,2 และ 3 และ x คือจุดทดสอบข้อมูล จากนั้นจะสร้างหน้าต่าง (Window) ขนาดที่ครอบคลุมข้อมูลจำนวน 1 ตัว, 2 ตัว, 3 ตัว ตามลำดับ จากนั้นจะพิจารณาว่าข้อมูลที่เราพิจารณานั้นอยู่ในกลุ่มข้อมูลใด เช่น 1-เนียร์เลส เนชเบอร์จะอยู่ในกลุ่ม “–” ส่วน 2-เนียร์เลส เนชเบอร์จะอยู่ในกลุ่ม “–” แต่ถ้ามีจำนวนเท่ากัน ก็จะวัดระยะห่างของข้อมูลว่าข้อมูลตัวใดใกล้ที่สุด ส่วน 3- เนียร์เลส เนชเบอร์ก็จะอยู่ในกลุ่ม “+” เป็นต้น

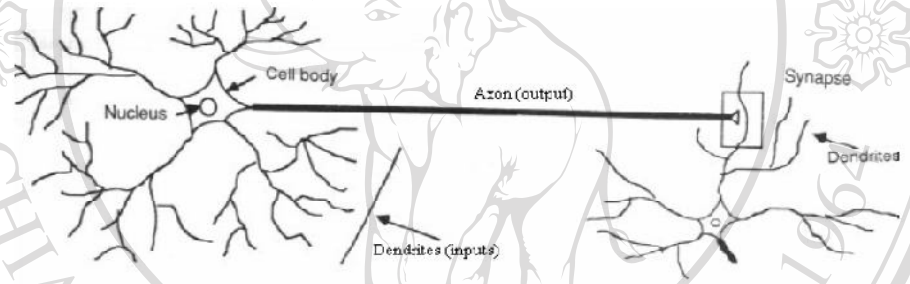
(Shakhnarovish et al., 2005)

3.2 โครงข่ายประสาท

โครงข่ายประสาทหรือข่ายงานประสาท (Neural network) คือแบบจำลองทางคณิตศาสตร์สำหรับประมวลผลข้อมูลสารสนเทศด้วยการคำนวณแบบคอนเนคชันนิสต์ (Connectionist) เพื่อจำลองการทำงานของเครือข่ายประสาทในสมองมนุษย์ ด้วยวัตถุประสงค์ คือสร้างเครื่องมือ ซึ่งมีความสามารถในการเรียนรู้และจดจำรูปแบบ (Pattern Recognition) และการอุปมานความรู้ (Knowledge deduction) เช่นเดียวกับความสามารถที่มีในสมองมนุษย์ แนวคิดเริ่มต้นของเทคนิคนี้ได้มาจากการศึกษาข่ายงานไฟฟ้าชีวภาพ (bioelectric network) ในสมอง ซึ่งประกอบด้วย เซลล์

ประสาท หรือ นิวรอน (neurons) และ จุดประสานประสาท (synapses) แต่ละเซลล์ประสาท ซึ่งประกอบด้วยปลายประสาทในการรับกระแสประสาท เรียกว่า เดนไดรท์ (Dendrite) ซึ่งเป็นข้อมูลเข้า (Input) และปลายในการส่งกระแสประสาทเรียกว่า แอคซอน (Axon) ซึ่งเป็นเหมือนข้อมูลออก (Output) ของเซลล์ ซึ่งเซลล์เหล่านี้จะทำงานด้วยปฏิกิริยาไฟฟ้าเคมี เมื่อมีการกระตุ้นด้วยสิ่งเร้าภายนอกหรือกระตุ้นด้วยเซลล์ด้วยกัน กระแสประสาทจะวิ่งผ่านเดนไดรท์เข้าสู่นิวเคลียสซึ่งจะเป็นตัวตัดสินว่าต้องกระตุ้นเซลล์อื่น ๆ หรือไม่ ถ้ากระแสประสาทแรงพอ นิวเคลียสก็จะกระตุ้นเซลล์อื่น ๆ ต่อไปผ่านทางแอคซอน

ตามโมเดลนี้ข่ายงานประสาทเกิดจากการเชื่อมต่อระหว่างเซลล์ประสาท จนเป็นเครือข่ายที่ทำงานร่วมกัน ดังรูป 3.3

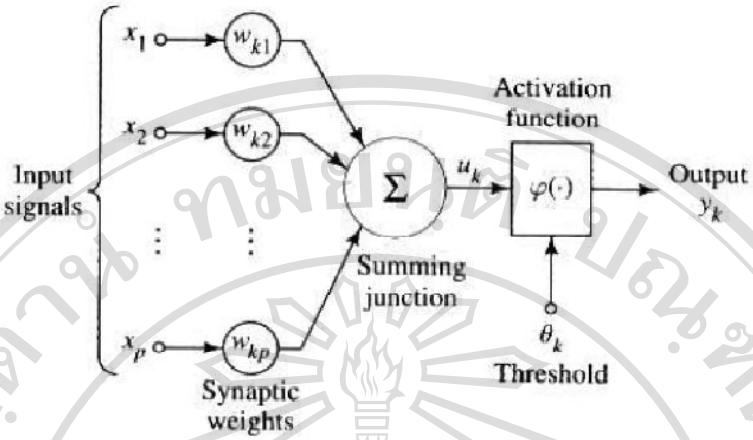


รูป 3.3 แบบจำลองของประสาทในสมองมนุษย์

(ที่มา: Du A., 2007)

3.2.1 โครงสร้างของโครงข่ายประสาท

โครงข่ายประสาทจะประกอบไปด้วยหน่วยประมวลผล เรียกว่านิวรอน (Neuron) ที่เชื่อมถึงกันหลายๆ ตัว ผ่านการเชื่อมต่อที่เรียกว่าลิงค์ (Link) แต่ละลิงค์จะมีค่าน้ำหนัก (Weight) กำกับอยู่ แต่ละนิวรอนจะมีค่าสถานะภายในที่เรียกว่า ระดับการกระตุ้น (Activity level) โดยสัญญาณข้อมูลออกของนิวรอนแรกที่ส่งไปยังนิวรอนถัดไปจะได้มาจากการพิจารณาผลรวมทั้งหมดของค่าน้ำหนักและคูณกับค่าระดับการกระตุ้น ดังแสดงในรูป 3.4



รูป 3.4 นิวรอนของโครงข่ายประสาทเทียม

(ที่มา: Lawrence et al., 1994)

จากรูป 3.4 จะเห็นว่าสัญญาณข้อมูลเข้าคือ $x_1, x_2, ..., x_n$ ถูกป้อนให้กับนิวรอน Y ซึ่งสัญญาณข้อมูลเข้านี้จะถูกนำไปคูณเข้ากับค่าน้ำหนักคือ $w_1, w_2, ..., w_n$ เมื่อรวมค่าที่ได้นำไปเปรียบเทียบกับค่าระดับ (Threshold) โดยอิงกับฟังก์ชันกระตุ้น (Activation function) ก็จะได้เป็นสัญญาณข้อมูลออกของนิวรอนออกมา

ค่าผลรวมของสัญญาณข้อมูลเข้าคูณกับน้ำหนักได้จากสมการ

$$Y = X_1W_1 + X_2W_2 + ... + X_nW_n \tag{3.11}$$

ค่าสัญญาณข้อมูลออกของนิวรอน Y ได้จากสมการ

$$Out = (Y) \tag{3.12}$$

โดยที่ f คือฟังก์ชันกระตุ้น ซึ่งกำหนดได้ดังนี้

$$f(Y) = \frac{1}{1+e^{-Y}} \tag{3.13}$$

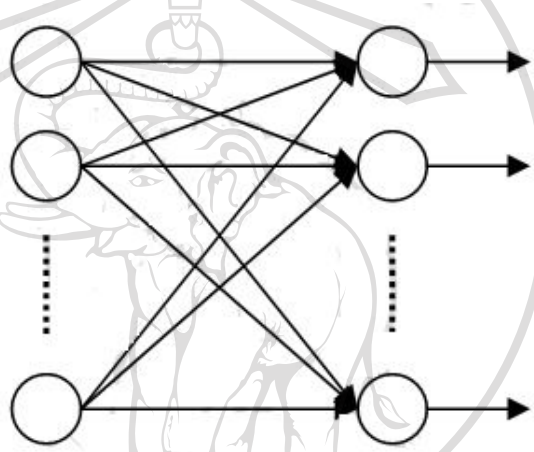
(จุพารัตน์ ตันประเสริฐ และ ธิติพงศ์ ตันประเสริฐ, 2541)

3.2.2 สถาปัตยกรรมโครงข่ายประสาท

สถาปัตยกรรมโครงข่ายประสาทสามารถจำแนกตามโครงสร้างของโครงข่ายได้ 2 ประเภท ดังนี้

1) โครงข่ายประสาทแบบชั้นเดียว (Single layer Neural Network)

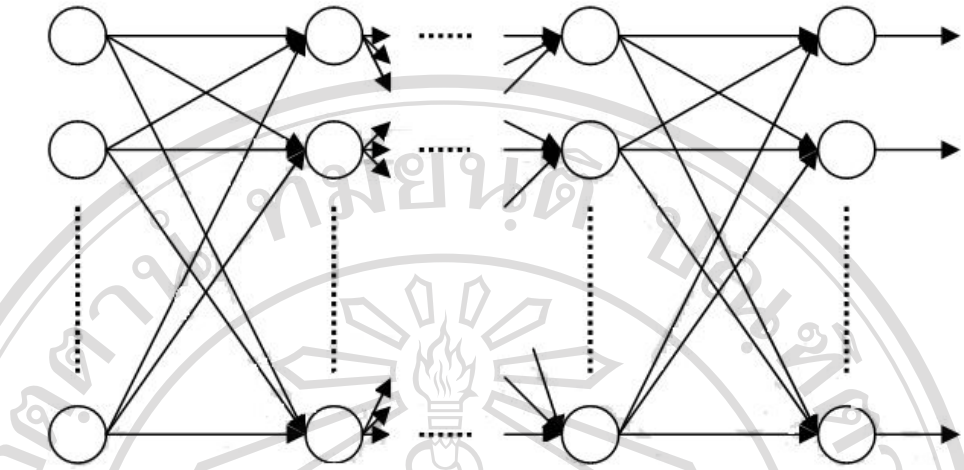
สถาปัตยกรรมโครงข่ายประสาทชนิดนี้ จะมีข้อมูลเข้าของนิวรอนเชื่อมต่อกับข้อมูลออกของนิวรอน และมีค่าน้ำหนักเป็นตัวปรับระดับสัญญาณข้อมูลเข้าเพียงชั้นเดียว ซึ่งค่าน้ำหนักแต่ละค่าจะเป็นอิสระต่อกัน ไม่ส่งผลกระทบต่อการปรับค่าน้ำหนักตัวอื่น จึงมักใช้กับการประมวลผลที่ไม่ซับซ้อนมากนัก โดยลักษณะของโครงข่ายสามารถแสดงได้ดังรูป 3.5



รูป 3.5 โครงข่ายประสาทแบบชั้นเดียว
(ที่มา: สมรัฐ แถลงการณ์, 2540)

2) โครงข่ายประสาทแบบหลายชั้น (Multilayer Neural Network)

สถาปัตยกรรมโครงข่ายประสาทชนิดนี้ จะมีชั้นของนิวรอนคั่นอยู่ระหว่างชั้นข้อมูลเข้าและชั้นข้อมูลออกซึ่งอาจจะมีหนึ่งหรือหลายชั้นก็ได้ ชั้นที่คั่นอยู่เรียกว่าชั้นซ่อน (Hidden layer) ดังนั้นจำนวนของค่าน้ำหนักก็จะมีมากขึ้นตามจำนวนชั้นซ่อนที่เพิ่มเข้าไป โดยทั่วไป โครงข่ายแบบหลายชั้นสามารถแก้ปัญหาที่มีความซับซ้อนได้ดีกว่าโครงข่ายแบบชั้นเดียว แต่การเรียนรู้จะยากกว่า โดยลักษณะโครงข่ายแสดงได้ดังรูป 3.6



รูป 3.6 โครงข่ายประสาทแบบหลายชั้น
(ที่มา: สมรัฐ แถลงการณ์, 2540)

3.2.3 การเรียนรู้ของโครงข่ายประสาท

การเรียนรู้ของโครงข่ายประสาท คือการหาค่าน้ำหนักที่เหมาะสมให้กับโครงข่าย เราสามารถสอนให้โครงข่ายเกิดการเรียนรู้ได้ โดยการป้อนข้อมูลตัวอย่างให้กับเซลล์แต่ละเซลล์ของโครงข่าย ซึ่งแต่ละโครงข่ายจะมีวิธีการปรับเปลี่ยนค่าน้ำหนักภายในให้สอดคล้องกับข้อมูลตัวอย่างและให้ผลลัพธ์ที่ถูกต้องออกมา จากการศึกษางานวิจัยที่เกี่ยวข้องพบว่าการให้ตัวอย่างสำหรับฝึกสอนโครงข่ายมากๆ โครงข่ายก็จะมีความแม่นยำสูง แต่ก็ใช้เวลาในการฝึกสอนเพิ่มมากขึ้นเช่นกัน

ดังนั้นการเรียนรู้ของโครงข่ายจะมีประสิทธิภาพมากเพียงใด ย่อมขึ้นอยู่กับค่าน้ำหนักของโครงข่ายที่ได้ทำการฝึกสอนแล้ว กระบวนการฝึกสอนโครงข่ายสามารถจำแนกได้เป็น 2 กลุ่มใหญ่

1) การฝึกสอนแบบควบคุม

การฝึกสอนแบบนี้จะมีการกำหนดข้อมูลอินพุทที่สอดคล้องกับเป้าหมายที่ถูกต้องให้กับโครงข่าย แล้วให้โครงข่ายทำการเรียนรู้ข้อมูลออกที่คำนวณได้แต่ละครั้งจะนำมาเปรียบเทียบกับเป้าหมายที่กำหนดไว้ ค่าผิดพลาดที่เกิดระหว่างข้อมูลออกกับเป้าหมายจะถูกนำกลับไปพิจารณาปรับค่าน้ำหนักในรอบต่อๆไป โครงข่ายจะทำการเรียนรู้จนได้ข้อมูลออกตรงตามเป้าหมายที่กำหนดหรือมีค่าความผิดพลาดน้อยพอที่จะยอมรับได้ สำหรับโครงข่ายแบบหลายชั้นนั้น

การเรียนรู้ที่เหมาะสมคือกระบวนการเรียนรู้แบบแพร่กลับ (Back propagation) และแบบเพอร์เซ็พตรอน (Perceptron)

2) การฝึกสอนแบบอิสระ

การฝึกสอนแบบนี้จะมีการกำหนดเพียงข้อมูลเข้าให้กับโครงข่ายแต่ไม่กำหนดเป้าหมายให้ โครงข่ายต้องหาวิธีการเรียนรู้ด้วยตนเองและปรับค่าน้ำหนักด้วยการจัดกลุ่มของข้อมูลออกที่มีลักษณะเดียวกันไว้ด้วยกัน

(สมรัฐ แถลงการณ์, 2540)

3.3 ซัพพอร์ตเวกเตอร์แมชชีน (SVM)

ซัพพอร์ตเวกเตอร์แมชชีน คือ โครงข่ายประสาทเทียมรูปแบบใหม่ที่สร้างขึ้นเพื่อแก้ไขปัญหาของโครงข่ายประสาทแบบดั้งเดิม ซึ่งซัพพอร์ตเวกเตอร์แมชชีนจะอาศัยพื้นฐานความรู้มาจากทฤษฎีการเรียนรู้ทางสถิติ และกระบวนการลดความเสี่ยงเชิงโครงสร้างให้ต่ำที่สุด (Statistical Learning Theory and Structure Risk Minimization Principle) มาใช้เป็นกระบวนการเลือกแบบจำลองที่เหมาะสมที่สุดโดยเป็นรูปแบบในการเรียนรู้ ซึ่งก็จะทำให้ได้คำตอบที่เป็นค่าที่เหมาะสมที่สุดของระบบ ด้วยเหตุนี้ซัพพอร์ตเวกเตอร์แมชชีน จึงมีข้อดีที่เหนือกว่าโครงข่ายประสาท ในด้านของโครงสร้างและการจัดการกับข้อมูลออกหรือผลลัพธ์ทำให้ซัพพอร์ตเวกเตอร์แมชชีนจึงเป็นที่นิยมและเริ่มนำไปใช้งานด้านการจดจำรูปแบบซึ่งจะเลือกใช้ซัพพอร์ตเวกเตอร์แมชชีนแบบแบ่งกลุ่ม (Support Vector Classification) และงานด้านการประมาณค่าฟังก์ชัน ซึ่งจะเลือกใช้ซัพพอร์ตเวกเตอร์แมชชีนแบบถดถอย (Support Vector Regression) ในการจัดการกับปัญหาการจำแนกชนิดข้อมูล เป็นต้น

3.3.1 ซัพพอร์ตเวกเตอร์แมชชีนสำหรับการแบ่งกลุ่ม (Support Vector Classification)

ซัพพอร์ตเวกเตอร์แมชชีนสำหรับการแบ่งกลุ่มนั้นได้ถูกพัฒนาขึ้นเพื่อใช้งานทางด้านการแบ่งกลุ่มข้อมูล ซึ่งซัพพอร์ตเวกเตอร์แมชชีนจะใช้ระนาบเกินที่เหมาะสมที่สุด (Optimal Hyperplane) ในการแบ่งกลุ่มเพื่อเพิ่มสมรรถนะให้สามารถใช้งานได้ดีขึ้น ซึ่งในการสร้างระนาบเกินที่ใช้ในการแบ่งกลุ่มข้อมูลสามารถสร้างได้หลายแบบด้วยกัน แต่จะมีระนาบเกินที่เหมาะสมที่สุดเพียงระนาบเดียวเท่านั้น ที่สามารถรักษาระยะห่างมากที่สุดระหว่างข้อมูล 2 กลุ่มที่ใกล้กันมากที่สุดได้

เมื่อเราพิจารณาข้อมูลที่ประกอบด้วยข้อมูล 2 กลุ่มดังสมการที่ 3.14

$$D = \{(x_1, y_1), \dots, (x_l, y_l)\}, x_i \in \mathbf{R}^n, y_i \in \{-1, 1\} \quad (3.14)$$

เมื่อ l คือ จำนวนของข้อมูล

i คือ ข้อมูลลำดับที่ $1,2,\ldots,l$

โดยจะแบ่งข้อมูล x ออกเป็น 2 กลุ่มข้อมูลภายใต้เงื่อนไขคือ

$$y_i = \begin{cases} 1 : x_i \text{ is a "+"point} \\ -1 : x_i \text{ is a " }-\text{ "point} \end{cases} \tag{3.15}$$

เมื่อพิจารณาชุดของกลุ่มข้อมูล x โดยกำหนดให้กลุ่มข้อมูล $x1$ เป็นข้อมูล x_i ที่มีค่าเป็นบวกและ $x2$ เป็นข้อมูล x_i ที่มีค่าเป็นลบ และกำหนดให้ γ คือค่าที่มากกว่า 0 $(\gamma > 0)$ แล้วจะสามารถเขียนสมการระนาบเกินได้ดังสมการที่ 3.16

$$\left.\begin{aligned} \langle w, x1\rangle + b &= \gamma \\ \langle w, x2\rangle + b &= -\gamma \end{aligned}\right\} \tag{3.16}$$

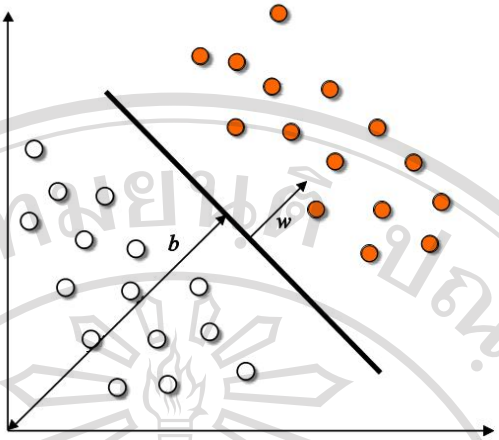
เมื่อ w คือเวกเตอร์น้ำหนัก

$x1$ คือเวกเตอร์ข้อมูลที่มีค่าเป็นบวก

$x2$ คือเวกเตอร์ข้อมูลที่มีค่าเป็นลบ

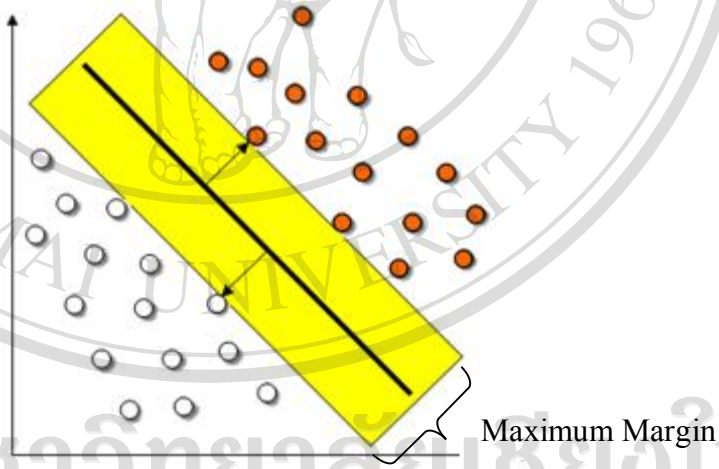
b คือค่าไบอัส

ในการหาระนาบเกินที่เหมาะสมที่สุดของซัพพอร์ตเวกเตอร์แมชชีนนั้น โดยซัพพอร์ตเวกเตอร์แมชชีนจะทำการหาตำแหน่งของซัพพอร์ตเวกเตอร์ (Support Vector) เพื่อใช้เป็นตัวแทนของกลุ่มข้อมูลทั้งชุด เพื่อใช้เป็นข้อมูลในการแบ่งกลุ่มโดยหลักการ คือ จะใช้ระนาบเกินที่เป็นระยะห่างมากที่สุดระหว่างข้อมูล 2 กลุ่ม ที่อยู่ใกล้กันมากที่สุด เพียงระนาบเดียวเท่านั้น ซึ่งระนาบเกินนั้นสามารถทำการแบ่งกลุ่มของข้อมูลที่อยู่ในรูปของเวกเตอร์และใช้ระนาบเกินนี้ในการพิจารณากลุ่มของข้อมูลที่เราสนใจ



รูป 3.7 กระบวนการในการหาระนาบเกินที่เหมาะสมในการแบ่งข้อมูล
(ที่มา: สุชาวดี ตันติศักดิ์, 2550)

ในทางทฤษฎี กล่าวว่าจะต้องไม่มีข้อมูลเกินเข้ามาในระหว่างขอบระนาบทั้งสอง จากนั้นจึงหาระนาบที่รักษาระยะห่างจากขอบมากที่สุด (Maximum Margin) และถือว่าระนาบดังกล่าวคือระนาบเกินสำหรับการแบ่งกลุ่มที่เหมาะที่สุด แสดงดังรูป 3.8



รูป 3.8 ระนาบเกินที่เหมาะสมที่สุดซึ่งระยะห่างจากระยะขอบมากที่สุด
(ที่มา: สุชาวดี ตันติศักดิ์, 2550)

ซึ่งเมื่อกำหนดเงื่อนไขของสมการระนาบเกินดังสมการที่ 3.17

$$\min_i |\langle w, x_i \rangle + b| = 1 \qquad (3.17)$$

ซึ่งทำให้สามารถวิเคราะห์สมการของระนาบเกินในการแบ่งข้อมูลได้ตามสมการที่ 3.18

$$\langle w, x_i \rangle + b \begin{cases} \geq 1 \; ; \; y_i = +1 \\ \leq 1 \; ; \; y_i = -1 \end{cases} \tag{3.18}$$

และสามารถหาค่าระยะห่าง (Distance; $d(w, b; x)$) จากตำแหน่งของซัพพอร์ตเวกเตอร์ x ถึงระนาบ (w, b) ได้ดังสมการ 3.19

$$d(w, b; x) = \frac{|\langle w, x_i \rangle + b|}{\|w\|} \tag{3.19}$$

โดยการหาระนาบเกินที่ดีที่สุดนั้น วิเคราะห์ได้จากค่าของระยะขอบที่มากที่สุด (Maximum Margin) ซึ่งค่าของระยะขอบนั้น สามารถหาได้ตามสมการ 3.20

$$\begin{aligned} \rho(w, b) &= \min_{x_i : y_i = -1} d(w, b; x_i) + \min_{x_i : y_i = +1} d(w, b; x_i) \\ &= \min_{x_i : y_i = -1} \frac{|\langle w, x_i \rangle + b|}{\|w\|} + \min_{x_i : y_i = +1} \frac{|\langle w, x_i \rangle + b|}{\|w\|} \\ &= \frac{1}{\|w\|} \Big(\min_{x^i : y^i = -1} |\langle w, x_i \rangle + b| + \min_{x^i : y^i = +1} |\langle w, x_i \rangle + b| \Big) \\ &= \frac{2}{\|w\|} \end{aligned} \tag{3.20}$$

เมื่อ ρ คือระยะขอบห่างที่มากที่สุด โดยเมื่อพิจารณาค่าเป็นไปได้ของค่าระยะขอบที่มากที่สุด ที่สามารถเป็นไปได้นั้น จะพิจารณาได้จาก ค่าที่น้อยที่สุดของ $\frac{2}{\|w\|}$ ดังนั้นค่าของระนาบเกินที่เหมาะสมที่สุดจึงสามารถพิจารณาได้จากค่าที่น้อยที่สุดในสมการที่ 3.21

$$\Phi(w) = \frac{1}{2} \|w\|^2 \tag{3.21}$$

เมื่อ $\Phi(w)$ คือ ฟังก์ชันที่ใช้ในพิจารณาหาค่าน้อยที่สุดเพื่อหาระนาบเกินที่เหมาะสมที่สุด ซึ่งอยู่ภายใต้เงื่อนไข $y_i(\langle w, x_i \rangle + b) \geq 1 \; ; i = 1,2, \ldots, l$ โดยเมื่อพิจารณาจากสมการที่ 3.21 แล้วพบว่าตัวแปร b หรือค่าไบอัสนั้น ไม่มีผลต่อสมการระนาบเกินดังแสดงในสมการข้างต้น ซึ่งพบว่าค่าของระนาบเกินที่เหมาะที่สุด ที่ได้แปรผันกับค่าของระยะขอบเท่านั้น แม้ว่าระยะของระนาบเกินที่แบ่งกลุ่มข้อมูลทั้งสองกลุ่มมีการเปลี่ยนแปลง ซึ่งอาจมีระยะขอบที่มีค่าคงที่แต่ระยะ

จากระนาบเกินถึงชุดของกลุ่มข้อมูลมีระยะห่างจากแต่ละชุดของกลุ่มข้อมูลที่ไม่เท่ากัน ซึ่งมีผลให้ระนาบเกินที่ใช้ไม่เป็นค่าของระนาบเกินที่เหมาะสมที่สุดอีกต่อไป ดังนั้นเพื่อแก้ไขปัญหาดังกล่าว จึงได้มีการกำหนดค่าระยะห่างจากกลุ่ม โดยแสดงไว้ดังสมการที่ 3.22

$$\|w\| < A \tag{3.22}$$

โดยเมื่อพิจารณาสมการที่ 3.17-3.19 จะทำให้ได้สมการที่ 3.23

$$d(w, b; x) = \frac{1}{A} \tag{3.23}$$

เมื่อพิจารณาเงื่อนไขดังกล่าว พบว่าระนาบเกินที่สร้างขึ้นนั้นจะไม่สามารถเข้าใกล้กลุ่มข้อมูลได้มากเกินกว่าค่าที่กำหนดไว้ดังสมการที่ 3.22

ในการหาค่าของ w และ b สำหรับสมการแบบหลายตัวแปร (Dual Problem) นั้น จะใช้หลักการสร้างฟังก์ชันลากรานจ์ (Lagrange Function) ซึ่งมีรูปแบบดังแสดงไว้ในสมการที่ 3.24

$$\boldsymbol{\mathcal{L}}(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} \alpha_i (y_i[\langle w, x_i \rangle + b] - 1) \tag{3.24}$$

โดยที่ $\boldsymbol{\mathcal{L}}$ คือ ฟังก์ชันลากรานจ์ (Lagrange Function)

α_i คือ ตัวคูณลากรานจ์ (Lagrange Multipliers)

ภายใต้เงื่อนไขการสร้างระนาบเกินที่เหมาะสมที่สุดนั้น ต้องพิจารณาค่าที่น้อยที่สุดของฟังก์ชันลากรานจ์ข้างต้น ซึ่งการหาค่าที่น้อยที่สุดของฟังก์ชันลากรานจ์นั้น จะพิจารณาเทอมของ w, b เป็นค่าน้อยที่สุดและเทอมของตัวคูณลากรานจ์ (α) จะพิจารณาเป็นค่าที่มากที่สุด โดยจะสามารถเขียนสมการการหาระนาบเกินที่เหมาะสมที่สุดภายใต้เงื่อนไขดังกล่าวได้ด้วยสมการที่ 3.25

$$\max_{\alpha} W(\alpha) = \max_{\alpha}\left(\min_{w,b} \boldsymbol{\mathcal{L}}(w, b, \alpha)\right) \tag{3.25}$$

ค่าน้อยที่สุดในเทอมของ w และ b ในสมการสามารถหาได้จากสมการที่ 3.26 และ 3.27 โดยค่าที่น้อยที่สุดในเทอมของ b สามารถหาได้จาก สมการ 3.26

$$\frac{\partial \boldsymbol{\mathcal{L}}}{\partial b} = 0 \Rightarrow \sum_{i=1}^{l} \alpha_i y_i = 0 \tag{3.26}$$

โดยค่าที่น้อยที่สุดในเทอมของ w สามารถหาได้จากสมการ 3.27

$$\frac{\partial \mathcal{L}}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{l} \alpha_i y_i \, x_i \tag{3.27}$$

จากเงื่อนไขของสมการที่ 3.24-3.27 จะเมื่อแทนค่าของ w ในสมการที่ 3.24 จะสามารถเขียนสมการได้คือ

$$\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle - \sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + b \sum_{i=1}^{l} \alpha_i y_i + \sum_{i=1}^{l} \alpha_i$$

เมื่อ $\sum_{i=1}^{l} \alpha_i y_i = 0$ จะได้

$$\max_{\alpha} W(\alpha) = \max_{\alpha} \left(-\frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_{i=1}^{l} \alpha_i \right) \tag{3.28}$$

จากสมการที่ 3.28 จะสามารถหาตัวคูณลากรานจ์สำหรับระนาบเกินที่เหมาะสมที่สุดได้จากสมการที่ 3.29

$$\alpha^* = \arg\min_{\alpha} \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle - \sum_{i=1}^{l} \alpha_i \tag{3.29}$$

เมื่อ α^* คือ ตัวคูณลากรานจ์สำหรับระนาบเกินที่เหมาะสมที่สุด ภายใต้เงื่อนไข

$$\left.\begin{aligned} \alpha_i \geq 0 \, ; i = 1,2, \ldots, l \\ \sum_{i=1}^{l} \alpha_i y_i = 0 \end{aligned}\right\} \tag{3.30}$$

และสามารถหาค่า w^* และ b^* ของระนาบเกินที่เหมาะที่สุด ได้จาก

$$\frac{\partial}{\partial w}\left[\frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} \alpha_i^* (y_i[\langle w, x_i \rangle + b] - 1)\right] = 0 \tag{3.31}$$

$$w - \sum_{i=1}^{l} \alpha_i^* y_i \, x_i = 0 \tag{3.32}$$

$$\sum_{i=1}^{l} \alpha_i^* y_i \, x_i = w^* \tag{3.33}$$

เมื่อ w^* คือ ค่าน้ำหนักสำหรับสร้างระนาบเกินที่เหมาะสมที่สุด

จากเงื่อนไขของ Kuhn-Tucker จะได้ความสัมพันธ์ α^*, w^*, b^* มีความสัมพันธ์กันคือ

$$\left.\begin{aligned} \frac{\partial}{\partial w}\left[\frac{1}{2}\langle w^*, w^*\rangle - \sum_{i=1}^{l}\alpha_i^*(y_i[\langle w, x_i\rangle + b] - 1)\right] = 0 \\ \frac{\partial}{\partial b}\left[\frac{1}{2}\langle w^*, w^*\rangle - \sum_{i=1}^{l}\alpha_i^*(y_i[\langle w, x_i\rangle + b] - 1)\right] = 0 \end{aligned}\right\} \quad (3.34)$$

เมื่อ $\alpha^* = 0$ หรือ $y_i[\langle w, x_i\rangle + b] - 1 = 0$

จากสมการที่ 3.30-3.31 ได้กำหนดเงื่อนไขของตัวคูณลากรานจ์ไว้ว่า $\alpha_i > 0$ ดังนั้นจึงสามารถหาค่า b^* ได้จากเงื่อนไขข้างต้น คือ

$$y_i[\langle w, x_i\rangle + b] - 1 = 0$$

$$1 - \langle w^*, x_i\rangle = b^* \quad (3.35)$$

กำหนดให้ $y_i = 1$

เมื่อ b^* คือ ค่าไบอัสสำหรับสร้างระนาบเกินที่เหมาะสมที่สุด

โดยในการแบ่งกลุ่มข้อมูลจะได้

$$f(x) = sign(\langle w^*, x\rangle + b) \quad (3.36)$$

เมื่อ $sign$ คือ ค่าของ Signum function ซึ่งมีรูปแบบในการหาค่า $f(x)$ จาก x ด้วยเงื่อนไข

$$sign(x) = \begin{cases} 1 : x > 0 \\ -1 : x < 0 \end{cases} \quad (3.37)$$

หรืออาจใช้แนวทางใหม่ในการแบ่งกลุ่มข้อมูลด้วยสมการ

$$f(x) = h(\langle w^*, x\rangle + b) \quad (3.38)$$

ซึ่งจะมีความหยืดหยุ่นกว่า โดยสามารถแบ่งกลุ่มข้อมูลได้ตามเงื่อนไข

$$h(z) = \begin{cases} -1 \ ; & z < -1 \\ z \ ; & -1 < z < 1 \\ +1 & z > 1 \end{cases} \quad (3.39)$$

ซึ่งพบว่าการแบ่งกลุ่มข้อมูลด้วยการใช้สมการที่ 3.37 นั้นจะได้ผลลัพธ์ของการแบ่งกลุ่มเป็นค่า -1 หรือ 1 เท่านั้น ซึ่งแตกต่างกับการแบ่งกลุ่มข้อมูลด้วยเงื่อนไขในสมการที่ 3.39 ซึ่งจะได้ผลลัพธ์จากการแบ่งกลุ่มข้อมูลที่อยู่ในช่วง -1 หรือ 1 ซึ่งมีข้อดีสำหรับข้อมูลที่มี ตัวรบกวน (Noise) ที่สูงซึ่งไม่สามารถทำการแบ่งกลุ่มข้อมูลได้อย่างชัดเจนด้วยระนาบเกินที่สร้างขึ้น รวมทั้งมีประโยชน์ในการกำจัดข้อมูลที่ไม่เหมาะสมสำหรับการแบ่งกลุ่มด้วยระนาบเกินที่สร้างขึ้นได้อีกด้วย ซึ่งจากเงื่อนไขของ Kuhn-Tucker จากสมการที่ 3.33 เมื่อแทนค่าของ w^* แล้วจะทำให้ได้ผลลัพธ์ตามสมการที่ 3.40

$$y_i[\langle w^*, x_i\rangle + b^*] = y_i\left[\sum_{i=1}^{l} \alpha_i^* y_i \langle x_i, x_j\rangle + b^*\right] = 1 \quad (3.40)$$

ดังนั้นเมื่อ x_i เป็นข้อมูลของซัพพอร์ตเวกเตอร์ (Support vector) แล้วจะสามารถคำนวณหาค่าของ $\|w\|^2$ ได้ดังสมการที่ 3.41และ 3.42

$$\|w\|^2 = \langle w^*, w^*\rangle = \sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i^*\alpha_j^* y_i y_j \langle x_i, x_j\rangle \quad (3.41)$$

$$= \sum_{i\in SV}^{l} \alpha_i^* y_i \sum_{j\in SV}^{l} \alpha_j^* y_j \langle x_i, x_j\rangle \quad (3.42)$$

เมื่อ $w^* = \sum_{i=1}^{l} \alpha_i^* y_i x_i$

และ $b^* = 1 - \langle w^*, x_i\rangle$

จากสมการที่ 3.41 จะได้

$$y_i b^* = 1 - \left[\sum_{j=1}^{l} \alpha_j^* y_j \langle x_i, x_j\rangle\right] \quad (3.43)$$

ทำให้ได้ $$\|w\|^2 = \sum_{i\in SV} \alpha_i^*(1 - y_i b^*) \quad (3.44)$$

$$\|w\|^2 = \sum_{i\in SV} \alpha_i^* \quad (3.45)$$

ด้วยเงื่อนไขจากสมการที่ 3.45

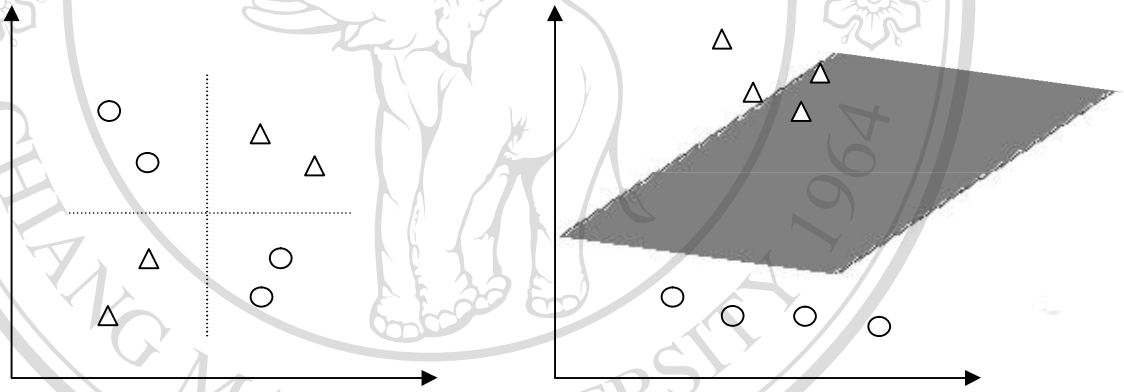
$$\frac{\partial}{\partial b}\left[\frac{1}{2}\langle w^*, w^*\rangle - \sum_{i=1}^{l} \alpha_i^*\left(y_i[\langle w, x_i\rangle + b] - 1\right)\right] = 0 \quad (3.46)$$

$$\sum_{i=1}^{l} \alpha_i^* y_i = 0 \quad (3.47)$$

ดังนั้นค่าของระนาบเกินที่เหมาะสมที่สุด จึงสามารถพิจารณาได้จากค่าที่มากที่สุดของระยะขอบซึ่งเป็นค่าที่มากที่สุดของ $\sum_{i \in SV} \alpha_i^*$ นั่นเอง

3.3.2 ซัพพอร์ตเวกเตอร์แมชชีนแบบไม่เชิงเส้น (Non-linear Support Vector Machine)

ในการใช้ซัพพอร์ตเวกเตอร์แมชชีนกับข้อมูลที่ไม่อาจแบ่งแบบเชิงเส้นได้ ต้องแก้ปัญหานี้โดยการส่งข้อมูลตัวอย่างไปยังปริภูมิอันดับสูง โดยมีสมมติฐานว่าข้อมูลที่มีความสัมพันธ์ไม่เป็นเชิงเส้นในปริภูมิอันดับต่ำ เมื่อส่งไปสู่ปริภูมิอันดับสูงจะมีความสัมพันธ์เป็นเชิงเส้นได้ อย่างไรก็ตามการส่งไปสู่ปริภูมิอันดับสูง อาจต้องการการคำนวณที่สูง ดังนั้นเคอร์เนลฟังก์ชันจึงเป็นวิธีการที่ช่วยหลีกเลี่ยงปัญหาการคำนวณหาฟังก์ชันในการส่งได้ โดยยอมให้คำนวณจากผลคูณสเกลาร์ของตัวแปรสองตัวในปริภูมิอันดับสูงได้โดยไม่ต้องคำนวณหาฟังก์ชันที่ใช้ส่ง



รูป 3.9 แนวคิดการส่งแบบไม่เชิงเส้นไปสู่ปริภูมิอันดับสูง (higher dimensional space)

โดยเลือกใช้ฟังก์ชันการส่งที่ไม่เป็นเชิงเส้นนั่นคือ เลือกฟังก์ชันในการส่ง $\Phi : R^N \rightarrow F$ เมื่อปริภูมิของ F มีอันดับสูงกว่าปริภูมิอันดับ N จะสามารถค้นหาระนาบที่แบ่งแยกที่ดีสุดในปริภูมิอันดับสูงที่เทียบเท่ากับการแยกที่ไม่เป็นเชิงเส้นใน R^N แล้วจึงใช้ขั้นตอนวิธีเชิงเส้นนี้ใน F ซึ่งสิ่งที่ต้องทำก็เพียงการหาค่าของผลคูณสเกลาร์ตามสมการที่ 3.48

$$k(x, y) = (\Phi(x) \cdot \Phi(y)) \tag{3.48}$$

ถ้า F มีอันดับสูงย่อมเท่ากับว่าพจน์ด้านขวามือของสมการ 3.48 จะคำนวณได้ยากมากอย่างไรก็ตาม ถ้าเรามีฟังก์ชันเคอร์เนลที่คำนวณแล้ว สามารถใช้ฟังก์ชันนี้แทน $x \cdot y$ ทุกๆ ที่ใน การคำนวณและ

ไม่จำเป็นต้องรู้ฟังก์ชันที่ใช้ส่ง Φ จริงๆ โดยฟังก์ชันเคอร์เนลถูกใช้ในปริภูมิข้อมูลเข้าไปยังปริภูมิลักษณะที่ซึ่งมีมิติที่สูงกว่าและทำให้นำไปสู่การคัดแยกข้อมูลเข้า แบบไม่เป็นเชิงเส้นได้

3.3.3 การฝึกสอนซัพพอร์ตเวกเตอร์แมชชีน ในปริภูมิลักษณะมิติสูง

เมื่อพิจารณากรณีที่ขอบเขตการจำแนกของข้อมูล ซึ่งข้อมูลในที่นี้จะอยู่ในรูปแบบ ของเวกเตอร์ โดยวิธีซัพพอร์ตเวกเตอร์แมชชีนสามารถส่ง (Map) ข้อมูล x ไปยังปริภูมิลักษณะมิติสูง (High dimensional feature space) ได้ ดังนั้นซัพพอร์ตเวกเตอร์แมชชีนไม่จำเป็นจะต้องผูกติดอยู่กับระนาบเกินแบบเชิงเส้นอีกต่อไป โดยการเลือกการส่งแบบไม่เป็นเชิงเส้น (Non-linear mapping) ที่เหมาะสมกับวิธีซัพพอร์ตเวกเตอร์แมชชีน จะสามารถหาระนาบเกินแบบไม่เชิงเส้นในปริภูมิ ที่มีมิติสูงขึ้นอย่างเหมาะที่สุดได้ ดังรายละเอียดต่อไปนี้

พิจารณาข้อมูลเวกเตอร์บนระนาบ 1 มิติในรูป 3.10 (ก) ที่ซึ่งข้อมูลสามารถแบ่งแยกได้แบบเชิงเส้นซัพพอร์ตเวกเตอร์แมชชีน สามารถหาระนาบเกินแบบเชิงเส้นที่สามารถจำแนกข้อมูลทั้ง 2 คลาสได้อย่างไม่มีปัญหา แต่โดยปกติแล้ว ข้อมูลส่วนใหญ่จะไม่ได้ถูกจำแนกได้อย่างสะดวกเช่นนั้น ยกตัวอย่างข้อมูลเวกเตอร์ฝึกสอนในรูป 3.10 (ข) สำหรับข้อมูลชุดดังกล่าววิธีซัพพอร์ต-เวกเตอร์แมชชีน ไม่สามารถที่จะหาระนาบเกินสำหรับแบ่งแยกแบบเชิงเส้นใดๆ ได้ที่ทำให้การจำแนกข้อมูลทั้ง 2 คลาสเป็นไปอย่างถูกต้องได้

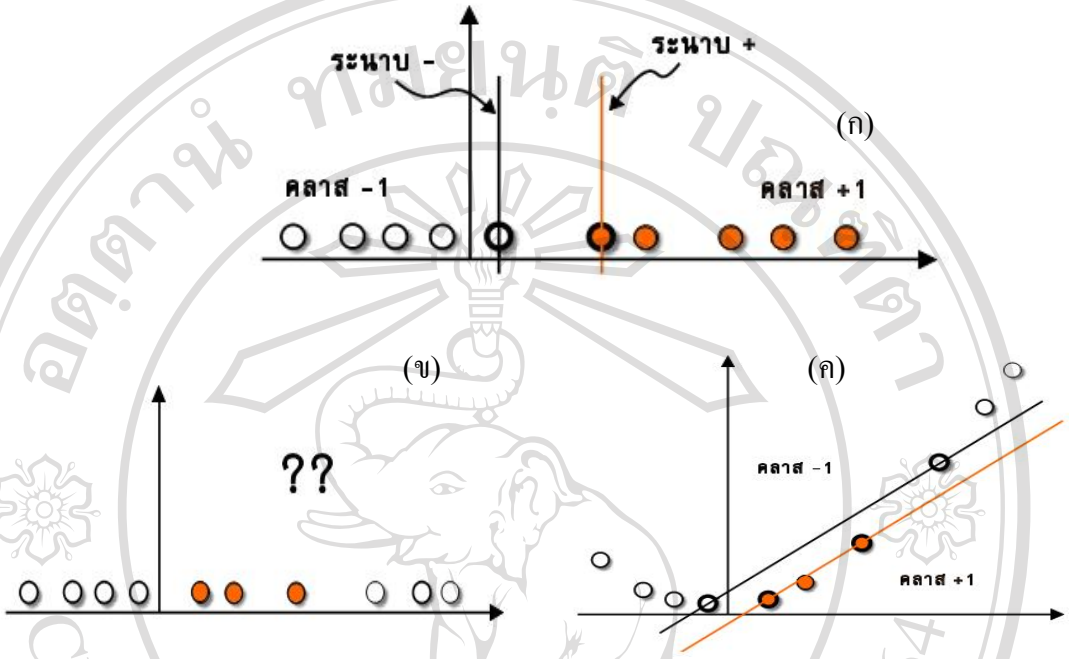
แนวคิดในการแก้ปัญหาดังกล่าวแสดงในรูป 3.10 (ค) โดยที่ข้อมูลถูกส่งไปยังปริภูมิมิติสูงขึ้นในที่นี้เป็นการส่งจากระนาบ 1 มิติไปยังระนาบ 2 มิติ ทำให้การจัดเรียงข้อมูลเปลี่ยนไป และซัพพอร์ตเวกเตอร์แมชชีน สามารที่จะหาระนาบเกินแบบเชิงเส้นสำหรับจำแนกข้อมูลเวกเตอร์ได้อย่างถูกต้อง ปริภูมิที่เป็นเวกเตอร์ของข้อมูลเข้าถูกส่งไป ก่อนที่จะถูกจำแนกเรียกว่าปริภูมิลักษณะ (Feature space) ตามแผนผังดังแสดงในรูป 3.8

ถึงแม้ว่าการใช้ฟังก์ชันส่งแบบไม่เป็นเชิงเส้นนี้มีข้อจำกัด แต่ฟังก์ชันต่างๆ ที่ใช้เพื่อการส่งนี้ให้ผลที่ยอมรับได้ และไม่ว่าจะเป็นฟังก์ชันโพลีโนเมียล (Polynomial function) ฟังก์ชันฐานรัศมี (Radial basis function) หรือฟังก์ชันซิกมอยด์ (Sigmoid function) เมื่อใช้ฟังก์ชันส่งเหล่านี้แล้วสมการสำหรับการหาค่าเหมาะที่สุดจาก 3.49 จะกลายเป็น

$$\max_{\alpha} \sum_{i=1}^{Q} \alpha^i + \sum_{i=1}^{Q} \sum_{j=1}^{Q} \alpha^i \alpha^j y^i y^j K\left(x^i, x^j\right) \qquad (3.49)$$

โดยที่ $K\left(x^i, x^j\right)$ คือฟังก์ชันเคอร์เนล (Kernel function) ที่ทำหน้าที่ในการส่งแบบไม่เป็นเชิงเส้นไปยังปริภูมิลักษณะ ตัวอย่างฟังก์ชันเคอร์เนลที่แสดงในรูป 3.10 คือ $K(x, x^2)$ ซึ่งเป็นการ

ส่งผ่านความสัมพันธ์ในรูปพหุนามกำลังสองหรือแบบโพลีโนเมียล 2 มิติ โดยรายละเอียดของฟังก์ชันเคอร์เนลจะได้กล่าวถึงในหัวข้อต่อไป



รูป 3.10 (ก) ตัวอย่างข้อมูลเวกเตอร์ฝึกสอนที่แบ่งแยกได้แบบเชิงเส้น
(ข) ข้อมูลเวกเตอร์ฝึกสอนที่มาสามารถแบ่งแยกได้แบบเชิงเส้น
(ค) การส่งเวกเตอร์ฝึกสอนไปยังปริภูมิมิติสูงขึ้น (ในที่นี้คือ 2 มิติ) ทำให้ซัพพอร์ตเวกเตอร์แมชชีนสามารถแบ่งแยกข้อมูลได้
(ที่มา: สุชาวดี ตันติศักดิ์, 2550)

3.3.4 ฟังก์ชันเคอร์เนล

สำหรับฟังก์ชันเคอร์เนลหรือฟังก์ชันแก่นจะเป็นการส่งแบบไม่เป็นเชิงเส้นไปยังปริภูมิลักษณะต่างๆ ตามฟังก์ชันที่เราต้องการตามรูป 3.11 โดยฟังก์ชันเคอร์เนลที่ใช้สำหรับซัพพอร์ตเวกเตอร์แมชชีน ได้แก่

รูป 3.11 แนวคิดการส่งในเคอร์เนลฟังก์ชัน
(ที่มา: สุชาวดี ตันติศักดิ์, 2550)

- ฟังก์ชันโพลีโนเมียล (Polynomial function) เป็นฟังก์ชันเคอร์เนลที่นิยมใช้สำหรับการจำลองแบบไม่เป็นเชิงเส้นฟังก์ชันดังกล่าวอยู่ในรูปต่อไปนี้

$$K(x.x') = \langle x.x' \rangle^d$$

$$K(x.x') = (\langle x.x' \rangle + 1)^d \quad (3.50)$$

- ฟังก์ชันฐานรัศมีแบบเกาส์เซียน (Gaussion radial basis function) เป็นฟังก์ชันที่อยู่ในรูปเกาส์เซียน

$$K(x.x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right) \quad (3.51)$$

- ฟังก์ชันฐานรัศมีแบบเลขชี้กำลัง (Exponential radial basis function) อยู่ในรูป

$$K(x.x') = \exp\left(-\frac{\|x-x'\|}{2\sigma^2}\right) \quad (3.52)$$

- ฟังก์ชันซิกมอยด์ (Sigmoid function) เป็นฟังก์ชันจากเครือข่ายเพอร์เซ็พตรอนแบบหลายชั้นที่อยู่ในรูป

$$K(x.x') = \tanh(\rho\langle x, x' \rangle + \varrho) \quad (3.53)$$

โดยที่ ρ คือมาตราส่วนและ ϱ คือค่าออฟเซต ซึ่งเป็นพารามิเตอร์ที่ต้องเลือก

โดยสามารถสรุปได้ว่าซัพพอร์ตเวกเตอร์แมชชีนเป็นกระบวนการทางคณิตศาสตร์แนวทางใหม่ที่ใช้ในการแก้ไขข้อด้อยของโครงข่ายประสาทแบบดั้งเดิม และทำให้ผู้ใช้สามารถออกแบบอัลกอริทึมได้ง่ายกว่าเดิม เนื่องจากซัพพอร์ตเวกเตอร์แมชชีน ไม่จำเป็นต้องกำหนดจำนวนชั้นซ่อน จำนวนนิวรอน ค่าอัตราการเรียนรู้ ค่าน้ำหนักเริ่มต้น ฯลฯ โดยซัพพอร์ตเวกเตอร์แมชชีนมีชั้นซ่อนเพียงชั้นเดียว ซึ่งสามารถสร้างระนาบเกินที่เหมาะสมที่สุดในการแบ่งกลุ่มหรือใช้แทนกลุ่มข้อมูลในการวิเคราะห์แบบถดถอย ซึ่งสามารถเลือกใช้งานได้ตามความเหมาะสม นอกจากนี้ซัพพอร์ตเวกเตอร์แมชชีนยังมีการใช้ซัพพอร์ตเวกเตอร์น้ำหนักที่ใช้ในโครงข่ายประสาทแบบดั้งเดิม เป็นการช่วยลดเวลาที่ใช้เรียนรู้ และมีผลลัพธ์ที่ถูกต้องมากขึ้น ส่วนในด้านการออกแบบซัพพอร์ตเวกเตอร์แมชชีน ยังมีการออกแบบอย่างมีหลักการเพิ่มมากขึ้น โดยอาศัยหลักกระบวนการลดความเสี่ยงเชิงโครงสร้างให้ต่ำที่สุด จึงทำให้ผลลัพธ์ที่ได้มีความน่าเชื่อถือมากขึ้นและใช้เวลาในการเรียนรู้ที่รวดเร็ว

3.3.5 การจำแนกข้อมูลแบบหลายกลุ่มด้วยซัพพอร์ตเวกเตอร์แมชชีน

ถึงแม้ว่าซัพพอร์ตเวกเตอร์แมชชีนจะเป็นตัวจำแนกข้อมูลแบบ 2 กลุ่มเราสามารถปรับโครงสร้างของระบบ เพื่อให้ซัพพอร์ตเวกเตอร์แมชชีนสามารถเรียนรู้และจำแนกข้อมูลแบบ *N* กลุ่มได้ โดยเลือกใช้ซัพพอร์ตเวกเตอร์แมชชีนจำนวน *N* ชุดเพื่อเรียนรู้ข้อมูลแต่ละกลุ่มด้วยรูปแบบต่อไปนี้

- SVM #1 เรียนรู้ข้อมูล “ชุดที่ 1” และ “ไม่ใช่ชุดที่ 1”
- SVM #2 เรียนรู้ข้อมูล “ชุดที่ 2” และ “ไม่ใช่ชุดที่ 2”
- SVM #3 เรียนรู้ข้อมูล “ชุดที่ 3” และ “ไม่ใช่ชุดที่ 3”
- ⋮
- SVM #N เรียนรู้ข้อมูล “ชุดที่ N” และ “ไม่ใช่ชุดที่ N”

หลังจากนั้นเราสามารถป้อนข้อมูลเข้าชุดใหม่ให้แต่ละซัพพอร์ตเวกเตอร์แมชชีนแล้วทำการคำนวณหาว่าซัพพอร์ตเวกเตอร์แมชชีนชุดใดที่จำแนกข้อมูลนำเข้านี้ได้ดีที่สุด โดยสามารถเลือกได้จากเงื่อนไขระยะห่างของข้อมูลเข้าที่ห่างจากระนาบเกิน + มากที่สุด โดยทั่วไปจะกำหนดให้ระนาบเกิน + แทนชุดข้อมูลที่ใช่ และระนาบเกิน – แทนชุดข้อมูลที่ไม่ใช่

(สุชาวดี ตันติศักดิ์, 2550)

3.4 สตริงเคอร์เนล

สตริงเคอร์เนลเป็นเครื่องมือทางคณิตศาสตร์ที่ใช้กันอย่างกว้างขวาง ในการวิเคราะห์เหมืองข้อมูล ซึ่งเป็นข้อมูลสายลำดับที่มีการจัดกลุ่มหรือแบ่งกลุ่ม โดยเฉพาะอย่างยิ่งในงานวิจัยที่เกี่ยวข้องกับข้อความและการวิเคราะห์กลุ่มยีน

โดยกลวิธีในการจัดจำแนกกลุ่มของสิ่งมีชีวิตและไม่มีชีวิตต่างๆ ที่มีจำนวนมากนั้น กลวิธีนี้เป็นกลวิธีหนึ่งที่ได้รับการยอมรับให้เป็นกลวิธีสำหรับการแก้ปัญหาในการจัดจำแนกประเภทหรือกลุ่มของยีนหรือโปรตีน รวมทั้งการหาความคล้ายของข้อมูลสายลำดับ ซึ่งเรียกว่าสตริงเคอร์เนล

ซึ่งกลวิธีในการจัดจำแนกนี้ ประสบความสำเร็จอย่างมากที่สุด คือการจัดจำแนกของกลุ่มโปรตีน (Jaakkola, 2004) โดยพวกเขาเริ่มต้นด้วยการฝึกสอนให้เครื่องรู้จำเกี่ยวกับการสร้างแบบจำลองมาร์คอฟ สำหรับชนิดโปรตีนในกลุ่มเดียวกัน จากนั้นจึงใช้รูปแบบการจัดจำแนกกลุ่มข้อมูลสายลำดับ โดยอาศัยคุณสมบัติของเวกเตอร์เข้ามาช่วยในขั้นตอนการเรียนรู้ด้วยเครื่องส่วนในการจำแนกชนิดของข้อมูลนั้นจะใช้วิธีซัพพอร์ตเวกเตอร์แมชชีน ซึ่งในซัพพอร์ตเวกเตอร์-แมชชีนนั้น จะมีฟังก์ชันหนึ่งที่เรียกว่าสเปกตรัมเคอร์เนล โดยเป็นวิธีการหนึ่งที่ใช้ในการเปรียบเทียบความเหมือนหรือความคล้ายของข้อมูล ซึ่งมีความยืดหยุ่นในการทำงานสูง และเหมาะสำหรับข้อมูลที่อยู่ในลักษณะเป็นอักขระหรือสายลำดับ

3.4.1 สเปกตรัมเคอร์เนล

ในปัจจุบันมีวิธีการในการจัดจำแนกข้อมูลสายเป็นจำนวนมาก ซึ่งวิธีการที่นิยมใช้กันมากคือฟังก์ชันเคอร์เนล ตัวอย่างเช่น สตริงเคอร์เนล ซึ่งเป็นกลวิธีที่ใช้การนับความเหมือนของอักขระที่มีความยาว n ที่แน่นอนระหว่างอักขระแต่ละอักขระ และยังมีวิธีการอื่นๆ ที่อนุญาตให้มีการใช้ช่องว่าง (Gaps) ซึ่งวิธีการนี้จะใช้ในการตรวจสอบความเหมือนหรือความคล้ายกันของแต่ละสายอักขระทั้งหมด โดยการใช้ช่องว่างเข้ามาช่วยในการหาความเหมือน ซึ่งจะมีการให้คะแนนของแต่ละสายอักขระเมื่อมีการใช้ช่องว่าง โดยคะแนนรวมของแต่ละสายอักขระนั้นจะเป็นตัววัดค่าความเหมือนของแต่ละสายอักขระ เป็นต้น ซึ่งวิธีการที่เราจะนำเสนอคือสเปกตรัมเคอร์เนล โดยจะเน้นเกี่ยวกับการใช้ฟังก์ชันเคอร์เนลสำหรับจัดจำแนกประเภทของโมเลกุลเอชแอลเอ

สำหรับงานวิจัยเกี่ยวกับฟังก์ชันเคอร์เนลที่ใช้กับข้อมูลสายลำดับนั้นสามารถศึกษาได้จากงานวิจัยของ (Watkins., 1990) และมีงานวิจัยอีกชิ้นหนึ่งที่นำเสนอเคอร์เนลฟังก์ชันใหม่และมีการประยุกต์วิธีการนี้เพื่อใช้กับงานประเภทจัดหมวดหมู่ของข้อความโดยใช้สตริงเคอร์เนล ที่ใช้การหาความเหมือนของข้อมูลสายลำดับย่อย เป็นต้น

ซึ่งสตริงเคอร์เนลมีความคล้ายคลึงอย่างมากกับวิธีการของถุงอักขระ (Bag of words) แต่จะต่างกันคือวิธีการที่นำเสนอนี้จะไม่นับความถี่ของแต่ละตัวอักขระ แต่จะนับความถี่เป็นช่วงความถี่เท่ากับ n (n-grams) ยกตัวอย่างเช่น การพิจารณาจำนวนของอักขระย่อยที่มีความยาว n ที่ปรากฏในเอกสารนั้น เป็นต้น แต่ต่อมาได้มีผู้นำเสนอวิธีการใช้สเปกตรัมที่มีความถี่เท่ากับสอง ซึ่งเป็นการนำเสนอในลักษณะหาความเหมือนของข้อมูลสายลำดับสำหรับการใช้ในการจำแนกข้อมูลประเภทโปรตีน (Leslie et al., 2002) โดยหลักการนี้ เริ่มต้นจะใช้หลักการของสเปกตรัมเคอร์เนลที่หาความเหมือนหรือความคล้ายคลึงกันของแต่ละข้อมูลสายลำดับ ซึ่งขึ้นอยู่กับจำนวนของข้อมูลย่อย (Subsequences) เป็นต้น

จากงานวิจัยของ Saunders et al., 2002 ได้ศึกษาถึงข้อดีในการคำนวณของสเปกตรัมเคอร์เนล เนื่องจากวิธีการนี้เป็นวิธีการที่คำนวณได้รวดเร็วและง่าย เพราะข้อมูลที่นำมาใช้ในการทำนายมีลักษณะที่เป็นรูปแบบที่แน่นอน จึงทำการทำนายได้ในเวลาเชิงเส้น (Linear time)

โดยวิธีการของสเปกตรัมเคอร์เนลเป็นวิธีการที่ใช้สำหรับการจัดจำแนกข้อมูลสายลำดับประเภทยีนหรือโปรตีน ซึ่งวิธีการนี้เป็นเคอร์เนลหนึ่งของวิธีการของสตริงเคอร์เนล ซึ่งเรียกว่าเคอร์เนลนี้ว่าสเปกตรัมเคอร์เนล และมี x เป็นข้อมูลสายลำดับเข้าที่มีความยาว l ซึ่งมีอักขระทั้งหมด A อักขระ โดยกำหนดได้ดังนี้

$$|A| = l \tag{3.54}$$

ซึ่งสเปกตรัมเคอร์เนลเป็นเคอร์เนลที่มีลักษณะพิเศษสำหรับการแก้ไขปัญหาในการเปรียบเทียบกันระหว่างข้อมูลสายลำดับแต่ละสาย โดยจะมีฟังก์ชันหนึ่งที่เรียกว่า เค-สเปกตรัม (k-spectrum) โดยกำหนดให้ $k \geq 1$ ซึ่งสายลำดับของเค-สเปกตรัมนั้นคือ เซตของชุดข้อมูลสายลำดับที่อยู่ติดกันที่มีขนาดเท่ากับเค (k-mers) ของกลุ่มย่อยในข้อมูลสายลำดับที่ประกอบขึ้นมา และมีการปรับคุณลักษณะ (Feature map) นั้นคือการแจกแจงทุกกรณีที่เป็นไปได้ของกลุ่มย่อยที่มีความยาวเค จากอักขระทั้งหมด A ดังนั้นเราจึงกำหนดการปรับคุณลักษณะจาก x to $\mathbb{R}^{l^k}$ โดย

$$\Phi_{\mathrm{k}}(x) = (\phi_a(x))_{a \in A^k} \tag{3.55}$$

โดยที่

$\phi_a(x)$ = จำนวนความถี่ที่ปรากฏ a ในข้อมูลสายลำดับ x

ดังนั้นรูปแบบหรือลักษณะของข้อมูลสายลำดับ x ภายใต้เงื่อนไขที่ระบุไว้แล้วตามสมการที่ 3.55 คือการแทนค่าด้วยน้ำหนัก (Weight) ของเค-สเปกตรัม ซึ่งสมการที่ใช้ในการคำนวณฟังก์ชันของเค-สเปกตรัมนี้แสดงได้ดังนี้

$$K_k(x, y) = \langle \Phi_{\mathrm{k}}(\mathrm{x}), \Phi_{\mathrm{k}}(\mathrm{y}) \rangle \tag{3.56}$$

โดยสังเกตว่าขณะที่ปริภูมิลักษณะทั้งหมดของข้อมูลมีอยู่หลายลักษณะ ซึ่งจะทำการปรับให้มีขนาดเท่ากับ เค เพื่อทำการคำนวณนั้นจะทำการปรับให้อยู่ในรูปของเวกเตอร์ (Feature vectors) โดยขอบเขตของจำนวนข้อมูลจะถูกกำหนดให้มีความยาวเท่ากับ $(\mathrm{length}(x)) - k + 1)$ ซึ่งหลักการนี้มีผลต่อประสิทธิภาพในการคำนวณค่าเคอร์เนล เช่นกัน

วิธีการที่มีประสิทธิภาพสำหรับการคำนวณฟังก์ชันเคอร์ล $K_k(x, y)$ นี้คือการสร้างต้นไม้ (Suffix tree) สำหรับเป็นการรวบรวมลักษณะของข้อมูลสายลำดับย่อยที่มีความยาวเท่ากับ เค ของข้อมูลสายลำดับ x และ y ซึ่งเป็นการกำหนดของวิธีการ เค-สเปกตรัม ที่มีการเลื่อนสไลด์ วินโดว์ (Slide window) ด้วยความยาวเท่ากับ เค ไปเรื่อยๆในแต่ละข้อมูลสายลำดับของ x และ y โดยในแต่ละระดับของ เค นั้นจะดูได้จากใบหรือโหนดของต้นไม้

ซึ่งในการอธิบายการใช้เวลาของการคำนวณด้วยวิธีการ เค-สเปกตรัมที่มีความยาวเท่ากับ เค ในข้อมูลสายลำดับ x และ y จนถึงโหนดสุดท้ายนั้น จะเห็นได้ว่าใช้เวลาคำนวณใกล้เคียงกัน โดยจะสังเกตว่าการสร้างต้นไม้นี้จะมีค่าการคำนวณทั้งหมดคือ $O(kn)$ ดังรูป 3.15

ส่วนขั้นตอนวิธีในการสร้างต้นไม้นั้นจะอยู่ในรูปของแบบเวลาเชิงเส้น ซึ่งวิธีการนี้สามารถสร้างและกำหนดค่า เค ให้กับต้นไม้ที่สร้างได้ เช่นกัน โดยการคำนวณของเคอร์เนลฟังก์ชันนี้สามารถใช้การข้าม (Traversing) ของต้นไม้ได้และการคำนวณค่าของผลรวมต่างๆ จะถูกจัดเก็บค่าทั้งหมดไว้ที่โหนดสุดท้ายเพื่อทำการคำนวณต่อไป

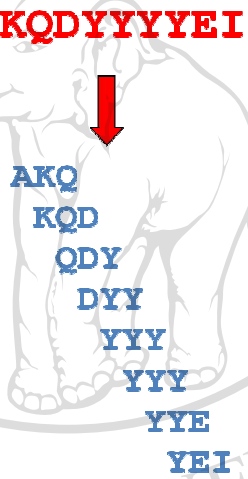
ดังนั้นวิธีการนี้ จึงเป็นวิธีการที่จำเป็นอย่างยิ่งในการคำนวณเคอร์เนลสำหรับการทดลอง โดยวิธีการที่เรานำเสนอนี้จะใช้ฟังก์ชันย้อนกลับ (Recursive function) ซึ่งจะค่อนข้างชัดเจนกว่าโครงสร้างที่เป็นแบบต้นไม้ ดังรูป 3.15

โดยทั่วไปมีวิธีการหลายวิธีในการคำนวณค่าเคอร์เนลว่ามีประสิทธิภาพมากน้อยเพียงใด ซึ่งง่ายต่อการพัฒนา โดยวิธีการที่นำเสนอนี้จะแสดงออกด้วยค่าเลขฐานสอง โดยการทำการปรับคุณลักษณะและใช้ค่าการนับ (Count-valued) ที่มีความคล้ายคลึงหรือเหมือนกันกัน สำหรับแต่ละข้อมูลสายลำดับ x และมีการเก็บรวบรวมคะแนนที่คำนวณได้เป็นเซต โดยเราจะระบุขนาดของความยาวของข้อมูลย่อยในการพิจารณาเท่ากับ เค หรือที่เรียกว่า เค-สเปกตรัม ซึ่งค่าหรือคะแนนที่

ได้คำนวณแล้วจะจัดเก็บคะแนนไว้ในอาเรย์ A_x ในขณะเดียวกันก็ได้ทำการคูณแบบอินเนอร์โปรดัค (Inner product) ของฟังก์ชัน $K_k(x, y)$ ซึ่งสามารถคำนวณในลักษณะของเชิงเส้นได้เหมือนกันคือ ฟังก์ชันการคำนวณของข้อมูลสายลำดับ $(x + y)$ ดังนั้นการคำนวณที่ได้จะมีความซับซ้อนมากขึ้นและได้ค่าใช้จ่าย (Cost) ในการคำนวณเคอร์เนลคือ $O(n \log(n))$ ของข้อมูลสายลำดับเข้าทั้งหมด เป็นต้น

3.4.2 เค-สเปกตรัม (k-Spectrum) สำหรับข้อมูลสายลำดับ

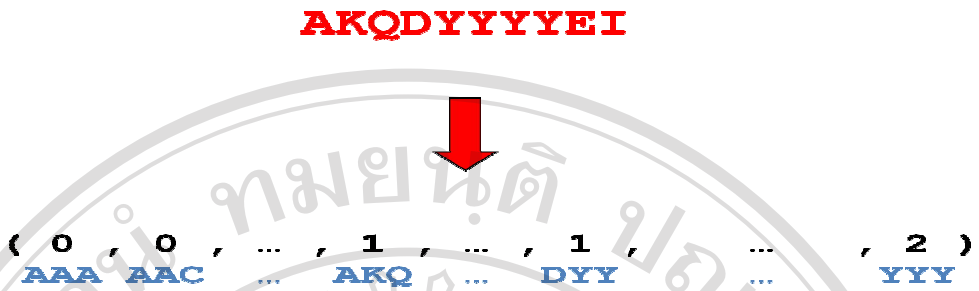
เค-สเปกตรัมสำหรับข้อมูลสายลำดับเป็นเซตของความยาวทั้งหมดขนาด เค ที่ติดกันของข้อมูลสายลำดับย่อย ซึ่งจะเรียกว่า เค-เมอร์ (k-mers) ซึ่งจะอยู่ในรูปของกรดอะมิโน โดยมิติของรูปแบบนั้นสามารถแบ่งได้ถึง 20^k



รูป 3.12 เซตของความยาวทั้งหมดที่มีขนาด เค เท่ากับ 3
(ที่มา: Leslie et al., 2002)

จากรูปจะเห็นได้ว่าข้อมูลสายลำดับที่แสดงนั้น เราจะใช้เค-สเปกตรัม เท่ากับ 3 ซึ่งจะอยู่ในลักษณะที่เรียกว่า โคดอน ที่ได้อธิบายในบทที่ 2 โดยจะเห็นว่ามิติของรูปแบบที่สามารถแบ่งได้จะเท่ากับ $20^3 = 8{,}000$ แบบ เป็นต้น

ซึ่งคุณลักษณะของสเปกตรัมเคอร์เนลสามารถกำหนดได้ตามสมการที่ 3.55 โดยจะมีการกำหนดค่าให้กับข้อมูลสายลำดับย่อยหรือค่า เค ซึ่งในที่นี้จะกำหนดให้เซตของความยาวทั้งหมดที่มีขนาด เค เท่ากับ 3 โดยหลักการของวิธีการนี้จะให้ค่าเท่ากับ 1 เมื่อพบว่าข้อมูลสายลำดับย่อย เกิดขึ้นในข้อมูลสายลำดับทั้งหมด แต่จะให้ค่าเท่ากับ 0 เมื่อพบว่าข้อมูลสายลำดับย่อยไม่เกิดขึ้นในข้อมูลสายลำดับทั้งหมด ดังที่แสดงในรูป 3.13



รูป 3.13 การให้ค่าของความเหมือนหรือแตกต่างให้กับข้อมูลสายลำดับ

(ที่มา: Leslie et al., 2002)

3.4.3 การกำหนดคุณลักษณะแบบ (เค, เอ็ม)-มิสแมช

(เค, เอ็ม)- มิสแมช เป็นรูปแบบการคำนวณเพื่อหาความเหมือนของข้อมูลสายลำดับทั้งหมด ซึ่งใช้มีการอนุญาตให้ใช้ ช่องว่าง เข้ามาช่วย โดยสามารถกำหนดค่าต่างๆ ได้ดังนี้

ถ้า s คือเค-เมอร์ จะกำหนดได้ว่า

$$F_{(k,m)}(s) = (F_t(s))_{\{k-mer_{s,t}\}} \tag{3.57}$$

โดยที่

$F_t(s) = 1$ ถ้า s พบว่ามีชุดข้อมูล m ที่ไม่เหมือนกันในชุดข้อมูลที่ได้จากชุดข้อมูลที่ t

$F_t(s) = 0$ ถ้า s พบว่ามีชุดข้อมูล m ที่เหมือนกันในชุดข้อมูลที่ได้จากชุดข้อมูลที่ t

รูป 3.14 การพิจารณานิยามความเหมือนที่กำหนดไว้

(ที่มา: Leslie et al., 2002)

จากรูป 3.14 จะเห็นได้ว่าการหาความเหมือนของสายลำดับนั้น จะพิจารณาทีละอักขระ โดยจะพิจารณาทุกกรณีที่เป็นไปได้

3.4.4 กลวิธีของสเปกตรัมเคอร์เนล

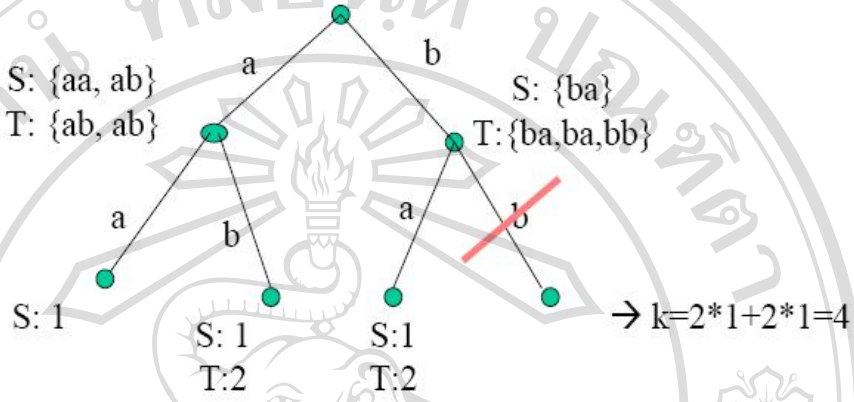
ในการเรียนรู้ด้วยเครื่องโดยใช้ซัพพอร์ตเวกเตอร์แมชชีนสามารถใช้สเปกตรัมเคอร์เนลเข้ามาช่วยในการจำแนกชนิดข้อมูล โดยการส่งคุณลักษณะสำหรับตัวแปร x และ y ภายใต้คุณสมบัติคือ

$$K_k(x,y) = \langle \Phi_k(x), \Phi_k(y) \rangle \tag{3.58}$$

โดยคะแนนความคล้ายคลึงหรือความเหมือนของข้อมูลสายลำดับ จะสามารถคำนวณผ่านทางเส้นทาง (Path) แต่ละเส้นของโครงสร้างข้อมูล (Trie) ซึ่งในการคำนวณของสเปกตรัมเคอร์เนลนั้นสามารถทำได้ โดยใช้โครงสร้างของข้อมูลที่เป็นแบบต้นไม้ เนื่องจากในการจัดรูปแบบของตัวอักขระทั้งหมดนั้นจะอยู่ในรูปแบบต้นไม้ ซึ่งสามารถกำหนดขนาดของการพิจารณาด้วย เค-สเปกตรัม ที่มีความยาวเท่ากับ เค ได้

ในข้อมูลสายลำดับของการสร้างเส้นทาง ที่เชื่อมต่อไปยังใบแต่ละใบของการสร้างต้นไม้ในแต่ละต้นนั้น จะเห็นว่าข้อมูลสายลำดับทั้งหมดจะมีความยาวที่จำกัด โดยวิธีการที่นำเสนอนี้จะมีการแจกแจงรูปแบบทุกกรณีที่เป็นไปได้ของชุดข้อมูลสายลำดับเข้าแต่ละตัว ซึ่งจะสามารถกำหนดได้ตามค่า เค จากนั้นเมื่อกำหนดขนาดของเส้นทางและค่า เค เรียบร้อยแล้ว จะเป็นขั้นตอนของการสร้างต้นไม้คือพิจารณาค่าที่เกิดขึ้นในแต่ละเส้นทาง ถ้าพบว่าเส้นทางใด ที่พบอักขระที่ได้จากการแจกแจงทุกกรณีแล้ว มีการเกิดขึ้นในเส้นทางนั้นจะให้ค่าเพิ่มขึ้นที่ละ 1 ในเซตของการคำนวณในแต่ละเส้นทาง แต่ถ้าไม่พบจะให้ค่าเท่ากับ 0 ซึ่งจะทำเช่นนี้จนครบทุกกรณีและทุกเส้นทาง สุดท้ายค่าที่คำนวณได้ทั้งหมดจะได้จากผลคูณของคะแนนในแต่ละเส้น จากนั้นจะนำมาบวกกันทั้งหมดจึงจะได้ผลลัพท์ ซึ่งเป็นคะแนนของค่าความเหมือนของข้อมูลสายลำดับ โดยในกรณีเป็นการวัดความเหมือนของข้อมูลสายลำดับ 2 สายลำดับคือ ข้อมูลสายลำดับ S และข้อมูลสายลำดับ T ดังรูป 3.15

- $k = 2$
- S=a a b a → {aa, ab, ba}
- T=b a b a b b → {ba, ab, ba, ab, bb}



รูป 3.15 การคำนวณค่าคะแนนของเค-สเปกตรัม

(ที่มา: Leslie et al., 2002)

ซึ่งวิธีการที่นำเสนอข้างต้นนี้จะมีข้อดีคือจะรวดเร็วและง่ายต่อการคำนวณ โดยใช้หน่วยความจำน้อย ทำให้วิธีการนี้มีประสิทธิภาพในการทำงานค่อนข้างสูง