บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

เอสวีเอ็มเป็นเทคนิคการเรียนรู้เชิงสถิติที่คิดค้นโดย Vapnik [10] ตั้งแต่ช่วงปี 1960 แต่ยังไม่เป็นที่นิยมจนถึงช่วงทศวรรษที่ผ่านมาเริ่มได้รับความสนใจมาก เนื่องจากมีการนำไปใช้แก้ปัญหาและพบว่าได้ผลดีมาก เอสวีเอ็มเกิดจากแนวคิดพื้นฐานดังนี้

1. การลดความเสี่ยงเชิงโครงสร้างให้ต่ำสุด (Structural Risk Minimization) เป็นแนวคิดที่แสดงขอบเขตของของความเสี่ยงของเครื่องเรียนรู้ ซึ่งขึ้นกับความเสี่ยงเชิงทดลองที่มาจากความผิดพลาดในการสอน กับช่วงความเชื่อมั่นที่เป็นฟังก์ชันของมิติ VC (VC Dimension) ที่แสดงถึงว่าฟังก์ชันที่ใช้จำแนกมีลักษณะทั่วไปเพียงใด ในทางปฏิบัติเราไม่สามารถลดความเสี่ยงที่แท้จริงได้ จึงพยายามลดความเสี่ยงจากความผิดพลาดให้ต่ำสุดแทน
2. ระนาบหลายมิติที่ใช้แยกดีที่สุด (Optimal Separating Hyperplane) ระนาบหลายมิตินี้จะต่างจากของโครงข่ายประสาทเทียมตรงที่เป็นระนาบที่แบ่งคอนเวกซ์ฮัล (Convex Hull) ของข้อมูลสองกลุ่มออกจากกันด้วยระยะห่างที่กว้างที่สุด
3. ฟังก์ชันเคอร์เนล (Kernel Function) เป็นเทคนิคที่ช่วยขยายความสามารถของเอสวีเอ็มให้จัดการกับปัญหาที่ไม่สามารถแบ่งแยกแบบเชิงเส้นได้ โดยการแมปข้อมูลในปริภูมิรับเข้า ไปสู่ปริภูมิอันดับสูงขึ้นไป ซึ่ง ณ ปริภูมิอันดับสูงนี้ จะสามารถใช้ระนาบหลายมิติแบบเชิงเส้นในการแยกข้อมูลสองกลุ่มออกจากกันได้

ส่วนต่อไปจะอธิบายแนวคิดพื้นฐานที่เกี่ยวข้องกับเอสวีเอ็ม ซึ่งได้กล่าวถึงข้างต้นอย่างละเอียด

2.1.1 การลดความเสี่ยงเชิงโครงสร้างให้ต่ำสุด (Structural Risk Minimization)

ในปัญหาการรู้จำแบบ เราต้องการหาฟังก์ชันที่มาประมาณตัวจำแนกที่แท้จริง ซึ่งค่าความผิดพลาดคือผลต่างระหว่างค่าที่ได้จากฟังก์ชันประมาณกับค่าที่ได้จากฟังก์ชันที่แท้จริง ซึ่งโดยปกติเราจะต้องการลดความเสี่ยงจากการจำแนกกลุ่มผิดให้ต่ำที่สุด แต่ในทางปฏิบัติเราไม่รู้ฟังก์ชันที่แท้จริง จึงไม่สามารถคำนวณหาค่าความผิดพลาดที่แท้จริงได้

ทางหนึ่งที่ทำได้คือ ในการสอนเราจะพิจารณาค่าความผิดพลาดจากกลุ่มตัวอย่างแทน และลดค่าของความเสี่ยงเชิงโครงสร้างซึ่งประกอบด้วย ความเสี่ยงเชิงทดลอง (Empirical Risk) กับ ช่วงความเชื่อมั่น (Confidence Interval) ให้ต่ำสุดแทน ความเสี่ยงเชิงโครงสร้างนี้มีพื้นฐานมาจากเรื่องขอบเขตของความผิดพลาดในการรู้จำแบบของเครื่อง กับมิติ VC

2.1.1.1 ขอบเขตของความผิดพลาดในการรู้จำแบบของเครื่อง

Vapnik ได้เสนอขอบเขตของความผิดพลาดในการรู้จำแบบของเครื่อง โดยสมมติให้เรามีข้อมูลอยู่ ℓ ตัว ซึ่งแต่ละข้อมูลประกอบด้วยคู่ของ เวกเตอร์ x กับ ชื่อกลุ่ม y โดยที่

$$(x_1,y_1), \ldots, (x_\ell,y_\ell) \in R^N \times \{\pm 1\}$$

ภายใต้การแจกแจงความน่าจะเป็น P(x,y) ตัวอย่างเช่น x อาจหมายถึงค่าสีของแต่ละจุดภาพของรูปตัวอักษรที่ต้องการรู้จำ ส่วน y แสดงถึงกลุ่มของรูปภาพว่าเป็นตัวอักษรที่สนใจหรือไม่ ถ้า y เป็น 1 หมายถึงว่าเป็นตัวอักษรที่สนใจ ส่วนถ้า y เป็น -1 ก็แสดงว่าไม่ใช่ตัวอักษรที่สนใจ

-1	1	-1	1	-1
-1	1	-1	1	-1
-1	1	1	1	-1
-1	1	-1	1	-1
-1	1	-1	1	-1

รูปที่ 1 ตัวอย่างข้อมูลบิตแมปของตัวอักษร H

เช่นกรณีที่เราสนใจตัวอักษร H โดยสมาชิกของ x มีค่าเป็น -1 เมื่อจุดนั้นมีสีขาว และมีค่าเป็น 1 เมื่อจุดนั้นมีสีดำ จากตัวอย่างในรูปที่ 1 คู่ลำดับนี้คือ ([-1,1,-1,1,-1,-1,1,-1,1,-1,-1,-1,1,1,1,-1,-1,1,-1,1,-1,-1,1,-1,1,-1,1,-1]^T,1)

ลองพิจารณาในแง่ของการรู้จำแบบ สมมติว่าเราต้องการเครื่องที่จะเรียนรู้การแมป (Mapping) $x_i \Rightarrow y_i$ ซึ่งเราจะนิยามให้เป็นเซตของการแมปที่เป็นไปได้ $x \Rightarrow f(x,\alpha)$ ซึ่ง α นี้สามารถแปรค่าได้ และเครื่องที่ได้รับการสอนแล้วก็จะมีค่า α ที่แน่นอนของตัวเอง อย่างเช่นในโครงข่ายประสาทเทียม ค่า α หมายถึงน้ำหนัก (Weight) และขีดแบ่ง (Threshold)

เรานิยามค่าความผิดพลาดสำหรับเครื่องที่ได้รับการสอนนี้เป็น

$$R(\alpha) = \int \frac{1}{2} |y - f(x,\alpha)| dP(x,y) \quad (1)$$

โดยที่ค่า $R(\alpha)$ เป็นความผิดพลาดที่แท้จริง และค่า $1/2 \left| y - f(x,\alpha) \right|$ จะมีค่าเป็น 0 หรือ 1 ส่วน $R_{emp}(\alpha)$ เป็นค่าความผิดพลาดโดยเฉลี่ยในข้อมูลสอนเท่านั้น และสามารถหาค่าได้ดังนี้

$$R_{emp}(\alpha) = \frac{1}{2\ell} \sum^{\ell} |y - f(x,\alpha)| \quad (2)$$

ค่า $R_{emp}(\alpha)$ จะคงที่สำหรับ α และข้อมูลสอน $\{(x_i,y_i)\}$ ชุดหนึ่งๆ เราเรียก $R_{emp}(\alpha)$ ว่าความเสี่ยงเชิงทดลอง

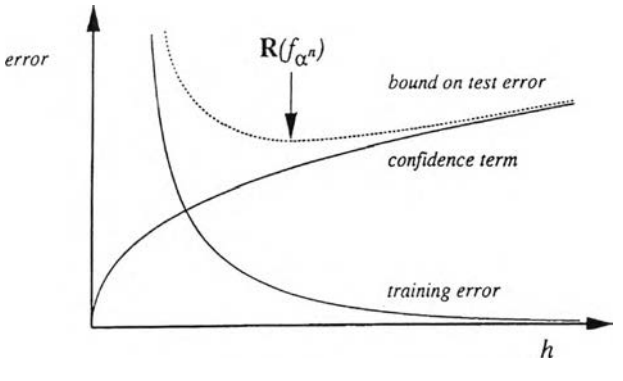
เมื่อเลือก η ตัวหนึ่ง โดยที่ $0 \leq \eta \leq 1$ จะได้ว่า ที่ระดับความเชื่อมั่น 1-η ความผิดพลาดที่แท้จริงจะมีขอบเขตดังนี้

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\frac{h\left(\log\left(\frac{2\ell}{h}\right)+1\right) - \log\left(\frac{\eta}{4}\right)}{\ell}} \qquad (3)$$

เราเรียก h ซึ่งเป็นจำนวนเต็มบวกหรือศูนย์ ว่ามิติ VC (Vapnik Chervonenkis Dimension) และเรียกพจน์ขวามือว่าขอบเขตความเสี่ยง (Risk bound) ซึ่งพบว่าขอบเขตนี้ไม่ขึ้นกับการแจกแจงความน่าจะเป็น P(x,y)

โดยทั่วไป เราไม่สามารถคำนวณค่าของพจน์ทางซ้ายมือได้ แต่ถ้ารู้ h จะสามารถคำนวณพจน์ทางขวามือได้ ดังนั้นโดยหลักการแล้ว ในการเรียนรู้ เราจึงกำหนดค่า η เป็นค่าคงที่ต่ำๆ แล้วเลือกเครื่องที่ลดค่าทางขวามือให้ต่ำที่สุด เราก็จะได้เครื่องที่ให้ขอบเขตความเสี่ยงต่ำที่สุด ซึ่งแนวคิดนี้เป็นแนวคิดที่สำคัญของการลดความเสี่ยงเชิงโครงสร้างให้ต่ำที่สุด

จากสมการ (3) เราเรียกพจน์ที่สองทางขวามือว่าช่วงความเชื่อมั่น พบว่าช่วงความเชื่อมั่นนี้เป็นฟังก์ชันของมิติ VC ที่แทนด้วย h ซึ่งการลดค่า h ให้ต่ำสุด จะเป็นการลดความเสี่ยงเชิงโครงสร้างด้วย แต่โดยปกติเมื่อค่า h ลดลงความเสี่ยงเชิงทดลองจะสูงขึ้น ดังนั้นในการลดความเสี่ยงเชิงโครงสร้างจึงต้องหาจุดที่ผลจากทั้งความเสี่ยงเชิงทดลองและช่วงความเชื่อมั่นต่ำที่สุด โดยเลือกฟังก์ชันในเครื่องเรียนรู้ $f(x,\alpha)$ ที่มีขอบเขตความเสี่ยงต่ำสุดดังในรูปที่ 2

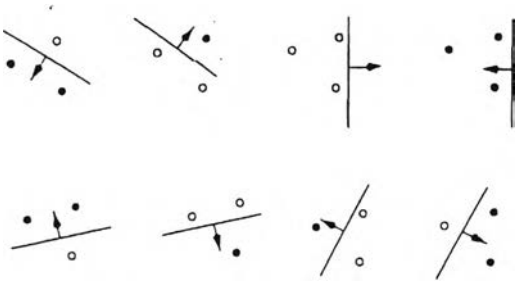


รูปที่ 2 กราฟแสดงความสัมพันธ์ระหว่าง VC (h) กับความผิดพลาด จากสมการ (3) [2]

2.1.1.2 มิติ VC (VC Dimension)

มิติ VC เป็นคุณสมบัติของเซตของฟังก์ชัน {f(α)} ที่แสดงถึงความสามารถของฟังก์ชันในการแบ่งแยกกลุ่มของจุดข้อมูลออกจากกัน โดยทั่วไปฟังก์ชันที่มีมิติ VC สูงจะมีความสามารถในการแบ่งแยกสูง ในที่นี้จะขออธิบายเฉพาะฟังก์ชันสำหรับกรณีการรู้จำแบบที่มี 2 กลุ่มเท่านั้น คือ $f(x,\alpha) \in \{-1,1\}\ \forall x,\alpha$

กำหนดให้มีเซตของจุด ℓ จุด ซึ่งสามารถกำหนดว่าจุดแต่ละจุดอยู่กลุ่มใดได้ทั้งหมด 2^{ℓ} แบบ และถ้าในแบบแต่ละแบบมีฟังก์ชันอย่างน้อยหนึ่งฟังก์ชันใน $\{ f(\alpha) \}$ ที่สามารถกำหนดกลุ่มให้กับจุดแต่ละจุดได้อย่างถูกต้อง จะเรียกได้ว่าเซตของจุดเหล่านี้ถูกแยก (Shattered) โดยเซตของฟังก์ชันนี้ ดังแสดงในรูปที่ 3



รูปที่ 3 เซตของจุด 3 จุดใน R^2 ถูกแยกโดยฟังก์ชันซึ่งเป็นเส้นที่มีทิศทาง [2]

เรานิยามมิติ VC ของเซตของฟังก์ชัน$\{ f(\alpha) \}$ ให้เป็นจำนวนสูงสุดของจุดในข้อมูลสอนที่สามารถแยกด้วย $\{ f(\alpha) \}$ ได้ โดยมีข้อสังเกตคือ ถ้าให้มิติ VC มีค่าเป็น h แล้ว จะมีอย่างน้อยหนึ่งเซตของจุดจำนวน h จุดที่ถูกแยกได้ แต่โดยทั่วไปไม่จำเป็นว่าทุกเซตของจุดจำนวน h จุดจะถูกแยกได้เสมอไป

<u>ทฤษฎีบทที่ 1</u> พิจารณาเซตที่ประกอบด้วยจุด m จุดใน R^n ถ้าเลือกจุดใดจุดหนึ่งเป็นจุดกำเนิด จะได้ว่าจุด m จุดนี้จะถูกแยกโดยระนาบหลายมิติแบบมีทิศทาง (Oriented Hyperplane) ได้ก็ต่อเมื่อเวกเตอร์บอกตำแหน่งของจุดที่เหลือเป็นอิสระเชิงเส้นต่อกัน

ผลที่ตามมาคือมิติ VC ของเซตของฟังก์ชันระนาบหลายมิติแบบมีทิศทางใน R^n มีค่าเป็น n+1 ซึ่งพิสูจน์ได้ดังนี้ เนื่องจากเราสามารถเลือกจุดจำนวน n+1 จุด แล้วให้จุดหนึ่งเป็นจุดกำเนิด ดังนั้นจุดที่เหลือ n จุดย่อมต้องเป็นอิสระเชิงเส้นต่อกันแน่นอน (เช่น ให้จุดแต่ละจุดอยู่บนแกนแต่ละแกนใน R^n)

มีข้อสังเกตว่าฟังก์ชันที่มีพารามิเตอร์มากไม่จำเป็นต้องมีมิติ VC มาก และในทางกลับกัน ฟังก์ชันที่มีพารามิเตอร์เพียงหนึ่งเดียว อาจมีมิติ VC เป็นอนันต์ได้ แต่แม้ฟังก์ชันจะมีมิติ VC เป็นอนันต์ก็อาจจะไม่สามารถแยกจุดเพียงไม่กี่จุดได้

ในทางปฏิบัติมิติ VC ที่ต่ำๆ ย่อมทำให้ขอบเขตของความผิดพลาดแท้จริงต่ำด้วย แต่ก็ไม่ได้หมายความว่าฟังก์ชันที่มีมิติ VC สูงๆ จะใช้งานได้ไม่ดี เช่นเอสวีเอ็มแบบอาร์บีเอฟ (RBF-Radial Basis Function) ซึ่งมีมิติ VC เป็นอนันต์ก็เป็นฟังก์ชันที่ใช้งานได้ดีฟังก์ชันหนึ่ง อย่างไรก็ตามเราสามารถควบคุมขอบเขตของมิติ VC ของฟังก์ชัน โดยแปรค่าความห่าง (σ) ได้ ดังนั้นการเลือกเคอร์เนลฟังก์ชันสำหรับเครื่องเรียนรู้ จึงไม่ได้มีรูปแบบที่แน่นอน หากแต่ต้องอ้างอิงจากผลการทดลอง เนื่องจากยังไม่มีข้อสรุปในทางทฤษฎีว่าเคอร์เนลแบบใดเหมาะสมกับปัญหาแบบใด [3]

แนวคิดเรื่องขอบเขตความเสี่ยง และมิติ VC นำไปสู่การคิดค้นเทคนิคการสอนเครื่องเรียนรู้แบบเอสวีเอ็ม ที่แตกต่างจากเทคนิคแบบเดิมอย่างโครงข่ายประสาทเทียม โดยความแตกต่างของทั้งสองเทคนิคจะอยู่ที่ลำดับของการลดความเสี่ยงระหว่าง ความเสี่ยงเชิงทดลองกับช่วงความเชื่อมั่น ซึ่งเป็นพจน์ทางขวาในสมการ (3)

2.1.1.3 วิธีการลดความเสี่ยงเชิงโครงสร้างให้ต่ำที่สุด

ขอบเขตความผิดพลาดอันอาจเกิดจากการทำให้มีลักษณะทั่วไปสามารถเขียนให้อยู่ในรูป

$$R(\alpha) \le R_{emp}(\alpha) + \Phi\left(\frac{\ell}{h}\right) \tag{4}$$

ซึ่งพจน์แรกคือ ความเสี่ยงเชิงทดลอง ส่วนพจน์หลังคือ ช่วงความเชื่อมั่นซึ่งเป็นฟังก์ชันของมิติ VC (h) ในการลดความเสี่ยงเชิงโครงสร้างให้ต่ำที่สุด (Structural Risk Minimization) เราจะต้องคำนึงถึงทั้งสองพจน์ของสมการข้างต้น โดยมีอยู่สองแนวทางดังนี้

1. แนวทางแรก คงค่าช่วงความเชื่อมั่นให้คงที่ แล้วลดความเสี่ยงเชิงทดลองให้ต่ำสุด

แนวทางนี้จะลดพจน์แรกของสมการให้ต่ำสุดโดยการออกแบบเครื่องเรียนรู้ให้ซับซ้อน แต่ถ้าออกแบบเครื่องที่ซับซ้อนเกินไป จะทำให้ช่วงความเชื่อมั่นกว้างขึ้น ซึ่งแม้จะสามารถลดความเสี่ยงเชิงทดลองลงจนหมดสิ้นไปได้ แต่ความผิดพลาดที่เกิดในการใช้งานจริงยังอาจสูงอยู่ ปรากฏการณ์ที่เกิดขึ้นนี้เรียกว่าโอเวอร์ฟิตติง [10] อย่างไรก็ตามถ้าเราเลือกเครื่องที่มีความซับซ้อนต่ำ เพื่อให้ช่วงความเชื่อมั่นแคบ เราก็จะประสบปัญหาในการหาฟังก์ชันที่จะนำมาใช้ประมาณปัญหาได้ยาก อันจะทำให้เกิดความผิดพลาดเชิงทดลองสูง เพื่อที่จะลดปัญหาการ

ประมาณที่ไม่ดีและโอเวอร์ฟิตติง เราจะต้องเลือกสถาปัตยกรรมของเครื่องที่เหมาะสม โดยอาศัยความรู้เกี่ยวกับลักษณะของปัญหา แล้วจึงหาฟังก์ชันในเครื่องนี้ที่สามารถลดจำนวนความผิดพลาดในข้อมูลสอนให้ต่ำที่สุดได้ วิธีการนี้มีการนำไปใช้ในโครงข่ายประสาทเทียม

2. แนวทางที่สอง คงค่าของความเสี่ยงเชิงทดลองให้คงที่(อาจเป็นศูนย์) แล้วลดช่วงความเชื่อมั่นให้ต่ำสุด

โดยการกำหนดขอบเขตความเสี่ยงเชิงทดลองสูงสุดที่ยอมรับได้ แล้วเปลี่ยนรูปแบบของฟังก์ชันเพื่อลดช่วงความเชื่อมั่นลงให้ต่ำสุด แนวทางนี้พบในเอสวีเอ็ม โดยเอสวีเอ็มจะแบ่งกลุ่มของฟังก์ชันออกเป็นเซตย่อย แล้วหาเซตของเคอร์เนลที่ให้ความเสี่ยงเชิงทดลองต่ำสุด จากนั้นจึงหาฟังก์ชันในเซตนั้นที่ให้มิติ VC ต่ำที่สุด แล้วบันทึกค่าขอบเขตความเสี่ยงที่ได้ ทำการพิจารณาเช่นเดิมกับเคอร์เนลกลุ่มถัดมาที่ให้ความเสี่ยงเชิงทดลองต่ำสุด ทำวนซ้ำจนได้ฟังก์ชันที่ให้ขอบเขตความผิดพลาดที่แท้จริงต่ำสุด

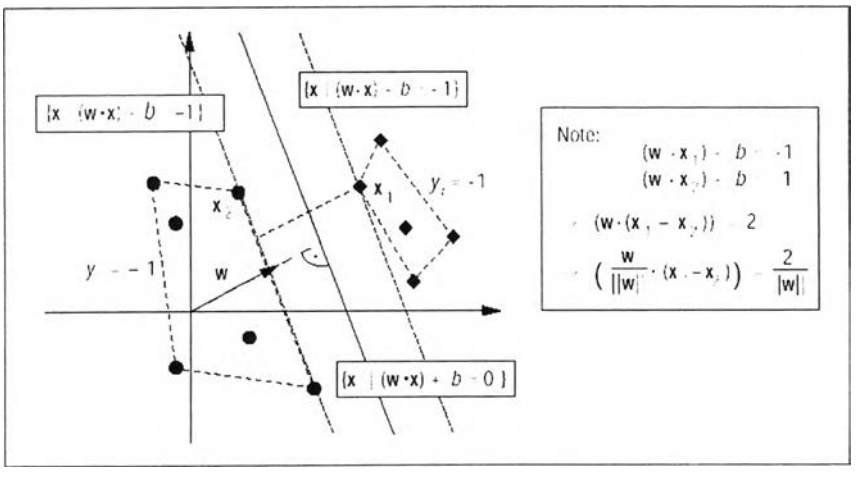
2.1.2 ระนาบหลายมิติที่ใช้แยกดีที่สุด (Optimal Separating Hyperplane)

ในการออกแบบขั้นตอนวิธีเรียนรู้ (Learning Algorithm) จะต้องใช้ฟังก์ชันที่สามารถคำนวณขีดความสามารถได้ ตัวจำแนกเอสวีเอ็มมีพื้นฐานจากฟังก์ชันประเภทระนาบหลายมิติ (Hyperplane) โดยที่ระนาบหลายมิติทำหน้าที่คล้ายเป็นตัวแยกเขตแดนระหว่างกลุ่มสองกลุ่ม ดังแสดงในสมการ (5)

$$(w \cdot x) + b = 0, \quad w \in R^N, b \in R \tag{5}$$

และสอดคล้องกับฟังก์ชันตัดสินใจ

$$f(x) = \text{sign}((w \cdot x) + b) \tag{6}$$



รูปที่ 4 ระนาบหลายมิติที่ใช้แยกดีที่สุด จะให้ระยะห่างระหว่างกลุ่มทั้งสองกลุ่มเป็น 2/||w||

ระนาบหลายมิติที่ใช้แยกดีที่สุดมีนิยามเป็นระนาบที่มีระยะห่างของการแบ่งแยกระหว่างกลุ่มสองกลุ่มมากที่สุด เราพบว่าระยะห่างระหว่างกลุ่มทั้งสองคือ $2/\|w\|$ ดังที่แสดงในรูปที่ 4 โดยที่ $y_i(w \cdot x_i + b) \geq 1$ ต้องเป็นจริงด้วย ปัญหานี้มีลากรองเกียนคือ

$$L(w, b, \alpha) = \tfrac{1}{2} w \cdot w - \sum \alpha_i((w \cdot x_i + b)y_i - 1) \quad (7)$$

ซึ่ง α คือตัวคูณลากรองก์ โดยต้องหาค่าต่ำสุดเมื่อเทียบกับ w, b และหาค่าสูงสุดเมื่อเทียบกับ $\alpha_i \geq 0$

ที่จุดที่ดีที่สุด คำตอบ w_0, b_0 และ α^0_i จะสอดคล้องกับคุณสมบัติของระนาบหลายมิติที่ใช้แยกดีที่สุดดังนี้

$$\sum \alpha^0_i y_i = 0, \quad \alpha^0_i \geq 0, \ i = 1,\ldots,\ell \quad (8)$$

$$w_0 = \sum \alpha^0_i y_i x_i, \quad \alpha^0_i \geq 0, \ i = 1,\ldots,\ell \quad (9)$$

โดยที่จริงแล้วเฉพาะสัมประสิทธิ์ α^0_i ของซัพพอร์ตเวกเตอร์เท่านั้นที่ไม่เป็นศูนย์ ดังนั้นในการหาค่าของเวกเตอร์ w_0 เราจึงหาจากผลรวมเชิงเส้นของข้อมูลตัวที่เป็นซัพพอร์ตเวกเตอร์เท่านั้น

$$b_0 = \frac{1}{2} [(w_0 \cdot x^*(1)) + (w_0 \cdot x^*(-1))] \quad (10)$$

โดยที่ $x^*(1)$ คือซัพพอร์ตเวกเตอร์ใดๆ ที่อยู่ในกลุ่ม 1 และ $x^*(-1)$ คือซัพพอร์ตเวกเตอร์ใดๆ ที่อยู่ในกลุ่ม –1 และจะได้ฟังก์ชันตัดสินใจในรูปแบบ

$$f(x) = \text{sign}(\textstyle\sum_{\text{Support Vector}} \nu_i (x_i \cdot x) + b_0), \qquad \nu_i = y_i \alpha^0_i$$

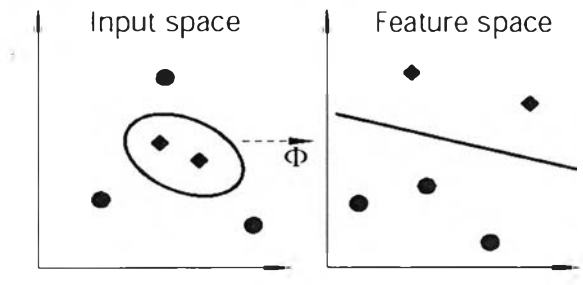
คำตอบนี้ใช้ได้เฉพาะกรณีที่สามารถแบ่งแยกแบบเชิงเส้นได้เท่านั้น แต่สำหรับกรณีที่แบ่งแยกแบบเชิงเส้นไม่ได้ ต้องมีการปรับข้อจำกัดเล็กน้อย คือ α_i จะมีขอบเขต $0 \leq \alpha_i \leq C$ โดยเราสามารถแปรค่า C ได้ โดยเป็นการแลกเปลี่ยนระหว่างความถูกต้องกับความเร็วในการสอน

สังเกตได้ว่า ทั้งการแก้ปัญหาการหาฟังก์ชันของระนาบหลายมิติและตัวฟังก์ชันการตัดสินใจต่างก็ขึ้นอยู่กับผลคูณเชิงสเกลาร์ระหว่างเวกเตอร์ สิ่งนี้เองที่ทำให้เราสามารถขยายขั้นตอนวิธีสำหรับกรณีที่ไม่เป็นเชิงเส้นได้ในปริภูมิอันดับสูง

2.1.3 ปริภูมิอันดับสูงและเคอร์เนล (Feature spaces and kernels)

การแมปข้อมูลเข้าไปสู่ปริภูมิอันดับสูงขึ้น จะช่วยให้สามารถแยกข้อมูลสองกลุ่มออกจากกันได้โดยใช้ฟังก์ชันเชิงเส้น โดยมีสมมติฐานว่าข้อมูลที่มีความสัมพันธ์ไม่เป็นเชิงเส้นใน

ปริภูมิอันดับต่ำ เมื่อแมปไปสู่ปริภูมิอันดับสูงจะมีความสัมพันธ์เป็นเชิงเส้นได้ อย่างไรก็ตามการแมปไปสู่ปริภูมิอันดับสูงอาจต้องการการคำนวณที่สูงเกินไป เคอร์เนลฟังก์ชันช่วยหลีกเลี่ยงปัญหาการคำนวณหาฟังก์ชันในการแมปได้ โดยยอมให้คำนวณผลคูณสเกลาร์ของตัวแปรสองตัวในปริภูมิอันดับสูงได้ โดยไม่ต้องคำนวณหาฟังก์ชันที่ใช้แมป การแสดงด้วยสมการจะช่วยให้เข้าใจถึงแนวคิดนี้ได้ง่ายขึ้น



รูปที่ 5 แนวคิดการแมปแบบไม่เป็นเชิงเส้นไปสู่ปริภูมิอันดับสูง

แนวคิดเบื้องต้นของเอสวีเอ็ม ดังแสดงในรูปที่ 5 เป็นการแมปข้อมูลไปสู่ปริภูมิผลคูณสเกลาร์อันดับสูงขึ้นไป (Feature Space, F) ผ่านการแมปแบบไม่เป็นเชิงเส้น

$$\Phi : R^N \rightarrow F \tag{11}$$

แล้วจึงใช้ขั้นตอนวิธีเชิงเส้นนี้ใน F ซึ่งสิ่งที่ต้องทำก็เพียงการหาค่าของผลคูณสเกลาร์

$$k(x,y) = (\Phi(x) \cdot \Phi(y)) \tag{12}$$

ถ้า F มีอันดับสูง ย่อมเท่ากับว่าพจน์ด้านขวามือของสมการ (12) จะคำนวณได้ยากมาก อย่างไรก็ตามในบางกรณีจะมีเคอร์เนล k ที่ง่ายต่อการคำนวณ ตัวอย่างเช่นเคอร์เนลแบบพหุนาม

$$k(x,y) = (x \cdot y)^d \tag{13}$$

ซึ่งสามารถแสดงให้เห็นว่าสอดคล้องกับการแมป Φ ไปสู่ปริภูมิที่สแปนโดยผลคูณทั้งหมดของอันดับ d ใน R^N ตัวอย่างเช่นในกรณีที่ d=2 และ $x, y \in R^2$ เรามี

$$\begin{aligned} (x \cdot y)^2 &= ((x_1,x_2) \cdot (y_1,y_2))^2 \\ &= ((x_1^2, \sqrt{2}\, x_1x_2, x_2^2) \cdot (y_1^2, \sqrt{2}\, y_1y_2, y_2^2)) \\ &= (\Phi(x) \cdot \Phi(y)) \end{aligned} \tag{14}$$

ที่นิยามให้ $\Phi(x) = (x_1^2, \sqrt{2}\, x_1x_2, x_2^2)$

นอกเหนือจากสมการ (13) แล้ว เอสวีเอ็มยังสามารถนำมาใช้กับเคอร์เนลแบบอาร์บีเอฟ

$$k(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}} \quad (15)$$

และเคอร์เนลแบบซิกมอยด์ (ที่มีเกน κ และออฟเซต Θ)

$$k(x,y) = \tanh(\kappa(x \cdot y) + \Theta) \quad (16)$$

2.1.4 ซัพพอร์ตเวกเตอร์แมชชีน (เอสวีเอ็ม)

เอสวีเอ็มเป็นเทคนิคการเรียนรู้ที่มีขีดความสามารถสูงสำหรับการจำแนกแบบสองกลุ่ม โดยสามารถลดความเสี่ยงเชิงโครงสร้างให้ต่ำสุด เอสวีเอ็มใช้ประโยชน์จากการแมปและเคอร์เนลฟังก์ชัน ซึ่งระนาบหลายมิติที่ใช้แยกกลุ่มข้อมูลจะอยู่ในปริภูมิอันดับสูง ถึงจุดนี้จะแทนที่ข้อมูลแต่ละตัวในชุดสอน x_i ด้วย $\Phi(x_i)$ และหาระนาบหลายมิติที่ใช้แยกดีที่สุด ใน F เนื่องจากเราใช้เคอร์เนล ดังนั้นจึงได้ผลเป็นฟังก์ชันตัดสินใจในรูปแบบ

$$f(x) = \text{sign}(\sum_{\text{Support Vector}} \nu_i \cdot k(x,x_i) + b_0),\ \nu_i = y_i \alpha^0_i \quad (17)$$

โดยที่พารามิเตอร์ ν_i สามารถคำนวณได้โดยเป็นคำตอบของปัญหา QP คล้ายแบบที่กล่าวมาแล้วดังนี้

ลดค่าของฟังก์ชันวัตถุประสงค์ให้ต่ำสุด

$$\frac{1}{2}\sum \alpha_i Q_j \alpha_j - \sum \alpha_i \quad (18)$$

โดยขึ้นกับข้อกำหนดต่อไปนี้

$$0 \le \alpha_i \le C$$

$$\sum y_i \alpha_i = 0$$

Q อยู่ในรูปของ $y_i\ y_j\ K(x_i, x_j)$ เป็นเมทริกซ์มิติ N×N ซึ่งขึ้นอยู่กับขนาดของข้อมูลสอน x_i และชื่อกลุ่ม y_i กับรูปแบบของฟังก์ชันที่จะใช้ ส่วน C เป็นค่าคงที่ที่สามารถกำหนดได้

ในปริภูมิอันดับสูงจะได้ระนาบหลายมิติเป็นแบบเชิงเส้น ส่วนในปริภูมิอันดับต่ำระนาบหลายมิติจะสอดคล้องกับฟังก์ชันการตัดสินใจแบบไม่เป็นเชิงเส้นซึ่งรูปแบบจะถูกกำหนดโดยเคอร์เนล และโดยการเปลี่ยนเคอร์เนล เราจะได้สถาปัตยกรรมที่ต่างออกไป ดังเช่นตัวจำแนกแบบพหุนาม ตัวจำแนกแบบอาร์บีเอฟ และโครงข่ายประสาทเทียมแบบสามระดับ เป็นต้น ในปัญหาต่างกัน เราต้องการเคอร์เนลที่อาจไม่เหมือนกัน แต่โดยปกติแล้ว ไม่ว่าจะใช้เคอร์เนลชนิดใด

ก็มักได้ซัพพอร์ตเวกเตอร์ชุดที่ใกล้เคียงกัน สิ่งนี้สนับสนุนแนวคิดที่ว่าซัพพอร์ตเวกเตอร์เป็นคุณลักษณะเฉพาะสำหรับปัญหาหนึ่งๆ [4]

2.2 งานวิจัยที่เกี่ยวข้อง

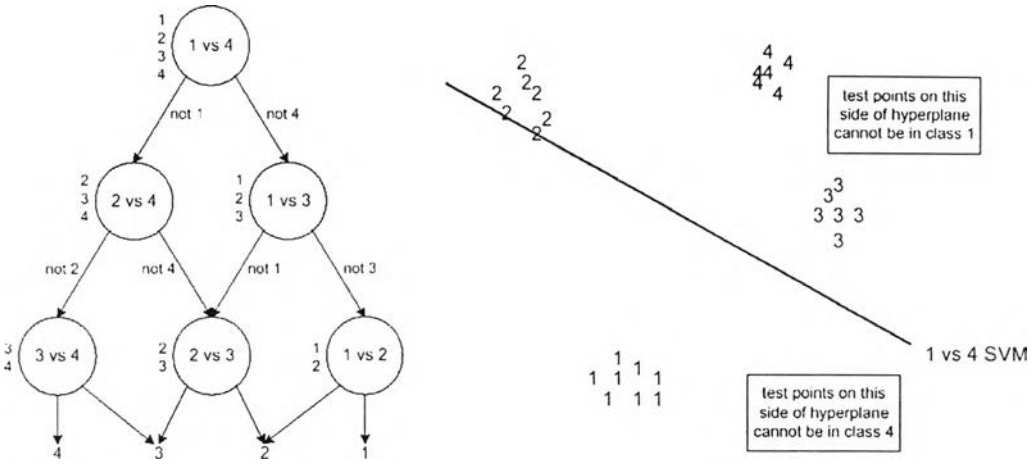
บทนี้กล่าวถึงงานวิจัยที่เกี่ยวกับวิทยานิพนธ์ฉบับนี้ ได้แก่ ดีเอจีเอสวีเอ็ม และ การใช้ซัพพอร์ตเวกเตอร์แมชชีนในการรู้จำเสียงภาษาไทย

2.2.1 ดีเอจีเอสวีเอ็ม (DAGSVM) [7]

งานวิจัยนี้เสนอเทคนิคการจำแนกแบบหลายกลุ่ม โดยใช้โทโปโลยีของกราฟมาวิเคราะห์ Platt และคณะเสนอสถาปัตยกรรมในการเรียนรู้แบบใหม่ที่ใช้ในการรวมตัวจำแนกแบบสองกลุ่มหลายตัวเข้าด้วยกันเป็นตัวจำแนกแบบหลายกลุ่ม สำหรับปัญหาที่ประกอบด้วยกลุ่มจำนวน N กลุ่มนั้น กราฟไม่มีวงแบบมีทิศทาง - ดีเอจี (Directed Acyclic Graph – DAG) จะประกอบด้วยตัวจำแนก N(N-1)/2 ตัว ตัวจำแนกแต่ละตัวใช้สำหรับคู่ของกลุ่มแต่ละคู่ นอกจากนี้งานวิจัยนี้เสนอผลการวิเคราะห์ถึงขอบเขตของความผิดพลาด โดยสรุปว่าขึ้นกับจำนวนกลุ่มและมาร์จินที่โนด (Node) แต่ไม่ขึ้นกับอันดับของปริภูมิ ผลที่ได้นี้ก่อให้เกิดขั้นตอนวิธีดีเอจีเอสวีเอ็มที่ใช้เคอร์เนล และใช้ระนาบหลายมิติที่ใช้แยกดีที่สุดสำหรับโนดตัดสินใจแต่ละโนดตัดสินใจของดีเอจี ดีเอจีเอสวีเอ็มให้ความถูกต้องที่เทียบได้กับของขั้นตอนวิธีมาตรฐานหรือขั้นตอนวิธีแมกซ์วิน แต่ให้ความเร็วที่สูงกว่าในขั้นตอนการจำแนกกลุ่ม

งานวิจัยนี้คำนึงถึงวิธีจำแนกแบบเดิมคือ 1-v-r ซึ่งพบว่ามีข้อจำกัดคือไม่มีขอบเขตของความผิดพลาดจากการทำให้เป็นลักษณะทั่วไป รวมทั้งเวลาที่ใช้ในการสอนยังเพิ่มขึ้นแบบเชิงเส้นตามจำนวนกลุ่มอีกด้วย ส่วนอีกวิธีคือ 1-v-1 นั้น Platt และคณะพบว่ามีข้อเสียคือถ้าตัวจำแนกแต่ละตัวไม่ถูกเรกูลาไรซ์ อย่างดีพอ เมื่อรวมเป็นระบบจำแนกแบบหลายกลุ่มแล้วอาจเกิดปัญหาโอเวอร์ฟิตติงได้ นอกจากนี้ยังไม่มีขอบเขตของความผิดพลาดอีกเช่นกัน รวมทั้งเวลาในการจำแนกกลุ่มยังเพิ่มขึ้นแบบมากกว่าเชิงเส้นเมื่อจำนวนกลุ่มเพิ่มขึ้นอีกด้วย

แนวคิดนี้เสนอตัวตัดสินใจโดยใช้โทโปโลยีกราฟไม่มีวงแบบมีทิศทางหรือดีดีเอจี โดยที่ราก (Root) ของดีดีเอจีเป็นโนดเดียวที่ไม่มีเส้นเชื่อม (edge) จากโนดอื่นชี้มา โนดแต่ละโนดจะมีเส้นเชื่อมชี้ออกไปไม่ศูนย์ก็สองเส้น โนดแต่ละโนดคือตัวจำแนกแบบสองกลุ่มสำหรับคู่ของกลุ่มแต่ละคู่ โดยในระดับของกราฟแต่ละชั้นจะขจัดกลุ่มที่เชื่อว่าไม่ใช่กลุ่มที่ถูกต้องออกไปทีละหนึ่งกลุ่ม โดยกลุ่มที่ยังเหลืออยู่จะถูกเปรียบเทียบกับกลุ่มอื่นที่เหลือต่อไปเรื่อยๆ จนเหลือกลุ่มเดียว ซึ่งจำนวนครั้งของการเปรียบเทียบทั้งหมดน้อยกว่าจำนวนกลุ่มอยู่หนึ่ง



รูปที่ 6 รูปซ้ายแสดงโทโปโลยีของดีเอจี และรูปขวาแสดงถึงการตัดสินใจในโนดของดีเอจี [7]

ในการจำแนกโดยใช้ดีดีเอจี ข้อมูลเข้าจะเริ่มจากโนดราก แล้วฟังก์ชันแบบทวิภาคที่โนดจะถูกนำมาคำนวณ ถ้าได้ค่าของฟังก์ชันน้อยกว่าศูนย์จะออกที่ขอบซ้าย ถ้าได้ค่ามากกว่าหรือเท่ากับศูนย์จะออกที่ขอบขวา หลังจากนั้นฟังก์ชันแบบทวิภาคที่โนดต่อมาก็จะถูกคำนวณ ค่าของฟังก์ชันตัดสินใจคือค่าของโนด เราเรียกโนดที่ใช้แยกระหว่างกลุ่ม i กับกลุ่ม j ว่าโนด ij สมมติให้จำนวนของใบเท่ากับจำนวนกลุ่ม โนดนี้คือโนดที่ i ในระดับที่ (N-j+i) โดยให้ i < j ดังในรูปที่ 6

วิธีการของดีดีเอจีนี้สามารถนำไปใช้ในรูปแบบของลิสต์ (List) ได้ ในตอนเริ่มต้นสมาชิกแต่ละตัวของลิสต์นี้จะแทนกลุ่มแต่ละกลุ่ม การจำแนกในโนดจะกระทำโดยเปรียบเทียบสมาชิกตัวแรกกับตัวสุดท้ายในลิสต์ ถ้าตัวจำแนกในโนดนั้นเลือกกลุ่มไหน กลุ่มอีกกลุ่มจะถูกขจัดออกไปจากลิสต์ ทำให้สมาชิกของลิสต์ลดลงทีละหนึ่งตัว ดำเนินการเช่นนี้ไปเรื่อยจนเหลือสมาชิกตัวสุดท้ายในลิสต์ซึ่งเป็นคำตอบ สำหรับปัญหาแบบ N กลุ่ม จะใช้โนดตัดสินใจ N-1 โนด

ผลการวิเคราะห์พบว่า การใช้ระนาบหลายมิติที่แบ่งกลุ่มสองกลุ่มให้อยู่ห่างกันมากที่สุดในแต่ละโนด จะลดขอบเขตของความผิดพลาดให้ต่ำที่สุดได้ โดยขอบเขตนี้เป็นอิสระจากมิติของข้อมูลเข้า เวลาที่ใช้ในการสอนและประเมินค่าต่างก็ดีกว่าขั้นตอนวิธีที่มีอยู่ดังแสดงในตารางที่ 1 โดยที่ c เป็นค่าคงที่, m เป็นขนาดชุดสอน, N เป็นจำนวนกลุ่ม

ตารางที่ 1 เวลาในการสอนเอสวีเอ็มแบบหลายกลุ่มตามขั้นตอนวิธีทั้งสาม [7]

ขั้นตอนวิธี	เวลาในการสอนโดยประมาณ	เวลาในการสอนโดยประมาณ ($\gamma \approx 2$)
1-v-r (1 class)	cm^{γ}	cm^2
1-v-r (N classes)	cNm^{γ}	Ncm^2
1-v-1 (N classes)	$2^{\gamma-1}cN^{2-\gamma}m^{\gamma}$	$2cm^2$

พบว่าในกรณีทั่วไปที่ $\gamma \approx 2$ เวลาที่ใช้ในการสอนสำหรับ 1-v-1 ไม่ขึ้นกับจำนวนกลุ่มและเป็นเพียง 2 เท่าของเวลาที่ใช้ในการสอนตัวจำแนก 1-v-r ตัวเดียว ดังนั้นการใช้ 1-v-1 ด้วยขั้นตอนวิธีนี้จึงดีกว่าในแง่เวลาที่ใช้ในขั้นตอนสอน

นอกจากนี้ Platt และคณะยังได้ทดลองเปรียบเทียบกับขั้นตอนวิธีดีดีเอจีกับ 1-v-r และแม็กซ์วิน (Max Wins) ซึ่งเป็นขั้นตอนวิธี 1-v-1 ประเภทหนึ่ง โดยทำการทดลองกับชุดข้อมูลมาตรฐานคือ USPS และ UCI Letter ซึ่งผลการทดลองดังแสดงในตารางที่ 2 สนับสนุนผลการวิเคราะห์ข้างต้น

ตารางที่ 2 ผลการทดลองเปรียบเทียบการจำแนกแบบหลายกลุ่ม โดยใช้ชุดข้อมูลมาตรฐาน 2 ชุด

	Kernel Chosen	σ	C	Number of Errors	Number of Kernel Evaluations	Training CPU Time (sec)
USPS						
1-v-r	Gaussian	3.58	100	92/2007	2,994	3,968
Max Wins	Gaussian	5.06	100	91/2007	1,887	326
DAGSVM	Gaussian	5.06	100	88/2007	828	326
UCI Letter						
1-v-r	Gaussian	0.447	100	86/4000	8,254	3,775
Max Wins	Gaussian	0.632	100	96/4000	7,320	744
DAGSVM	Gaussian	0.447	10	90/4000	3,844	1,738

งานวิจัยดังกล่าวได้เสนอขั้นตอนวิธีที่เรียบง่าย และมีประสิทธิภาพในการนำไปใช้จริง อย่างไรก็ตามงานวิจัยดังกล่าวยังมีประเด็นที่สามารถพัฒนาได้อีก โดย Platt และคณะสรุปว่าลำดับของตัวจำแนกสำหรับโนดแต่ละโนดไม่มีผลกระทบต่อความถูกต้องในการจำแนก ดังนั้นในการเลือกลำดับของสมาชิกในลิสต์จึงเสนอให้ใช้วิธีสุ่มเลือกลำดับของสมาชิกในลิสต์มาใช้

2.2.2 การใช้ซัพพอร์ตเวกเตอร์แมชชีนในการรู้จำเสียงภาษาไทย [9]

ณัฐกร และบุญเสริม ได้วิจัยโดยใช้เอสวีเอ็มในการรู้จำเสียงภาษาไทย โดยทดลองกับปัญหา 2 อย่าง คือ การรู้จำโทนเสียง 5 กลุ่ม และการรู้จำเสียงสระ 12 กลุ่ม โดยใช้เอสวีเอ็มแบบหลายกลุ่ม 3 ประเภทคือแบบ 1-v-r 1-v-1 และแบบดีดีเอจี เปรียบเทียบกับเพอร์เซปตรอนแบบหลายชั้น (Multi-Layer Perceptron) ผลที่ได้พบว่าเอสวีเอ็มแบบหลายกลุ่มทุกประเภทให้ผลได้ดีกว่าเพอร์เซปตรอนแบบหลายชั้น ทั้งในด้านเวลาในการสอนและผลการรู้จำที่ได้

ข้อมูลที่ใช้ในการทดลองได้จากเสียงพูดของคนไทย 24 คน เป็นชาย 8 คน (M1-M8) และเป็นหญิง 16 คน (F1-F16) การทดลองแบ่งออกเป็น 2 กลุ่มได้แก่ แบบภายใน (Inside) และแบบภายนอก (Outside) แบบภายในจะใช้ข้อมูลเสียงพูดชุดเดียวกันทั้งในชุดสอนและชุดทดสอบ โดยเป็นข้อมูลสอนจำนวน 12,384 ตัว และข้อมูลทดสอบ 3,096 ตัว ส่วนแบบภายนอกจะใช้ข้อมูลจากผู้พูดที่แตกต่างกันในการสอนและการทดสอบ โดยเป็นข้อมูลสอนจำนวน 6,192 ตัว และข้อมูลทดสอบ 3,096 ตัว

ในการทดลองแรกคือการรู้จำโทนเสียงภาษาไทย มีโทนเสียง 5 กลุ่ม คือ สามัญ เอก โท ตรี และจัตวา โดยข้อมูลมีลักษณะสำคัญที่ใช้ในการสอนและทดสอบมีลักษณะสำคัญ 6 ตัว ผลการทดสอบปรากฏว่าดีดีเอจีให้ผลการรู้จำที่ดีที่สุดในการทดลองส่วนใหญ่ และผลการรู้จำในแบบภายนอกด้อยกว่าแบบภายในประมาณ 5% ในด้านประสิทธิภาพทางเวลาในการจำแนกพบว่าดีดีเอจีเป็นขั้นตอนวิธีที่เร็วที่สุดในทุกการทดลอง

อีกการทดลองหนึ่งซึ่งซับซ้อนกว่าปัญหาแรกคือการรู้จำเสียงสระภาษาไทย เลือกเสียงสระ 12 กลุ่ม โดยข้อมูลที่ใช้ในการสอนและทดสอบมีลักษณะสำคัญ 72 ตัว ผลการทดสอบปรากฏว่าขั้นตอนวิธี 1-v-r และแมกซ์วินให้ผลการรู้จำที่ดีที่สุดในการทดลองส่วนใหญ่ และผลการรู้จำในแบบภายนอกด้อยกว่าแบบภายในประมาณ 8% ในด้านประสิทธิภาพทางเวลาในการจำแนก พบว่าดีดีเอจีเป็นขั้นตอนวิธีที่เร็วที่สุดในทุกการทดลอง ซึ่งสอดคล้องกับที่พบในปัญหาการรู้จำโทนเสียง