บทที่ 3

เทคนิคการจำแนกข้อมูลแบบแตกครึ่งตามสารสนเทศ

บทนี้จะกล่าวถึงรายละเอียดของเทคนิคการจำแนกข้อมูลแบบแตกครึ่งตามสารสนเทศซึ่งแบ่งเนื้อหาออกเป็นสองส่วน ส่วนแรกกล่าวถึงแนวคิดเบื้องต้นในการสร้างต้นไม้สำหรับการจำแนกแบบหลายประเภท และส่วนที่สองกล่าวถึงขั้นตอนในการสร้างต้นไม้สำหรับการจำแนกแบบหลายประเภท

3.1 แนวคิดเบื้องต้นในการสร้างต้นไม้สำหรับการจำแนกแบบหลายประเภท

แนวคิดเบื้องต้นของเทคนิคนี้สร้างต้นไม้โดยพิจารณาจากความน่าจะเป็นของข้อมูลแต่ละประเภทซึ่งสามารถอธิบายได้เช่นเดียวกับปัญหาการเข้ารหัสข้อมูล กรณีตัวอย่างจากปัญหาการเข้ารหัสข้อมูลของอักขระ 4 ตัวคือ A, B, C และ D โดยสมมติให้ความน่าจะเป็นของการเกิดข้อมูลแต่ละประเภทมีเท่ากัน จะได้ว่าจำนวนบิตที่น้อยที่สุดในการแทนอักขระเหล่านี้คือ 2 บิตโดยแทนแต่ละอักขระดังนี้

A	แทนด้วย	00
B	แทนด้วย	01
C	แทนด้วย	10
D	แทนด้วย	11

ตัวอย่าง หากต้องการเข้ารหัสข้อมูลของ ABACADAB จะได้ 0001001000110001 จำนวนบิตที่ใช้ คือ 16 เมื่อพิจารณาข้อมูลจากตัวอย่าง ตลอดจนข้อมูลที่เป็นข้อความทั่วไปในชีวิตประจำวันจะมีความน่าจะเป็นในการเกิดข้อมูลของแต่ละอักขระไม่เท่ากัน อาทิ อักขระ A, E, I, O และ U ซึ่งเป็นสระในภาษาอังกฤษจะมีความน่าจะเป็นในการเกิดมากกว่าอักขระอื่นๆ

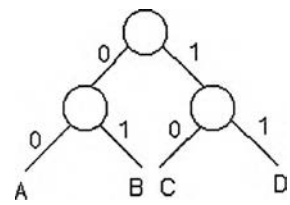
จากตัวอย่างเดิมสมมติว่าความน่าจะเป็นในการเกิดแต่ละอักขระเป็นดังนี้

A	มีความน่าจะเป็นคือ 1/2 แทนรหัสเดิมด้วย	0
B	มีความน่าจะเป็นคือ 1/4 แทนรหัสเดิมด้วย	10
C	มีความน่าจะเป็นคือ 1/8 แทนรหัสเดิมด้วย	110
D	มีความน่าจะเป็นคือ 1/8 แทนรหัสเดิมด้วย	111

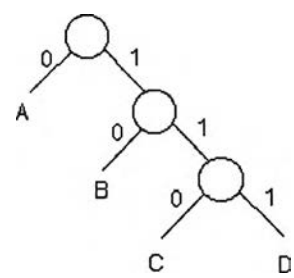
หากต้องการเข้ารหัสข้อมูลชุดเดิมคือ ABACADAB จะได้ 01001100111010 จำนวนบิตที่ใช้ คือ 14 จะเห็นว่าเมื่อนำความน่าจะเป็นในการเกิดของแต่ละอักขระ มาพิจารณา

ด้วย จะทำให้จำนวนบิตเฉลี่ยในการเข้ารหัสข้อมูลลดลง โดยอาศัยหลักพื้นฐานที่ว่าหากอักขระใดใช้บ่อย ก็ให้แทนด้วยจำนวนบิตน้อยๆ ส่วนอักขระใดไม่ค่อยใช้ก็ยอมให้แทนด้วยจำนวนบิตที่ยาวกว่าได้

เมื่อพิจารณาในกรณีแรกหากความน่าจะเป็นในการเกิดข้อมูลของแต่ละอักขระมีค่าเท่ากันก็เหมาะสมแล้วที่จะแทนอักขระแต่ละตัวด้วยรหัสไบนารีจำนวน 2 บิต เพื่อให้ง่ายแก่การเข้าใจจึงแสดงวิธีการแทนอักขระกรณีที่ความน่าจะเป็นในการเกิดข้อมูลของแต่ละอักขระเท่ากันและไม่เท่ากันด้วยรูปที่ 12 (ก) และ (ข) ตามลำดับ



รูปที่ 12 (ก) การแทนอักขระด้วยต้นไม้แบบไบนารี กรณีที่ความน่าจะเป็นในการเกิดของแต่ละอักขระเท่ากัน



รูปที่ 12 **(ข)** การแทนอักขระด้วยต้นไม้แบบไบนารี กรณีที่ความน่าจะเป็นในการเกิดของแต่ละอักขระไม่เท่ากัน

จากรูปที่ 12 (ก) และ (ข) ซึ่งแสดงการแทนอักขระด้วยต้นไม้แบบไบนารีจะเห็นว่าในรูปที่ 12 (ก) กรณีที่ความน่าจะเป็นในการเกิดของแต่ละอักขระเท่ากัน ต้นไม้แบบไบนารีที่สมดุลจะให้จำนวนเส้นเชื่อม (edge) จากรากไปถึงแต่ละใบนั้นเท่ากัน แต่ในรูปที่ 12 (ข) กรณีที่ความน่าจะเป็นในการเกิดของแต่ละอักขระไม่เท่ากันจะให้จำนวนเส้นเชื่อมจากรากไปถึงใบแต่ละใบด้วยจำนวนเส้นเชื่อมมากน้อยต่างกันขึ้นกับความน่าจะเป็นในการเกิดของอักขระนั้น เช่น อักขระ A มีความน่าจะเป็นสูงสุดจะให้จำนวนเส้นเชื่อมจากรากไปถึงใบต่ำที่สุด

ส่วนอักขระ C และ D จะมีจำนวนเส้นเชื่อมจากรากไปถึงใบมากที่สุดเนื่องจากมีความน่าจะเป็นในการเกิดต่ำสุด

หากพิจารณาในแง่ของการค้นหาข้อมูลของรูปที่ 12 (ข) อักขระ A เกิดบ่อย จึงแทนด้วยจำนวนบิตต่ำ หรือกล่าวในเชิงการค้นหาข้อมูลคือ จำนวนครั้งในการค้นหาต่ำ ส่วนอักขระ C และ D มีโอกาสเกิดน้อย จึงแทนด้วยจำนวนบิตที่ยาวได้หรือจำนวนครั้งในการค้นหาสูงได้ เมื่อพิจารณาการค้นหาโดยรวมจึงทำให้เวลาการค้นหาเฉลี่ยต่ำสุด

จากข้างต้นการสร้างต้นไม้สำหรับการจำแนกแบบหลายประเภทซึ่งเทียบได้กับการเข้ารหัสข้อมูลโดยแต่ละโนดคือ ตัวจำแนกแบบสองประเภทที่เลือกมา เมื่อพิจารณาความลึกของต้นไม้เฉลี่ยซึ่งขึ้นกับความน่าจะเป็นของการเกิดข้อมูลในแต่ละประเภท ต้นไม้สำหรับการจำแนกที่ได้ย่อมเป็นต้นไม้ที่มีความลึกเฉลี่ยสั้นที่สุดหรือให้จำนวนครั้งในการจำแนกเฉลี่ยต่ำสุดด้วยนั่นเอง

ผู้วิจัยได้เสนอวิธีการใหม่เรียกว่า เทคนิคการจำแนกข้อมูลแบบแตกครึ่งตามสารสนเทศ ซึ่งมีแนวคิดในการจำแนกข้อมูลด้วยระนาบหลายมิติที่แบ่งข้อมูลออกเป็นสองส่วนได้ดีที่สุดโดยพิจารณาจากความน่าจะเป็นในการเกิดข้อมูลแต่ละประเภท

3.2 ขั้นตอนในการสร้างต้นไม้สำหรับการจำแนกแบบหลายประเภทด้วยการแตกครึ่งตามสารสนเทศ

- **ขั้นตอนที่ 1 สร้างตัวจำแนกแบบสองประเภททุกแบบที่เป็นไปได้**

พิจารณาข้อมูลสอนในรอบนั้นๆ ว่าอยู่ในประเภทใดบ้าง แล้วจึงสร้างตัวจำแนกแบบสองประเภททุกแบบที่เป็นไปได้ซึ่งจะมีจำนวนเป็น $k(k-1)/2$ แบบ เมื่อจำนวนประเภทของข้อมูลสอนในรอบนั้นๆ เป็น k ประเภท ตัวอย่างกรณีข้อมูลสอนในรอบที่พิจารณามีจำนวน 4 ประเภทได้แก่ ประเภทที่ 1, 2, 3 และ 4 จะได้ว่า ตัวจำแนกแบบสองประเภททุกแบบที่เป็นไปได้จะมีจำนวนเป็น 4*(4-1)/2 =6 แบบได้แก่ 1-2, 1-3, 1-4, 2-3, 2-4 และ 3-4

- **ขั้นตอนที่ 2 การเลือกตัวจำแนกแบบสองประเภทเพื่อนำมาสร้างโนดของต้นไม้สำหรับการจำแนกแบบหลายประเภท**

สิ่งที่วิธีนี้ใช้ในการเลือกตัวจำแนกแบบสองประเภทมาใช้ในการสร้างโนดของต้นไม้คือ การเลือกตัวจำแนกที่สามารถแบ่งข้อมูลสอนแล้วให้ข้อมูลสอนที่ตกอยู่แต่ละด้านของระนาบมีจำนวนประเภทปนกันน้อยที่สุดหรือกล่าวอีกนัยหนึ่งว่าจะเลือกตัวจำแนกที่แบ่งข้อมูลแล้วข้อมูลที่เป็นประเภทเดียวกันก็ตกไปอยู่ในด้านเดียวกันของระนาบหรือถ้าหากมีประเภทอื่นปนอยู่บ้างก็จะเลือกตัวจำแนกที่แบ่งแล้วมีประเภทอื่นปนอยู่น้อยที่สุดโดยในที่นี้จะใช้ค่าเอนโทรปีต่ำสุดในการพิจารณาจากสูตรดังนี้

$$\text{ค่าเอนโทรปีของตัวจำแนก} = \sum_{i=1}^{n} \frac{|t_i|}{|T|} I(t_i)$$

$$I(t_i) = \sum - P(m_{ij}) \log_2 P(m_{ij}) \quad (19)$$

โดยที่ n เป็นจำนวนกิ่งของต้นไม้ในที่นี้จะมีค่าเป็น 2 ซึ่งได้แก่ กิ่งบวกและกิ่งลบ

k เป็นจำนวนประเภทของข้อมูลทั้งหมด

$P(m_{ij})$ เป็นความน่าจะเป็นในการเกิดข้อมูลประเภท j ในกิ่ง t_i

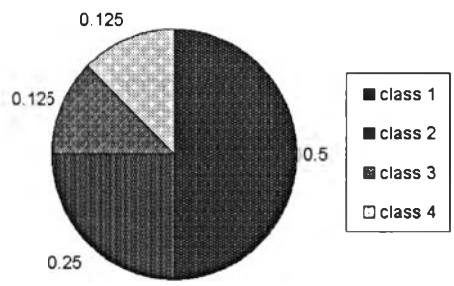
- **ขั้นตอนที่ 3**

ทำการแบ่งประเภทข้อมูลสอนในตอนเริ่มต้นด้วยตัวจำแนกแบบสองประเภทที่ได้มาจากขั้นตอนที่ 2 ซึ่งจะได้ข้อมูลสอน 2 ส่วน คือ ข้อมูลสอนที่เป็นบวกและข้อมูลสอนที่เป็นลบ จากนั้นจึงพิจารณาข้อมูลสอนแต่ละส่วนว่าในข้อมูลสอนส่วนนั้นๆ มีกี่ประเภท

- กรณีพบว่าเป็นประเภทเดียวแล้วก็ไม่ต้องดำเนินการต่อ
- กรณีพบว่าข้อมูลสอนที่พิจารณามีจำนวนประเภทเพียง 2 สองประเภทจึงเลือกตัวจำแนกแบบสองประเภทดังกล่าวมาสร้างโนดได้ทันที
- กรณีพบว่าข้อมูลสอนที่พิจารณามีจำนวนประเภทมากกว่าสองประเภทจึงกลับไปทำซ้ำในขั้นตอนที่ 1

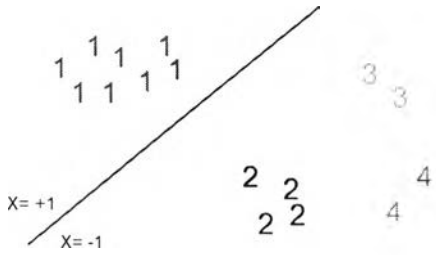
ตัวอย่าง

กรณีปัญหาการจำแนก 4 ประเภทซึ่งมีความน่าจะเป็นในการเกิดข้อมูลของแต่ละประเภทเป็นดังรูปที่ 13

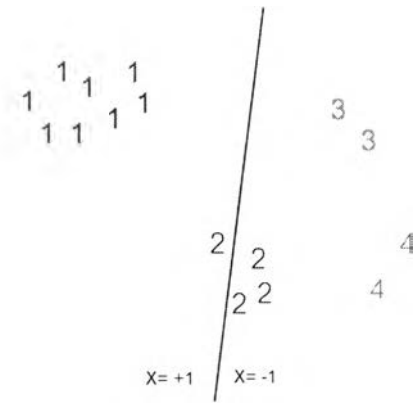


รูปที่ 13 ความน่าจะเป็นในการเกิดข้อมูลแต่ละประเภท ของปัญหา 4 ประเภท

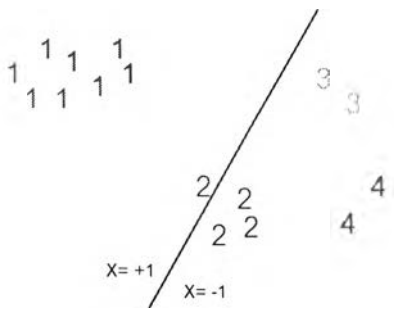
ขั้นตอนที่ 1 สร้างตัวจำแนกแบบสองประเภททุกแบบที่เป็นไปได้ซึ่งจะได้ว่า ตัวจำแนกแบบสองประเภททุกแบบที่เป็นไปได้จะมีจำนวนเป็น k (k -1)/2=4*(4-1)/2 =6 แบบ ได้แก่ 1-2, 1-3, 1-4, 2-3, 2-4 และ 3-4 ซึ่งสามารถจำลองตัวจำแนกที่ได้ดังรูปที่ 14 (ก-ฉ)



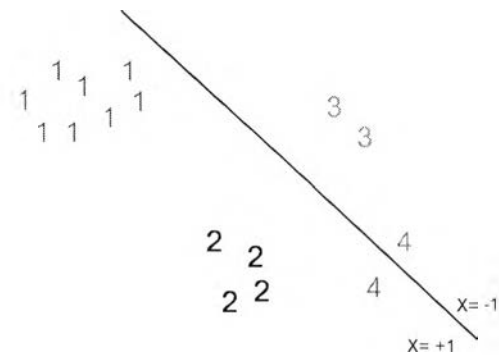
(ก) ตำแหน่งของข้อมูลประเภทอื่นๆ
เมื่อเทียบกับตัวจำแนกประเภท 1-2



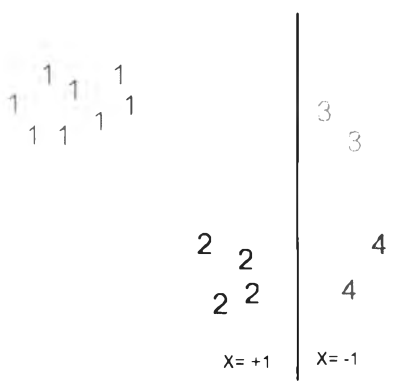
(ข) ตำแหน่งของข้อมูลประเภทอื่นๆ
เมื่อเทียบกับตัวจำแนกประเภท 1-3



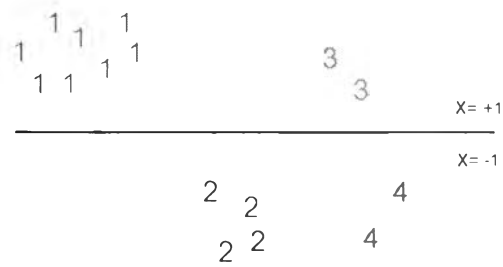
(ค) ตำแหน่งของข้อมูลประเภทอื่นๆ
เมื่อเทียบกับตัวจำแนกประเภท 1-4



(ง) ตำแหน่งของข้อมูลประเภทอื่นๆ
เมื่อเทียบกับตัวจำแนกประเภท 2-3



(จ) ตำแหน่งของข้อมูลประเภทอื่นๆ
เมื่อเทียบกับตัวจำแนกประเภท 2-4



(ฉ) ตำแหน่งของข้อมูลประเภทอื่นๆ
เมื่อเทียบกับตัวจำแนกประเภท 3-4

รูปที่ 14 การแบ่งข้อมูลด้วยตัวจำแนกแบบสองประเภททุกแบบที่เป็นไปได้ กรณีปัญหา 4 ประเภท

ขั้นตอนที่ 2 เลือกตัวจำแนกแบบสองประเภทเพื่อนำมาสร้างโนดของต้นไม้สำหรับการจำแนกแบบหลายประเภทโดยเลือกตัวจำแนกที่ให้ค่าเอนโทรปีต่ำสุดในรอบนั้นๆ

สำหรับในรอบแรกนั้นสามารถคำนวณค่าเอนโทรปีของตัวจำแนกต่างๆ ได้ดังนี้

ค่าเอนโทรปีของตัวจำแนก 1-2

$$= \frac{8}{16}\left(-\frac{8}{8}\log_2\frac{8}{8}\right) + \frac{8}{16}\left(-\frac{4}{8}\log_2\frac{4}{8} - \frac{2}{8}\log_2\frac{2}{8} - \frac{2}{8}\log_2\frac{2}{8}\right)$$

$$= 0.75$$

ค่าเอนโทรปีของตัวจำแนก 1-3

$$= \frac{9}{16}\left(-\frac{8}{9}\log_2\frac{8}{9} - \frac{1}{9}\log_2\frac{1}{9}\right) + \frac{7}{16}\left(-\frac{3}{7}\log_2\frac{3}{7} - \frac{2}{7}\log_2\frac{2}{7} - \frac{2}{7}\log_2\frac{2}{7}\right)$$

$$= 0.879$$

ค่าเอนโทรปีของตัวจำแนก 1-4

$$= \frac{9}{16}\left(-\frac{8}{9}\log_2\frac{8}{9} - \frac{1}{9}\log_2\frac{1}{9}\right) + \frac{7}{16}\left(-\frac{3}{7}\log_2\frac{3}{7} - \frac{2}{7}\log_2\frac{2}{7} - \frac{2}{7}\log_2\frac{2}{7}\right)$$

$$= 0.879$$

ค่าเอนโทรปีของตัวจำแนก 2-3

$$= \frac{13}{16}\left(-\frac{8}{13}\log_2\frac{8}{13} - \frac{4}{13}\log_2\frac{4}{13} - \frac{1}{13}\log_2\frac{1}{13}\right) + \frac{3}{16}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right)$$

$$= 1.179$$

ค่าเอนโทรปีของตัวจำแนก 2-4

$$= \frac{12}{16}\left(-\frac{8}{12}\log_2\frac{8}{12} - \frac{4}{12}\log_2\frac{4}{12}\right) + \frac{4}{16}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right)$$

$$= 0.939$$

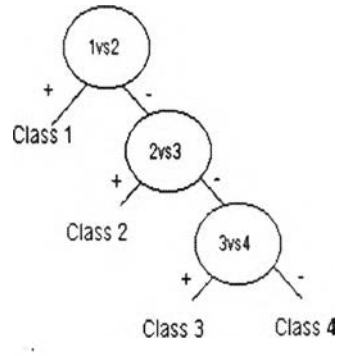
ค่าเอนโทรปีของตัวจำแนก 3-4

$$= \frac{10}{16}\left(-\frac{8}{10}\log_2\frac{8}{10} - \frac{2}{10}\log_2\frac{2}{10}\right) + \frac{6}{16}\left(-\frac{4}{4}\log_2\frac{4}{4} - \frac{2}{4}\log_2\frac{2}{4}\right)$$

$$= 0.796$$

ดังนั้นตัวจำแนกที่ให้ค่าเอนโทรปีต่ำสุดคือ 1-2 จึงได้ว่าโนดแรกหรือโนดรากของต้นไม้สำหรับการจำแนกแบบหลายประเภทของปัญหานี้ คือ ตัวจำแนกแบบสองประเภท1-2 นั่นเอง

ในรอบต่อๆ มาก็ใช้วิธีการเดียวกันโดยต้นไม้สำหรับการจำแนกแบบหลายประเภทที่สร้างจากค่าเอนโทรปีของข้อมูลชุดดังกล่าวกรณีที่ดีที่สุดแสดงดังรูปที่ 15



รูปที่ 15 ต้นไม้สำหรับการจำแนกแบบหลายประเภทที่สร้างจากค่าเอนโทรปี กรณีดีที่สุดสำหรับปัญหา 4 ประเภท

นอกจากนี้เราสามารถคำนวณค่าจำนวนครั้งในการจำแนกที่คาดหวังซึ่งเป็นกรณีดีที่สุดได้จากสูตร

$$\sum_{i=1}^{k} -P(m_i)\log_2 P(m_i) \tag{20}$$

สำหรับ ปัญหาการจำแนก k ประเภท และ $P(m_i)$ คือความน่าจะเป็นในการเกิดข้อมูลของประเภท m_i

กรณีตัวอย่างจากรูปที่ 13 สามารถคำนวณค่าจำนวนครั้งในการจำแนกที่คาดหวังได้ดังนี้

$$\sum_{i=1}^{4} -P(m_i)\log_2 P(m_i)$$

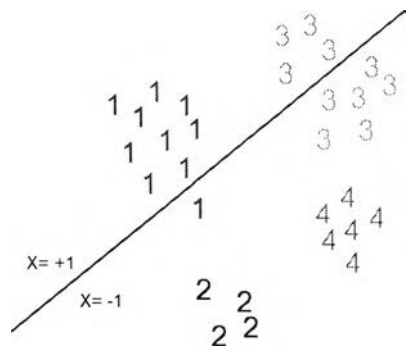
$= [-(0.5)^* \log_2(0.5)]+[-(0.25)^* \log_2(0.25)]$

$\quad + [-(0.125)^* \log_2(0.125)] + [-(0.125)^* \log_2(0.125)]$

$= 1.75$

3.3 การตัดเล็มและการกำหนดขอบเขตความผิดพลาด

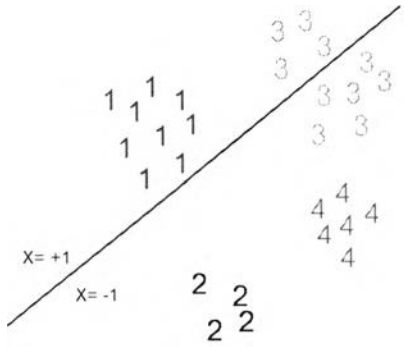
การใช้งานเทคนิคการแตกครึ่งตามสารสนเทศให้ค่าจำนวนครั้งในการจำแนกที่คาดหวังเพียง $\sum_{i=1}^{k} -P(m_i)\log_2 P(m_i)$ สำหรับ ปัญหาการจำแนก k ประเภท เมื่อ $P(m_i)$ คือความน่าจะเป็นในการเกิดข้อมูลของประเภท m_i โดยกรณีนี้จะเกิดได้กรณีระนาบที่สร้างได้จำแนกข้อมูลแล้วให้ผลลัพธ์ข้อมูลทุกตัวที่อยู่ในประเภทเดียวกันตกอยู่ในด้านใดด้านหนึ่งของระนาบเท่านั้น ซึ่งในความเป็นจริงกรณีนี้เป็นไปได้ยาก ดังนั้นจึงจำเป็นต้องทำการตัดเล็มโครงสร้างของการจำแนกโดยการกำหนดค่าร้อยละในตัดเล็ม (P : Pruning Percentage) เพื่อกำจัดประเภทข้อมูลที่มีตำแหน่งอยู่ทั้งสองด้านของระนาบในด้านที่มีจำนวนร้อยละน้อยกว่าที่เรากำหนดไว้ทิ้งไป เพื่อให้ระนาบสามารถจำแนกข้อมูลได้รวดเร็วขึ้น โดยการทำให้ระนาบแต่ละตัวไม่นำประเภทข้อมูลที่อยู่ในด้านที่มีจำนวนร้อยละน้อยกว่าค่าร้อยละในการตัดเล็มมาใช้ในการคำนวณหาค่าเอนโทรปีของข้อมูลที่เหลือในแต่ละด้าน ตัวอย่างจากรูปที่ 16 แสดงให้เห็นว่ามีประเภทข้อมูลที่มีตำแหน่งอยู่ทั้งสองด้านของระนาบอยู่สอง ประเภทคือ ประเภทที่ 1 และประเภทที่ 3 โดยที่มีจำนวนร้อยละของข้อมูลในด้านที่น้อยกว่าของประเภทที่ 1 เป็น 10% และของประเภทที่ 3 เป็น 40%



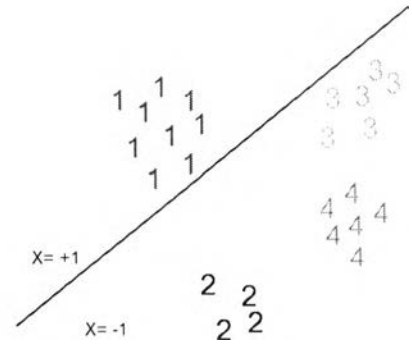
รูปที่ 16 ตัวอย่างระนาบแบ่งประเภทข้อมูลที่ทำให้ตำแหน่งของข้อมูลประเภทเดียวกันตกอยู่ทั้งสองด้านของระนาบ

ถ้ากำหนดการกำหนดค่าร้อยละในตัดเล็มอยู่ที่ 10 % ข้อมูลตำแหน่งของประเภทที่ 1 ในด้านที่น้อยกว่าจะถูกกำจัดทิ้ง (ดูรูปที่ 17 (ก)) แต่ถ้าเรากำหนดค่าร้อยละอยู่ที่ 40 % ข้อมูลตำแหน่งของทั้งประเภทที่ 1 และประเภทที่ 3 ในด้านที่น้อยกว่าจะถูกกำจัดทิ้งทั้งสองประเภท ทำให้เหลือประเภทข้อมูลในแต่ละด้านที่หากคำนวณค่าเอนโทรปีแล้วให้ค่าที่ต่ำ (ดูรูปที่ 17 (ข))

ซึ่งเป็นตัวบ่งชี้ว่าหากนำระนาบนี้ไปใช้แล้วจะสามารถแยกประเภทข้อมูลที่เป็นประเภทเดียวกันไปอยู่ในด้านเดียวกันได้ดีโดยมีข้อมูลอื่นที่อยู่คนละประเภทปนอยู่น้อย



รูปที่ 17 (ก) การตัดเล็มกรณีค่าร้อยละของการตัดเล็มเป็น 10



รูปที่ 17 (ข) การตัดเล็มกรณีค่าร้อยละของการตัดเล็มเป็น 40

อีกหนึ่งปัญหาที่ต้องทำการกำหนดค่าพารามิเตอร์ควบคุมก็คือ ค่าความผิดพลาดของการจำแนก เนื่องจากเทคนิคการแตกครึ่งตามสารสนเทศจะเลือกระนาบจากการจำแนกแบบหนึ่งต่อหนึ่งทุกตัวมาทำการเลือก ซึ่งแต่ระนาบแต่ละตัวนั้นมีค่าความผิดพลาดของการจำแนกมากน้อยแตกต่างกัน ทำให้ระนาบที่เราเลือกอาจให้ค่าความผิดพลาดอยู่ในช่วงสูง เมื่อนำมาจำแนกแม้ว่าจะจำแนกข้อมูลได้อย่างรวดเร็วแต่ก็ทำให้เกิดความผิดพลาดสูงตามมาด้วย

ค่าประสิทธิภาพโดยนัยทั่วไป (Generalization Performance) [11] ของซัพพอร์ตเวกเตอร์แมชชีนสามารถบอกขอบเขตความผิดพลาดของระนาบที่ใช้จำแนกข้อมูลแต่ละตัวได้ โดยกำหนดชนิดของฟังก์ชันเป็นค่าตัวเลขจำนวนจริงภายในลูกบอลที่มีรัศมี R ค่ารัศมีไม่เกิน $\Re^n$ เป็นฟังก์ชันดังนี้ $F = \{\mathbf{x} \mapsto \mathbf{w} \cdot \mathbf{x} : \|\mathbf{w}\| \le 1, \|\mathbf{x}\| \le R\}$ จะมีค่าคงที่ c ที่สำหรับการกระจายทุกประเภทด้วยความน่าจะเป็นอย่างน้อย 1-δ บนตัวอย่าง $\mathbf{z}$ ที่เลือกมาอย่างอิสระจากกัน m ตัว ถ้าตัวจำแนก $h = \mathrm{sgn}(f) \in \mathrm{sgn}(F)$ มีค่าระยะห่างระหว่างระนาบหลายมิติกับข้อมูลที่ใช้สอนอย่างน้อยเป็น γ สำหรับตัวอย่างทุกตัวใน $\mathbf{z}$ แล้ว ความผิดพลาดของตัวจำแนก h จะไม่เกิน

$$\frac{c}{m}\left(\frac{R^2}{\gamma^2}\log^2 m + \log\left(\frac{1}{\delta}\right)\right). \qquad (21)$$

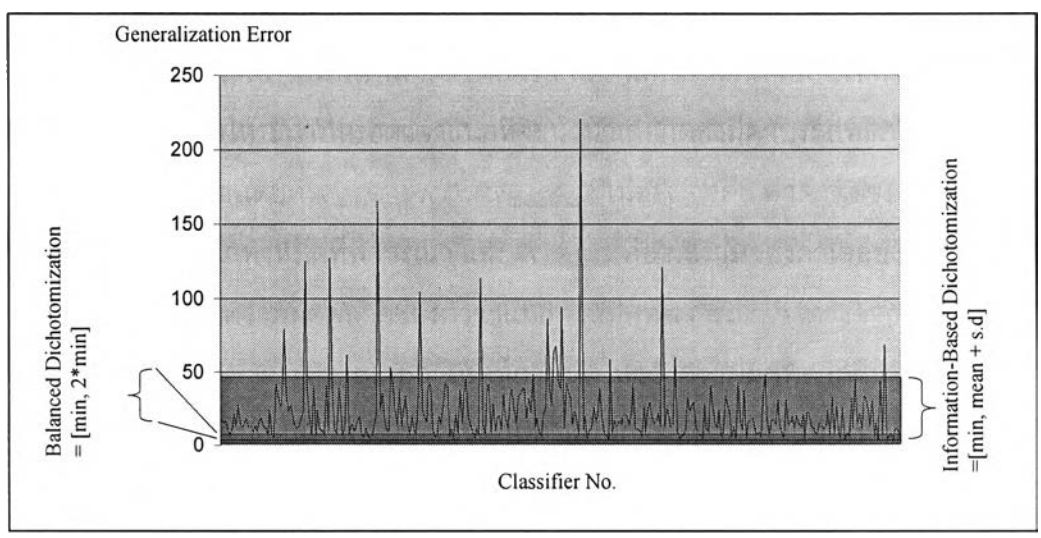
นอกเหนือจากนี้แล้ว ด้วยความน่าจะเป็นอย่างน้อย $1-\delta$ สำหรับทุกตัวจำแนก $h \in \mathrm{sgn}(F)$ จะมีความผิดพลาดไม่เกิน

$$\frac{k}{m} + \sqrt{\frac{c}{m}\left(\frac{R^2}{\gamma^2}\log^2 m + \log\left(\frac{1}{\gamma}\right)\right)} \tag{22}$$

เมื่อ k คือจำนวนตัวอย่างข้อมูลใน z ที่มีระยะห่างระหว่างระนาบหลายมิติกับซัพพอร์ตเวกเตอร์น้อยกว่า γ

เมื่อรู้ค่าความผิดพลาดของระนาบแต่ละตัวแล้วในวิธีการแตกครึ่งตามสารสนเทศจะมีการกำหนดขอบเขตความผิดพลาดของการจำแนกลักษณะคล้ายกับเทคนิคการแตกครึ่งแบบสมดุลโดยกำหนดให้ระนาบหรือตัวจำแนกที่จะนำมาเลือกต้องมีค่าประสิทธิภาพโดยนัยทั่วไปอยู่ในช่วง R ที่กำหนด โดยที่ $x_{min} \leq R \leq x_{mean+sd}$ เมื่อ x_{min} คือ ค่าต่ำสุดของขอบเขตความผิดพลาดของตัวจำแนกทั้งหมดที่พิจารณา และ $x_{mean+sd}$ คือ ผลรวมของค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของขอบเขตความผิดพลาดของตัวจำแนกทั้งหมดที่พิจารณา ซึ่งต่างจากแบบสมดุลที่จะกำหนดให้ $x_{min} \leq R \leq 2*x_{min}$ โดยสาเหตุที่กำหนดให้ช่วงความผิดพลาดกว้างขึ้นได้มากกว่าแบบสมดุลแต่ไม่ให้สูงจนเกินไปเพื่อเป็นการเพิ่มโอกาสในการเลือกตัวจำแนกให้มีมากขึ้นซึ่งหากมีตัวจำแนกให้เลือกน้อยเกินไปโอกาสที่จะได้ต้นไม้ที่สั้นย่อมน้อยลงด้วย

ตัวอย่างในรูปที่ 18 แสดงตัวอย่างค่าประสิทธิภาพโดยนัยทั่วไปของระนาบที่ใช้จำแนกข้อมูล 325 ระนาบในปัญหา 26 ประเภทและช่วงของขอบเขตความผิดพลาดระหว่างแบบสมดุลและแบบแตกครึ่งตามสารสนเทศ



รูปที่ 18 ตัวอย่างช่วงค่าประสิทธิภาพโดยนัยทั่วไปของตัวจำแนกข้อมูล 325 ตัว

อย่างไรก็ตามในการกำหนดค่าพารามิเตอร์ของทั้งร้อยละการตัดเล็ม (P) และช่วงค่าของความผิดพลาด (R) นั้น จำเป็นต้องมีการคำนวณหาค่าที่เหมาะสมสำหรับพารามิเตอร์แต่ละตัว เนื่องจากถ้ากำหนดร้อยละการตัดเล็มมากจนเกินไปก็จะทำให้มีข้อมูลที่มีประโยชน์ถูกกำจัดทิ้งไปด้วยและทำให้การจำแนกผิดมีสูงขึ้น หรือถ้าเรากำหนดช่วงค่าของความผิดพลาดน้อยจนเกินไปก็จะทำให้โอกาสได้ต้นไม้ที่สั้นน้อยลงด้วย โดยวิธีการเลือกหาค่าที่เหมาะสมของพารามิเตอร์ทั้งสองตัวนี้ จะใช้การแบ่งข้อมูลสอน (Training Data) ออกเป็นข้อมูลสอนจริง (Actual Training Data) และข้อมูลทดสอบความถูกต้อง (Validation Data) แล้วนำข้อมูลสอนจริงไปสร้างระนาบ จากนั้นนำระนาบที่ได้มาทำการจำแนกกับข้อมูลทดสอบความถูกต้องโดยกำหนดค่าพารามิเตอร์ต่างๆ กันไป แล้วเลือกค่าพารามิเตอร์ที่ทำให้มีความผิดพลาดน้อยสุดในข้อมูลทดสอบความถูกต้องมาเป็นค่าที่เหมาะสมที่จะใช้กับข้อมูลทดสอบ (Test Data) ต่อไป