บทที่ 5

สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

เทคนิคการแตกครึ่งตามสารสนเทศซึ่งมีหลักการพื้นฐานในการสร้างต้นไม้สำหรับการจำแนกแบบหลายประเภทโดยเลือกโนดจากตัวจำแนกที่ให้ค่าเอนโทรปีต่ำที่สุด สามารถเพิ่มประสิทธิภาพของการจำแนกประเภทของซัพพอร์ตเวกเตอร์แมชชีนแบบหลายประเภทโดยลดจำนวนครั้งในการจำแนกลงได้จากวิธีอื่นๆ ทุกวิธี โดยที่ยังให้ค่าความถูกต้องใกล้เคียงกับวิธีอื่น

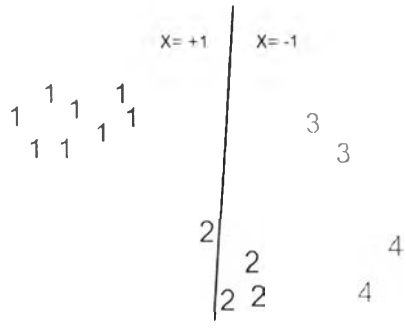
การเลือกกำหนดช่วงขอบเขตความผิดผลาด(R)เทคนิคนี้กำหนดค่าโดย $x_{min} \leq R \leq x_{mean+sd}$ เมื่อ x_{min} คือ ค่าต่ำสุดของขอบเขตความผิดพลาดของตัวจำแนกทั้งหมดที่พิจารณา และ $x_{mean+sd}$ คือ ผลรวมของค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของขอบเขตความผิดพลาดของตัวจำแนกทั้งหมดที่พิจารณา โดยสาเหตุที่กำหนดให้ช่วงความผิดพลาดกว้างขึ้นได้มากกว่าแบบสมดุลเพื่อเป็นการเพิ่มโอกาสในการเลือกตัวจำแนกให้มีมากขึ้นซึ่งหากมีตัวจำแนกให้เลือกน้อยเกินไปโอกาสที่จะได้ต้นไม้ที่สั้นย่อมน้อยลงด้วย แต่ในทางตรงข้ามหากช่วงขอบเขตความผิดพลาดสูงเกินไปก็จะส่งผลต่อค่าความถูกต้องของการจำแนกด้วย ดังนั้นเมื่อใช้เทคนิคการแตกครึ่งแบบสมดุลโดยเลือกให้ $x_{min} \leq R \leq x_{mean+sd}$ จะทำให้กรณีที่บางชุดข้อมูล ตัวจำแนกหรือระนาบที่สร้างได้ส่วนใหญ่มีค่าความผิดพลาดอยู่ในช่วงที่สูงก็จะทำให้ขอบเขตความผิดพลาดสูงด้วยอันจะส่งผลต่อค่าความถูกต้องของการจำแนกตามมาได้

ความสามารถในการลดจำนวนครั้งของการจำแนกข้อมูลจะขึ้นกับระนาบของตัวจำแนกแบบสองประเภทที่สร้างได้จากค่า P และ R ที่ดีที่สุด โดยหากระนาบที่ได้สามารถแบ่งประเภทของข้อมูลได้ดีนั่นคือ แต่ละด้านของระนาบมีประเภทของข้อมูลปนกันน้อยหรือถ้าหากด้านใดด้านหนึ่งของระนาบมีเพียงประเภทเดียว ค่าจำนวนครั้งเฉลี่ยของการจำแนกที่ได้จริงยิ่งให้ค่าใกล้เคียงกับค่าจำนวนครั้งในการจำแนกที่คาดหวังจากสูตร $\sum_{i=1}^{k} -P(m_i)\log_2 P(m_i)$ ในทางตรงกันข้ามหากระนาบที่ได้ไม่มีระนาบใดที่แบ่งประเภทข้อมูลได้ดีพอก็ยิ่งส่งผลให้ตัวจำแนกแบบหลายประเภทที่ได้มีจำนวนครั้งในการจำแนกสูงและคลาดเคลื่อนจากค่าจำนวนครั้งในการจำแนกที่คาดหวังมากขึ้นเท่านั้น

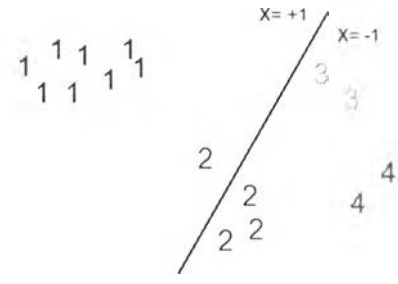
อย่างไรก็ดีแม้ว่าวิธีการแตกครึ่งตามสารสนเทศจะสามารถลดจำนวนครั้งในการจำแนกลงได้จากวิธีอื่นๆ ทุกวิธี โดยที่ยังให้ค่าความถูกต้องใกล้เคียงกับวิธีอื่น แต่จากผลการทดลองจะพบว่าวิธีนี้จะมีความเหมาะสมกับลักษณะข้อมูลบางแบบโดยเฉพาะกรณีที่ค่าของเขตความผิดพลาดของตัวจำแนกแบบสองประเภทที่นำมาสร้างต้นไม้โดยส่วนใหญ่อยู่ในช่วงที่ต่ำก็จะได้ต้นไม้สำหรับการจำแนกแบบหลายประเภทที่มีแนวโน้มให้ค่าความถูกต้องในการจำแนกที่สูงรวมถึงหากชุดข้อมูลดังกล่าวมีจำนวนประเภทเป็นจำนวนมากและเป็นงานในระบบแบบทันที (Real time System) วิธีนี้จะยิ่งมีประสิทธิภาพสูงกว่าวิธีอื่นๆ ซึ่งนอกจากให้ค่าความถูกต้องสูงแล้วยังทำงานได้อย่างรวดเร็วอีกด้วย

5.2 ข้อเสนอแนะ

1. งานวิจัยในอนาคตเพื่อเพิ่มความถูกต้องของเทคนิคนี้อาจทำได้โดยการพิจารณาหลักการเบื้องต้นของซัพพอร์ตเวกเตอร์แมชชีนซึ่งยึดหลักในการหาระนาบสำหรับการจำแนกแบบสองประเภทที่ให้ค่าระยะห่างระหว่างระนาบกับประเภทข้อมูลสอนที่อยู่ใกล้ที่สุดให้มีค่ามากที่สุด ค่าระยะห่างดังกล่าวจะแปรผกผันกับค่าขอบเขตความผิดพลาด นั่นหมายความว่ายิ่งข้อมูลสอนกับระนาบอยู่ไกลกันมากเท่าไรยิ่งดีมากขึ้นเท่านั้น เมื่อพิจารณาขั้นตอนของการสร้างต้นไม้สำหรับจำแนกแบบหลายประเภทของการแตกครึ่งตามสารสนเทศมีเกณฑ์ในการเลือกระนาบซึ่งอยู่ในช่วงขอบเขตความผิดพลาดที่กำหนดจากค่า R ที่ได้จากการสอน แล้วจึงเลือกตัวจำแนกที่ให้ค่าเอนโทรปีต่ำสุด ซึ่งค่า R ที่ได้ในที่นี้จะบ่งบอกให้ทราบเพียงช่วงความผิดพลาดของประเภทข้อมูลสองประเภทเท่านั้นโดยยังมิได้นำเอาตำแหน่งของข้อมูลประเภทอื่นมาเทียบว่าหากเลือกระนาบนั้นมาใช้แล้วระยะห่างของข้อมูลประเภทอื่นกับระนาบที่เลือกมานั้นเป็นเช่นไร ตัวอย่างจากรูปที่ 28 ซึ่งตัวจำแนก 1-3 และ 1-4 ทั้งคู่ต่างให้ค่าเอนโทรปีที่เท่ากันแต่เมื่อพิจาณาในส่วนของระยะห่างระหว่างระนาบกับข้อมูลสอนประเภทที่ 3 เทียบกับระนาบที่หาได้ทั้งคู่จะพบว่าหากเลือกตัวจำแนก 1-4 ไปใช้ในการสร้างโนดย่อมมีแนวโน้มจะให้ค่าความถูกต้องต่ำกว่าตัวจำแนก 1-3 เนื่องจากตัวจำแนก 1-4 ให้ค่าระยะห่างของตัวอย่างสอนที่อยู่ใกล้ระนาบมากที่สุดมีค่าน้อยกว่าตัวจำแนก 1-3 นั่นเอง โดยการพิจารณาตัวอย่างสอนที่ใกล้ที่สุดนี้จะเลือกพิจาณาเฉพาะตัวอย่างสอนประเภทที่ตกอยู่ด้านใดด้านหนึ่งของระนาบเท่านั้น กรณีอื่นไม่นำมาพิจารณาด้วยเหตุว่าไม่มีผลต่อค่าความถูกต้องของต้นไม้แต่อย่างใด



(ก) ตำแหน่งของข้อมูลประเภทอื่นๆ เมื่อเทียบกับตัวจำแนกประเภท 1-3



(ข) ตำแหน่งของข้อมูลประเภทอื่นๆ เมื่อเทียบกับตัวจำแนกประเภท 1-4

รูปที่ 28 การแบ่งข้อมูลด้วยตัวจำแนกแบบสองประเภทของตัวจำแนก 1-3 และ 1-4

2. การคำนวณหาค่า P และ R งานวิจัยในอนาคตอาจเพิ่มฟังก์ชันฮิวริสติกในการคำนวณหาค่าที่ดีที่สุดของพารามิเตอร์ตัวนี้ เพื่อช่วยลดเวลาในการคำนวณและโดยเฉพาะอย่างยิ่งการหาค่า R ที่เหมาะสมซึ่งมีความสำคัญอย่างยิ่งต่อค่าความถูกต้องของการจำแนกโดยที่ไม่ไปตัดโอกาสการได้ต้นไม้การจำแนกแบบหลายประเภทของซัพพอร์ตเวกเตอร์แมชชีนที่สั้นที่สุดและให้ค่าความถูกต้องสูงที่สุด

3. จากหลักการในการเลือกตัวจำแนกที่อยู่ในช่วงความผิดพลาดที่กำหนดของเทคนิคการแตกครึ่งตามสารสนเทศโดย $x_{min} \le R \le x_{mean+sd}$ เพื่อช่วยในการกำจัดตัวจำแนกที่อาจมีความผิดพลาดสูงออกไปได้ส่วนหนึ่ง งานวิจัยในอนาคตอาจนำเอาหลักการตรงนี้ไปใช้ร่วมกับวิธีการของแมกซ์วินตามปกติเพื่อช่วยลดเวลาของการสอนและลดจำนวนครั้งของการจำแนกลงจาก $k*(k-1)/2$ ครั้ง กรณีปัญหา k ประเภท