บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ซัพพอร์ตเวกเตอร์แมชชีนเป็นเทคนิคการเรียนรู้ที่ใช้คุณสมบัติเชิงเรขาคณิตในการคำนวณหาระนาบหลายมิติ (Hyperplane) ที่ดีที่สุดในการแยกข้อมูลออกจากกัน โดยในช่วงแรกเริ่มนั้นเทคนิคนี้มีข้อจำกัดอยู่ที่สามารถจำแนกข้อมูลได้เพียงสองประเภทเท่านั้น แต่ในปัญหาการจำแนกส่วนใหญ่มักเป็นแบบหลายประเภท ต่อมาจึงมีผู้พัฒนาเทคนิคนี้ให้สามารถใช้ได้กับการจำแนกแบบหลายประเภท วิธีการส่วนใหญ่มักจะเป็นการนำฟังก์ชันที่จำแนกแบบสองประเภทหลายฟังก์ชันมาใช้ร่วมกัน เช่น การจำแนกแบบหนึ่งต่อที่เหลือ (One-against-the rest) เป็นการเปรียบเทียบประเภทข้อมูลหนึ่งกับประเภทอื่นที่เหลือทั้งหมดและการจำแนกแบบหนึ่งต่อหนึ่ง (One-against-one) ซึ่งเป็นการเปรียบเทียบประเภทข้อมูลหนึ่งกับประเภทข้อมูลอื่นทีละประเภท [1]

อัลกอริทึมแมกซ์วินเป็นหนึ่งในวิธีการจำแนกแบบหนึ่งต่อหนึ่งซึ่งให้ค่าความถูกต้องสูงกว่าการจำแนกแบบหนึ่งต่อที่เหลือ แต่ใช้เวลาในการจำแนกนานกว่า [2] ในขณะที่วิธีดีดีเอจี (Decision Directed Acyclic Graph (DDAG)) ที่เสนอโดย Platt และคณะสามารถลดเวลาในการจำแนกลงจากวิธีแมกซ์วินได้โดยที่ยังให้ค่าความถูกต้องใกล้เคียงวิธีแมกซ์วิน [3] แต่วิธีดีดีเอจีมีข้อเสียคือ มีการขึ้นต่อกันของลำดับของการจำแนกประเภทที่ถูกต้อง โดยหากประเภทที่ถูกต้องถูกจำแนกกับประเภทอื่นหลายครั้งก็มีโอกาสที่จะจำแนกผิดมากขึ้น ต่อมา Ussivakul และ Kijsirikul ได้เสนอวิธีเอดีเอจี (Adaptive Directed Acyclic Graph (ADAG)) ซึ่งเป็นวิธีที่ดัดแปลงมาจากวิธีดีดีเอจีซึ่งสามารถลดการขึ้นต่อกันของลำดับของการจำแนก และลดจำนวนครั้งที่ประเภทที่ถูกต้องถูกจำแนกกับประเภทอื่น จึงให้ความถูกต้องและความน่าเชื่อถือในการจำแนกข้อมูลสูงกว่าวิธีดีดีเอจี [4] แต่วิธีเอดีเอจียังคงมีค่าความถูกต้องไม่แน่นอนขึ้นกับลำดับของโนด (Node) ด้วย ต่อมา Phetkaew, et al. ได้เสนอวิธีอาร์เอดีเอจี (Reordering Adaptive Directed Acyclic Graph (RADAG)) [5] ซึ่งวิธีนี้จะมีการเลือกลำดับของโนดที่เหมาะสมซึ่งมีโอกาสที่จะจำแนกผิดพลาดน้อย โดยใช้อัลกอริทึมของการจับคู่สมบูรณ์แบบน้ำหนักน้อยสุด (Minimum-Weight Perfect Matching) [6] และมีการจัดเรียงโนดใหม่ในทุกชั้นของการจำแนกทำให้มีค่าความผิดพลาดลดลงจากวิธีเอดีเอจี อย่างไรก็ตามในวิธีที่กล่าวมาข้างต้นสำหรับการ

จำแนกข้อมูล k ประเภทใดๆ วิธีอาร์เอดีเอจีต้องใช้จำนวนครั้งของการจำแนกข้อมูลเป็น k -1 ครั้ง เนื่องจากในการจำแนกข้อมูลแต่ละครั้งสามารถตัดข้อมูลที่ไม่ถูกต้องออกไปได้เพียง 1 ประเภทต่อครั้งเท่านั้น ต่อมา Kijsirikul, et al. ได้เสนอการสร้างต้นไม้สำหรับการจำแนกแบบหลายประเภทด้วยวิธีการแตกครึ่งแบบสมดุล (Balanced Dichotomization Classification) [7] ที่ทำการจำแนกข้อมูลโดยค้นหาระนาบที่แบ่งข้อมูลในตำแหน่งที่สมดุลที่สุดของข้อมูลทั้งหมดในแต่ละรอบมาจำแนก เพื่อให้การจำแนกข้อมูลแต่ละครั้งสามารถตัดประเภทที่ไม่ถูกต้องออกไปได้มากกว่า 1 ประเภท ทั้งนี้ในการค้นหาระนาบนั้นจะเลือกระนาบที่แบ่งข้อมูลแล้วให้ค่าจำนวนประเภทซึ่งตกอยู่ในแต่ละด้านของระนาบมีค่าใกล้เคียงกันมากที่สุด ภายใต้สมมติฐานที่ว่าความน่าจะเป็นในการเกิดข้อมูลในแต่ละประเภทของข้อมูลสอนมีค่าใกล้เคียงกับข้อมูลทดสอบ วิธีนี้จะมีความเหมาะสมกรณีที่ข้อมูลแต่ละประเภทมีความน่าจะเป็นในการเกิดที่เท่ากัน โดยวิธีนี้จะพยายามสร้างต้นไม้ให้ทั้งสองกิ่งมีความสมดุลที่สุด ซึ่งส่งผลให้ค่าจำนวนครั้งในการจำแนกข้อมูลแต่ละประเภทจะมีค่าใกล้เคียงกัน ในความเป็นจริงข้อมูลที่นำมาจำแนกอาจไม่ได้มีความน่าจะเป็นในการเกิดเท่ากัน หากยึดวิธีสร้างต้นไม้ที่ให้จำนวนครั้งในการจำแนกข้อมูลแต่ละประเภทที่ใกล้เคียงกันโดยไม่พิจารณาความน่าจะเป็นในการเกิดข้อมูลแต่ละประเภทประกอบอาจไม่ได้ต้นไม้สำหรับจำแนกที่ให้จำนวนครั้งในการจำแนกเฉลี่ยต่ำที่สุด ซึ่งเป็นการสะสมความผิดพลาดในการจำแนกโดยไม่จำเป็น

วิทยานิพนธ์ฉบับนี้จะนำเสนอวิธีการใหม่ในการจำแนกข้อมูลโดยสร้างต้นไม้สำหรับการจำแนกแบบหลายประเภทด้วยเทคนิคการแตกครึ่งตามสารสนเทศโดยการนำความน่าจะเป็นในการเกิดข้อมูลแต่ละประเภทมาใช้ในการค้นหาระนาบที่แบ่งข้อมูล ในทางทฤษฎีวิธีนี้จะสามารถลดจำนวนครั้งในการจำแนกลงเหลือประมาณ $\log_2 k$ ครั้ง ซึ่งเป็นการลดเวลาของการจำแนกเฉลี่ยลงได้โดยเฉพาะอย่างยิ่งกรณีปัญหาการจำแนกที่มีจำนวนประเภทเป็นจำนวนมาก

1.2 วัตถุประสงค์ของการวิจัย

ออกแบบและพัฒนาวิธีการเพื่อเพิ่มประสิทธิภาพในการเรียนรู้ของซัพพอร์ตเวกเตอร์แมชชีนแบบหลายประเภทให้มีจำนวนครั้งของการจำแนกเฉลี่ยลดลงจากวิธีแตกครึ่งแบบสมดุลโดยที่ยังคงให้ค่าความถูกต้องใกล้เคียงกัน

1.3 ขอบเขตของการวิจัย

1. การเปรียบเทียบประสิทธิภาพ จะพิจารณาจากจำนวนครั้งของการจำแนกและความถูกต้องของการจำแนกข้อมูล
2. ข้อมูลที่นำมาใช้ทดสอบจะใช้ข้อมูลของ UCI Machine Learning Repository [8] จำนวน 8 ชุด และ Thai printed character recognition dataset [9] จำนวน 2 ชุด

1.4 ขั้นตอนและวิธีดำเนินการวิจัย

1. ศึกษาและทดสอบประสิทธิภาพของอัลกอริทึมต่างๆ ของซัพพอร์ตเวกเตอร์แมชชีนแบบหลายประเภท เช่น แมกซ์วิน ดีดีเอจี เอดีเอจี อาร์เอดีเอจี และการแตกครึ่งแบบสมดุล
2. ทดสอบประสิทธิภาพของเทคนิคการแตกครึ่งตามสารสนเทศเพื่อเปรียบเทียบกับวิธีเดิมข้างต้น
3. วิเคราะห์ผลการเปรียบเทียบประสิทธิภาพของวิธีการแตกครึ่งตามสารสนเทศกับวิธีอื่นๆ
4. สรุปผลการวิจัยและจัดทำรายงานวิทยานิพนธ์

1.5 ประโยชน์ที่คาดว่าจะได้รับ

ได้เทคนิคใหม่ในการเรียนรู้ของซัพพอร์ตเวกเตอร์แมชชีนแบบหลายประเภทที่มีจำนวนครั้งของการจำแนกลดลงจากวิธีแตกครึ่งแบบสมดุล โดยที่ยังคงให้ค่าความถูกต้องใกล้เคียงกัน

1.6 ลำดับการจัดเรียงเนื้อหาในวิทยานิพนธ์

วิทยานิพนธ์นี้แบ่งเนื้อหาออกเป็น 5 บทดังต่อไปนี้ บทที่ 1 เป็นบทนำซึ่งกล่าวถึงความเป็นมาและความสำคัญของปัญหา รวมถึงวัตถุประสงค์ของการวิจัย บทที่ 2 กล่าวถึงทฤษฎีพื้นฐานและงานวิจัยที่เกี่ยวข้องในงานวิจัยนี้ บทที่ 3 กล่าวถึงรายละเอียดทั้งหมดของเทคนิคการแตกครึ่งตามสารสนเทศของสำหรับซัพพอร์ตเวกเตอร์แมชชีนแบบหลายประเภท บทที่ 4 แสดงรายละเอียดของการทดลองและผลการทดลอง และบทที่ 5 เป็นข้อสรุปและข้อเสนอแนะจากการวิจัย

1.7 ผลงานที่ตีพิมพ์จากวิทยานิพนธ์

ส่วนหนึ่งของวิทยานิพนธ์นี้ได้รับการตีพิมพ์เป็นบทความทางวิชาการในหัวข้อเรื่อง "เทคนิคการแตกครึ่งตามสารสนเทศสำหรับซัพพอร์ตเวกเตอร์แมชชีนแบบหลายประเภท" โดย ปทุมศิริ สงศิริ, บุญเสริม กิจศิริกุล และ ฐิมาพร เพชรแก้ว ในงานประชุมวิชาการ "The 9th National Computer Science and Engineering Conference (NCSEC 2005)" ณ มหาวิทยาลัยหอการค้าไทย กรุงเทพมหานคร ระหว่างวันที่ 27-28 ตุลาคม 2548