บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของงานวิจัย

ต้นไม้ตัดสินใจ (Decision Tree) เป็นหนึ่งในวิธีการหาข้อสรุปจากข้อมูล ที่เป็นที่รู้จักกันอย่างกว้างขวาง มีการทำวิจัยและพัฒนาเกี่ยวกับต้นไม้ตัดสินใจมาอย่างต่อเนื่องเพื่อให้ได้ข้อสรุปที่มีความถูกต้องสูงขึ้นหรือ เพื่อให้มีขนาดของต้นไม้ที่เล็กลงแต่ยังคงความถูกต้องที่ใกล้เคียงเดิมหรือดีขึ้นกว่าเดิม หนึ่งในการพัฒนาที่สำคัญคือ การสรุปผลจากต้นไม้ตัดสินใจหลายต้นหรือการรวมตัวจำแนก แทนที่การสรุปผลจากต้นไม้ตัดสินใจเพียงต้นเดียว เรียกการทำงานนี้ว่า การรวม (Ensemble of Classifiers) คือ การมีกลุ่มต้นไม้ตัดสินใจที่แต่ละต้นมีข้อสรุปเป็นของตนเอง แล้วนำข้อสรุปที่ได้มาประมวลรวมกันเพื่อหาข้อสรุปสุดท้ายให้กับตัวอย่างใหม่ จากการวิจัยที่ผ่านมาพบว่า เทคนิคการรวมนี้ได้ข้อสรุปที่มีความถูกต้องสูงขึ้นกว่าการใช้ข้อสรุปจากต้นไม้ตัดสินใจต้นเดียว เมื่อการรวมนี้เป็นการประมวลรวมกันของต้นไม้ตัดสินใจที่แต่ละต้นให้ความผิดพลาดในข้อมูลที่ต่างกันและยังค้นพบอีกว่าการรวมนี้ ต้นไม้ตัดสินใจแต่ละต้นจะต้องให้ความผิดพลาดน้อยกว่า 50% มิฉะนั้นการรวมนั้นอาจส่งผลที่ไม่ดีต่อความถูกต้องในภายหน้าได้

เทคนิคการรวมต้นไม้ตัดสินใจ หรือการรวมการจำแนกเริ่มจากนำชุดข้อมูลตั้งต้นมาทำเป็นชุดทดสอบสำหรับต้นไม้แต่ละต้นแล้วนำผลของต้นไม้แต่ละต้นมารวมกัน โดยวิธีการรวมที่เป็นที่รู้จักได้แก่ วิธีแบ็กกิง (Bagging) และวิธีบูสต์ (Boosting Classifier) โดยวิธีแบ็กกิงใช้วิธีการบูตสแตรป (Bootstrapping) (Bauer and Kohavi, 1999, pp 105-139) ในการสุ่มตัวอย่างจากชุดตัวอย่างสำหรับเรียนรู้ (Training Set) เพื่อนำมาใช้เป็นชุดข้อมูลฝึกต้นไม้แต่ละต้น ในขณะที่วิธีบูสต์จะสร้างชุดทดสอบโดยอิงกับประสิทธิภาพของต้นไม้ก่อนหน้า จะเห็นได้ว่าทั้งสองวิธีใช้วิธีต่างกันในการเตรียมชุดข้อมูลทดสอบสำหรับต้นไม้แต่ละต้น แต่ชุดข้อมูลสำหรับทดสอบเหล่านี้ล้วนแล้วมาจากข้อมูลตั้งต้นเดียวกันทั้งสิ้น จากจุดนี้ผู้วิจัยเห็นว่า การที่ชุดทดสอบแต่ละชุดมาจากชุดข้อมูลตั้งต้นเดียวกัน กฎที่ได้จากต้นไม้แต่ละต้นน่าจะมีความสัมพันธ์กันจนถึงระดับที่สามารถถือได้ว่าเป็นกฎเดียวกันหรือนำมาใช้ร่วมกันได้ และเมื่อนำกฎเหล่านี้มาใช้ร่วมกับการรวมต้นไม้น่าจะให้ค่าความถูกต้องที่สูงขึ้น เพื่อทดสอบสมมติฐานนี้ผู้วิจัยได้พัฒนา การลงคะแนนโดยกฎ$^{+}$

(Majority Rule^{+}) และ การลงคะแนนโดยตัวอย่างเรียนรู้$^{+}$ (Majority Class^{+}) อันเป็นการพัฒนาโดยอิงกับการรวมต้นไม้ตัดสินใจด้วยวิธีการลงคะแนนแบบทั่วไปคือ การลงคะแนนแบบทั่วไป (Simple Majority Rule) หรือแบ็กกิง (Bagging) และการลงคะแนนแบบทั่วไปโดยตัวอย่างเรียนรู้ (Simple Majority Class) ตามลำดับ โดยการนำกฎที่มีค่าความใกล้เคียงตามแบบการทดลอง (ในการทำวิจัยนี้ทำการทดลองที่ค่าความใกล้เคียง 0.6, 0.7, 0.8, และ 0.9) มาร่วมลงคะแนนเพื่อให้ได้ข้อสรุปสุดท้าย แล้วเปรียบเทียบค่าความถูกต้องกับค่าความถูกต้องจากการลงคะแนนแบบทั่วไปพื้นฐาน

วิทยานิพนธ์นี้ได้ถูกเขียนตามลำดับโครงบทดังต่อไปนี้ ในบทที่ 2 กล่าวถึงทฤษฎีที่เกี่ยวข้องได้แก่ ต้นไม้ตัดสินใจ, การเตรียมชุดทดสอบแบบบูธสแตปปิง (Bootstrapping), วิธีการรวมแบ็กกิงที่ใช้การลงคะแนนแบบทั่วไปในการรวมต้นไม้และ การลงคะแนนแบบทั่วไปโดยตัวอย่างเรียนรู้ (Simple Majority Class) และงานวิจัยที่เกี่ยวข้องได้แก่ งานวิจัยอัลกอริทึม เพื่อนบ้านที่ใกล้ที่สุด *k* ตัว (k-Nearest Neighbor Algorithm, k-NN), การนำกฎที่ใกล้มากที่สุดมาใช้และการหาลดจำนวนกฎ บทที่ 3 อธิบายกระบวนการการทดลอง บทที่ 4 เสนอผลการทดลองและบทสุดท้าย บทที่ 5 สรุปผลการศึกษาและเสนอแนะแนวทางการพัฒนาต่อไปในอนาคต

1.2 วัตถุประสงค์ของงานวิจัย

ศึกษาการรรวมต้นไม้ตัดสินใจโดยนำกฎใกล้เคียงมาร่วมในการลงคะแนน เพื่อหาคำตอบให้กับตัวอย่างใหม่ และสังเกตหาค่าความใกล้เคียงที่น้อยที่สุดที่ทำให้ค่าความถูกต้องสูงขึ้นในกรณีที่การนำกฎใกล้เคียงมาใช้นี้ทำให้ความถูกต้องสูงขึ้นจริง เพื่อนำไปอ้างอิงในงานวิจัยต่อไปในอนาคต

1.3 ขอบเขตของงานวิจัย

1.3.1 พัฒนาวิธีการรวมต้นไม้ตัดสินใจวิธีการลงคะแนนโดยกฎ$^{+}$ และ การลงคะแนนโดยตัวอย่างเรียนรู้$^{+}$ โดยใช้วิธีบูตสแตรปในการเตรียมข้อมูลชุดทดสอบสำหรับต้นไม้แต่ละต้น

1.3.2 เปรียบเทียบความถูกต้องจากข้อสรุปคำตอบที่ได้จากวิธี การลงคะแนนโดยกฎ$^{+}$ กับวิธีแบ็กกิง และ การลงคะแนนโดยตัวอย่างเรียนรู้$^{+}$ กับการลงคะแนนแบบทั่วไปโดยตัวอย่างเรียนรู้

1.3.3 สังเกตหาค่าความใกล้เคียงที่น้อยที่สุดที่ยังคงให้ค่าความถูกต้องสูงขึ้นในกรณีที่ให้ความถูกต้องสูงขึ้นจริง