Journal of Engineering and Digital Technology (JEDT)
Vol.11 No.1 January - June 2023

การประยุกต์ใช้ขั้นตอนวิธีการหาค่าเหมาะสมที่สุดแบบไบนารีวาฬสำหรับแก้ปัญหาข้อมูลไม่สมดุล

จักรกฤษณ์ พลรบ[1*] เบญจวรรณ โรจนดิษฐ์[2] เจษฎา ตัณฑนุช[3] Eckart Schulz[4]

[1*,2,3,4]*สาขาวิชาคณิตศาสตร์ สำนักวิชาวิทยาศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี, นครราชสีมา, ประเทศไทย*

*ผู้ประพันธ์บรรณกิจ อีเมล : jakkrit.polrob@gmail.com

รับต้นฉบับ : 7 มกราคม 2566; รับบทความฉบับแก้ไข : 1 กุมภาพันธ์ 2566; ตอบรับบทความ : 28 กุมภาพันธ์ 2566
เผยแพร่ออนไลน์ : 29 มิถุนายน 2566

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อสร้างขั้นตอนวิธีการสุ่มตัวอย่างลดแบบใหม่ โดยใช้แนวคิดของขั้นตอนวิธีการหาค่าเหมาะสมที่สุดแบบวาฬและไบนารีวาฬร่วมกับเทคนิคเคเพื่อนบ้านใกล้ที่สุดเพื่อใช้ในการแก้ปัญหาข้อมูลไม่สมดุล ซึ่งขั้นตอนวิธีที่นำเสนอนี้จะกำหนดค่าพารามิเตอร์เคเท่ากับหนึ่ง ทั้งนี้ได้เลือกชุดข้อมูลทดสอบจำนวน 12 ชุดจาก KEEL และ imbalanced-learn ซึ่งชุดข้อมูลจะมีอัตราส่วนความไม่สมดุลอยู่ในช่วง 1.82 ถึง 42.01 เพื่อใช้ในการประเมินขั้นตอนวิธีใหม่เปรียบเทียบกับการแก้ปัญหาข้อมูลไม่สมดุลด้วยวิธีการลดจำนวนตัวอย่างข้อมูล 3 ขั้นตอนวิธี ได้แก่ การลดจำนวนตัวอย่างข้อมูลแบบสุ่ม คลัสเตอร์เซนทรอยด์ และเนียร์มิส งานวิจัยนี้เริ่มจากการนำชุดข้อมูลมาแบ่งเป็นสองชุด คือ ชุดข้อมูลฝึกสอนและชุดข้อมูลทดสอบ สำหรับชุดข้อมูลฝึกสอนคลาสข้อมูลกลุ่มน้อยจะใช้ข้อมูลชุดเดิม ในขณะที่คลาสข้อมูลกลุ่มมากจะถูกวิเคราะห์เพื่อดึงชุดข้อมูลย่อยที่เป็นตัวแทนที่ดีที่สุดด้วยขั้นตอนวิธีที่นำเสนอ และประเมินประสิทธิภาพด้วยเทคนิคซัพพอร์ตเวกเตอร์แมชชีน จากผลการวิจัยพบว่าขั้นตอนวิธีที่ได้นำเสนอมีประสิทธิภาพการทำงานโดยรวมสูงที่สุด เมื่อเทียบกับขั้นตอนวิธีการสุ่มตัวอย่างลดทั้งสามที่นำมาเปรียบเทียบ ซึ่งมีผลการวัดประสิทธิภาพโดยเฉลี่ยดังนี้ Accuracy = 0.8596, F1 score = 0.6255, G-mean = 0.8941, AUROC = 0.9363, AUPRC = 0.6978, Sensitivity = 0.9444, Precision = 0.5271, MCC = 0.6204, และ Kappa = 0.5695

คำสำคัญ **:** ขั้นตอนวิธีการหาค่าเหมาะสมที่สุดแบบไบนารีวาฬ ปัญหาข้อมูลไม่สมดุล เคเพื่อนบ้านใกล้ที่สุด ซัพพอร์ตเวกเตอร์แมชชีน ขั้นตอนวิธีการสุ่มตัวอย่างลด

Application of Binary Whale Optimization Algorithm for Solving Imbalanced Data Problems

Jakkrit Polrob[1*] Benjawan Rodjanadid[2] Jessada Tanthanuch[3] Eckart Schulz[4]

[1*,2,3,4]*School of Mathematics, Institute of Science, Suranaree University of Technology, Nakhon Ratchasima, Thailand*

*Corresponding Author. E-mail address: jakkrit.polrob@gmail.com

Received: 7 January 2023; Revised: 1 February 2023; Accepted: 28 February 2023
Published online: 29 June 2023

Abstract

This research is aimed at developing a novel undersampling algorithm by combining the ideas of the whale and binary whale optimization algorithms with K- nearest neighbor classification, in order to solve imbalanced data problems. Twelve datasets of varying imbalance ratios ranging from 1.82 to 42.01 were selected from the Knowledge Extraction based on Evolutionary Learning (KEEL) repository and also the imbalanced-learn repository, to be used in the evaluation of the novel algorithm. This research work started by splitting each dataset into two parts, the training set and the testing set. Whereas the minority class of each training set remained untouched, its majority class was processed by the proposed algorithm with the parameter in K-nearest neighbor classification fixed to K = 1, to obtain an optimal representative subset of the majority class. Then a support vector machine classifier was trained with the new and reduced training set for performance assessment. It was found that the proposed algorithm had best overall performance when compared with another three undersampling methods, namely random undersampling, cluster centroid, and near-miss algorithms, showing average efficiency measurement results as follows: Accuracy = 0.8596, F1 score = 0.6255, G-mean = 0.8941, AUROC = 0.9363, AUPRC = 0.6978, Sensitivity = 0.9444, Precision = 0.5271, MCC = 0.6204, and Kappa = 0.5695.

***Keywords*:** Binary whale optimization algorithm, Imbalanced data problem, K-nearest neighbor, Support vector machine, Undersampling algorithm

1) บทนำ

ข้อมูลไม่สมดุลบนปัญหาการจำแนกข้อมูล (classification) นั้น หมายถึงความไม่เท่ากันของจำนวนกลุ่มตัวอย่างในแต่ละคลาส (class) ซึ่งสถานการณ์เช่นนี้มักพบบ่อยในหลายสายงาน เช่น การวินิจฉัยเพื่อจำแนกประเภทผู้ป่วยโรคมะเร็ง เมื่อจำนวนของผู้ที่ป่วยมีจำนวนน้อยกว่าผู้ที่ไม่ป่วยมาก [1] การตรวจจับธุรกรรมทางการเงินที่ผิดปกติ เมื่อจำนวนผู้ทุจริตมีจำนวนน้อยกว่าผู้ที่ไม่ทุจริต [2] การจำแนกประเภทผู้ป่วยโรคเบาหวาน เมื่อจำนวนข้อมูลที่มีพบกว่าผู้ที่ป่วยมีจำนวนน้อยกว่าผู้ที่ไม่ป่วยอย่างเห็นได้ชัด [3] และ อีกในหลาย ๆ สายงานที่มักจะพบเจอกับปัญหาข้อมูลไม่สมดุล หนึ่งในปัญหาที่มักพบสำหรับปัญหาการจำแนกประเภทเมื่อข้อมูลเกิดความไม่สมดุลขึ้น คือ ถ้านำข้อมูลไม่สมดุลไปสร้างตัวแบบโดยใช้การเรียนรู้ของเครื่อง (machine learning) โดยตรงจะทำให้ผลลัพธ์ที่ได้มีความไม่ถูกต้องและมีความเอนเอียง (bias) ในการทำนายผล ที่เป็นเช่นนี้เพราะการเรียนรู้ของเครื่องมักจะถูกออกแบบมาเพื่อปรับปรุงความถูกต้องและลดความคลาดเคลื่อน (error) ดังนั้น ตัวแบบที่ได้จะทำการทำนายผลลัพธ์ไปทางคลาสกลุ่มมาก (majority class) โดยจะไม่เน้นการทำนายไปที่คลาสกลุ่มน้อย (minority class)

เทคนิคการสุ่มตัวอย่าง (resampling technique) เป็นเทคนิคที่สามารถแก้ปัญหาของข้อมูลที่ไม่สมดุลได้ และเมื่อข้อมูลมีการปรับให้มีความสมดุลแล้วก็จะสามารถนำข้อมูลเข้าสู่การสร้างตัวแบบโดยใช้การเรียนรู้ของเครื่องได้อย่างมีประสิทธิภาพมากขึ้น เทคนิคการสุ่มตัวอย่างสามารถแบ่งออกได้เป็น 3 วิธีหลัก ๆ คือ วิธีการสุ่มตัวอย่างลด (undersampling methods) วิธีการสุ่มตัวอย่างเพิ่ม (oversampling methods) และ วิธีการแบบผสม (hybrid methods) [4] ซึ่งจะเห็นได้ว่ามีวิธีการสุ่มตัวอย่างหลายแบบโดยหนึ่งในวิธีที่มีประสิทธิภาพคือการสุ่มตัวอย่างลด วิธีนี้จะช่วยลดจำนวนข้อมูลของคลาสกลุ่มมากลงมาเพื่อให้มีขนาดเท่ากับจำนวนข้อมูลของคลาสกลุ่มน้อย ถึงแม้ว่าวิธีการสุ่มตัวอย่างลดอาจจะทำให้สูญเสียข้อมูลที่มีความสำคัญไป แต่อย่างไรก็ตามวิธีนี้จะช่วยลดเวลาในการประมวลผลเพื่อสร้างตัวแบบของการเรียนรู้ของเครื่อง และ สามารถลดปัญหาการเกิดปัญหาเกินพอดี (overfitting) ในการสร้างตัวแบบได้

ในปัจจุบันมีการใช้ขั้นตอนวิธีที่ได้รับแรงบันดาลใจจากธรรมชาติ (nature-inspired algorithms) กับการแก้ปัญหาข้อมูลไม่สมดุล ยกตัวอย่างเช่น ขั้นตอนวิธีการหาค่าเหมาะสมที่สุดแบบฝูงมด (ant colony optimization algorithms) ขั้นตอนวิธีการวิวัฒนาการ (evolutionary algorithms) ขั้นตอนวิธีเชิงพันธุกรรม (genetic algorithms) และขั้นตอนวิธีการปรับสมดุลกลุ่มปรับตัว (adaptive swarm balancing algorithms) โดยในปี ค.ศ.2013 Yu *et al.* [5] ได้นำเสนอขั้นตอนวิธีการสุ่มตัวอย่างลดแบบใหม่ที่มีแนวคิดมาจากการหาค่าเหมาะสมแบบฝูงมด ซึ่งเรียกว่า ACO Sampling โดยขั้นตอนวิธีนี้จะเป็นหาชุดข้อมูลย่อยที่เหมาะสมสำหรับคลาสกลุ่มมาก และได้ทำการประเมินขั้นตอนวิธีด้วยชุดข้อมูล DNA microarray 4 ชุด ที่เป็นข้อมูลไม่สมดุล โดยตัวแยกประเภทเทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) และผลลัพธ์แสดงให้เห็นว่าวิธีการที่นำเสนอนั้นมีประสิทธิภาพที่ดีกว่าขั้นตอนวิธีอื่น โดยฟังก์ชันวัตถุประสงค์ที่ใช้จะมีส่วนประกอบเป็น 3 ตัววัดประสิทธิภาพ ได้แก่ F1 score, G-mean และ ค่าพื้นที่ใต้กราฟ Receiver operating characteristic curve (AUROC) ต่อมาในปี ค.ศ. 2014 López *et al.* [6] ได้เสนอการใช้ขั้นตอนวิธีแบบ Iterative Instance Adjustment for Imbalanced Domains (IPADE-ID) ที่มีกรอบการทำงานแบบเชิงวิวัฒนาการ (evolutionary framework) ซึ่งผลลัพธ์ที่ได้พบว่าวิธีการที่นำเสนอมีประสิทธิภาพกว่าวิธีการอื่น โดยฟังก์ชันวัตถุประสงค์ใช้ตัววัดประสิทธิภาพ AUROC สำหรับการวิวัฒนาการผลเฉลย ในปี ค.ศ. 2016 Kim *et al.* [7] ได้เสนอแนวทางการปรับให้เหมาะสมของการสุ่มตัวอย่างตามคลัสเตอร์เพื่อเลือกตัวแทนที่เหมาะสม ซึ่งวิธีการนี้สามารถแก้ปัญหาข้อมูลไม่สมดุลได้ และได้ทำการทดสอบประสิทธิภาพของวิธีผสมระหว่างเทคโนโลยีคลัสเตอร์และขั้นตอนวิธีเชิงพันธุกรรมบนตัวแบบโครงข่ายประสาทเทียม โดยวิธีการที่นำเสนอสามารถใช้แก้ปัญหาการทำนายการล้มละลายของสถาบันการเงินได้สำเร็จโดยที่ใช้ตัววัด G-mean ในฟังก์ชันวัตถุประสงค์ และใน ปี ค.ศ. 2017 Li *et al.* [8] ได้นำเสนอขั้นตอนวิธีการปรับสมดุลกลุ่มปรับตัว ซึ่งพบว่าสามารถปรับปรุงประสิทธิภาพของข้อมูลขนาดใหญ่อีกทั้งยังพบว่ามีความสอดคล้องกับชุดข้อมูลทางการแพทย์ที่ไม่สมดุลขนาดใหญ่โดยทั่วไปอีกด้วย ซึ่งวิธีการที่นำเสนอทำให้ตัวแบบจำแนกประเภทมีความน่าเชื่อถือมากขึ้นและลดเวลาการทำงานให้สั้นลงเมื่อเทียบกับวิธี brute-force โดยที่ใช้ตัววัดประสิทธิภาพ kappa และ ค่าความถูกต้อง ในการประเมินฟังก์ชันวัตถุประสงค์

และเมื่อไม่นานมานี้ได้มีการสร้างขั้นตอนวิธีใหม่ที่ชื่อว่าขั้นตอนวิธีการหาค่าเหมาะสมที่สุดแบบไบนารีวาฬ (binary whale optimization algorithms) [9] ซึ่งเป็นขั้นตอนวิธีที่ได้รับความ

นิยมในการประยุกต์นำไปใช้กับหลากหลายศาสตร์ เช่น การแก้ปัญหาการเลือกคุณลักษณะที่เหมาะสมสำหรับตัวแบบ (feature selection) [10] การแก้ปัญหาทางด้านวิศวกรรมอิเล็กทรอนิกส์ [9] และ การแก้ปัญหาการหาค่าพารามิเตอร์ที่เหมาะสมสำหรับตัวแบบ ซึ่งจะเห็นได้ว่าขั้นตอนวิธีนี้ค่อนข้างน่าสนใจเป็นอย่างมากถ้านำมาประยุกต์ใช้เพื่อแก้ปัญหาข้อมูลไม่สมดุล โดยในปี ค.ศ. 2019 Hussien *et al.* ได้เสนอขั้นตอนวิธีการหาค่าเหมาะสมที่สุดแบบวาฬแบบใหม่โดยมีการใช้ไบนารีมาประยุกต์เข้ากับการเลือกคุณลักษณะย่อยที่เหมาะสมเพื่อลดจำนวนมิติของชุดข้อมูลบนปัญหาการจำแนกประเภท [11] โดยวิธีการใหม่นี้จะใช้ฟังก์ชันการถ่ายโอนซิกมอยด์ (S shape) ในการหาคุณลักษณะที่เหมาะสม ผลการวิจัยพบว่าขั้นตอนวิธีใหม่นี้มีประสิทธิภาพในการค้นหาคุณลักษณะที่เหมาะสมได้เป็นอย่างดี ในปี ค.ศ. 2020 V. Kumar และ D. Kumar ได้ปรับปรุง WOA ให้กลายเป็น BWOA โดยปรับเปลี่ยนจากเวกเตอร์ผลเฉลยที่อยู่ในปริภูมิของจำนวนจริง ให้กลายเป็นเวกเตอร์ผลเฉลยที่อยู่ในปริภูมิของไบนารี [9] และใช้ฟังก์ชันการถ่ายโอนซิกมอยด์เพื่อใช้ในการปรับปรุงตำแหน่งของผลเฉลย ผลการทดสอบเมื่อเทียบกับขั้นตอนวิธีอื่น ๆ พบว่าขั้นตอนวิธีนี้มีประสิทธิภาพที่เหนือกว่า อีกทั้งยังทดสอบกับปัญหาการสั่งการเดินเครื่องโรงไฟฟ้า (unit commitment) ซึ่งเป็นปัญหาทางวิศวกรรมไฟฟ้า ผลการทดสอบพบว่าขั้นตอนวิธีที่นำเสนอมีประสิทธิภาพที่ดีกว่าขั้นตอนวิธีอื่น ๆ ในแง่ของต้นทุนการผลิตที่ต่ำลง และในปีเดียวกันนี้ Sayed *et al.* ได้นำเสนอตัวแบบอัจฉริยะแบบผสมที่ใช้ขั้นตอนวิธีการวิเคราะห์คลัสเตอร์ร่วมกับขั้นตอนวิธีแบบไบนารีเวอร์ชันของ WOA และ Moth flam optimization [12] เพื่อใช้ในการเลือกคุณลักษณะที่สำคัญสำหรับวิเคราะห์ข้อมูลมะเร็งเต้านมทางคลินิก ผลการทดลองแสดงให้เห็นว่าขั้นตอนวิธีที่นำเสนอสามารถสร้างคุณลักษณะของข้อมูลที่มีความหมายและมีนัยสำคัญได้ และงานวิจัยสุดท้ายในปี ค.ศ. 2020 Hussien *et al.* ได้เสนอ BWOA ที่ใช้ฟังก์ชันการถ่ายโอนแบบ S shape และ V shape [13] ผลลัพธ์พบว่าขั้นตอนวิธีที่นำเสนอมีประสิทธิภาพดีกว่าขั้นตอนวิธีอื่น ๆ และได้นำขั้นตอนวิธีนี้ไปแก้ปัญหาทางวิศวกรรม และปัญหาการเดินทางของพนักงาน (travelling salesman problem) จากการศึกษาพบว่าได้ผลลัพธ์ที่ดีกว่าขั้นตอนวิธีอื่นทั้งด้านความแม่นยำและความเร็ว

จากงานวิจัยต่าง ๆ จะพบว่าขั้นตอนวิธีการหาค่าเหมาะสมแบบไบนารี่วาฬนั้น มีความสามารถที่โดดเด่นในการตัดสินใจว่าจะเลือกหรือไม่เลือกสิ่งใดตามฟังก์ชันวัตถุประสงค์ที่กำหนด เช่น การเลือกคุณลักษณะที่เหมาะสมสำหรับการสร้างตัวแบบและตัดคุณลักษณะที่ไม่เหมาะสมทิ้งไป การเลือกสั่งการเดินเครื่องโรงไฟฟ้าว่าเครื่องใดควรทำงานเพื่อให้ได้ประโยชน์สูงสุด เป็นต้น จะเห็นได้ว่าขั้นตอนวิธีการหาค่าเหมาะสมแบบไบนารี่วาฬมีประสิทธิภาพที่ดีจึงทำให้หลายงานวิจัยมีการประยุกต์ใช้ขั้นตอนวิธีนี้กันอย่างแพร่หลาย ด้วยความสามารถของขั้นตอนวิธีที่ได้กล่าวมาข้างต้น สามารถนำมาประยุกต์ใช้ในการเลือกกลุ่มตัวอย่างข้อมูลย่อยที่มีแนวโน้มที่จะเป็นตัวแทนที่ดีของกลุ่มตัวอย่างข้อมูลขนาดใหญ่ได้ ดังนั้น งานวิจัยนี้จึงได้นำเสนอขั้นตอนวิธีใหม่ที่อิงการสุ่มตัวอย่างลดโดยเกิดจากการรวมการทำงานของ ขั้นตอนวิธีการหาค่าเหมาะสมที่สุดแบบไบนารีวาฬ และขั้นตอนวิธีเคเพื่อนบ้านใกล้ที่สุด (K-nearest neighbor) [14] และเนื่องจากการทำงานของขั้นตอนวิธีไบนารี่วาฬมีการวนซ้ำเพื่อคำนวณฟังก์ชันวัตถุประสงค์จำนวนมาก ในงานวิจัยนี้จึงได้เลือกขั้นตอนวิธีเคเพื่อนบ้านใกล้ที่สุดในการสร้างตัวแบบเพื่อประเมินค่าของฟังก์ชันวัตถุประสงค์ เพราะง่ายและไม่ซับซ้อน หลังจากที่ได้ทำการวิเคราะห์เพื่อดึงชุดข้อมูลย่อยที่เป็นตัวแทนที่ดีที่สุดของคลาสกลุ่มมากด้วยขั้นตอนวิธีที่นำเสนอเสร็จแล้ว จะนำข้อมูลที่มีการปรับจำนวนทั้งสองคลาสให้มีความสมดุลแล้วนั้น เข้าสู่การสร้างตัวแบบโดยใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีน [15] เนื่องจากเทคนิคซัพพอร์ตเวกเตอร์แมชชีนเป็นเทคนิคที่มีประสิทธิภาพที่ดีในการแก้ปัญหาการจำแนกประเภทข้อมูล เหมาะสำหรับข้อมูลที่มีขนาดไม่ใหญ่มากและทำงานได้ดีแม้ว่าข้อมูลจะมีจำนวนคุณลักษณะที่มาก [5] ถึงอย่างไรก็ตามเทคนิคซัพพอร์ตเวกเตอร์ก็มีข้อจำกัดเมื่อนำไปใช้กับข้อมูลไม่สมดุลซึ่งทำให้ประสิทธิภาพการทำนายลดลงอย่างมาก [16] หากขั้นตอนวิธีใดสามารถปรับปรุงสมดุลของข้อมูลได้ดี ก็จะสามารถดึงประสิทธิภาพในการสร้างตัวแบบของเทคนิคซัพพอร์ตเวกเตอร์แมชชีนได้ดีเช่นกัน ดังนั้นจึงเลือกใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนเพื่อประเมินประสิทธิภาพการทำงานของขั้นตอนวิธีที่นำเสนอ โดยในงานวิจัยนี้ได้มีการเปรียบเทียบผลลัพธ์ที่ได้กับขั้นตอนวิธีการสุ่มตัวอย่างลดอีก 3 วิธี ซึ่งได้แก่ การลดจำนวนตัวอย่างข้อมูลแบบสุ่ม (random undersampling) [17] คลัสเตอร์เซนทรอยด์ (cluster centroid) [18] และ เนียร์มิส (near-miss) [19] สำหรับตัววัดที่ใช้ในการประเมินประสิทธิภาพของขั้นตอนวิธีข้างต้นมีดังต่อไปนี้ Accuracy, F1score, G-mean, AUROC, AUPRC, Sensitivity, Precision, Matthew' s correlation coefficient (MCC), และ Cohen's Kappa Coefficient (Kappa)

2) วัตถุประสงค์ของงานวิจัย

เพื่อพัฒนาขั้นตอนวิธีใหม่ในการแก้ปัญหาข้อมูลไม่สมดุลอิงวิธีการสุ่มตัวอย่างลด โดยรวมการทำงานของขั้นตอนวิธีการหาค่าเหมาะสมที่สุดแบบไบนารีวาฬ และขั้นตอนวิธีเคเพื่อนบ้านใกล้ที่สุด

3) ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

3.1) อัตราส่วนความไม่สมดุล (Imbalance Ratio)

อัตราส่วนความไม่สมดุลเป็นตัววัดที่เอาไว้บ่งบอกระดับความไม่สมดุลของข้อมูลที่มีลักษณะสองคลาส แสดงในสมการที่ 1 [20]

$$\text{ImbalanceRatio}(IR) = \frac{n^-}{n^+} \quad (1)$$

โดยที่ n^-คือ จำนวนของกลุ่มตัวอย่างในคลาสลบ (คลาสกลุ่มมาก) และ n^+คือ จำนวนของกลุ่มตัวอย่างในคลาสบวก (คลาสกลุ่มน้อย) ซึ่งตัววัดนี้สามารถใช้เพื่อเรียงลำดับความแตกต่างของความไม่สมดุลของข้อมูลแต่ละชุดได้ หากค่าอัตราส่วนความไม่สมดุลมีค่ามากกว่าหนึ่งหรือน้อยกว่าหนึ่งอาจบ่งบอกได้ว่าชุดข้อมูลนั้นเป็นข้อมูลที่ไม่สมดุล ถึงอย่างไรก็ตามตัววัดนี้ก็อาจจะไม่ใช่ตัววัดความไม่สมดุลของข้อมูลที่ดีเสมอไป ทั้งนี้ขึ้นอยู่กับความซับซ้อนของข้อมูลที่นำมาวิเคราะห์ด้วย [4]

3.2) วิธีการสุ่มตัวอย่างลด (Undersampling Methods)

เป็นวิธีการที่ใช้เพื่อเลือกหาชุดข้อมูลย่อยเพื่อเป็นตัวแทนของข้อมูลคลาสกลุ่มมาก หลังจากนั้นจะนำข้อมูลที่ได้เข้าสู่การสร้างตัวแบบในปัญหาการจำแนกประเภทต่อไป โดยวิธีการสุ่มตัวอย่างลดที่งานวิจัยนี้นำมาเปรียบเทียบกับขั้นตอนวิธีที่นำเสนอมีทั้งหมด 3 ขั้นตอนวิธี ได้แก่ การลดจำนวนตัวอย่างข้อมูลแบบสุ่ม คลัสเตอร์เซนทรอยด์ และ เนียร์มิส

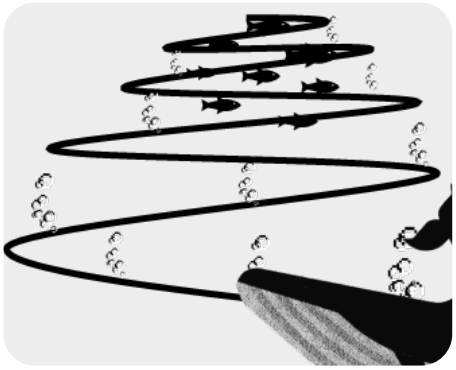
3.2.1) การลดจำนวนตัวอย่างข้อมูลแบบสุ่ม จะทำการสุ่มเลือกหาชุดข้อมูลย่อยจากคลาสกลุ่มมาก เนื่องจากเป็นวิธีการแบบสุ่มทำให้ในแต่ละครั้งที่ทำการสุ่มก็จะได้ชุดข้อมูลย่อยที่แตกต่างกันออกไป [17]

3.2.2) คลัสเตอร์เซนทรอยด์ จะทำการสร้างจุดข้อมูลใหม่หรือชุดข้อมูลย่อยใหม่ ที่มีจำนวนเท่ากับขนาดของคลาสกลุ่มน้อย ซึ่งเบื้องหลังการทำงานของวิธีนี้คือใช้ K-means algorithm โดยทำการระบุจำนวนกลุ่มหรือคลัสเตอร์ (cluster) ที่ต้องการจะแบ่ง ซึ่งจำนวนกลุ่มที่ว่านี้จะต้องมีขนาดเท่ากับจำนวนของคลาสกลุ่มน้อย หลังจากนั้น จะทำการสร้างเซนทรอยด์บนข้อมูลคลาสกลุ่มมาก และจะทำการปรับค่าตำแหน่งเซนทรอยด์ไปจนกว่าเซนทรอยด์จะไม่มีการเปลี่ยนแปลงตำแหน่ง สุดท้ายเซนทรอยด์ที่ถูกสร้างจะกลายเป็นชุดข้อมูลย่อยที่เป็นตัวแทนของคลาสกลุ่มมาก [18]

3.2.3) เนียร์มิส จะทำการเลือกชุดข้อมูลย่อยจากคลาสกลุ่มมาก โดยในแต่ละจุดข้อมูลของคลาสกลุ่มมากจะทำการหาจุดข้อมูลของคลาสกลุ่มน้อย 3 จุดที่อยู่ใกล้ที่สุด ทำการหาระยะทางระหว่าง 3 จุดนั้น แล้วคำนวณระยะทางเฉลี่ย ทำแบบนี้ทุกจุดข้อมูลของคลาสกลุ่มมาก หลังจากนั้นนำระยะทางเฉลี่ยที่คำนวณได้มาเรียงจากน้อยไปมาก และจะทำการเลือกชุดข้อมูลย่อยของคลาสกลุ่มมากจากจุดข้อมูลที่มีระยะทางเฉลี่ยน้อยที่สุด จำนวนเท่ากับขนาดของคลาสกลุ่มน้อย [19]

3.3) ขั้นตอนวิธีการหาค่าเหมาะสมที่สุดแบบวาฬ (Whale Optimization Algorithm: WOA)

WOA เป็นหนึ่งในขั้นตอนวิธีการหาค่าเหมาะสมที่สุดแบบเมตาฮิวริสติก (meta heuristic optimization algorithm) ที่ได้แรงบันดาลใจในการสร้างมาจากพฤติกรรมการล่าเหยื่อของวาฬหลังค่อม (humpback whale) โดยพฤติกรรมการล่าเหยื่อของวาฬชนิดนี้จะแบ่งออกเป็นสองเฟส เฟสแรกวาฬหลังค่อมจะทำการสำรวจหาเหยื่อ (exploration phase) โดยเหยื่อของวาฬหลังค่อมจะเป็นปลาตัวเล็กหรือแพลงก์ตอน หลังจากที่เจอเหยื่อแล้วจะเข้าสู่เฟสที่สอง โดยวาฬหลังค่อมจะทำการปิดล้อมเหยื่อและโจมตีเหยื่อ (exploitation phase) โดยวิธีการโจมตีเหยื่อวาฬหลังค่อมจะว่ายวนรอบเหยื่อพร้อมกับปล่อยฟองอากาศมาล้อมรอบเหยื่อซึ่งจะมีลักษณะคล้ายเลข ‘9’ แสดงดังรูปที่ 1 ด้วย พฤติกรรมการล่าเหยื่อที่น่าสนใจนี้จึงทำให้ Mirjalili และ Lewis [21] ทำการศึกษาและพัฒนา WOA ขึ้นมาเพื่อประยุกต์ใช้ในการแก้ปัญหาต่าง ๆ ได้อย่างมากมาย



รูปที่ 1 : การปล่อยฟองอากาศของวาฬหลังค่อม

พิจารณาฟังก์ชัน $f(x)$ ที่นิยามบนเซต $W \subset \mathbb{R}^d$ ซึ่งเป็นเซตที่มีขอบเขต โดยในที่นี้จะเรียกฟังก์ชัน f ว่าฟังก์ชันวัตถุประสงค์ (fitness function) โดยเป้าหมายของฟังก์ชันนี้คือหา $x_0 \in W$ ที่ทำให้ฟังก์ชัน $f(x)$ มีค่ามากที่สุด (หรือน้อยที่สุด) เริ่มต้นด้วยการระบุจำนวนจุด $x_1, \ldots, x_m$ ใน W โดยจุดเหล่านี้จะเปรียบเสมือนเวกเตอร์บอกตำแหน่งของวาฬแต่ละตัว ซึ่งมีทั้งหมด m ตัว (หรือ เปรียบเสมือนผลเฉลยของฟังก์ชันทั้งหมด m ผลเฉลย) ซึ่งจุดเหล่านี้จะถูกปรับปรุงตำแหน่งไปตามรอบการวนซ้ำสูงสุดที่กำหนดไว้ (maximum iteration) และสุดท้ายจะได้ x_0 ที่ทำให้ฟังก์ชัน $f(x)$ มีค่ามากที่สุด (หรือน้อยที่สุด)

เนื่องจากวาฬทั้งหมด m ตัว จะทำงานโดยอิสระจากวาฬตัวอื่น เราจึงได้ให้คำอธิบายสำหรับวาฬแต่ละตัวและกำหนดสัญลักษณ์ไว้ดังนี้ ให้ $\overrightarrow{X_j}(t)$ เป็นเวกเตอร์บอกตำแหน่งของวาฬตัวที่ j ณ รอบการวนซ้ำ t (นั่นคือ $\overrightarrow{X_j}(0) = \boldsymbol{x_j}$โดยที่ $j = 1, \ldots, m$ และกำหนดให้ตำแหน่งในตอนเริ่มต้น ($t = 0$) จะถูกสร้างขึ้นมาด้วยการสุ่มตัวเลขในจำนวนจริง) และ $\overrightarrow{X^*}(t)$ จะบ่งบอกถึงตำแหน่งของวาฬตัวที่ดีที่สุดที่ทำให้ฟังก์ชันวัตถุประสงค์มีค่ามากที่สุด (หรือน้อยที่สุด) ตั้งแต่เริ่มการวนซ้ำจนจบรอบการวนซ้ำ

ในการปรับปรุงเวกเตอร์บอกตำแหน่งของวาฬในทั้งสองเฟส ตัวแปรที่สำคัญแสดงดังสมการที่ 2

$$a(t) = 2\left(1 - \frac{t}{MaxIter}\right) \quad (2)$$

โดยที่ t คือ ครั้งการวนรอบปัจจุบัน และ $MaxIter$ คือ การวนซ้ำสูงสุดก่อนการหยุดการทำงาน โดยค่า $a(t)$ นี้จะลดลงแบบเชิงเส้นตั้งแต่ 2 ถึง 0 โดยขึ้นอยู่กับครั้งการวนรอบปัจจุบัน และเปรียบเสมือนตัวแปรที่เอาไว้กำหนดขนาดของการค้นหาเหยื่อของวาฬ ยิ่งในช่วงแรกของการค้นหาวาฬ วาฬจะออกไปได้ไกลมากแต่เมื่อถึงช่วงท้ายวาฬจะเริ่มไม่ไปห่างจากบริเวณที่หาอาหาร และเพื่อให้เข้าใจได้ง่ายค่า $a(t)$ จะถูกเขียนแทนด้วย a

3.3.1) ตัวแบบการปิดล้อมและโจมตี (Exploitation Phase) อันดับแรกในการล่าเหยื่อของวาฬหลังค่อมเมื่อเจอเหยื่อแล้ววาฬหลังค่อมจะทำการเข้าไปปิดล้อมเหยื่อ (Encircling method) โดยลักษณะพฤติกรรมสามารถจำลองและสร้างตัวแบบทางคณิตศาสตร์ดังสมการที่ (3) และ (4)

$$\overrightarrow{D_E} = \overrightarrow{D_E}(t,j) = |C \cdot \overrightarrow{X^*}(t) - \vec{X}_j(t)| \quad (3)$$

$$\vec{X}_j(t+1) = \overrightarrow{X^*}(t) - A \cdot \overrightarrow{D_E} \quad (4)$$

โดยที่ $|\cdot|$ เป็นการหาค่าสมบูรณ์ในแต่ละองค์ประกอบ (elementwise absolute), $A = A(t,j) = a \cdot (2r - 1)$ และ $C = C(t,j) = 2r$ เมื่อ $r = r(t,j)$ เป็นตัวแปรแบบสุ่มที่มีการแจกแจงแบบยูนิฟอร์ม (Uniform Distribution) อยู่ในช่วง [0,1] ดังนั้นค่า *A* และ *C* จึงมีการแจกแจงแบบยูนิฟอร์มซึ่งอยู่ในช่วง $[a, -a]$ และ [0,2] ตามลำดับ และ $\overrightarrow{D_E}$ บ่งบอกถึงระยะทางจากเวกเตอร์บอกตำแหน่งที่ดีที่สุดที่ถูกคูณด้วยตัวแปรสุ่ม C ถึงเวกเตอร์บอกตำแหน่งตัวที่ j ณ รอบการวนซ้ำที่ t

สมการที่ (4) เป็นสมการที่แสดงการอัพเดตเวกเตอร์บอกตำแหน่งของวาฬให้ปิดล้อมเข้าใกล้วาฬตัวที่ทำให้ค่าของฟังก์ชันวัตถุประสงค์ดีที่สุด หรือเข้าใกล้แหล่งอาหาร โดยจะเห็นว่าสมการการปรับปรุงเวกเตอร์บอกตำแหน่งแบบปิดล้อมจะขึ้นอยู่กับตัวแปร *A* และ *C* และพฤติกรรมการปิดล้อมที่ค่อย ๆ หดตัวเข้าหาเหยื่อก็ถูกจำลองด้วยการลดลงของค่าพารามิเตอร์ a

เมื่อวาฬหลังค่อมทำการปิดล้อมเหยื่อเรียบร้อยแล้ววาฬหลังค่อมจะทำการว่ายวนรอบเหยื่อพร้อมกับปล่อยฟองอากาศออกมา (bubble-net attacking method) โดยสมการการปรับปรุงเวกเตอร์บอกตำแหน่งของพฤติกรรมนี้ แสดงในสมการที่ (5) และ (6)

$$\overrightarrow{D_B} = \overrightarrow{D_B}(t,j) = |\overrightarrow{X^*}(t) - \vec{X}_j(t)| \quad (5)$$

$$\vec{X}_j(t+1) = \overrightarrow{X^*}(t) + \overrightarrow{D_B} e^{bl} cos(2\pi l) \quad (6)$$

โดยที่ $\overrightarrow{D_B}$ บ่งบอกถึงระยะทางจากเวกเตอร์บอกตำแหน่งที่ดีที่สุดถึงเวกเตอร์บอกตำแหน่งตัวที่ j ณ รอบการวนซ้ำที่ t และ $l = l(t,j)$ เป็นตัวแปรแบบสุ่มที่มีการแจกแจงแบบยูนิฟอร์มอยู่ในช่วง [-1,1] และ b คือค่าคงที่ที่ใช้ในการกำหนดรูปร่างการหมุนวนของวาฬ

3.3.2) ตัวแบบการสำรวจหาเหยื่อ (Exploration Phase) เพื่อให้การค้นหาเหยื่อหรือการค้นหาผลเฉลยมีความหลากหลาย ดังนั้นต้องมีการปรับเปลี่ยนตำแหน่งที่จะให้วาฬไปหา ซึ่งจากเดิมวาฬจะต้องปรับปรุงเวกเตอร์บอกตำแหน่งไปหาวาฬตัวที่มีค่าฟังก์ชันวัตถุประสงค์ที่ดีที่สุด เปลี่ยนเป็นให้ปรับปรุงเวกเตอร์บอกตำแหน่งไปหาวาฬตัวอื่น ๆ เพื่อทำให้ได้ผลเฉลยที่มีความหลากหลายและมีโอกาสที่จะเกิดตำแหน่งที่มีค่าของฟังก์ชันวัตถุประสงค์ที่ดีกว่าเดิมได้ ซึ่งสามารถสร้างตัวแบบทางคณิตศาสตร์ได้ดังสมการที่ (7) และ (8)

$$\overrightarrow{D_R} = \overrightarrow{D_R}(t,j) = |C \cdot \vec{X}_{rand}(t) - \vec{X}_j(t)| \quad (7)$$

$$\vec{X}_j(t+1) = \vec{X}_{rand}(t) - A \cdot \overrightarrow{D_R} \quad (8)$$

โดยที่ $\overrightarrow{D_R}$ บ่งบอกถึงระยะทางจากเวกเตอร์บอกตำแหน่งที่ถูกสุ่มมาหนึ่งตัวจากทั้งหมด m ตัว โดยคูณด้วยตัวแปรสุ่ม C ถึงเวกเตอร์บอกตำแหน่งตัวที่ j ณ รอบการวนซ้ำที่ t และ $\vec{X}_{rand}(t)$ คือ เวกเตอร์บอกตำแหน่งของวาฬตัวอื่น ๆ ที่ไม่ใช่ตัวที่ดีที่สุดซึ่งจะถูกเลือกมาแบบสุ่ม ณ เวลา t

3.3.3) การเปลี่ยนเฟส (Switching Phase) เนื่องจากมีเวกเตอร์บอกตำแหน่งของวาฬ (หรือผลเฉลย) หลายตัวในแต่ละรอบการวนซ้ำ ทำให้วาฬแต่ละตัวต้องมีการสุ่มเลือกพฤติกรรมว่าจะถูกปรับปรุงเวกเตอร์บอกตำแหน่งด้วยสมการใด ซึ่งเกณฑ์การเปลี่ยนเฟสจะแสดงดังสมการที่ (9)

$$\vec{X}_j(t+1) = \begin{cases} \overrightarrow{X^*}(t) - A \cdot \overrightarrow{D_E} & \text{ถ้า} p < 0.5 \text{และ} |A| < 1; \\ \overrightarrow{X^*}(t) + \overrightarrow{D_B} e^{bl} cos(2\pi l) & \text{ถ้า} p \geq 0.5; \\ \vec{X}_{rand}(t) - A \cdot \overrightarrow{D_R} & \text{ถ้า} p < 0.5 \text{และ} |A| \geq 1 \end{cases} \quad (9)$$

โดยที่ $p = p(t,j)$ เป็นตัวแปรแบบสุ่มที่มีการแจกแจงแบบยูนิฟอร์มบนช่วง [0,1]

3.4) ขั้นตอนวิธีการหาค่าเหมาะสมที่สุดแบบไบนารีวาฬ (Binary Whale Optimization Algorithm: BWOA)

BWOA ได้ถูกพัฒนามาจาก WOA [9] ซึ่งสามารถนำมาใช้ในการหาผลเฉลยแบบไบนารีเวกเตอร์ (binary vectors) โดยโดเมนของฟังก์ชันวัตถุประสงค์ คือ ปริภูมิ

$$\mathcal{X} = \{0,1\}^d = \prod_{i=1}^{d} \{0,1\}$$

จากส่วนประกอบของเวกเตอร์คือ 0 และ 1 ดังนั้นสามารถนำไปประยุกต์ใช้กับหลากหลายงานที่ต้องใช้การตัดสินใจแบบไบนารี (binary)

โดยที่ในแต่ละองค์ประกอบของเวกเตอร์บอกตำแหน่งจะถูกปรับปรุง โดยการเปรียบเทียบระหว่างค่าตัวแปรสุ่มและค่าของฟังก์ชันการถ่ายโอนซิกมอยด์ (sigmoid transfer function) ซึ่งฟังก์ชันการถ่ายโอนซิกมอยด์จะแสดงดังสมการที่ (10)

$$g(s) = \frac{1}{1 + e^{-10(s-0.5)}} \quad (10)$$

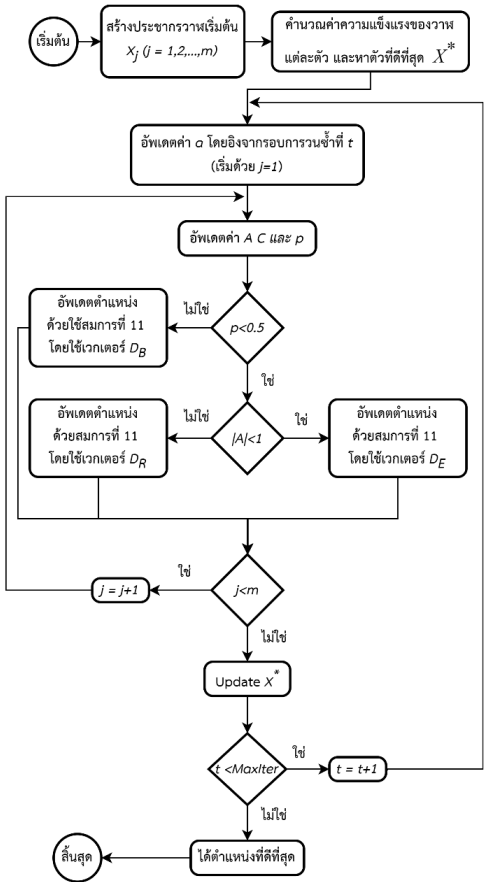
โดยที่ s เป็นค่าคงที่ที่เพิ่มขึ้นจาก -2 ถึง 2 ซึ่งส่งผลให้ฟังก์ชัน $g(s)$ จะมีค่าเพิ่มขึ้นจาก $\frac{1}{1+e^{25}} \approx 0$ ถึง $\frac{1}{1+e^{-15}} \approx 1$ และสมการการปรับปรุงเวกเตอร์บอกตำแหน่งจะแสดงดังสมการที่ (11)

$$\vec{X}_j(t+1,i) = \begin{cases} 1 - \vec{X}_j(t,i), & \text{ถ้า} r \leq g(A\vec{D}_0(i)) \\ \vec{X}_j(t,i), & \text{ถ้า} r > g(A\vec{D}_0(i)) \end{cases} \quad (11)$$

โดยที่ r เป็นตัวแปรสุ่มที่มีการแจกแจงแบบยูนิฟอร์มอยู่ในช่วง [0,1] และ i เป็นดัชนีของเวกเตอร์ $\vec{D}_0$ โดยที่เวกเตอร์ $\vec{D}_0$ จะมีเงื่อนไขในการเลือกดังสมการที่ (12) ซึ่งจะเห็นได้ว่าการปรับปรุงเวกเตอร์บอกตำแหน่งของวาฬในแต่ละองค์ประกอบจะอาศัยการเปรียบเทียบกับค่าของตัวแปรสุ่ม r

$$\vec{D}_0 = \begin{cases} \overrightarrow{D_E} & \text{ถ้า} p < 0.5 \text{ และ } |A| < 1; \\ \overrightarrow{D_B} & \text{ถ้า} p \geq 0.5; \\ \overrightarrow{D_R} & \text{ถ้า} p < 0.5 \text{ และ } A| \geq 1 \end{cases} \quad (12)$$

จากสมการที่ (12) จะเห็นได้ว่าเงื่อนไขในการปรับปรุงเวกเตอร์บอกตำแหน่งของ BWOA จะเหมือนกับ WOA ในสมการที่ (9) เพียงแต่แตกต่างกันที่สมการการปรับปรุงเวกเตอร์บอกตำแหน่งเพียงเท่านั้น ภาพขั้นตอนการทำงานของ BWOA แสดงดังรูปที่ 2



รูปที่ 2 : ขั้นตอนการทำงานของ BWOA

3.5) ตัวแบบการเรียนรู้ของเครื่อง

การเรียนรู้ของเครื่องสามารถแบ่งออกเป็น 3 ประเภทใหญ่ ๆ คือ การเรียนรู้แบบมีผู้สอน (supervised learning) การเรียนรู้แบบไม่มีผู้สอน (unsupervised learning) และ การเรียนรู้แบบเสริมแรง (reinforcement learning) ซึ่งในงานวิจัยจะใช้การเรียนรู้ของเครื่องประเภทการเรียนรู้แบบมีผู้สอนและใช้สำหรับงานการจำแนกประเภท (classification task) เพียงเท่านั้น

3.5.1) เคเพื่อนบ้านใกล้ที่สุด (K-nearest neighbors: K-NN) เป็นขั้นตอนวิธีที่มีพื้นฐานมาจากการหาระยะทาง โดยเริ่มจากการนำจุดข้อมูลใหม่มาพล็อตลงปริภูมิเดียวกับข้อมูลชุดฝึกสอน หลังจากนั้นหาระยะทางเทียบกับจุดข้อมูลชุดฝึกสอนว่าจุดข้อมูลใหม่นี้มีจุดข้อมูลชุดฝึกสอนใกล้จุดข้อมูลใหม่กี่จุด โดยที่จำนวนจุดที่ล้อมรอบจุดข้อมูลใหม่จะถูกกำหนดจำนวนด้วยพารามิเตอร์ K (โดยส่วนมากหากเป็นปัญหาสองคลาสมักจะกำหนด K เป็นจำนวนคี่) ในขั้นตอนสุดท้ายจะเป็นการระบุว่าจุดข้อมูลใหม่นี้จะถูกจำแนกว่าเป็นคลาสไหน โดยดูจากจุดข้อมูลชุดฝึกสอนที่ใกล้กับจุดข้อมูลใหม่มากที่สุดจำนวน K ตัว และดูว่าคลาสใดมีจำนวนมากที่สุด คลาสนั้นจะถูกติดฉลากให้กับจุดข้อมูลใหม่โดยทันที [14]

3.5.2) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machines: SVM) คือขั้นตอนวิธีที่อาศัยหลักการการสร้างไฮเปอร์เพลน (Hyperplane) เพื่อตัดแบ่งข้อมูลให้ออกเป็นกลุ่ม ๆ อย่างชัดเจน โดยจะมีการปรับสัมประสิทธิ์ของสมการวัตถุประสงค์ให้เหมาะสม เพื่อที่จะทำให้ไฮเปอร์เพลนสามารถแบ่งกลุ่มข้อมูลให้ดีที่สุด [15]

3.6) การประเมินตัวแบบ (Model Evaluation)

หลังจากทำการสร้างตัวแบบในการจำแนกประเภทข้อมูลแล้ว จะต้องมีการวัดประสิทธิภาพว่าตัวแบบมีว่าความเหมาะสมมากน้อยเพียงใด โดยตัววัดประสิทธิภาพโดยส่วนใหญ่จะมีพื้นฐานการสร้างมาจากเมทริกซ์ความสับสน (confusion matrix) [4] แสดงดังรูปที่ 3

		ผลทำนาย (Prediction)	
		Negative	Positive
ผลจริง (Actual)	Negative	TN	FP
	Positive	FN	TP

รูปที่ 3 : Confusion matrix

จากรูปที่ 3 Positive, Negative, TN, FN, FP, และ TP มีความหมายดังต่อไปนี้

- Positive คือ ค่าทางบวก ซึ่งจะถูกระบุให้กับคลาสที่สนใจที่จะทำนายผล
- Negative คือ ค่าทางลบ ซึ่งจะถูกระบุให้กับคลาสที่สนใจรองลงมา
- TN คือ ผลทำนายตรงกับผลจริงในทางลบ
- FN คือ ผลทำนายเป็นทางลบแต่ผลจริงเป็นทางบวก
- FP คือ ผลทำนายเป็นทางบวกแต่ผลจริงเป็นทางลบ
- TP คือ ผลทำนายตรงกับผลจริงในทางบวก

ตัววัดประสิทธิภาพของตัวแบบที่ใช้ในงานวิจัยนี้มีดังต่อไปนี้

- ค่าความถูกต้อง (Accuracy) แสดงดังสมการที่ (13)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (13)$$

- ค่าความแม่นยำ (Precision) แสดงดังสมการที่ (14)

$$\text{Precision} = \frac{TP}{TP + FP} \quad (14)$$

- ค่าความไว (Sensitivity) หรือเรียกอีกอย่างว่า ความระลึก (Recall) แสดงดังสมการที่ (15)

$$\text{Sensitivity}(\text{orRecall}) = \frac{TP}{TP + FN} \quad (15)$$

- ค่าเฉลี่ยเลขคณิตระหว่างค่าความแม่นยำและความไว (F1 score) แสดงดังสมการที่ (16)

$$\text{F1 score} = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \quad (16)$$

- ค่าเฉลี่ยเรขาคณิตระหว่างค่าความจำเพาะและค่าความไว (G-mean) [22] แสดงดังสมการที่ (17)

$$\text{G-mean} = \sqrt{Sensitivity \times Specificity} \quad (17)$$

โดยที่ Specificity คือ ค่าความจำเพาะ สามารถคำนวณได้ดังสมการที่ (18)

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (18)$$

- ค่า Matthew's correlation coefficient (MCC) [23] เป็นตัววัดที่ใช้วัดคุณภาพของปัญหาที่มีสองคลาส ซึ่งเป็นหน่วยวัดที่สมดุลแม้ขนาดของคลาสจะแตกต่างกัน แสดงดังสมการที่ (19)

$$\text{MCC} = \frac{(TN \times TP) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (19)$$

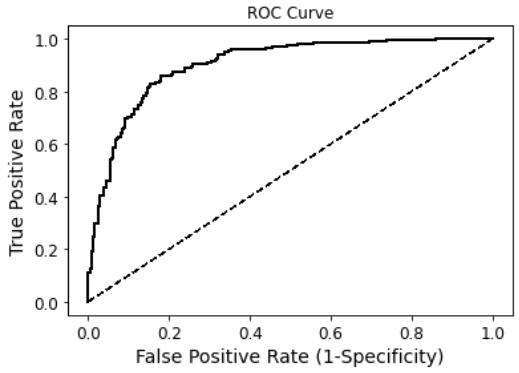
- ค่า Cohen's Kappa coefficient (kappa) [24] เป็นค่าตัวชี้วัดทางสถิติ ระหว่างผู้ให้ความเห็นสองฝ่ายว่ามีความสอดคล้องกันมากเพียงใด เมื่อเปรียบเทียบกับ เมทริกซ์ความสับสน ผู้ให้ความเห็นฝ่ายแรกก็คือ ผลจริง และผู้ให้ความเห็นฝ่ายที่สอง คือ ผลการทำนาย ซึ่งแสดงการคำนวณดังสมการที่ (20)

$$kappa = \frac{p_a - p_e}{1 - p_e} = 1 - \frac{1 - p_a}{1 - p_e} \quad (20)$$

โดยที่ p_a คือ ค่าความถูกต้อง และ

$$p_e = \frac{(TN+FP)(TN+FN)+(FN+TP)(FP+TP)}{(TP+TN+FP+FN)^2}$$

- ค่าพื้นที่ใต้กราฟ Receiver operating characteristic curve (AUROC) กราฟ ROC [25] คือกราฟที่พล็อตระหว่างอัตราผลบวกจริง (True Positive Rate : TPR) และ อัตราผลบวกเท็จ (False Positive Rate : FPR) แสดงดังรูปที่ 4

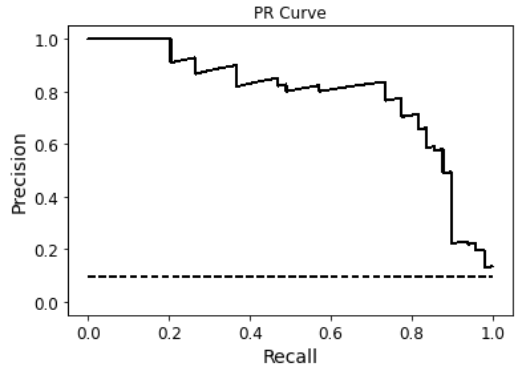


รูปที่ 4 : กราฟ ROC

โดยที่ TPR คือ ค่าความไว และ FPR สามารถคำนวณได้ดังนี้

$$\text{FPR} = \frac{FP}{FP+TN}$$

- ค่าพื้นที่ใต้กราฟ Precision-Recall curve (AUPRC) กราฟ PR [26] คือกราฟที่พล็อตระหว่างค่าความแม่นยำและค่าความระลึก แสดงดังรูปที่ 5



รูปที่ 5: กราฟ PR

3.7) K-fold Cross Validation

เป็นวิธีที่นิยมในการใช้ทดสอบประสิทธิภาพของตัวแบบ ซึ่งมักจะใช้ร่วมกับการฝึกสอนตัวแบบการเรียนรู้ของเครื่อง โดยจะแบ่งข้อมูลชุดฝึกสอนออกเป็น k ส่วน ซึ่งแต่ละส่วนจะมีจำนวนข้อมูลที่เท่ากัน หลังจากนั้นข้อมูลหนึ่งส่วนจะใช้เป็นตัวทดสอบประสิทธิภาพของโมเดลและทำวนไปเช่นนี้จนครบจำนวนที่แบ่งไว้ [27]

4) วิธีดำเนินการวิจัย

4.1) ชุดข้อมูลที่ใช้ในงานวิจัย

ในงานวิจัยนี้เลือกใช้ข้อมูลจำนวน 12 ชุดข้อมูล ซึ่งเป็นชุดข้อมูลที่ตัวแปรตามแบ่งเป็นสองคลาส (Binary class) และค่าอัตราส่วนความไม่สมดุล (IR) มีค่าที่แตกต่างกัน ซึ่งจะเรียงชุดข้อมูลตาม IR จากน้อยไปมากดังแสดงในตารางที่ 1 โดยชุดข้อมูลเหล่านี้นำมาจากฐานข้อมูล KEEL [28] และ imbalanced-learn [29] ซึ่งเป็นฐานข้อมูลที่ได้รับความนิยมและมีความน่าเชื่อถือสูง เพราะงานวิจัยโดยส่วนใหญ่นิยมนำข้อมูลจากฐานข้อมูลนี้ไปใช้ทดสอบประสิทธิภาพของตัวแบบหรือขั้นตอนวิธีที่ได้ถูกสร้างขึ้นมาใหม่ ซึ่งจากตารางที่ 1 จะเห็นได้ว่าข้อมูลทั้ง 12 ชุดที่เลือกนำมาใช้ในงานวิจัยนี้ มีความหลากหลายทั้งทางด้านของจำนวนคุณลักษณะ ขนาดข้อมูล ขนาดข้อมูลกลุ่มน้อย ขนาดข้อมูลกลุ่มมาก และ ค่าอัตราส่วนความไม่สมดุล (IR)

ตารางที่ 1: รายละเอียดชุดข้อมูลที่ใช้ในงานวิจัย

ชื่อชุดข้อมูล	จำนวนคุณ-ลักษณะ	ขนาดข้อมูล	ขนาดข้อมูลกลุ่มน้อย	ขนาดข้อมูลกลุ่มมาก	IR
glass1	9	214	76	138	1.82
Iris0	4	150	50	100	2.00
glass-0-1-2-3_VS_4-5-6	9	214	51	163	3.20
ecoli2	7	336	52	284	5.46
glass6	9	214	29	185	6.38
ecoli	7	336	35	301	8.60
pen_digits	16	10,992	1,055	9,937	9.42
abalone	10	4,177	391	3,786	9.68
Libras_move	90	360	24	336	14.00
solar_flare_m0	32	1,389	68	1,321	19.43
yeast_me2	8	1,484	51	1,433	28.10
mammography	6	11,183	260	10,923	42.01

4.2) วิธีการทำงานของขั้นตอนวิธีที่นำเสนอ

เราได้นำเสนอการประยุกต์รวมการทำงานของขั้นตอนวิธีการหาค่าเหมาะสมที่สุดแบบไบนารีวาฬ และ ขั้นตอนวิธีเคเพื่อนบ้านใกล้ที่สุดเพื่อแก้ปัญหาข้อมูลไม่สมดุล โดยขั้นตอนวิธีที่นำเสนอนี้จะอิงวิธีการสุ่มตัวอย่างลด ซึ่งวิธีการทำงานของขั้นตอนวิธีที่เสนอสามารถอธิบายได้ดังต่อไปนี้

ให้ D เป็นเซตของข้อมูลฝึกสอน จากนั้นแบ่งเซต D ออกเป็น 2 เซต คือ D^-เป็นเซตของคลาสกลุ่มมาก และ D^+แทนเซตของคลาสกลุ่มน้อย และให้ $d = |D^-|$ แทนจำนวนของตัวอย่างในคลาสกลุ่มมาก และ $n^+ = |D^+|$ แทนจำนวนของตัวอย่างในคลาสกลุ่มน้อย

วัตถุประสงค์ของงานวิจัยนี้คือ ต้องการสุ่มตัวอย่างลดข้อมูลคลาสกลุ่มมาก โดยเราต้องการหาเซตย่อย $D_{red}^- \subset D^-$ซึ่ง $|D_{red}^-| \approx |D^+|$ ในขณะเดียวกันเซตย่อยที่ถูกเลือกนี้จะถูกนำมาเป็นข้อมูลในการสร้างตัวแบบ ซึ่งตอนนี้ $D_{red}^- \cup D^+$ จะกลายเป็นข้อมูลชุดฝึกสอนแล้ว และฟังก์ชันวัตถุประสงค์ คือ

$$f = f(A) := (1 - \text{F1 score})^2 + (1 - \text{AUROC})^2 + (1 - \text{Sensitivity})^2 + \beta(n^- - n^+)^2 \quad (21)$$

โดยที่ $A \subseteq D^-$ เป็นเซตย่อยของคลาสกลุ่มมาก, $n^- = n^-(A) = |A|$ เป็นจำนวนของกลุ่มตัวอย่างที่อยู่ใน A, β เป็นพารามิเตอร์ที่ไม่ติดลบ ซึ่งเปรียบเสมือนกับค่า Penalty หากเลือกเซตย่อยที่มีจำนวนของตัวอย่างไม่เท่ากับจำนวนของคลาสกลุ่มน้อย กำหนดให้ ค่า F1 score, AUROC และ Sensitivity จะเป็นค่าที่ได้จากการทำ 10-fold cross validation ที่ใช้ข้อมูลฝึกสอน $A \cup D^+$ เรียบร้อยแล้ว

ฟังก์ชันวัตถุประสงค์ที่เราได้สร้างขึ้นมานั้น เราได้ฟังก์ชัน

$$f : 2^{D^-} \to [0, \infty)$$

นิยามบนเซตกำลัง 2^{D^-}ของเซต D^- (ซึ่งเป็นโอกาสของการเกิดเซตย่อยทั้งหมดที่เป็นไปจากการเลือกเซตย่อยจากคลาสกลุ่มมาก และส่งไปยังเซต $[0, \infty)$) ซึ่งเราต้องการหาเซตย่อยที่ทำให้ฟังก์ชันนี้มีค่าต่ำที่สุด

จากฟังก์ชันวัตถุประสงค์เราต้องการเลือกเซตย่อยออกมาจากคลาสกลุ่มมาก และเรามี $D^- = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_d\}$ เป็นกลุ่มตัวอย่างของคลาสกลุ่มมาก โดยเราจะกำหนดฟังก์ชันที่มีคุณสมบัติเป็นฟังก์ชันหนึ่งต่อหนึ่งคือ

$$\phi : \{0,1\}^d \to 2^{D^-}$$

ซึ่งนิยามโดย

$$\phi(\vec{X}) = \{\boldsymbol{x}_i \in D^- : \overrightarrow{X}(i) = 1\} \quad (i = 1, \dots, d)$$

โดยเงื่อนไขในการเลือกเซตย่อย คือ เมื่อพบว่าองค์ประกอบของ $\vec{X}$ ณ ดัชนีที่ $\boldsymbol{i}$ ใด ๆ ที่มีค่าเท่ากับ 1 จะถูกกำหนดให้ไปเลือกตัวอย่างตัวที่ $\boldsymbol{i}$ ในD^-มาเป็นเซตย่อยของคลาสกลุ่มมาก และจะได้ฟังก์ชันประกอบ คือ

$$f \circ \phi : \{0,1\}^d \to [0, \infty)$$

ซึ่งจะเป็นฟังก์ชันประกอบที่เราต้องการที่จะหาค่าต่ำสุด เนื่องจากโดเมนของฟังก์ชันประกอบนี้เป็นปริภูมิของไบนารีเวกเตอร์ ด้วยคุณสมบัตินี้ทำให้สามารถนำขั้นตอนวิธีการหาค่าเหมาะสมที่สุดแบบไบนารีวาฬเข้ามาช่วยหาค่าที่ต่ำที่สุดของฟังก์ชันวัตถุประสงค์ $f \circ \phi$ และตัวแบบที่เราใช้ในการประเมินค่าฟังก์ชันวัตถุประสงค์จะเป็นตัวแบบที่เข้าใจง่ายและไม่ซับซ้อน นั้นคือ เคเพื่อนบ้านใกล้ที่สุด ซึ่งได้กำหนดพารามิเตอร์ $K = 1$ เพื่อลดความซับซ้อนของการทำงานและเวลาในการคำนวณ

4.3) การทดสอบประสิทธิภาพของขั้นตอนวิธีที่เสนอ

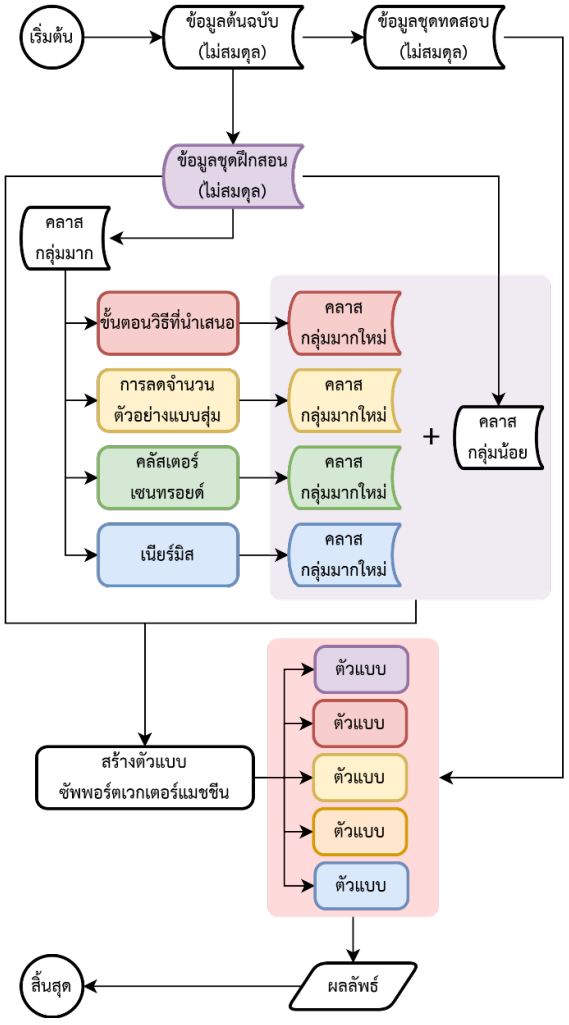
ขั้นตอนการทดสอบประสิทธิภาพมีดังต่อไปนี้

1. ในแต่ละชุดข้อมูล (12 ชุดข้อมูล) จะถูกแบ่งออกเป็น 2 ชุด คือ ข้อมูลชุดฝึกสอน (training set) และ ข้อมูลชุดทดสอบ (testing set) ในอัตราส่วน 80:20 ซึ่งการแบ่งในครั้งนี้อัตราส่วนความไม่สมดุลระหว่างสองคลาสของข้อมูลชุดฝึกสอนและข้อมูลชุดทดสอบยังคงมีความใกล้เคียงกับข้อมูลต้นฉบับ
2. ข้อมูลชุดฝึกสอนแต่ละชุดจะถูกแบ่งออกเป็น 2 กลุ่ม คือ คลาสกลุ่มมาก D^- และคลาสกลุ่มน้อย D^+
3. ทำการสุ่มตัวอย่างลดข้อมูลคลาสกลุ่มมาก D_{red}^- โดยใช้ขั้นตอนวิธีทั้ง 4 วิธี ได้แก่ ขั้นตอนวิธีที่นำเสนอ การลดจำนวนตัวอย่างข้อมูลแบบสุ่ม คลัสเตอร์เซนทรอยด์ และเนียร์มิส ในขณะที่ข้อมูลคลาสกลุ่มน้อยยังคงไว้เหมือนเดิม ในกรณีของขั้นตอนวิธีที่นำเสนอ เราต้องการหาไบนารีเวกเตอร์ $\overrightarrow{X^*}$ ของ $f \circ \phi$ ซึ่งเซตย่อยของคลาสกลุ่มมากคือ $D_{red}^- = \{\boldsymbol{x}_i \in D^- : \overrightarrow{X^*}(i) = 1\}$ เราได้กำหนดจำนวนผลเฉลยเท่ากับ 20 ผลเฉลย ($m = 20$) และกำหนด β เท่ากับ 100 และเงื่อนไขในการหยุดการทำงานของขั้นตอนวิธีที่นำเสนอ คือ
   - หากค่าของฟังก์ชันวัตถุประสงค์เท่ากับศูนย์
   - หากค่าของฟังก์ชันวัตถุประสงค์ที่ดีที่สุดไม่มีการเปลี่ยนแปลงภายใน 350 รอบการวนซ้ำ

- หากครบรอบการวนซ้ำสูงสุด คือ 1,000 รอบการวนซ้ำ

4. นำข้อมูลฝึกสอน $D_{red}^{-} \cup D^{+}$ ที่ผ่านการสุ่มตัวอย่างลดในแต่ละขั้นตอนวิธี มาสร้างตัวแบบด้วยเทคนิคซัพพอร์ตเวกเตอร์แมชชีน โดยใช้ 10-fold cross validation สำหรับการหาค่าพารามิเตอร์ที่เหมาะสมที่สุด และใช้ F1 score เป็นตัววัดประสิทธิภาพ
5. ประเมินประสิทธิภาพด้วยข้อมูลชุดทดสอบ และวัดผลของตัววัดประสิทธิภาพต่าง ๆ ที่ได้กล่าวในหัวข้อ 3.6

การทำงานโดยภาพรวมของงานวิจัยนี้ แสดงแผนผังการทำงานดังรูปที่ 6



รูปที่ 6 : กระบวนการทำงานโดยภาพรวม

5) ผลการวิจัย

การหาเซตย่อยของคลาสกลุ่มมากที่เป็นตัวแทนที่เหมาะสมที่สุดด้วยขั้นตอนวิธีที่นำเสนอ สามารถแสดงออกมาในรูปแบบของกราฟการลู่เข้าของฟังก์ชันวัตถุประสงค์ของแต่ละชุดข้อมูล ซึ่งสามารถแสดงดังตารางที่ 2 โดยที่ แกน X และ Y ของกราฟ คือ ค่าวัตถุประสงค์และรอบการวนซ้ำ ตามลำดับ ซึ่งจากตารางที่ 2 จะเห็นได้ว่า แต่ละชุดข้อมูลจะมีค่าฟังก์ชันวัตถุประสงค์ที่แตกต่างกันออกไป อีกทั้งยังมีรอบการวนซ้ำที่หยุดทำงานที่แตกต่างกันด้วย

ผลการวิจัยพบว่าเมื่อนำข้อมูลที่ไม่ได้ผ่านการทำสมดุลข้อมูลและข้อมูลที่ผ่านการทำสมดุลข้อมูลด้วยขั้นตอนวิธีที่นำเสนอและขั้นตอนวิธีอื่นอีก 3 วิธี ได้แก่ การลดจำนวนตัวอย่างข้อมูลแบบสุ่ม คลัสเตอร์เซนทรอยด์ และ เนียร์มิส ผลลัพธ์จากการนำชุดข้อมูลทดสอบมาทดสอบผ่านตัวแบบซัพพอร์ตเวกเตอร์แมชชีนที่ได้หาค่าพารามิเตอร์อย่างเหมาะสมจากข้อมูลชุดฝึกสอนเรียบร้อยแล้ว ผลลัพธ์ของตัววัดประสิทธิภาพต่าง ๆ ในแต่ละชุดข้อมูล (12 ชุดข้อมูล) จะถูกนำมาหาค่าเฉลี่ยเลขคณิตซึ่งจะแสดงดังตารางที่ 3 และได้กำหนดตัวย่อที่อยู่บนหัวตารางที่ 3 ให้มีความหมายดังต่อไปนี้

- None หมายถึง การใส่ข้อมูลต้นฉบับลงไปโดยไม่มีการทำสมดุลข้อมูล
- CC หมายถึง วิธีการสุ่มแบบคลัสเตอร์เซนทรอยด์
- NM หมายถึง วิธีเนียร์มิส
- RUS หมายถึง วิธีการลดจำนวนตัวอย่างข้อมูลแบบสุ่ม
- PA หมายถึง ขั้นตอนวิธีที่รวมการทำงานของขั้นตอนวิธีการหาค่าเหมาะสมที่สุดแบบไบนารีวาฬ และขั้นตอนวิธีเคเพื่อนบ้านใกล้ที่สุด(ขั้นตอนวิธีที่นำเสนอ)

6) สรุปและอภิปรายผล

โดยปกติแล้วถ้าเรานำข้อมูลที่ไม่สมดุลมาสร้างตัวแบบทันทีจะทำให้ได้ผลลัพธ์ของการทำนายมีค่า Accuracy ที่สูงมากและส่งผลให้ Precision สูงตามไปด้วย แต่ในขณะเดียวกันจะเห็นได้ว่าค่า Sensitivity มีค่าที่ต่ำมาก นั้นหมายความว่า ตัวแบบซัพพอร์ตเวกเตอร์แมชชีนที่ถูกสร้างจากข้อมูลไม่สมดุลนั้นไม่สามารถทำนายคลาสกลุ่มน้อยได้ดีเท่าที่ควร จากผลลัพธ์ของประสิทธิภาพการทำงานของขั้นตอนวิธีที่นำเสนอข้างต้น จะเห็นได้ชัดเจน ว่าขั้นตอนวิธีที่เสนอมีค่าเฉลี่ยของตัววัดประสิทธิภาพที่ค่อนข้างสูงหลายตัววัด โดยตัววัดที่มีค่าเฉลี่ยสูงสุดมาเป็นอันดับหนึ่ง คือ G-mean, AUROC, AUPRC, Sensitivity และ MCC ตัววัดที่มีค่าเฉลี่ยสูงสุดอับดับที่สอง คือ Accuracy, F1 score, Precision และ kappa ที่เป็นเช่นนี้เพราะขั้นตอนวิธีที่นำเสนอมี

การเลือกชุดข้อมูลกลุ่มย่อยที่เป็นตัวแทนที่ดีของคลาสข้อมูลกลุ่มมากโดยอิงการเลือกจากฟังก์ชันวัตถุประสงค์ ซึ่งฟังก์ชันวัตถุประสงค์จะเป็นตัวบ่งบอกว่าชุดข้อมูลกลุ่มย่อยที่เลือกมาดีมากน้อยเพียงใด ซึ่งในการทำงานได้ทำการหาชุดข้อมูลกลุ่มย่อยที่ทำให้ค่าฟังก์ชันวัตถุประสงค์มีค่าต่ำที่สุดเท่าที่เป็นไปได้ ผลลัพธ์คือ กลุ่มข้อมูลย่อยนั้นจะมีขนาดของคลาสข้อมูลกลุ่มน้อยและคลาสข้อมูลกลุ่มมากใหม่ที่มีขนาดที่ใกล้เคียงกันมาก อีกทั้งค่าของ F1 score, AUROC และ Sensitivity ก็จะดีตามไปด้วย ซึ่งกระบวนการทำงานจะแตกต่างกับวิธีการสุ่มตัวอย่างลดอีก 3 วิธี โดยวิธีการลดจำนวนตัวอย่างข้อมูลแบบสุ่มจะเน้นไปที่การสุ่มตัวอย่างแบบไม่มีเป้าหมายหรือฟังก์ชันวัตถุประสงค์ วิธีคลัสเตอร์เซนทรอยด์จะเน้นไปที่การหาจุดกึ่งกลางของกลุ่มเพื่อเป็นตัวแทนของคลาสข้อมูลกลุ่มมาก และ วิธีเนียร์มิสจะเน้นการหาระยะระหว่างจุดข้อมูลเพื่อหาชุดข้อมูลย่อยของคลาสข้อมูลกลุ่มมาก

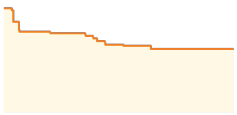
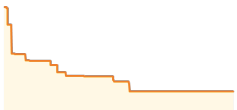
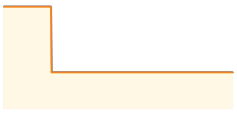
สรุปแล้วขั้นตอนวิธีที่นำเสนอที่เกิดจากการรวมการทำงานของขั้นตอนวิธีการหาค่าเหมาะสมที่สุดแบบไบนารีวาฬ และขั้นตอนวิธีเคเพื่อนบ้านใกล้ที่สุด มีประสิทธิภาพในการแก้ปัญหาข้อมูลไม่สมดุลจากข้อมูล 12 ชุดข้อมูลที่เหนือกว่าขั้นตอนวิธีอื่นที่นำมาเปรียบเทียบอีก 3 วิธี ได้แก่ การลดจำนวนตัวอย่างข้อมูลแบบสุ่ม คลัสเตอร์เซนทรอยด์ และ เนียร์มิส

7) ข้อเสนอแนะ

1. ขั้นตอนวิธีที่นำเสนอได้ใช้เทคนิคเคเพื่อนบ้านใกล้ที่สุดในการประเมินค่าของฟังก์ชันวัตถุประสงค์ ในการพัฒนาต่อไปอาจจะเปลี่ยนเทคนิคการเรียนรู้ของเครื่องเป็นวิธีอื่น ที่มีประสิทธิภาพมากกว่านี้ได้ ถึงอย่างไรก็ตามควรคำนึงถึงการกำหนดค่าพารามิเตอร์ของตัวแบบให้เหมาะสม หากตัวแบบมีความซับซ้อนมากก็จะทำให้การกำหนดพารามิเตอร์มีความยากและท้าทายตามไปด้วย

2. สามารถประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่องแบบอื่น ๆ เพื่อประเมินประสิทธิภาพของขั้นตอนวิธีที่นำเสนอ ซึ่งการทำเช่นนี้จะทำให้เห็นว่า ขั้นตอนวิธีที่นำเสนอนั้นสามารถใช้ร่วมกับเทคนิคการเรียนรู้ของเครื่องแบบใดถึงทำให้เกิดประสิทธิภาพสูงสุด อย่างไรก็ตามประสิทธิภาพของขั้นตอนวิธีก็ขึ้นอยู่กับชุดข้อมูลที่นำมาใช้ด้วยเช่นกัน

ตารางที่ 2 : ค่าวัตถุประสงค์ที่ดีที่สุดจากขั้นตอนวิธีที่นำเสนอ

ชื่อชุดข้อมูล	กราฟลักษณะการลู่เข้าของค่าวัตถุประสงค์	ค่าวัตถุประสงค์ที่ดีที่สุด (รอบที่หยุดการทำงาน)
glass1		0.0047557 (631)
Iris0	**หยุดตั้งแต่รอบการวนซ้ำแรก	0.0000000 (1)
glass-0-1-2-3_VS_4-5-6		0.0006235 (369)
ecoli2		0.0157653 (423)
glass6		0.0000000 (332)
ecoli		0.0017889 (410)
pen_digits		0.0000189 (814)
abalone		0.1069821 (993)
libras_move		0.0000000 (35)
solar_flare_m0		0.1135251 (779)
yeast_me2		0.0074181 (884)
mammography		0.0453169 (462)

ตารางที่ 3: ค่าเฉลี่ยของตัววัดประสิทธิภาพ

ตัววัดประสิทธิภาพ	**วิธีการสุ่มตัวอย่างลด**				
	None	**CC**	**NM**	**RUS**	**PA**
Accuracy	**0.9430** (0.0616)	0.8192 (0.1557)	0.6915 (0.2892)	0.8391 (0.1420)	0.8596 (0.1321)
F1 score	**0.6370** (0.3680)	0.5957 (0.3147)	0.5047 (0.3369)	0.5996 (0.2773)	0.6255 (0.2962)
G-mean	0.6939 (0.3670)	0.8552 (0.1155)	0.7459 (0.2336)	0.8711 (0.1186)	**0.8941** (0.1042)
AUROC	0.8673 (0.1348)	0.8992 (0.0980)	0.7962 (0.2369)	0.9171 (0.1030)	**0.9363** (0.0812)
AUPRC	0.6807 (0.3338)	0.6708 (0.3093)	0.5739 (0.3845)	0.6554 (0.3155)	**0.6978** (0.3283)
Sensitivity	0.6250 (0.3798)	0.9126 (0.1008)	0.8639 (0.1694)	0.9239 (0.0846)	**0.9444** (0.0765)
Precision	**0.6685** (0.3662)	0.5240 (0.3539)	0.4236 (0.3463)	0.4973 (0.3046)	0.5271 (0.3289)
MCC	0.6119 (0.3674)	0.5741 (0.3025)	0.4279 (0.4046)	0.5769 (0.2711)	**0.6204** (0.2783)
kappa	**0.6072** (0.3676)	0.5233 (0.3449)	0.4069 (0.3792)	0.5253 (0.3089)	0.5695 (0.3194)

หมายเหตุ : ตัวหนาขีดเส้นใต้และตัวธรรมดาขีดเส้นใต้ คือ ขั้นตอนวิธีที่มีค่าของตัววัดประสิทธิภาพสูงสุดอันดับที่ 1 และ 2 ตามลำดับ และค่าที่อยู่ในวงเล็บ คือ ค่าส่วนเบี่ยงเบนมาตรฐาน

REFERENCES

[1] S. Fotouhi, S. Asadi, and M. W. Kattan, "A comprehensive data level analysis for cancer diagnosis on imbalanced data," *J. Biomed. Inform.*, vol. 90, Feb. 2019, Art. no. 103089, doi: 10.1016/j.jbi.2018.12.003.

[2] N. M. Mqadi, N. Naicker, and T. Adeliyi, "Solving misclassification of the credit card imbalance problem using near miss," *Math. Probl. Eng.*, vol. 2021, Jul. 2021, Art. no. 7194728, doi: 10.1155/2021/7194728.

[3] W. Kesornsit, V. Lorchirachoonkul, and J. Jitthavech, "Imbalanced data problem solving in classification of diabetes patients," (in Thai), *KKU Res. J. (Graduate Studies)*, vol. 18, no. 3, pp. 11–21, 2018.

[4] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets.* Cham, Switzerland: Springer, 2018.

[5] H. Yu, J. Ni, and J. Zhao, "ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data," *Neurocomputing*, vol. 101, pp. 309–318, 2013.

[6] V. López, I. Triguero, C. J. Carmona, S. García, and F. Herrera, "Addressing imbalanced classification with instance generation techniques: IPADE-ID," *Neurocomputing*, vol. 126, pp. 15–28, 2014.

[7] H.-J. Kim, N.-O. Jo, and K.-S. Shin, "Optimization of cluster-based evolutionary undersampling for the artificial neural networks in corporate bankruptcy prediction," *Expert Syst. Appl.*, vol. 59, pp. 226–234, 2016.

[8] J. Li *et al.* "Adaptive swarm balancing algorithms for rare-event prediction in imbalanced healthcare data," *PloS One*, vol. 12, no. 7, 2017, Art. no. e0180830, doi: 10.1371/journal.pone.0180830.

[9] V. Kumar and D. Kumar, "Binary whale optimization algorithm and its application to unit commitment problem," *Neural. Comput. Appl.,* vol. 32, no. 7, pp. 2095–2123, 2020.

[10] M. M. Mafarja and S. Mirjalili, "Hybrid whale optimization algorithm with simulated annealing for feature selection," *Neurocomputing*, vol. 260, pp. 302–312, 2017.

[11] A. G. Hussien, A. E. Hassanien, E. H. Houssein, S. Bhattacharyya, and M. Amin, "S-shaped binary whale optimization algorithm for feature selection," in *Recent Trends in Signal and Image Processing (Advances in Intelligent Systems and Computing)*, vol 727, S. Bhattacharyya, A. Mukherjee, H. Bhaumik, S. Das, K. Yoshida Eds., Singapore, Singapore: Springer, 2019, pp. 79–87.

[12] G. I. Sayed, A. Darwish, and A. E. Hassanien, "Binary whale optimization algorithm and binary moth flame optimization with clustering algorithms for clinical breast cancer diagnoses," *J. Classif.*, vol. 37, no. 1, pp. 66–96, 2020.

[13] A. G. Hussien, A. E. Hassanien, E. H. Houssein, M. Amin, and A. T. Azar, "New binary whale optimization algorithm for discrete optimization problems," *Eng. Optim.*, vol. 52, no. 6, pp. 945–959, 2020.

[14] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Mach. Learn.*, vol. 6, no. 1, pp. 37–66, 1991.

[15] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[16] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *Proc. Mach. Learn.: ECML 2004: 15th Eur. Conf. Mach. Learn.*, Pisa, Italy, Sep. 2004, pp. 39–50.

[17] S. Mishra, "Handling imbalanced data: SMOTE vs. random undersampling," *Int. Res. J. Eng. Technol*., vol. 4, no. 8, pp. 317–320, 2017.

[18] The Imbalanced-learn Developers. "ClusterCentroids." IMBALANCED-LEARN.org. https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.ClusterCentroids.html (accessed Mar. 3, 2022).

[19] J. Zhang and I. Mani, "kNN approach to unbalanced data distributions: A case study involving information extraction," presented at ICML'2003 Workshop on Learning from Imbalanced Data Sets (II), Washington, DC, USA, Aug. 21, 2003.

[20] A. Orriols-Puig and E. Bernadó-Mansilla, "Evolutionary rule-based systems for imbalanced data sets," *Soft Comput.* vol. 13, no. 3, pp. 213–225, 2009.

[21] S. Mirjalili and A. Lewis, "The whale optimization algorithm," *Adv. Eng. Softw.*, vol. 95, pp. 51–67, 2016.

[22] J. S. Akosa, "Predictive accuracy: A misleading performance measure for highly imbalanced data," presented at the SAS Global Forum 2017, Orlando, FL, USA, Apr. 2–5, 2017, Paper 942–2017.

[23] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–13, 2020.

[24] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Meas.*, vol. 20, no. 1, pp. 37–46, 1960.

[25] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[26] *scikit-learn 1.2.2: Precision-Recall.* (2023). [Online]. Available: https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html

[27] K. Battula, "Research of machine learning algorithms using K-fold cross validation," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 6S, pp. 215–218, 2021.

[28] *Imbalanced data sets*, KEEL, 2011. [Online]. Available: http://www.keel.es/

[29] *fetch_datasets*, The imbalanced-learn developers, 2018. [Online]. Available: https://imbalanced-learn.org/stable/references/generated/imblearn.datasets.fetch_datasets.html