



การเพิ่มประสิทธิภาพการสุ่มเพิ่มข้อมูลสำหรับข้อมูลไม่สมดุลในการจำแนกประเภท

Enhancing Random Oversampling for Imbalanced Classification

นพมาศ อัครจันทโชติ* ศิลา เต็มศิริฤกษ์กุล วรณυχ มีภูมิรู้ และ ยุวธิดา ชิวปรีชา

สาขาวิชาวิทยาการคำนวณและเทคโนโลยีดิจิทัล คณะวิทยาศาสตร์และเทคโนโลยี

มหาวิทยาลัยหัวเฉียวเฉลิมพระเกียรติ สมุทรปราการ 10540

Noppamas Akarachantachote*, Sila Temsiririkkul, Woranuch Meepoomroo,
and Yuwathida ChiwpreecharDivision of Computing Science and Digital Technology, Faculty of Science and Technology,
Huachiew Chalermprakiet University, Samutprakarn 10540

*Corresponding author: a_noppamas@yahoo.com

Received: 3 August 2024/ Revised: 29 October 2024/ Accepted: 11 November 2024

บทคัดย่อ

การดำเนินงานกับปัญหาการจำแนกประเภท ข้อมูลที่ไม่สมดุลเป็นความท้าทายที่พบได้บ่อย ซึ่งตัวจำแนกประเภทมีแนวโน้มที่จะจำแนกหน่วยตัวอย่างใหม่เป็นกลุ่มส่วนใหญ่ ทำให้การทำนายกลุ่มส่วนน้อยมีประสิทธิภาพต่ำ ดังนั้นจึงควรคำนึงถึงการแก้ปัญหาข้อมูลไม่สมดุล การสุ่มเพิ่มข้อมูลเป็นวิธีการหนึ่งที่ไม่ซับซ้อนในการจัดการกับข้อมูลไม่สมดุล แต่หน่วยตัวอย่างที่ถูกสุ่มซ้ำบางตัวอาจไม่ได้เป็นหน่วยตัวอย่างที่สำคัญต่อการจำแนกประเภท งานวิจัยนี้จึงได้นำเสนอวิธีการใหม่ในการจัดการกับข้อมูลที่ไม่สมดุลด้วย Hotelling Important Data Point Oversampling Algorithm (HIDPO) ซึ่งเป็นการปรับปรุงแนวทางการสุ่มเพิ่มข้อมูล วัตถุประสงค์ของงานวิจัยเพื่อเปรียบเทียบประสิทธิภาพการจำแนกประเภท เมื่อใช้ข้อมูลดั้งเดิม ข้อมูลจากการสุ่มเพิ่มข้อมูลระหว่างวิธีสุ่มเพิ่มข้อมูลกลุ่มส่วนน้อย (Random Oversampling: ROS) และวิธีที่นำเสนอใหม่ HIDPO บนข้อมูลจำลอง 96 สถานการณ์ของ 4 พารามิเตอร์ ได้แก่ 1) อัตราความไม่สมดุล (IR) 2) จำนวนตัวแปรทำนายที่เกี่ยวข้องกับการจำแนกประเภท (RelVar) 3) ความแตกต่างของค่าเฉลี่ยของตัวแปรทำนายที่เกี่ยวข้องกับการจำแนกระหว่างกลุ่มส่วนน้อยกับกลุ่มส่วนใหญ่ (ClassDif) และ 4) ขนาดตัวอย่าง (n) โดยจำแนกประเภทด้วยการสร้างตัวแบบการถดถอยลอจิสติก ผลการทดลองพบว่า วิธี HIDPO ให้ค่าการวัดประสิทธิภาพสูงสุดในสถานการณ์ที่มีความแตกต่างกันน้อยระหว่างกลุ่มส่วนน้อยและกลุ่มส่วนใหญ่ โดยเฉพาะเมื่อมีความไม่สมดุลกันมาก อันเป็นสถานการณ์ที่มีความยากในการจำแนกประเภท ส่วนอัตราความถูกต้องในการทำนายกลุ่มส่วนน้อย และอัตราความถูกต้องในการทำนายกลุ่มส่วนใหญ่ วิธี HIDPO ให้ค่าปานกลาง

คำสำคัญ: ข้อมูลที่ไม่สมดุล วิธีการสุ่มเพิ่มข้อมูล โฮเทลลิงทีสแควร์ การจำแนกประเภท

Abstract

In classification problems, imbalanced data is a common challenge, as classifiers often exhibit a tendency to assign new sample points to the majority class. This leads to suboptimal prediction performance for the minority class. Therefore, it is imperative to mitigate the imbalanced data problem.



Random oversampling is a simple employed technique to address class imbalance in datasets. Nevertheless, a subset of the sampled data points may prove inconsequential for the classification process. This research introduces a novel approach: the Hotelling Important Data Point Oversampling Algorithm (HIDPO), an improved version of Random Oversampling. This study aimed to compare classification performance using original data and data from oversampling techniques between Random Oversampling method and the proposed HIDPO method across 96 simulated scenarios. These scenarios varied in four parameters: 1) imbalance rate (IR), 2) the number of relevant predictor variables (RelVar), 3) the difference in means of predictor variables between the minority and majority groups (ClassDif), and 4) sample size (n). Logistic regression models were employed to perform classification tasks. The empirical findings demonstrated that the HIDPO method yields the highest F-measure in scenarios with minimal differences between the minority and majority classes, particularly in cases of severe imbalance, which present challenges for classification. Regarding true positive rate and true negative rate, the HIDPO method yielded moderate values.

Keywords: imbalanced data, random oversampling, Hotelling's T-squared, classification

บทนำ

ข้อมูลที่ไม่สมดุล (Imbalanced Data) หมายถึง ชุดข้อมูลที่มีการกระจายตัวของแต่ละกลุ่มไม่เท่ากัน โดยกลุ่มส่วนใหญ่ (Majority Class) มีจำนวนหน่วยตัวอย่างมากกว่ากลุ่มส่วนน้อย (Minority Class) อย่างมาก อันจะทำให้เกิดความยากในการสร้างตัวแบบจำแนกประเภท เนื่องจากตัวแบบอาจให้ความสำคัญกับกลุ่มส่วนใหญ่มากเกินไป ทำให้ประสิทธิภาพในการทำนายกลุ่มส่วนน้อยลดลง สาเหตุของข้อมูลไม่สมดุลอาจเกิดจากปรากฏการณ์ธรรมชาติ เช่น การตรวจพบการฉ้อโกง ซึ่งรายการที่มีการฉ้อโกงจะมีจำนวนน้อยกว่าการไม่ฉ้อโกง การเป็นโรคร้ายก็มักมีจำนวนคนที่เป็นโรคร้ายน้อยกว่าไม่เป็น ข้อมูลไม่สมดุลอาจเกิดจากข้อจำกัดในการรวบรวมข้อมูล มีค่าใช้จ่ายสูง หรือความยากลำบากในการเก็บรวบรวม ตัวอย่างเช่น การรวบรวมข้อมูลเกี่ยวกับสายพันธุ์ที่หายากนั้นมีต้นทุน และความยากกว่าการรวบรวมข้อมูลเกี่ยวกับสายพันธุ์ทั่วไป นอกจากนี้ข้อมูลไม่สมดุลอาจเกิดจากการให้คำจำกัดความของกลุ่มที่อาจนำไปสู่ความไม่สมดุลได้ เช่น การกำหนดกลุ่มเป็นผลิตภัณฑ์ที่มีข้อบกพร่อง และไม่มีข้อบกพร่อง ซึ่งผลิตภัณฑ์ในกลุ่มที่มีข้อบกพร่องจะมีหน่วยตัวอย่างน้อยกว่ากลุ่มที่ไม่มีข้อบกพร่อง ข้อมูลไม่สมดุลเป็นปัญหาที่พบบ่อยในสถานการณ์จริงหลายอย่าง และต้องใช้เทคนิคเฉพาะในการจัดการ [1-4]

ค่าที่ใช้วัดระดับความไม่สมดุลของข้อมูล วัดด้วยอัตราความไม่สมดุล (Imbalanced Ratio: IR) คำนวณจาก
$$IR = \frac{n_{pos}}{n_{neg}}$$
 โดยที่ n_{pos} แทนจำนวนหน่วยตัวอย่างที่อยู่ในกลุ่มส่วนน้อย และ n_{neg} แทนจำนวนหน่วยตัวอย่างที่อยู่ในกลุ่มส่วนมาก ถ้าอัตราความไม่สมดุลมีค่าเข้าใกล้ 0 แสดงว่ากลุ่มส่วนน้อยมีจำนวนน้อยกว่ากลุ่มส่วนใหญ่มาก และถ้าอัตราความไม่สมดุลมีค่าเข้าใกล้ 1 แสดงว่าทั้งสองกลุ่มมีจำนวนใกล้เคียงกัน ถ้า $IR < 0.2$ จะมีโอกาสที่จะเกิดปัญหา แต่ก็ไม่ได้มีเกณฑ์ตายตัวสำหรับระดับความไม่สมดุลที่เป็นปัญหา แต่อัตราความไม่สมดุลที่มีค่าเข้าใกล้ศูนย์มีแนวโน้มที่จะทำให้การจำแนกประเภททำได้ยากขึ้น [5]

แนวทางการจัดการข้อมูลไม่สมดุลสำหรับปัญหาการจำแนกประเภทโดยทั่วไปมี 2 แนวทาง ได้แก่ 1) แนวทางในระดับข้อมูล (Data-level Approach) และ 2) แนวทางในระดับของขั้นตอนวิธี (Algorithm-level Approach) แนวทางแรกเป็นแนวทางที่ปรับความสมดุลของแต่ละกลุ่มให้มีจำนวนใกล้เคียงกันในกระบวนการก่อน (Pre-processing step) ซึ่งเป็นขั้นตอนก่อนนำข้อมูลไปเรียนรู้เพื่อสร้างตัวแบบ แนวทางหลังเป็นแนวทางแก้ปัญหาข้อมูลไม่สมดุลในขั้นตอนการสร้างตัวแบบ โดยจะปรับแต่งตัวแบบเพื่อปรับปรุงการเรียนรู้ โดยเฉพาะเมื่อเทียบกับกลุ่มส่วนน้อย [6] การจัดการข้อมูลไม่สมดุลด้วยแนวทางในระดับข้อมูลเป็นแนวทางที่ไม่ซับซ้อน และสามารถนำไปใช้กับการสร้างตัวแบบแบบใด ๆ ได้ โดยสามารถดำเนินการ

ผ่านการสุ่มลดข้อมูลกลุ่มส่วนใหญ่ (Random Undersampling: RUS) หรือสุ่มเพิ่มข้อมูลกลุ่มส่วนน้อย (Random Oversampling: ROS) เพื่อที่จะลดอัตราความไม่สมดุล การสุ่มลดข้อมูลจะเป็นการนำหน่วยตัวอย่างของกลุ่มส่วนใหญ่ออกจากชุดข้อมูลแบบสุ่ม เหมาะกับข้อมูลที่มีจำนวนมาก สำหรับการสุ่มเพิ่มข้อมูลจะเป็นการเพิ่มข้อมูลของกลุ่มส่วนน้อยโดยการสุ่มซ้ำข้อมูลจากกลุ่มส่วนน้อย เหมาะกับข้อมูลที่มีขนาดเล็ก

ROS เป็นวิธีการแก้ปัญหาข้อมูลไม่สมดุลที่มีความง่าย ไม่ซับซ้อน เหมาะกับข้อมูลที่มีจำนวนไม่มาก Firuz และคณะ [7] ได้เปรียบเทียบประสิทธิภาพของ ROS กับวิธีการสุ่มตัวอย่างขั้นสูงอื่น ๆ ได้แก่ Synthetic Minority Oversampling Technique (SMOTE), Borderline-SMOTE และ Variational Autoencoder (VAE) ด้วยข้อมูลส่วนหนึ่งของชุดข้อมูล MNIST ซึ่งเป็นภาพของตัวเลขที่เขียนด้วยลายมือ และจำแนกประเภทด้วย Convolutional Neural Network (CNN) ผลการวิจัยพบว่า ROS มีประสิทธิภาพที่ดีกว่าวิธีอื่น โดยมีค่าความแม่นยำ ความเที่ยง และอัตราความถูกต้องในการทำนายกลุ่มส่วนน้อยที่สูงที่สุดในสถานการณ์ส่วนใหญ่ นอกจากนี้ Jagelid and Movin [8] ได้ทำนายการใช้บริการสตรีมเพลงดิจิทัลพอดแคสต์ และวิดีโอ โดยมีบริการแบบฟรี และบริการแบบเสียเงิน ในงานวิจัยต้องการทำนายการเปลี่ยนจากบริการฟรีเป็นบริการเสียเงิน โดยแบ่งลูกค้าเป็นกลุ่มที่เปลี่ยนและไม่เปลี่ยน กลุ่มที่เปลี่ยนมีเพียง 2% ของลูกค้าที่ใช้บริการฟรี มีลักษณะของข้อมูลไม่สมดุล จึงใช้วิธี ROS, RUS และ SMOTE เพื่อสุ่มตัวอย่างซ้ำ และจำแนกประเภทด้วยการถดถอยลอจิสติก ต้นไม้ตัดสินใจ และ Gradient Boosting Trees ผลวิจัยพบว่า การสุ่มตัวอย่างซ้ำช่วยเพิ่มประสิทธิภาพการจำแนก โดยวิธี ROS ร่วมกับการถดถอยลอจิสติกและ Gradient Boosting Trees มีประสิทธิภาพสูงสุด

เนื่องจาก ROS เป็นการสุ่มข้อมูลซ้ำจากข้อมูลเดิมอย่างสุ่ม ซึ่งบางหน่วยตัวอย่างอาจไม่ได้เป็นหน่วยตัวอย่างที่มีความสำคัญต่อการจำแนกประเภท คณะผู้วิจัยได้เสนออัลกอริทึมใหม่ที่ได้พัฒนาจากการสุ่มเพิ่มข้อมูลอย่างสุ่ม โดยปรับค่าน้ำหนักของแต่ละหน่วยตัวอย่างในการสุ่มเพิ่มเพื่อให้ได้หน่วยตัวอย่างที่เป็นประโยชน์ต่อการจำแนกประเภท (Hotelling Important Data Point Oversampling Algorithm: HIDPO) งานวิจัยนี้จึงมีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพการจำแนก เมื่อใช้ข้อมูลดั้งเดิม ข้อมูลจากการสุ่มเพิ่มข้อมูลระหว่างวิธี ROS และวิธี HIDPO บนข้อมูลจำลอง

วิธีดำเนินการวิจัย

การดำเนินงานวิจัยเพื่อเปรียบเทียบวิธีการสุ่มเพิ่มข้อมูลประกอบด้วยขั้นตอน ดังนี้

- 1) จำลองข้อมูลตามสถานการณ์ที่กำหนด
- 2) แบ่งข้อมูลเป็น 2 ส่วน ได้แก่ ข้อมูลฝึกฝน (Training Set) และข้อมูลทดสอบ (Test Set) ด้วยอัตรา 70:30
- 3) คัดเลือกตัวแปรทำนายด้วยสถิติทดสอบ t (Statistical t-test) ที่ระดับนัยสำคัญ 0.10 โดย

$$t_i = \frac{\bar{x}_{1i} - \bar{x}_{2i}}{\sqrt{\frac{s_{1i}^2}{n_1} + \frac{s_{2i}^2}{n_2}}}$$

เมื่อ t_i แทน ค่าสถิติทดสอบ t ของตัวแปรทำนายที่ i

$\bar{x}_{1i}$ และ $\bar{x}_{2i}$ แทน ค่าเฉลี่ยของกลุ่มส่วนใหญ่และกลุ่มส่วนน้อยของตัวแปรทำนายที่ i ตามลำดับ

s_{1i}^2 และ s_{2i}^2 แทน ค่าความแปรปรวนของกลุ่มส่วนใหญ่และกลุ่มส่วนน้อยของตัวแปรทำนายที่ i ตามลำดับ

- 4) สุ่มเพิ่มข้อมูลกลุ่มส่วนน้อยสำหรับข้อมูลฝึกฝน โดย
 - การสุ่มเพิ่มข้อมูลอย่างสุ่ม (Random Oversampling)
 - Hotelling Important Data Point Oversampling Algorithm (HIDPO Algorithm)
- 5) จำแนกประเภทกลุ่มส่วนน้อย และกลุ่มส่วนใหญ่ ด้วยตัวแบบการถดถอยลอจิสติก (Logistic Regression Model) สำหรับ



- ข้อมูลดั้งเดิม
 - ข้อมูลที่ได้รับการสุ่มเพิ่มข้อมูลโดย ROS สำหรับตัวแปรทำนายที่ถูกคัดเลือก
 - ข้อมูลที่ได้รับการสุ่มเพิ่มข้อมูลโดย HIDPO Algorithm สำหรับตัวแปรทำนายที่ถูกคัดเลือก
- 6) ประเมินประสิทธิภาพของการจำแนกจากวิธีการสุ่มแต่ละวิธีและข้อมูลดั้งเดิม บนข้อมูลทดสอบของข้อมูลแต่ละชุด ด้วยค่าความแม่นยำ (Accuracy) อัตราความถูกต้องในการทำนายกลุ่มส่วนน้อย (True Positive Rate) อัตราความถูกต้องในการทำนายกลุ่มส่วนใหญ่ (True Negative Rate) และค่าการวัดเอฟ (F-measure)

การจำลองข้อมูล

ขั้นตอนนี้เป็นกรสร้างข้อมูล โดยกำหนดพารามิเตอร์ที่ใช้ในการสร้างสถานการณ์จำลอง ได้แก่

- 1) อัตราความไม่สมดุล (Imbalanced Ratio: IR) ซึ่งคำนวณได้จากอัตราส่วนของจำนวนสมาชิกในกลุ่มส่วนน้อยต่อจำนวนสมาชิกในกลุ่มส่วนใหญ่ โดยกำหนด 4 ระดับ ได้แก่ 0.05, 0.07, 0.1 และ 0.2
- 2) จำนวนตัวแปรทำนายที่เกี่ยวข้องกับการจำแนก (Relevant Variable: RelVar) 2 ระดับ ได้แก่ 2 และ 5 ตัวแปร
- 3) ความแตกต่างของค่าเฉลี่ยของตัวแปรทำนายที่เกี่ยวข้องกับการจำแนก ระหว่างกลุ่มส่วนน้อยกับกลุ่มส่วนใหญ่ (Class Mean Difference: ClassDif) กำหนด 3 ระดับ ได้แก่ 1 (ยาก), 2 (ปานกลาง) และ 5 (ง่าย)
- 4) ขนาดตัวอย่าง 4 ระดับ ได้แก่ 80, 150, 210 และ 300

สถานการณ์จำลองมีจำนวนรวม 96 สถานการณ์

ขั้นตอนของการสร้างข้อมูลจำลอง ดังนี้

- 1) สร้างข้อมูลตัวแปรทำนาย
 - กลุ่มส่วนใหญ่ มีการแจกแจงปกติ โดยมีค่าเฉลี่ย 0 และส่วนเบี่ยงเบนมาตรฐาน 1
 - กลุ่มส่วนน้อย สำหรับตัวแปรทำนายที่เกี่ยวข้อง มีการแจกแจงปกติ โดยมีค่าเฉลี่ย μ ซึ่งกำหนดเป็น 3 ระดับ ได้แก่ 1, 2 และ 5 ตามที่กำหนดใน ClassDif และส่วนเบี่ยงเบนมาตรฐาน 1 สำหรับตัวแปรทำนายที่ไม่เกี่ยวข้อง มีการแจกแจงปกติ โดยมีค่าเฉลี่ย 0 และส่วนเบี่ยงเบนมาตรฐาน 1 (การกำหนดตัวแปรทำนายให้มีทั้งตัวแปรทำนายที่เกี่ยวข้อง และตัวแปรทำนายที่ไม่เกี่ยวข้อง เพื่อให้สอดคล้องกับสถานการณ์จริงที่ชุดข้อมูลอาจมีตัวแปรทำนายทั้งที่เกี่ยวข้องและไม่เกี่ยวข้องกับการจำแนกประเภท)
- 2) กำหนดตัวแปรตามของแต่ละกลุ่ม โดยกลุ่มส่วนน้อยกำหนดเป็น 1 และกลุ่มส่วนใหญ่ กำหนดเป็น 0
- 3) ในแต่ละสถานการณ์ทำซ้ำ 1,000 รอบ

การสุ่มเพิ่มข้อมูล

การสุ่มเพิ่มข้อมูล เป็นการแก้ปัญหาข้อมูลไม่สมดุลในระดับข้อมูล โดยการเพิ่มจำนวนตัวอย่างของกลุ่มส่วนน้อยให้มีจำนวนเท่ากับกลุ่มส่วนใหญ่ ซึ่งจะเลือกหน่วยตัวอย่างซ้ำอย่างสุ่มแบบคืนที่จากชุดข้อมูลเดิมในกลุ่มส่วนน้อย การสุ่มเพิ่มข้อมูลอย่างสุ่ม (Random Oversampling) นั้น อาจทำให้ได้หน่วยตัวอย่างที่ไม่เหมาะสมในการจำแนก จึงควรมีการวัดความสำคัญของแต่ละหน่วยตัวอย่างในการจำแนก และกำหนดโอกาสที่แต่ละหน่วยตัวอย่างจะได้รับการสุ่มให้สอดคล้องกับสำคัญของแต่ละหน่วยตัวอย่าง ซึ่งในงานวิจัยนี้ได้สร้างอัลกอริทึมใหม่สำหรับการสุ่มเพิ่มข้อมูลตามความสำคัญของแต่ละหน่วยตัวอย่าง โดยการวัดความสำคัญของแต่ละหน่วยตัวอย่างจะใช้วิธีการ Hotelling's T-squared ร่วมด้วย

Hotelling's T-squared เป็นวิธีการหนึ่งของการวิเคราะห์หลายตัวแปร (Multivariate Analysis) ที่ใช้ในการเปรียบเทียบค่าเฉลี่ย 2 กลุ่ม ในกรณีที่มีตัวแปรตามหลายตัว โดยมีข้อมูลที่น่าวิเคราะห์ต้องมีการแจกแจงปกติหลายตัวแปร ซึ่งกำหนดสมมติฐานว่าง และสมมติฐานทางเลือก ดังนี้ [9]

$$H_0: \begin{bmatrix} \mu_{11} \\ \mu_{21} \\ \vdots \\ \mu_{p1} \end{bmatrix} = \begin{bmatrix} \mu_{12} \\ \mu_{22} \\ \vdots \\ \mu_{p2} \end{bmatrix} \quad H_1: \begin{bmatrix} \mu_{11} \\ \mu_{21} \\ \vdots \\ \mu_{p1} \end{bmatrix} \neq \begin{bmatrix} \mu_{12} \\ \mu_{22} \\ \vdots \\ \mu_{p2} \end{bmatrix} \quad (1)$$

$$\text{หรือ} \quad H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 \neq \mu_2$$

สถิติทดสอบ T^2 เป็นดังนี้

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2) \quad (2)$$

เมื่อ T^2 แทน สถิติทดสอบ T^2

p แทน จำนวนตัวแปรทำนาย

n_i แทน ขนาดตัวอย่างของกลุ่ม i โดยที่ $i = 1, 2$

μ_i แทน เวกเตอร์ค่าเฉลี่ยของประชากรกลุ่ม i โดยที่ $i = 1, 2$

$\bar{x}_i$ แทน เวกเตอร์ค่าเฉลี่ยของตัวอย่างกลุ่ม i โดยที่ $i = 1, 2$

S แทน เมทริกซ์ความแปรปรวนรวมของตัวอย่าง (Pooled Variance)

ในกรณีที่ความแปรปรวนประชากรทั้งสองกลุ่มมีข้อสมมติว่าไม่แตกต่างกัน (ซึ่งสอดคล้องกับข้อมูลจำลอง) ประมาณความแปรปรวนรวมได้ดังนี้

$$S = \frac{(n_1 - 1)}{(n_1 + n_2 - 2)} S_1 + \frac{(n_2 - 1)}{(n_1 + n_2 - 2)} S_2 \quad (3)$$

S_1 แทน เมทริกซ์ความแปรปรวนร่วมของตัวอย่างกลุ่ม 1

S_2 แทน เมทริกซ์ความแปรปรวนร่วมของตัวอย่างกลุ่ม 2

โดยที่

$$\frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} \sim F(p, n_1 + n_2 - p - 1) \quad (4)$$

ในการทดสอบสมมติฐานทางสถิติ p-value จะถูกใช้เป็นการเปรียบเทียบกับระดับนัยสำคัญ (α) ที่กำหนด ถ้า $p\text{-value} < \alpha$ จะนำไปสู่การสรุปผลว่า มีหลักฐานเพียงพอที่จะสรุปผลตามสมมติฐานทางเลือก หรือในกรณีนี้ก็คือมีหลักฐานเพียงพอที่จะสรุปว่าเวกเตอร์ค่าเฉลี่ยของทั้งสองกลุ่มแตกต่างกันอย่างมีนัยสำคัญทางสถิติ มิฉะนั้นจะสรุปผลว่าไม่มีหลักฐานเพียงพอที่เวกเตอร์ค่าเฉลี่ยของทั้งสองกลุ่มแตกต่างกัน

p-value หรือ Probability Value คือ ระดับนัยสำคัญต่ำสุดที่ทำให้ค่าสถิติทดสอบที่สังเกตได้มีนัยสำคัญทางสถิติ [10] ถ้า p-value น้อย แสดงถึงการมีหลักฐานอันแข็งแกร่งในการสรุปผลว่าเกิดความต่างอย่างมีนัยสำคัญทางสถิติ ทำให้มั่นใจว่าผลลัพธ์ที่เกิดขึ้น ไม่ใช่เพียงการบังเอิญ ซึ่งปกติใช้ในการทดสอบสมมติฐาน สำหรับในงานวิจัยนี้ได้อาศัย p-value มาใช้เป็นแนวทางเพื่อเป็นค่าวัดความสำคัญของแต่ละหน่วยตัวอย่าง

Hotelling Important Data Point Oversampling Algorithm (HIDPO Algorithm) เป็นขั้นตอนวิธีใหม่ที่ใช้สำหรับการสุ่มเพิ่มข้อมูลสำหรับกลุ่มส่วนน้อย โดยกำหนดความน่าจะเป็นของแต่ละหน่วยตัวอย่างที่ไม่เท่ากัน ขึ้นอยู่กับความสำคัญของแต่ละหน่วยตัวอย่าง ซึ่งใช้กับปัญหาการจำแนกประเภท 2 กลุ่ม โดยอาศัย p-value ที่ได้จาก Hotelling's T-squared ในการเป็นค่าวัดความสำคัญของหน่วยตัวอย่าง แสดงขั้นตอนวิธีโดยมีรายละเอียดดังนี้

- 1) นำเข้าข้อมูลกลุ่มส่วนน้อยจำนวน n_1 และกลุ่มส่วนมากจำนวน n_0
- 2) กำหนดเกณฑ์จุดตัดความสำคัญ (threshold) โดยในงานวิจัยนี้กำหนดที่ 0.00001
- 3) ตัดหน่วยตัวอย่างที่ i ของกลุ่มส่วนน้อย เมื่อ $i = 1, 2, \dots, n_1$
- 4) ใช้วิธีการของ Hotelling's T-squared ในการเปรียบเทียบเวกเตอร์ค่าเฉลี่ยของตัวแปรทำนายระหว่างกลุ่มส่วนน้อยและกลุ่มส่วนใหญ่ ซึ่งกลุ่มส่วนน้อยจะเหลือจำนวน $n_1 - 1$ และหา p-value สำหรับการตัดหน่วยตัวอย่าง



ที่ i ซึ่ง p -value ที่น้อย (p -value < threshold) แสดงถึงข้อมูลที่เหลือหลังจากมีการตัดหน่วยตัวอย่าง i ยังคงสามารถจำแนกประเภทระหว่าง 2 กลุ่มได้อย่างถูกต้อง หรือก็คือการไม่มีหน่วยตัวอย่าง i ก็ยังคงทำให้สามารถพบความแตกต่างระหว่างกลุ่มได้ ดังนั้นหน่วยตัวอย่างนี้จึงไม่ได้มีความสำคัญต่อการจำแนกประเภท

- 5) ดำเนินการขั้นที่ 3 และ 4 จนครบทุกหน่วยตัวอย่างของกลุ่มส่วนน้อย
- 6) คำนวณน้ำหนักสำหรับหน่วยตัวอย่างที่ i คำนวณ ดังนี้

$$w_i = \begin{cases} \frac{p - value_i}{\sum_{j=1}^{m_1} p - value_j} & ; p - value_i \geq \alpha \\ 0 & ; p - value_i < \alpha \end{cases} \quad (5)$$

โดยที่ w_i แทน น้ำหนักของหน่วยตัวอย่างที่ i ของกลุ่มส่วนน้อย

m_1 แทน จำนวนหน่วยตัวอย่างของกลุ่มส่วนน้อยที่มี $p - value_i \geq \alpha$

ในกรณีที่ไม่มีพบหน่วยตัวอย่างที่มี $p - value_i \geq \alpha$ จะกำหนดความน่าจะเป็นที่แต่ละหน่วยตัวอย่างจะถูกสุ่มเพิ่มข้อมูล ด้วยความน่าจะเป็นที่เท่ากันซึ่งจะให้ผลเหมือนกับวิธี ROS

- 7) สุ่มเพิ่มข้อมูลกลุ่มส่วนน้อย จำนวนเท่ากับ $n_0 - n_1$ จากหน่วยตัวอย่างในกลุ่มส่วนน้อย ด้วยความน่าจะเป็นของแต่ละหน่วยตัวอย่างที่จะถูกสุ่ม ตามที่ได้ในขั้นที่ 6
- 8) ส่งออกข้อมูลที่เพิ่มจำนวนแล้ว

การจำแนกประเภท

ในงานวิจัยนี้กำหนดตัวจำแนกประเภทโดยใช้ตัวแบบการถดถอยลอจิสติก (Logistic Regression Model) ซึ่งเป็นเทคนิคทางสถิติในการสร้างตัวแบบความน่าจะเป็นของตัวแปรตามที่มีค่าแบบไม่ต่อเนื่อง เมื่อกำหนดค่าตัวแปรทำนาย โดยตัวแปรตามที่มีค่าไม่ต่อเนื่องนี้อาจมีสองค่า หรือมากกว่าสองค่าได้ แต่ที่ใช้ส่วนใหญ่มักเป็นลักษณะที่มีผลลัพธ์ได้สองค่า (Binary Outcome) เช่น เป็นโรค-ไม่เป็นโรค เป็นหนี้-ไม่หนี้ บกพร่อง-ไม่บกพร่อง เป็นต้น การถดถอยลอจิสติกเป็นเทคนิคที่มีประโยชน์สำหรับปัญหาการจำแนกประเภท สามารถสร้างตัวแบบในการทำนายกลุ่มของตัวอย่างใหม่ โดยการสร้างลอจิสติกฟังก์ชัน (Logistic Function) ในการหาค่าความน่าจะเป็นที่ตัวแปรตามจะอยู่ในกลุ่มที่สนใจ ($y=1$) เป็นดังนี้ [9]

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}} \quad (6)$$

p แทน ความน่าจะเป็นที่ผลลัพธ์ของตัวแปรตามจะตกอยู่ในกลุ่มที่สนใจ ($y = 1$)

x_i แทน ตัวแปรทำนายที่ i

k แทน จำนวนตัวแปรทำนาย

β_i แทน สัมประสิทธิ์การถดถอยของตัวแปรทำนายที่ i

ตัวแบบลอจิสติกจะแสดงในรูปของ log odds ดังนี้

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k \quad (7)$$

โดยทั่วไปจะใช้วิธีของ Maximum Likelihood ในการประมาณค่าสัมประสิทธิ์การถดถอย

ในปัญหาการจำแนกประเภทแบบสองกลุ่ม ตัวแบบถดถอยลอจิสติกเป็นวิธีหนึ่งที่ได้รับคามนิยมสำหรับการแก้ปัญหาในหลากหลายแขนง [8, 11-14] และมีการพัฒนาเพื่อใช้ในกรณีที่ข้อมูลไม่สมดุลอย่างมาก [15]



การประเมินประสิทธิภาพ

การวัดประสิทธิภาพของการจำแนกโดยสร้าง Confusion Matrix และคำนวณค่าความแม่นยำ (Accuracy) อัตราความถูกต้องในการทำนายกลุ่มส่วนน้อย (True Positive Rate) อัตราความถูกต้องในการทำนายกลุ่มส่วนใหญ่ (True Negative Rate) และค่าการวัดเอฟ (F-measure) บนข้อมูลทดสอบ โดยแสดงการสร้าง Confusion Matrix และคำนวณค่าต่าง ๆ แสดงดังตารางที่ 1 [16]

ตารางที่ 1 แสดง Confusion Matrix ของค่าจริงและผลการทำนาย ดังนี้

ค่าจริง	ผลการทำนาย	
	1 (กลุ่มส่วนน้อย)	0 (กลุ่มส่วนใหญ่)
1 (กลุ่มส่วนน้อย)	TP	FN
0 (กลุ่มส่วนมาก)	FP	TN

โดยที่ TP หมายถึง True Positive แสดงถึงจำนวนหน่วยตัวอย่างที่อยู่ในกลุ่มส่วนน้อย และทำนายว่าอยู่ในกลุ่มส่วนน้อย

FN หมายถึง False Negative แสดงถึงจำนวนหน่วยตัวอย่างที่อยู่ในกลุ่มที่ส่วนน้อย แต่ทำนายว่าอยู่ในกลุ่มส่วนมาก

FP หมายถึง False Positive แสดงถึงจำนวนหน่วยตัวอย่างที่อยู่ในกลุ่มส่วนมาก แต่ทำนายว่าอยู่ในกลุ่มส่วนน้อย

TN หมายถึง True Negative แสดงถึงจำนวนหน่วยตัวอย่างที่อยู่ในกลุ่มส่วนมาก และทำนายว่าอยู่ในกลุ่มส่วนมาก

1) ค่าความแม่นยำ (Accuracy) แสดงถึงความถูกต้องในการทำนายในภาพรวม ทั้งกลุ่มส่วนน้อยและกลุ่มส่วนมาก ซึ่งคำนวณได้ ดังนี้

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

2) อัตราความถูกต้องในการทำนายกลุ่มส่วนน้อย (True Positive Rate: TP Rate) แสดงถึงความถูกต้องในการทำนายกลุ่มส่วนน้อยว่าอยู่กลุ่มส่วนน้อย

$$True Positive Rate = \frac{TP}{TP + FN} \quad (9)$$

3) อัตราความถูกต้องในการทำนายกลุ่มส่วนใหญ่ (True Negative Rate: TN Rate) แสดงถึงความถูกต้องในการทำนายกลุ่มส่วนใหญ่ว่าอยู่กลุ่มส่วนใหญ่

$$True Negative Rate = \frac{TN}{TN + FP} \quad (10)$$

4) ค่าการวัดเอฟ (F-measure) เป็นการวัดความถูกต้องโดยการใช้ค่าเฉลี่ยฮาร์โมนิระหว่าง True Positive Rate และ Precision

$$F - measure = \frac{2 \times (TP rate \times Precision)}{TP rate + Precision} \quad (11)$$

โดยที่ Precision หมายถึง ความเที่ยง ซึ่งแสดงถึงความถูกต้องของการทำนายกลุ่มส่วนน้อย เมื่อเทียบกับผลการทำนายกลุ่มส่วนน้อยทั้งหมด ซึ่งคำนวณจาก $Precision = \frac{TP}{TP + FP}$

การวัดประสิทธิภาพด้วยค่าความแม่นยำเพียงอย่างเดียวอาจไม่สะท้อนความเป็นจริง โดยเฉพาะถ้าข้อมูลไม่สมดุล เนื่องจากค่าความแม่นยำจะเอนเอียงไปตามกลุ่มส่วนมาก ค่าการวัดเอฟเป็นตัววัดประสิทธิภาพที่เป็นทางเลือกที่ดีกว่า โดยค่าการวัดเอฟจะพิจารณาทั้งความเที่ยง และอัตราความถูกต้องในการทำนายกลุ่มส่วนน้อย มีแนวโน้มน้อยกว่าที่จะเอนเอียงไปทางกลุ่มส่วนใหญ่หรือกลุ่มส่วนน้อย [13]



ผลการวิจัย

ผลการเปรียบเทียบประสิทธิภาพการจำแนกประเภท เมื่อใช้ข้อมูลดั้งเดิม ข้อมูลจากการสุ่มเพิ่มข้อมูลระหว่างวิธี ROS และวิธี HIDPO บนข้อมูลจำลองโดยกำหนดพารามิเตอร์ ได้แก่ อัตราความไม่สมดุล จำนวนตัวแปรทำนายที่เกี่ยวข้อง ความแตกต่างของค่าเฉลี่ยของตัวแปรทำนายที่เกี่ยวข้องกับการจำแนกระหว่างกลุ่มส่วนน้อยกับกลุ่มส่วนใหญ่ และขนาดตัวอย่างรวม 96 สถานการณ์ จำแนกประเภทด้วยตัวแบบการถดถอยลอจิสติก โดยวัดประสิทธิภาพของการจำแนกด้วยค่าความแม่นยำ อัตราความถูกต้องในการทำนายกลุ่มส่วนน้อย อัตราความถูกต้องในการทำนายกลุ่มส่วนใหญ่ และค่าการวัดเอฟ ซึ่งผลการวิเคราะห์แสดงดังตารางที่ 2 – 5

จากตารางที่ 2 เป็นสถานการณ์ที่ $n = 80$ พบว่าเมื่อ $IR = 0.05$ ค่า Accuracy และ TN Rate ส่วนใหญ่จะสูงสุด เมื่อเป็นข้อมูลดั้งเดิม โดย Accuracy อยู่ในช่วงร้อยละ 95.62 – 100.00 และ TN Rate อยู่ในช่วงร้อยละ 99.63 – 100.00 สำหรับ TP Rate จะสูงสุด เมื่อใช้วิธี ROS โดยอยู่ในช่วงร้อยละ 44.80 – 100.00 และค่า F-measure สูงสุดอย่างมีนัยสำคัญทางสถิติ เมื่อใช้วิธี HIDPO ในสถานการณ์ ClassDif = 1 และ RelVar = 2 โดยมีค่าร้อยละ 56.03 ส่วนสถานการณ์อื่น วิธี HIDPO ให้ผลใกล้เคียงกับ ROS

เมื่อ $IR = 0.07$ ค่า Accuracy และ TN Rate ส่วนใหญ่จะสูงสุด เมื่อเป็นข้อมูลดั้งเดิม โดย Accuracy อยู่ในช่วงร้อยละ 95.37 – 100.00 และ TN Rate อยู่ในช่วงร้อยละ 99.25 – 100.00 สำหรับ TP Rate ส่วนใหญ่จะสูงสุด เมื่อใช้วิธี ROS โดยอยู่ในช่วงร้อยละ 50.20 – 100.00 และค่า F-measure สูงสุดอย่างมีนัยสำคัญทางสถิติ เมื่อใช้วิธี HIDPO ในสถานการณ์ ClassDif = 1, RelVar = 2 และสถานการณ์ ClassDif = 2, RelVar = 2 โดยมีค่าร้อยละ 18.26 และ 87.89 ตามลำดับ ส่วนสถานการณ์อื่น วิธี HIDPO ให้ผลใกล้เคียงกับ ROS

เมื่อ $IR = 0.10$ ค่า Accuracy และ TN Rate ส่วนใหญ่จะสูงสุด เมื่อเป็นข้อมูลดั้งเดิม โดย Accuracy อยู่ในช่วงร้อยละ 91.62 – 100.00 และ TN Rate อยู่ในช่วงร้อยละ 99.00 – 100.00 สำหรับ TP Rate วิธี ROS ให้ค่าสูงสุดในสถานการณ์ ClassDif = 1 และ 2 โดยมีค่าอยู่ในช่วงร้อยละ 54.80 – 95.00 ส่วนสถานการณ์อื่น วิธี ROS และวิธี HIDPO จะมีค่าใกล้เคียงกัน สำหรับค่า F-measure วิธี HIDPO จะสูงสุดอย่างมีนัยสำคัญทางสถิติเมื่อ ClassDif = 2 และ RelVar = 2 โดยมีค่าร้อยละ 67.88 ส่วนสถานการณ์อื่น วิธี HIDPO ให้ผลใกล้เคียงกับ ROS

เมื่อ $IR = 0.20$ ค่า Accuracy และ TN Rate ส่วนใหญ่จะสูงสุด เมื่อเป็นข้อมูลดั้งเดิม โดย Accuracy อยู่ในช่วงร้อยละ 84.53 – 100.00 และ TN Rate อยู่ในช่วงร้อยละ 96.66 – 100.00 สำหรับ TP Rate วิธี ROS จะสูงสุดในสถานการณ์ส่วนใหญ่ โดยมีค่าอยู่ในช่วงร้อยละ 63.43 – 100.00 สำหรับค่า F-measure วิธี HIDPO จะสูงสุดอย่างมีนัยสำคัญเมื่อ ClassDif = 1, RelVar = 5 และสถานการณ์ ClassDif = 2, RelVar = 2 โดยมีค่าร้อยละ 63.17 และ 78.30 ตามลำดับ ส่วนสถานการณ์อื่น วิธี HIDPO ให้ผลใกล้เคียงกับ ROS

จากตารางที่ 3 เป็นสถานการณ์ที่ $n = 150$ พบว่าเมื่อ $IR = 0.05$ ค่า Accuracy และ TN Rate ส่วนใหญ่จะสูงสุด เมื่อเป็นข้อมูลดั้งเดิม โดย Accuracy อยู่ในช่วงร้อยละ 95.41 – 100.00 และ TN Rate อยู่ในช่วงร้อยละ 99.64 – 100.00 สำหรับ TP Rate จะสูงสุด เมื่อใช้วิธี ROS โดยอยู่ในช่วงร้อยละ 52.45 – 100.00 และค่า F-measure สูงสุดอย่างมีนัยสำคัญทางสถิติ เมื่อใช้วิธี HIDPO ในสถานการณ์ ClassDif = 1 และสถานการณ์ ClassDif = 2, RelVar = 2 โดยมีค่าอยู่ในช่วงร้อยละ 19.92 – 59.93 ส่วนสถานการณ์อื่น วิธี HIDPO ให้ผลใกล้เคียงกับ ROS

เมื่อ $IR = 0.07$ ค่า Accuracy และ TN Rate ส่วนใหญ่จะสูงสุด เมื่อเป็นข้อมูลดั้งเดิม โดย Accuracy อยู่ในช่วงร้อยละ 93.28 – 100.00 และ TN Rate อยู่ในช่วงร้อยละ 99.34 – 100.00 สำหรับ TP Rate ส่วนใหญ่จะสูงสุด เมื่อใช้วิธี ROS โดยอยู่ในช่วงร้อยละ 59.93 – 100.00 และค่า F-measure สูงสุดอย่างมีนัยสำคัญทางสถิติ เมื่อใช้วิธี HIDPO ในสถานการณ์ ClassDif = 1 โดยมีค่าอยู่ในช่วงร้อยละ 28.98 – 48.30 ส่วนสถานการณ์อื่นส่วนใหญ่ วิธี HIDPO ให้ผลใกล้เคียงกับ ROS



ตารางที่ 2 ผลการเปรียบเทียบประสิทธิภาพเมื่อใช้ข้อมูลดั้งเดิม ข้อมูลที่ได้จากการสุ่มตัวอย่างวิธี ROS และวิธี HIDPO (หน่วยเป็น %) เมื่อ $n = 80$

IR	ClassDiff	RelVar	Accuracy			Stat test	TP Rate			Stat test	TN Rate			F-measure			Stat test
			original	ROS	HIDPO		original	ROS	HIDPO		original	ROS	HIDPO	original	ROS	HIDPO	
0.05	1	2	95.62	80.71	86.71	a, b	3.50	44.80	32.90	a, b	99.63	82.27	89.05	3.40	17.58	16.65	a
		5	95.95	87.55	92.04	a, b	11.20	52.00	40.30	a, b	99.64	89.10	94.29	11.03	27.70	27.69	a
	2	2	97.23	93.62	96.30	a, b	39.60	77.40	67.10	a, b	99.73	94.32	97.57	38.93	53.60	56.03	a, b
		5	98.71	98.82	99.15	a, b	70.80	91.40	88.80	a, b	99.92	99.14	99.60	70.33	86.07	86.20	a
	5	2	99.89	99.97	99.97	a	97.40	99.80	99.80	a	100.00	99.98	99.97	97.40	99.63	99.60	a
0.07	1	2	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	100.00	100.00	100.00	ไม่แตกต่าง
		5	95.37	78.36	86.04	a, b	6.20	50.20	37.80	a, b	99.25	79.59	88.14	5.90	16.81	18.26	a, b
	2	2	95.90	87.11	91.88	a, b	19.00	60.00	47.40	a, b	99.24	88.29	93.82	17.80	30.49	31.53	a
		5	97.35	93.15	96.11	a, b	49.00	80.30	72.20	a, b	99.46	93.71	97.15	47.03	53.02	58.69	a, b
	5	2	99.94	99.98	99.98	a	98.60	99.90	99.90	a	100.00	99.99	99.99	98.60	99.80	99.80	a
0.1	1	2	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	100.00	100.00	100.00	ไม่แตกต่าง
		5	91.62	77.23	83.78	a, b	10.45	54.80	41.50	a, b	99.00	79.26	87.62	12.29	29.41	29.77	a
	2	2	92.82	86.08	90.21	a, b	28.45	64.90	53.90	a, b	98.67	88.01	93.51	32.06	45.38	46.45	a
		5	95.55	92.38	94.47	a, b	54.50	80.70	72.95	a, b	99.28	93.44	96.43	60.01	65.82	67.88	a, b
	5	2	98.64	98.74	98.78	ไม่แตกต่าง	85.20	95.00	95.00	a	99.86	99.08	99.13	88.59	92.85	92.98	a
0.2	1	2	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	100.00	100.00	100.00	ไม่แตกต่าง
		5	84.53	74.78	78.98	a, b	23.90	63.43	53.33	a, b	96.66	77.05	84.11	29.68	45.85	45.17	a
	2	2	88.98	84.76	87.09	a, b	50.40	73.95	67.68	a, b	96.70	86.92	90.97	57.28	62.32	63.17	a, b
		5	93.46	91.57	91.89	a, b	70.35	85.95	85.50	a, b	98.09	92.70	93.17	76.22	77.88	78.30	a, b
	5	2	98.59	98.43	98.45	ไม่แตกต่าง	92.60	96.68	96.93	a, b	99.79	98.78	98.75	95.12	95.43	95.53	ไม่แตกต่าง
หมายเหตุ	1	2	99.94	99.99	99.98	a	99.65	99.95	99.93	a	100.00	100.00	100.00	99.80	99.96	99.95	a
		5	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	100.00	100.00	100.00	ไม่แตกต่าง
	2	2	84.53	74.78	78.98	a, b	23.90	63.43	53.33	a, b	96.66	77.05	84.11	29.68	45.85	45.17	a
		5	88.98	84.76	87.09	a, b	50.40	73.95	67.68	a, b	96.70	86.92	90.97	57.28	62.32	63.17	a, b
	5	2	93.46	91.57	91.89	a, b	70.35	85.95	85.50	a, b	98.09	92.70	93.17	76.22	77.88	78.30	a, b

หมายเหตุ IR – อัตราความไม่สมดุล RelVar – จำนวนตัวแปรทำนายที่เกี่ยวข้อง ClassDiff – ความแตกต่างของค่าเฉลี่ย a - HIDPO แตกต่างจาก original อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05

b - HIDPO แตกต่างจาก ROS อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05



ตารางที่ 3 ผลการเปรียบเทียบประสิทธิภาพเพื่อใช้ข้อมูลดั้งเดิม ข้อมูลที่ได้จากการสุ่มตัวอย่างวิธี ROS และวิธี HIDPO (หน่วยเป็น %) เมื่อ n = 150

IR	ClassDif	RelVar	Accuracy			Stat test	TP Rate			Stat test	TN Rate			Stat test	F-measure			Stat test
			original	ROS	HIDPO		original	ROS	HIDPO		original	ROS	HIDPO					
0.05	1	2	95.41	78.61	86.92	a, b	4.35	52.45	36.45	a, b	99.64	79.83	89.27	a, b	5.17	18.81	19.92	a, b
	1	5	95.94	87.62	92.95	a, b	19.85	62.40	48.00	a, b	99.48	88.79	95.04	a, b	23.15	33.66	37.29	a, b
	2	2	97.19	93.37	95.69	a, b	45.00	80.90	72.30	a, b	99.62	93.95	96.78	a, b	50.77	55.48	59.93	a, b
	2	5	99.02	98.90	98.90	a	79.15	94.30	94.25	a	99.95	99.11	99.12	a	84.10	89.36	89.34	a
	5	2	99.92	99.98	99.98	a	98.25	99.85	99.85	a	100.00	99.99	99.99	a	98.83	99.78	99.78	a
0.07	5	5	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	ไม่แตกต่าง
	1	2	93.28	76.96	85.25	a, b	8.53	59.93	44.47	a, b	99.34	78.17	88.16	a, b	11.29	26.64	28.98	a, b
	1	5	94.49	87.48	91.82	a, b	31.27	69.40	58.17	a, b	99.01	88.78	94.22	a, b	37.57	44.48	48.30	a, b
	2	2	96.48	92.65	93.43	a, b	55.93	84.80	82.90	a, b	99.37	93.21	94.19	a, b	63.50	62.88	64.51	b
	2	5	98.93	98.77	98.78	a	85.63	95.37	95.50	a	99.88	99.01	99.01	a	90.08	91.69	91.80	a
0.1	5	2	99.94	99.97	99.97	a	99.10	99.93	99.93	a	100.00	99.98	99.98	a	99.46	99.82	99.82	a
	5	5	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	ไม่แตกต่าง
	1	2	91.32	76.42	83.72	a, b	14.35	65.58	51.78	a, b	98.82	77.47	86.84	a, b	19.06	33.60	36.11	a, b
	1	5	93.25	86.88	89.30	a, b	41.83	75.05	69.10	a, b	98.26	88.04	91.27	a, b	48.46	51.61	53.86	a, b
	2	2	95.93	92.32	92.38	a	63.88	86.30	85.90	a, b	99.06	92.90	93.01	a	71.05	68.25	68.25	a
0.2	2	5	98.90	98.70	98.69	a	89.33	96.53	96.60	a	99.83	98.92	98.89	a	92.74	93.36	93.32	ไม่แตกต่าง
	5	2	99.95	100.00	100.00	a	99.48	99.98	99.98	a	100.00	100.00	100.00	ไม่แตกต่าง	99.70	99.97	99.97	a
	5	5	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	ไม่แตกต่าง
	1	2	86.08	75.43	78.45	a, b	30.71	70.04	63.87	a, b	96.28	76.42	81.14	a, b	38.65	47.24	47.94	a, b
	1	5	90.72	86.27	86.29	a	60.37	82.00	81.53	a, b	96.31	87.06	87.17	a	65.84	65.57	65.41	ไม่แตกต่าง
0.2	2	2	94.40	91.75	91.67	a	75.49	89.01	88.94	a	97.88	92.25	92.17	a	79.97	77.59	77.40	a
	2	5	98.85	98.54	98.57	a	94.71	97.34	97.36	a	99.62	98.76	98.79	a	96.11	95.52	95.60	a
	5	2	99.97	99.96	99.96	ไม่แตกต่าง	99.80	99.90	99.91	a	100.00	99.97	99.97	a	99.89	99.87	99.87	ไม่แตกต่าง
0.2	5	5	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	ไม่แตกต่าง

หมายเหตุ IR – อัตราความไม่สมดุล RelVar – จำนวนต้นปรมาณายที่เกี่ยวข้อง ClassDif – ความแตกต่างจาก original อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05

b - HIDPO แตกต่างจาก ROS อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05

ตารางที่ 4 ผลการเปรียบเทียบประสิทธิภาพเมื่อใช้ข้อมูลดั้งเดิม ข้อมูลที่ได้จากการสุ่มตัวอย่างวิธี ROS และวิธี HIDPO (หน่วยเป็น %) เมื่อ $n = 210$

IR	ClassDif	RelVar	Accuracy			Stat test	TP Rate			Stat test	TN Rate			F-measure			Stat test
			original	ROS	HIDPO		original	ROS	HIDPO		original	ROS	HIDPO	original	ROS	HIDPO	
0.05	1	2	95.15	78.07	87.24	a, b	5.43	60.63	43.93	a, b	99.64	78.94	89.41	7.43	21.79	25.24	a, b
	1	5	95.85	88.60	93.12	a, b	24.47	68.77	56.33	a, b	99.42	89.59	94.96	30.38	38.66	44.24	a, b
	2	2	97.23	93.28	94.00	a, b	50.20	84.47	83.37	a, b	99.58	93.72	94.53	58.14	57.31	59.46	b
	2	5	99.12	98.96	98.95	a	83.07	94.83	94.77	a	99.92	99.17	99.16	88.28	90.53	90.38	a
	5	2	99.95	99.98	99.98	a	98.97	99.90	99.90	a	100.00	99.99	99.99	99.38	99.81	99.84	a
0.07	5	5	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	100.00	100.00	100.00	ไม่แตกต่าง
	1	2	93.62	76.95	85.38	a, b	9.98	66.13	50.48	a, b	99.29	77.68	87.75	13.59	27.32	30.86	a, b
	1	5	94.97	87.43	89.91	a, b	36.00	75.68	69.53	a, b	98.97	88.23	91.29	43.96	44.81	47.90	a, b
	2	2	96.78	92.51	92.55	a	59.58	86.20	86.35	a	99.31	92.94	92.97	67.40	61.24	61.47	a
	2	5	99.12	98.82	98.81	a	87.55	95.98	95.95	a	99.90	99.02	99.01	91.77	91.75	91.70	ไม่แตกต่าง
0.1	5	2	99.97	99.98	99.98	ไม่แตกต่าง	99.48	99.98	99.98	a	100.00	99.98	99.98	99.70	99.84	99.83	ไม่แตกต่าง
	5	5	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	100.00	100.00	100.00	ไม่แตกต่าง
	1	2	90.77	76.65	82.85	a, b	14.05	67.72	55.40	a, b	98.85	77.59	85.74	19.97	36.06	38.17	a, b
	1	5	93.37	86.99	87.52	a, b	44.93	78.72	77.22	a, b	98.47	87.86	88.60	54.02	54.40	54.86	b
	2	2	95.66	92.19	92.10	a	64.42	87.68	87.55	a	98.94	92.67	92.58	72.30	69.32	69.09	a
0.2	2	5	98.97	98.75	98.74	a	91.12	96.85	96.97	a	99.79	98.95	98.93	94.02	93.93	93.89	ไม่แตกต่าง
	5	2	99.95	99.97	99.97	a	99.45	99.93	99.92	a	100.00	99.98	99.98	99.69	99.87	99.86	a
	5	5	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	100.00	100.00	100.00	ไม่แตกต่าง
	1	2	86.19	75.85	76.91	a, b	31.14	72.13	70.27	a, b	96.57	76.55	78.16	39.99	48.99	49.37	a, b
	1	5	90.86	86.46	86.44	a	61.59	82.99	82.76	a	96.38	87.11	87.13	67.43	66.54	66.35	a
	2	2	94.47	91.94	91.90	a	76.53	89.84	89.68	a	97.85	92.34	92.32	81.07	78.45	78.32	a
	2	5	98.86	98.55	98.57	a	94.86	97.43	97.44	a	99.61	98.76	98.78	96.24	95.61	95.67	a
	5	2	99.96	99.97	99.97	ไม่แตกต่าง	99.76	99.92	99.92	a	100.00	99.98	99.98	99.86	99.92	99.91	ไม่แตกต่าง
	5	5	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	100.00	100.00	100.00	ไม่แตกต่าง

หมายเหตุ IR – อัตราความไม่สมดุล RelVar – จำนวนตัวแปรทำนายที่เกี่ยวข้อง ClassDif – ความแตกต่างของค่าเฉลี่ย a - HIDPO แตกต่างจาก original อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05

b - HIDPO แตกต่างจาก ROS อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05



ตารางที่ 5 ผลการเปรียบเทียบประสิทธิภาพเมื่อใช้ข้อมูลดั้งเดิม ข้อมูลที่ได้จากการสุ่มตัวอย่างวิธี ROS และวิธี HIDPO (หน่วยเป็น %) เมื่อ $n = 300$

IR	ClassDif	RelVar	Accuracy			Stat test	TP Rate			Stat test	TN Rate			F-measure			Stat test
			original	ROS	HIDPO		original	ROS	HIDPO		original	ROS	HIDPO	original	ROS	HIDPO	
0.05	1	2	95.52	77.63	86.95	a, b	6.35	65.13	47.90	a, b	99.67	78.21	88.76	9.14	21.30	25.36	a, b
	1	5	96.27	87.99	90.21	a, b	30.53	74.58	69.00	a, b	99.32	88.62	91.20	38.16	37.58	40.37	a, b
	2	2	97.48	92.88	92.94	a	53.65	86.53	86.35	a	99.52	93.17	93.25	62.43	54.58	54.76	a
	2	5	99.31	98.97	98.98	a	86.28	95.80	95.83	a	99.92	99.12	99.12	90.73	90.03	90.08	ไม่แตกต่าง
	5	2	99.96	99.98	99.98	a	99.13	99.95	99.95	a	100.00	99.98	99.98	99.50	99.76	99.79	a
0.07	5	5	100.00	100.00	100.00	ไม่แตกต่าง	99.98	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	99.99	100.00	100.00	ไม่แตกต่าง
	1	2	93.48	76.45	83.24	a, b	10.68	68.90	56.22	a, b	99.40	76.99	85.17	15.85	28.60	31.56	a, b
	1	5	94.98	87.29	87.48	a, b	40.12	78.57	78.15	a	98.90	87.92	88.15	49.28	46.52	46.73	a
	2	2	96.74	92.33	92.38	a	61.02	88.33	88.30	a	99.29	92.62	92.68	69.85	62.13	62.18	a
	2	5	99.15	98.82	98.82	a	89.43	96.30	96.28	a	99.85	99.00	99.00	92.92	91.96	91.97	a
0.1	5	2	99.96	99.97	99.98	a, b	99.35	99.88	99.90	a	100.00	99.98	99.98	99.64	99.81	99.84	a, b
	5	5	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	100.00	100.00	100.00	ไม่แตกต่าง
	1	2	91.49	76.24	79.87	a, b	15.43	70.96	64.28	a, b	98.91	76.75	81.39	22.24	35.07	36.67	a, b
	1	5	93.90	86.94	87.00	a	47.74	81.39	81.34	a	98.41	87.49	87.56	56.56	53.48	53.55	a
	2	2	96.05	92.06	92.08	a	67.18	89.30	89.28	a	98.86	92.33	92.36	74.18	67.76	67.83	a
0.2	2	5	99.07	98.78	98.79	a	91.83	97.05	97.08	a	99.78	98.95	98.96	94.39	93.67	93.71	a
	5	2	99.98	99.98	99.98	ไม่แตกต่าง	99.74	99.99	99.99	a	100.00	99.98	99.98	99.86	99.91	99.92	ไม่แตกต่าง
	5	5	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	100.00	100.00	100.00	ไม่แตกต่าง
	1	2	85.63	75.48	75.42	a	30.93	73.31	73.03	a	96.57	75.91	75.89	40.56	50.05	49.88	a
	1	5	90.76	86.10	86.11	a	61.95	83.37	83.37	a	96.52	86.64	86.65	68.55	66.95	66.95	a
หมายเหตุ	2	2	94.30	91.83	91.78	a	77.51	90.43	90.58	a	97.66	92.11	92.03	81.62	78.97	78.91	a
	2	5	98.87	98.57	98.57	a	95.18	97.69	97.76	a	99.61	98.74	98.73	96.50	95.84	95.85	a
	5	2	99.97	99.98	99.98	ไม่แตกต่าง	99.85	99.99	99.98	a	100.00	99.98	99.98	99.91	99.94	99.94	ไม่แตกต่าง
หมายเหตุ	5	5	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	ไม่แตกต่าง	100.00	100.00	100.00	100.00	100.00	100.00	ไม่แตกต่าง

หมายเหตุ IR – อัตราความไม่สมดุล RelVar – จำนวนตัวแปรทำนายที่เกี่ยวข้อง ClassDif – ความแตกต่างของค่าเฉลี่ย a - HIDPO แตกต่างจาก original อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05

b - HIDPO แตกต่างจาก ROS อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05



เมื่อ $IR = 0.10$ ค่า Accuracy และ TN Rate ส่วนใหญ่จะสูงสุด เมื่อเป็นข้อมูลดั้งเดิม โดย Accuracy อยู่ในช่วงร้อยละ 91.32 – 100.00 และ TN Rate อยู่ในช่วงร้อยละ 98.82 – 100.00 สำหรับ TP Rate วิธี ROS ให้ค่าสูงสุดในสถานการณ์ ClassDif = 1 และสถานการณ์ ClassDif = 2, RelVar = 2 โดยมีค่าอยู่ในช่วงร้อยละ 65.58 – 86.30 ส่วนสถานการณ์อื่น วิธี ROS และวิธี HIDPO จะมีค่าใกล้เคียงกัน สำหรับค่า F-measure วิธี HIDPO จะสูงสุดอย่างมีนัยสำคัญทางสถิติเมื่อ ClassDif = 1 โดยมีค่าอยู่ในช่วงร้อยละ 36.11 – 53.86 ส่วนสถานการณ์อื่นส่วนใหญ่ วิธี HIDPO ให้ผลใกล้เคียงกับ ROS

เมื่อ $IR = 0.20$ ค่า Accuracy และ TN Rate ส่วนใหญ่จะสูงสุด เมื่อเป็นข้อมูลดั้งเดิม โดย Accuracy อยู่ในช่วงร้อยละ 86.08 – 100.00 และ TN Rate อยู่ในช่วงร้อยละ 96.28 – 100.00 สำหรับ TP Rate วิธี ROS ให้ค่าสูงสุดในสถานการณ์ ClassDif = 1 โดยมีค่าอยู่ในช่วงร้อยละ 70.04 – 89.01 ส่วนสถานการณ์อื่น วิธี ROS และวิธี HIDPO จะมีค่าใกล้เคียงกัน สำหรับค่า F-measure วิธี HIDPO จะสูงสุดอย่างมีนัยสำคัญทางสถิติเมื่อ ClassDif = 1 และ RelVar = 2 โดยมีค่าร้อยละ 47.94 และข้อมูลดั้งเดิมให้ค่าสูงสุดเมื่อ ClassDif = 2 ส่วนสถานการณ์อื่นทุกวิธีให้ผลไม่แตกต่างกัน

จากตารางที่ 4 เมื่อ $n = 210$ พบว่าเมื่อ $IR = 0.05$ ค่า Accuracy และ TN Rate ส่วนใหญ่จะสูงสุด เมื่อเป็นข้อมูลดั้งเดิม โดย Accuracy อยู่ในช่วงร้อยละ 95.15 – 100.00 และ TN Rate อยู่ในช่วงร้อยละ 99.64 – 100.00 สำหรับ TP Rate จะสูงสุด เมื่อใช้วิธี ROS โดยอยู่ในช่วงร้อยละ 60.63 – 100.00 และค่า F-measure สูงสุดอย่างมีนัยสำคัญทางสถิติ เมื่อใช้วิธี HIDPO ในสถานการณ์ ClassDif = 1 โดยมีค่าอยู่ในช่วงร้อยละ 25.24 – 44.24 ส่วนสถานการณ์อื่นส่วนใหญ่ วิธี HIDPO ให้ผลใกล้เคียงกับวิธี ROS

เมื่อ $IR = 0.07$ ค่า Accuracy และ TN Rate ส่วนใหญ่จะสูงสุด เมื่อเป็นข้อมูลดั้งเดิม โดย Accuracy อยู่ในช่วงร้อยละ 93.62 – 100.00 และ TN Rate อยู่ในช่วงร้อยละ 99.29 – 100.00 สำหรับ TP Rate ส่วนใหญ่จะสูงสุด เมื่อใช้วิธี ROS โดยอยู่ในช่วงร้อยละ 66.13 – 100.00 และค่า F-measure สูงสุดอย่างมีนัยสำคัญทางสถิติ เมื่อใช้วิธี HIDPO ในสถานการณ์ ClassDif = 1 โดยมีค่าอยู่ในช่วงร้อยละ 30.86 – 47.90 ส่วนสถานการณ์อื่นส่วนใหญ่แต่ละวิธีให้ผลไม่แตกต่างกัน

เมื่อ $IR = 0.10$ ค่า Accuracy และ TN Rate ส่วนใหญ่จะสูงสุด เมื่อเป็นข้อมูลดั้งเดิม โดย Accuracy อยู่ในช่วงร้อยละ 90.77 – 100.00 และ TN Rate อยู่ในช่วงร้อยละ 98.85 – 100.00 สำหรับ TP Rate วิธี ROS จะสูงสุดเมื่อ ClassDif = 1 โดยมีค่าอยู่ในช่วงร้อยละ 67.72 – 78.72 ส่วนสถานการณ์อื่นส่วนใหญ่ วิธี HIDPO จะให้ผลใกล้เคียงกับวิธี ROS สำหรับค่า F-measure วิธี HIDPO จะสูงสุดอย่างมีนัยสำคัญทางสถิติเมื่อ ClassDif = 1 โดยมีค่าอยู่ในช่วงร้อยละ 38.17 – 54.86 และข้อมูลดั้งเดิมมีค่าสูงสุดเมื่อ ClassDif = 2, RelVar = 2 ส่วนสถานการณ์อื่น แต่ละวิธีส่วนใหญ่ให้ผลไม่แตกต่างกัน

เมื่อ $IR = 0.20$ ค่า Accuracy และ TN Rate ส่วนใหญ่จะสูงสุด เมื่อเป็นข้อมูลดั้งเดิม โดย Accuracy อยู่ในช่วงร้อยละ 86.19 – 100.00 และ TN Rate อยู่ในช่วงร้อยละ 96.57 – 100.00 สำหรับ TP Rate วิธี ROS จะสูงสุดเมื่อ ClassDif = 1, RelVar = 2 โดยมีค่าร้อยละ 72.13 ส่วนสถานการณ์อื่นส่วนใหญ่ วิธี ROS และวิธี HIDPO จะสูงสุดโดยมีค่าใกล้เคียงกัน สำหรับค่า F-measure วิธี HIDPO จะสูงสุดอย่างมีนัยสำคัญเมื่อ ClassDif = 1, RelVar = 2 โดยมีค่าร้อยละ 49.37 เมื่อ ClassDif = 5 วิธีทั้งสามให้ผลไม่แตกต่างกัน ส่วนสถานการณ์อื่นข้อมูลดั้งเดิมให้ค่าสูงสุด โดยมีค่าอยู่ในช่วงร้อยละ 67.43 – 96.24

จากตารางที่ 5 เมื่อ $n = 300$ พบว่าเมื่อ $IR = 0.05$ ค่า Accuracy และ TN Rate ส่วนใหญ่จะสูงสุด เมื่อเป็นข้อมูลดั้งเดิม โดย Accuracy อยู่ในช่วงร้อยละ 95.52 – 100.00 และ TN Rate อยู่ในช่วงร้อยละ 99.67 – 100.00 สำหรับ TP Rate ส่วนใหญ่จะสูงสุด เมื่อใช้วิธี ROS โดยอยู่ในช่วงร้อยละ 65.13 – 100.00 และค่า F-measure สูงสุดอย่างมีนัยสำคัญทางสถิติ เมื่อใช้วิธี HIDPO ในสถานการณ์ ClassDif = 1 โดยมีค่าอยู่ในช่วงร้อยละ 25.36 – 40.37 ส่วนสถานการณ์อื่นส่วนใหญ่ วิธี HIDPO ให้ผลใกล้เคียงกับวิธี ROS



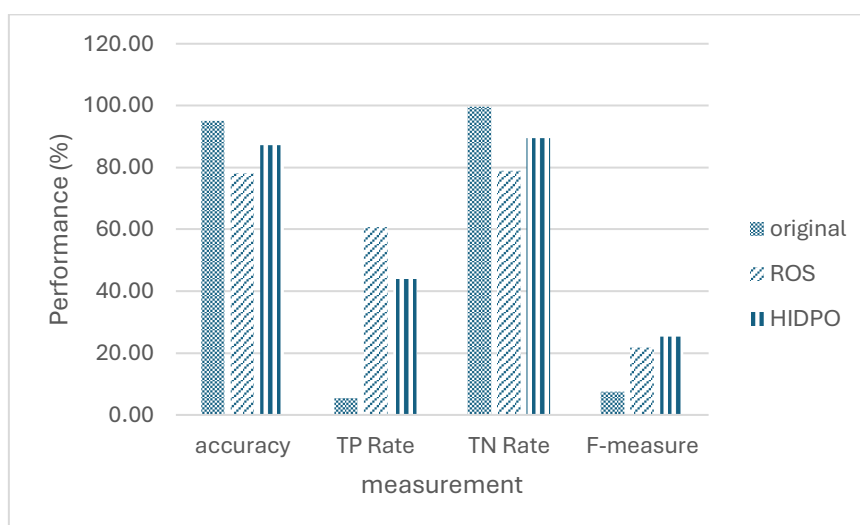
เมื่อ $IR = 0.07$ ค่า Accuracy และ TN Rate ส่วนใหญ่จะสูงสุดเมื่อเป็นข้อมูลดั้งเดิม โดย Accuracy อยู่ในช่วงร้อยละ 93.48 – 100.00 และ TN Rate อยู่ในช่วงร้อยละ 99.40 – 100.00 สำหรับ TP Rate ส่วนใหญ่จะสูงสุดเมื่อใช้วิธี ROS โดยอยู่ในช่วงร้อยละ 68.90 – 100.00 และค่า F-measure สูงสุดอย่างมีนัยสำคัญทางสถิติ เมื่อใช้วิธี HIDPO ในสถานการณ์ $ClassDif = 1$ และ $RelVar = 2$ โดยมีค่าร้อยละ 31.56 ส่วนสถานการณ์อื่นส่วนใหญ่ข้อมูลดั้งเดิมให้ค่าสูงสุด

เมื่อ $IR = 0.10$ ค่า Accuracy และ TN Rate ส่วนใหญ่จะสูงสุดเมื่อเป็นข้อมูลดั้งเดิม โดย Accuracy อยู่ในช่วงร้อยละ 91.49 – 100.00 และ TN Rate อยู่ในช่วงร้อยละ 98.91 – 100.00 สำหรับ TP Rate วิธี ROS ส่วนใหญ่จะสูงสุดโดยมีค่าอยู่ในช่วงร้อยละ 70.96 – 100.00 สำหรับค่า F-measure วิธี HIDPO จะสูงสุดอย่างมีนัยสำคัญทางสถิติเมื่อ $ClassDif = 1$ และ $RelVar = 2$ โดยมีค่าร้อยละ 36.67 ส่วนสถานการณ์อื่นส่วนใหญ่ข้อมูลดั้งเดิมให้ค่าสูงสุด

เมื่อ $IR = 0.20$ ค่า Accuracy และ TN Rate ส่วนใหญ่จะสูงสุดเมื่อเป็นข้อมูลดั้งเดิม โดย Accuracy อยู่ในช่วงร้อยละ 85.63 – 100.00 และ TN Rate อยู่ในช่วงร้อยละ 96.57 – 100.00 สำหรับ TP Rate วิธี HIDPO ส่วนใหญ่จะสูงสุดแต่ไม่มีนัยสำคัญทางสถิติ โดยมีค่าอยู่ในช่วงร้อยละ 73.03 – 100.00 และสำหรับค่า F-measure วิธี HIDPO จะสูงสุดแต่ไม่มีนัยสำคัญเมื่อ $ClassDif = 1$ และ $RelVar = 2$ โดยมีค่าร้อยละ 49.88 ส่วนสถานการณ์อื่นส่วนใหญ่ข้อมูลดั้งเดิมให้ค่าสูงสุด

อภิปรายและสรุปผลการวิจัย

จากการเปรียบเทียบประสิทธิภาพการจำแนก เมื่อใช้ข้อมูลดั้งเดิม ข้อมูลจากการสุ่มเพิ่มข้อมูลระหว่างวิธี ROS และวิธี HIDPO บนข้อมูลจำลอง พบว่าวิธี HIDPO มักให้ค่าการวัดเอฟสูงสุดอย่างมีนัยสำคัญทางสถิติในสถานการณ์ที่มีความแตกต่างกันน้อยระหว่างกลุ่มส่วนน้อยและกลุ่มส่วนใหญ่ ($ClassDif = 1$) อันเป็นสถานการณ์ที่มีความยากต่อการจำแนกประเภท โดยให้ผลดีเป็นพิเศษในสถานการณ์ที่มีความไม่สมดุลอย่างมาก (IR มีค่าน้อย) ซึ่งหากปัญหาการจำแนกประเภทนั้นต้องนำไปใช้กับข้อมูลไม่สมดุล ค่าวัดประสิทธิภาพที่เหมาะสมในการสะท้อนความเป็นจริงควรใช้ค่าการวัดเอฟ [17] ดังนั้นในสถานการณ์ดังกล่าววิธี HIDPO จึงเหมาะสมในการใช้เป็นอัลกอริทึมเพื่อสุ่มเพิ่มข้อมูลให้มีความสมดุล ภาพที่ 1 แสดงการเปรียบเทียบค่าวัดประสิทธิภาพทั้ง 4 ระหว่างข้อมูลดั้งเดิม ROS และ HIDPO สำหรับสถานการณ์ $IR = 0.05$, $ClassDif = 1$, $RelVar = 2$ และ $n = 210$ ซึ่งจะพบว่าวิธี HIDPO ให้ค่าการวัดเอฟสูงสุด ส่วนค่าวัดอื่น ๆ ให้ค่าปานกลาง



ภาพที่ 1 การเปรียบเทียบ Accuracy, TP Rate, TN Rate และ F-measure ระหว่างข้อมูลดั้งเดิม ROS และ HIDPO

สำหรับสถานการณ์ $IR = 0.05$, $ClassDif = 1$, $RelVar = 2$ และ $n = 210$

อย่างไรก็ตามเมื่อ $n = 80$ พบว่าวิธี HIDPO ไม่ได้ให้ค่าการวัดเอฟสูงสุดสำหรับสถานการณ์ที่ $\text{ClassDif} = 1$ ถึงแม้ $\text{IR} = 0.05$ ทั้งนี้เนื่องจากในสถานการณ์ดังกล่าว กลุ่มส่วนน้อยมีจำนวนเพียงประมาณ 4 หน่วยตัวอย่าง และเมื่อแบ่งข้อมูลเป็นข้อมูลฝึกฝนจะเหลือหน่วยตัวอย่างเพียงประมาณ 3 หน่วยตัวอย่าง โดยเป็นจำนวนที่น้อยมาก การสุ่มเพิ่มข้อมูลด้วยวิธี HIDPO จึงให้ผลลัพธ์ไม่ค่อยแตกต่างจากวิธี ROS ซึ่งเมื่อพิจารณาที่ $n = 150$ ในสถานการณ์ดังกล่าวแล้ว พบว่าขนาดตัวอย่างที่นำไปใช้ฝึกฝนไม่ควรมีจำนวนน้อยกว่า 5

กรณีที่กลุ่มส่วนน้อยและกลุ่มส่วนใหญ่มีค่าเฉลี่ยแตกต่างกันมาก หรือ $\text{ClassDif} = 5$ การวัดความสำคัญของแต่ละหน่วยตัวอย่างด้วยวิธี HIDPO จะมีโอกาสมากที่ไม่สามารถพบความสำคัญของแต่ละหน่วยตัวอย่างได้ เนื่องจากข้อมูลมีความแตกต่างกันอย่างชัดเจนระหว่างกลุ่ม ดังนั้นเมื่อมีการตัดหน่วยตัวอย่างหนึ่งตัวออกก็ยังคงพบความแตกต่างระหว่างกลุ่มได้ ทำให้ไม่สามารถแสดงอิทธิพลหรือความสำคัญของหน่วยตัวอย่างที่ถูกตัดทิ้งได้ และมีโอกาสมากที่จะพบว่าไม่มีหน่วยตัวอย่างใดเลยที่มีความสำคัญตามเกณฑ์ที่กำหนด ซึ่งในกรณีเช่นนี้ HIDPO จะถ่วงน้ำหนักแบบสุ่มด้วยความน่าจะเป็นที่เท่ากันเช่นเดียวกับ ROS จึงทำให้ได้ผลใกล้เคียงกันมาก

ข้อเสนอแนะ

วิธี HIDPO เป็นทางเลือกในการแก้ปัญหาการจำแนกประเภทสำหรับข้อมูลไม่สมดุล เนื่องจากให้ค่าการวัดเอฟที่สูงสุดในสถานการณ์ที่มีความแตกต่างกันน้อยระหว่างกลุ่มส่วนน้อยและกลุ่มส่วนใหญ่ ในสถานการณ์ที่มีความไม่สมดุลมาก ถึงแม้ค่าวัดประสิทธิภาพอื่นจะไม่ได้ให้ค่าสูงสุด แต่ในสถานการณ์ของการไม่สมดุลของกลุ่มส่วนน้อยและกลุ่มส่วนมาก ค่าวัดที่เหมาะสมควรใช้ค่าการวัดเอฟ

การคัดเลือกตัวแปร เป็นกระบวนการหนึ่งที่จะช่วยให้การจำแนกประเภทมีประสิทธิภาพมากขึ้น โดยเฉพาะเมื่อข้อมูลไม่สมดุล การระบุเฉพาะตัวแปรทำนายที่เกี่ยวข้องมาใช้ในการสร้างตัวแบบจะช่วยให้สามารถสร้างขอบเขตการตัดสินใจที่ชัดเจนเกี่ยวกับแต่ละกลุ่มได้ ซึ่งจะช่วยให้ตัวแบบสามารถจำแนกประเภทได้แม่นยำขึ้น [18] โดยกระบวนการนี้เป็นส่วนหนึ่งของวิธี HIDPO ซึ่งการคัดเลือกตัวแปรทำนายให้เหลือเฉพาะตัวแปรที่มีความเกี่ยวข้องกับการจำแนกประเภท จะทำให้สามารถวัดความสำคัญของแต่ละหน่วยตัวอย่างได้ถูกต้องมากขึ้น สำหรับงานวิจัยนี้ได้ใช้การคัดเลือกตัวแปรโดยใช้สถิติทดสอบที่ช่วยระดับนัยสำคัญ 0.10 จึงควรมีการศึกษาต่อไปถึงวิธีการคัดเลือกตัวแปรที่เหมาะสมในการใช้ร่วมกับวิธี HIDPO

เอกสารอ้างอิง

1. Sara F, Shahrokh A, Michael WK. A comprehensive data level analysis for cancer diagnosis on imbalanced data. J Biomed Inf 2019;90:103089.
2. Soh WW, Rika MY. Predicting Credit Card Fraud on an Imbalanced Data. Int J Data Sci Adv Anal 2019;1(1):12-7.
3. Meryem C, Mohamed H. A sight on defect detection methods for imbalanced industrial data. ITM Web of Conferences 2022;43:01012.
4. กิตติภาพ แซ่เตีย, จิรภัทร หยกรัตนศักดิ์. การจัดการข้อมูลไม่สมดุลของการทำกลยุทธ์เสนอขายประกันต่อยอดสำหรับผู้ถือบัตรเครดิต. ใน: เอกสารประกอบการประชุมวิชาการระดับชาติ ครั้งที่ 13 วันที่ 8 - 9 กรกฎาคม 2564. มหาวิทยาลัยราชภัฏนครปฐม. นครปฐม; 2564. หน้า 514-23.
5. Andrea B. Imbalanced Data Classification with Neural Networks and Classifiers. [Internet]. 2021 [cited 2024 Apr 14]. Available from: <https://aboskovic21.github.io/projects/thesis.pdf>



6. Aida A, Siti MS, Anca LR. Classification with class imbalance problem: A Review. *Int J Advance Soft Compu Appl* 2015;7(3):176-204.
7. Firuz K, Ho-Hon L, Aswani KC. Keep it simple: random oversampling for imbalanced data. In: *proceedings of Advances in Science and Engineering Technology International Conferences (ASET)*, February 20-23, 2023; Dubai, United Arab Emirates; 2023. p. 1-4.
8. Michelle J, Maria M. A Comparison of Resampling Techniques to Handle the Class Imbalance Problem in Machine Learning Conversion prediction of Spotify Users - A Case Study. [Internet]. 2017 [cited 2024 Apr 14]. Available from: <https://www.kth.se/social/files/5a5ad14056be5b323d61de10/MJagelid.MMovin.pdf>
9. กัลยา วานิชย์บัญชา. การวิเคราะห์ข้อมูลหลายตัวแปร. กรุงเทพฯ: บริษัทธรรมสารจำกัด; 2552.
10. Ronald EW, Raymond HM. *Probability and Statistics for Engineers and Scientists*. 5th ed. New York: Macmillan Publishing Company; 1993.
11. Qianyou M. Recent Applications and Perspectives of Logistic Regression Modelling in Healthcare. In: *proceedings of the 2nd International Conference on Mathematical Physics and Computational Simulation*, August 9, 2024; Glasgow, UK; 2024. p.185-90.
12. Mingze S. Research on Influencing Factors of Video Game Sales using Binary Logistic Regression. In: *3rd International Conference on Applied Mathematics, Modeling Simulation and Automatic Control (AMMSAC 2024)*, June 22-23, 2024; San Diego, USA; 2024. p. 65-70.
13. Abdulrashid S, Zahriya LH, Anas TB. A Logistic Regression-based Model for Identifying Credit Card Fraudulent Transactions. *Asian J Res Com Sci* 2024;17(7):41-54.
14. Prabhakaran N, Nedunchelian R. Combined Feature Set with Logistic Regression Model to Detect Credit Card Frauds in Real Time Applications. *J Mach Comput* 2024;4(3):804-12.
15. Art BO. Infinitely Imbalanced Logistic Regression. *J Mach Learn Res* 2007;8:761-73.
16. Pang-Ning T, Michael S, Vipin K. *Introduction to Data Mining*. Boston: Pearson Education, Inc.; 2006.
17. Lian Y, Nengfeng Z. Survey of Imbalanced Data Methodologies. [Internet]. 2021 [cited 2024 Apr 14]. Available from: <https://arxiv.org/pdf/2104.02240>
18. Himanshu T. What Is Balanced And Imbalanced Dataset?. [Internet]. 2019 [cited 2024 Apr 14]. Available from: <https://medium.com/analytics-vidhya/what-is-balance-and-imbalance-dataset-89e8d7f46bc5>