Science and Technology
RMUTT Journal
Vol.8 No.2 (2018): 125-142
www.sci.rmutt.ac.th/stj
Online ISSN2229-1547

วิธีการปรับข้อมูลตัวอย่างแบบผสมผสานเพื่อเพิ่มประสิทธิภาพการจำแนกข้อมูลที่มีจำนวนตัวอย่างในแต่ละคลาสไม่สมดุลกัน

A Hybrid Data Level Approach for Improving Classification Performance in Imbalanced Dataset

วันทนี ประจวบศุภกิจ

ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีและการจัดการอุตสาหกรรม
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ตำบลเนินหอม อำเภอเมือง จังหวัดปราจีนบุรี 25230
E-mail: wanthanee.p@fitm.kmutnb.ac.th

บทคัดย่อ

ปัญหาความไม่สมดุลกันของจำนวนตัวอย่างในแต่ละคลาสเป็นปัญหาที่เกิดจากจำนวนตัวอย่างในคลาสใดคลาสหนึ่งมีการกระจายตัวลักษณะลาดเอียงสูงกว่าอีกคลาสหนึ่งเป็นจำนวนมาก ซึ่งการจำแนกบนชุดข้อมูลที่มีปัญหาแบบนี้ส่งผลให้ตัวจำแนกพื้นฐานจะจำแนกให้ผลลัพธ์ที่ดีกับคลาสที่มีจำนวนมาก (คลาสลบ) แต่ให้ผลลัพธ์ในการจำแนกที่แย่ลงกับคลาสที่มีจำนวนน้อย (คลาสบวก) ดังนั้น งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาขั้นตอนวิธีสำหรับการจำแนกประเภทข้อมูลที่มีจำนวนตัวอย่างในแต่ละความไม่สมดุลกัน โดยนำเสนอแนวทางการปรับปรุงการจำแนกข้อมูลโดยใช้เทคนิคการจัดกลุ่ม (Clustering) เพื่อลดข้อมูลตัวอย่างในกลุ่มมากควบคู่เทคนิคการสุ่มเพิ่มข้อมูลตัวอย่างในกลุ่มน้อย ในชื่อ Clustering Switching Method for Sampling Imbalanced Data หรือ ClusIM ที่สามารถเพิ่มประสิทธิภาพการจำแนกข้อมูลในข้อมูลกลุ่มน้อยหรือคลาสบวกให้มีความแม่นยำมากขึ้น จากผลการทดลองจะแสดงให้เห็นว่า ClusIM ให้ประสิทธิภาพในการจำแนกได้ดีกว่าขั้นตอนวิธีอื่น ๆ ที่นำมาเปรียบเทียบโดยเฉพาะอย่างยิ่งการจำแนกในคลาสบวกซึ่ง ClusIM มีค่า F-measure และ G-mean เฉลี่ยที่ร้อยละ 90 ในทุกชุดข้อมูล นอกจากนี้ยังพบว่า ClusIM สามารถลดความซ้อนทับและลดอัตราความไม่สมดุลกันระหว่างคลาสบวกและคลาสลบได้อย่างมีประสิทธิภาพ

Received: June 14, 2018
Revised: November 12, 2018
Accepted: December 27, 2018

คำสำคัญ: ปัญหาจำนวนตัวอย่างไม่สมดุลกัน ขั้นตอนวิธีคลัสอิม วิธีการปรับข้อมูลแบบผสมผสาน การปรับข้อมูลตัวอย่าง การจำแนกข้อมูล การทำเหมืองข้อมูล

Abstract

The imbalanced problem occurs when the number of instance in the one class sharply outnumber another class. The classification on imbalanced data always brings about problems because the traditional classifiers tend to predict well on the majority class while the prediction based on the minority class is poor. Therefore, the aim of this research is to propose the hybrid data level approaches in order to improve the classification performance based on the two-class imbalanced dataset. This research introduces a new approach that combines the clustering approach of k-means algorithm and over-sampling techniques namely Clustering Switching Method for Sampling Imbalanced Data or ClusIM. The research's result shows that ClusIM has higher F-measure and G-mean results than the other methods especially on majority classes that ClusIM obtains the F-measure and the G-mean values about 90% on all dataset. Moreover, ClusIM reduced the overlap and imbalanced ratio between classes to get good performance.

Keywords: Imbalanced data, ClusIM algorithm, Hybrid data level approaches, Sampling data, Classification, Data mining

1. บทนำ

การจำแนกประเภทข้อมูล เป็นเทคนิคหนึ่งที่สำคัญในการทำเหมืองข้อมูล ซึ่งเทคนิคการจำแนกประเภทข้อมูล เป็นกระบวนการสร้างโมเดลเพื่อใช้ในการจำแนกข้อมูลให้อยู่ในคลาส (ประเภท) ที่กำหนดให้ จุดประสงค์ของการจำแนกประเภทข้อมูลคือการสร้างโมเดลของคลาส ซึ่งโมเดลที่ได้จากการจำแนกข้อมูลจะสามารถนำมาใช้ในการจำแนกหรือทำนายข้อมูลที่ต้องการในอนาคตได้ จากการสำรวจยังพบว่า [5, 10, 11, 13] ข้อมูลที่จะถูกนำมาใช้ในการจำแนกประเภท (คลาส) นั้น มีจำนวนไม่น้อยที่พบปัญหาความไม่สมดุลกันของจำนวนตัวอย่างในแต่ละคลาส (Imbalanced Data) ซึ่งเกิดจากจำนวนตัวอย่างในคลาสใดคลาสหนึ่งมีการกระจายตัวลักษณะลาดเอียงสูงกว่าอีกคลาสหนึ่งเป็นจำนวนมาก โดยในคลาสที่มีจำนวนข้อมูลตัวอย่างจำนวนมากจะถูกเรียกว่า ข้อมูลกลุ่มมาก (Majority class หรือคลาสลบ) ส่วนคลาสที่มีจำนวนตัวอย่างน้อยกว่าจะถูกเรียกว่า ข้อมูลกลุ่มน้อย (Minority class หรือคลาสบวก)

ปัญหาจากการที่ชุดข้อมูลมีจำนวนตัวอย่างในแต่ละคลาสไม่สมดุลกันนี้ส่งผลให้ตัวจำแนกพื้นฐานจะจำแนกให้ผลลัพธ์ที่ดีกับคลาสลบ แต่ให้ผลลัพธ์ในการจำแนกที่แย่ลงกับคลาสบวก ซึ่งในการจำแนกข้อมูลคลาสบวกนี้จะมีความสำคัญอย่างมากสำหรับการจำแนกข้อมูลบางประเภท ตัวอย่างเช่น การจำแนกข้อมูลเพื่อตรวจหาการบุกรุกทางเครือข่าย [4] การจำแนกข้อมูลเพื่อใช้ในการวินิจฉัยโรคทางการแพทย์ [7] การจำแนกประเภทข้อมูลผลสัมฤทธิ์ทางการเรียนของนักศึกษา [12] ซึ่ง

ข้อมูลเหล่านี้จะมีจำนวนตัวอย่างในคลาสที่มีความผิดปกติน้อยว่าข้อมูลปกติจำนวนมาก และตัวจำแนกจะจำแนกข้อมูลได้ไม่ดีในคลาสเหล่านี้ ส่งผลให้เกิดความผิดพลาดในการจำแนกข้อมูลที่ต้องให้ความสำคัญ

ตัวอย่างเช่น การจำแนกข้อมูลทางการแพทย์ที่ส่วนใหญ่พบว่าจำนวนข้อมูลตัวอย่างของคลาสผู้ป่วยทั่วไปจะมีจำนวนมากกว่าจำนวนผู้ป่วยวิกฤตอยู่มาก ซึ่งหากนำมาจำแนกด้วยขั้นตอนวิธีการจำแนกทั่วไปแล้วตัวจำแนกจะให้ค่าความถูกต้องในภาพรวม (Accuracy) ถึงร้อยละ 90 ซึ่งในความเป็นจริงค่านี้เป็นค่าความถูกต้องที่ไม่สามารถวัดประสิทธิภาพของตัวจำแนกได้จริง เพราะหากพิจารณาทีละคลาสแล้วพบว่าตัวจำแนกจะไม่สามารถจำแนกข้อมูลผู้ป่วยวิกฤต (คลาสบวก) ได้ถูกต้องอันเนื่องมาจากตัวจำแนกจะโน้มเอียงไปยังข้อมูลส่วนมาก ซึ่งผลลัพธ์การจำแนกถูกของคลาสนี้จะมีค่าที่ต่ำมาก บางทีตัวจำแนกอาจจะจำแนกได้ไม่ถูกต้องเลยนั่นคือมีค่าความถูกต้องเป็น 0 ซึ่งในความเป็นจริงข้อมูลในคลาสนี้จะเป็นคลาสที่ให้ความสำคัญมากและตัวจำแนกควรจำแนกได้ถูกต้อง

สำหรับวิธีการที่ใช้ในการแก้ปัญหาชุดข้อมูลที่มีจำนวนตัวอย่างในแต่ละคลาสไม่สมดุลกันนั้น สามารถแบ่งเป็น 2 วิธีคือ การแก้ปัญหาในระดับข้อมูล (data-level approach) และการแก้ปัญหาที่ขั้นตอนวิธี (algorithm-level approach) โดยในระดับข้อมูลเป็นวิธีการที่ปรับจำนวนตัวอย่างโดยการกระจายจำนวนตัวอย่างในแต่ละคลาสให้มีความสมดุลก่อนที่ตัวจำแนกจะเรียนรู้กับข้อมูลเหล่านั้น ส่วนวิธีขั้นตอนวิธีเป็นการเพิ่มความสามารถในการจำแนกให้กับตัวจำแนก โดยการปรับขั้นตอนวิธีให้คำนึงถึงความสำคัญของคลาสบวก สำหรับการแก้ปัญหาในระดับข้อมูลแบ่งออกเป็น 3 วิธีคือ

วิธีที่ 1) การเพิ่มข้อมูลตัวอย่างในคลาสบวก (Over-sampling) คือการทำให้จำนวนตัวอย่างในแต่ละคลาสมีความสมดุลกันโดยการเพิ่มข้อมูลตัวอย่างของคลาสบวกซึ่งมีหลายวิธี เช่น ROS (Random Over-sampling) เป็นเทคนิคในการเพิ่มตัวอย่างของคลาสบวก โดยการสุ่มจำนวนตัวอย่างของคลาสบวกเพิ่มขึ้นเรื่อย ๆ จนมีความสมดุลกับคลาสลบ ซึ่งวิธีการนี้จะทำให้เกิดปัญหา over-fitting ได้เพราะมีการทำซ้ำตัวอย่างของคลาสบวก นอกจากนี้ยังมีงานวิจัยของ Chawla *et al.* [3] ได้พัฒนาขั้นตอนวิธีชื่อ SMOTE ที่ใช้การเพิ่มข้อมูลตัวอย่างของคลาสบวกโดยการสร้างตัวอย่างสังเคราะห์ (synthetic samples) จากการสร้าง feature space และใช้ขั้นตอนวิธี K-Nearest Neighbor ในการหาเพื่อนบ้านใกล้เคียงตามจำนวน k ของข้อมูลตัวอย่างคลาสบวก x หลังจากนั้นจะสุ่มข้อมูลตัวอย่าง 1 ตัวคือ y ที่อยู่ในกลุ่ม k แล้วหาระยะห่างระหว่าง x และ y โดย SMOTE จะสุ่มค่า gap ขึ้นมาระหว่าง 0 กับ 1 เพื่อนำค่า gap นี้มาสร้างข้อมูลตัวอย่างสังเคราะห์ โดยนำมาคูณกับค่าระยะห่างแล้วบวกกับ x ในหลักการนี้ทำให้สามารถสร้างข้อมูลตัวอย่างสังเคราะห์ขึ้นมาได้โดยไม่ต้องสร้างคลาสบวกซ้ำ หลังจากการทดลองพบว่า หลักการนี้สามารถเพิ่มประสิทธิภาพในการจำแนกข้อมูลตัวอย่างในคลาสบวกได้ดี โดยไม่เกิดปัญหาโมเดลที่ถูกสร้างจากการเรียนรู้มีความสามารถในการจำแนกดีแต่ไม่สามารถนำไปจำแนกข้อมูลในอนาคตได้ดี (overfitting)

วิธีที่ 2) การลดข้อมูลตัวอย่างในคลาสลบ (Under-sampling) คือ การลดจำนวนข้อมูลตัวอย่างที่มีคำเฉลยเป็นคลาสลบ เพื่อให้จำนวน

ข้อมูลตัวอย่างทั้งสองคลาสมีความสมดุลกัน ตัวอย่างวิธีการลดข้อมูลตัวอย่างในคลาสลบเช่น RUS (Random Undersampling) เป็นเทคนิคในการลดจำนวนข้อมูลตัวอย่างโดยวิธีสุ่มลบข้อมูลตัวอย่างในคลาสลบจนกระทั่งการกระจายตัวของคลาสมีความสมดุลเท่ากับคลาสบวก แต่วิธีการนี้ก็มีข้อเสียคือในการลดข้อมูลตัวอย่างของคลาสบวกอาจจะทำให้ข้อมูลที่สำคัญ ๆ สูญหายจากชุดข้อมูลสอนระบบได้ นอกจากนี้ยังมีการลดข้อมูลตัวอย่างด้วยเทคนิคการจัดกลุ่มคืองานวิจัยของ Xiaoheng Deng และคณะ [5] ที่ได้นำเสนอวิธีการ ACUS ซึ่งเป็นขั้นตอนการลดข้อมูลตัวอย่างในคลาสลบโดยใช้วิธีการ ensemble ที่ให้เกิดการจัดกลุ่มแบบอัตโนมัติ ในเบื้องต้น ACUS ได้จัดกลุ่มโดยใช้ค่าน้ำหนักของข้อมูลตัวอย่างหลังจากนั้นจะค่อยๆ สร้างกลุ่มข้อมูลขึ้นมาใหม่โดยสร้างชุดข้อมูลที่สมดุลตามจำนวนอัตราส่วนของคลาสบวกและคลาสลบ และจำแนกข้อมูลทั้งหมดด้วยขั้นตอนวิธี Adaboost โดยทดลองกับชุดข้อมูลจาก UCI จำนวน 22 ชุดข้อมูล ซึ่งจากผลการทดลองพบว่าวิธีการดังกล่าวให้ประสิทธิภาพในการจำแนกข้อมูลที่มีจำนวนชุดข้อมูลตัวอย่างในแต่ละคลาสไม่สมดุลกันได้ถูกต้องเมื่อเทียบวิธีการอื่น ๆ ที่เปรียบเทียบ

วิธีที่ 3) การปรับข้อมูลตัวอย่างแบบผสมผสาน (Hybrid-sampling) คือ วิธีการรวมทั้งหลักการของ Over-sampling และ Under-sampling เพื่อปรับชุดข้อมูลตัวอย่างในแต่ละคลาสให้มีความสมดุลกัน ซึ่งงานวิจัยของ Gazzah et al. [9] ได้ใช้เทคนิคนี้ในการเพิ่มประสิทธิภาพในการจำแนกด้วยการใช้เทคนิค SMOTE star topology ควบคู่กับ การลดข้อมูลตัวอย่างคลาสบวกที่มีความเกี่ยวข้องต่ำ โดยทดลองกับชุดข้อมูลทางชีวภาพและเลือกใช้ตัวจำแนก SVM ซึ่งผลการทดลองแสดงให้เห็นว่าวิธีการนี้ให้ค่า true-positive ที่ดีและมีประสิทธิภาพ

ซึ่งจากการสำรวจพบว่าการจัดการปัญหาข้อมูลตัวอย่างที่มีจำนวนตัวอย่างในแต่ละคลาสไม่สมดุลกันนั้น วิธีการระดับข้อมูลเป็นวิธีการที่ได้รับความนิยมอย่างมาก เพราะเป็นเทคนิคที่ไม่ซับซ้อน ทั้งสามารถสุ่มเพิ่มหรือสุ่มลดข้อมูลตัวอย่างก่อนที่จะนำไปจำแนกกับขั้นตอนวิธีต่าง ๆ และสามารถนำไปประยุกต์ใช้กับขั้นตอนวิธีที่หลากหลาย

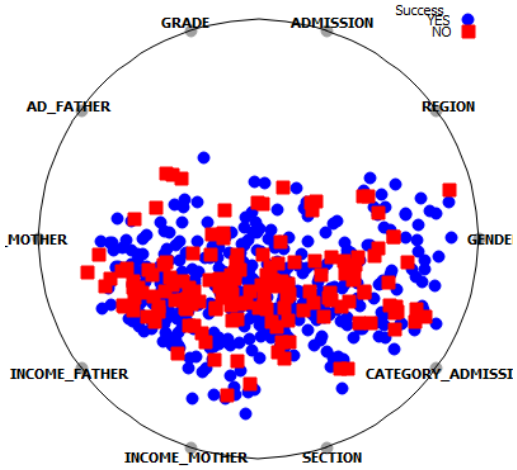
ดังนั้น ในงานวิจัยนี้ผู้วิจัยมุ่งเน้นการพัฒนาขั้นตอนวิธีในการสุ่มข้อมูลที่สามารถเพิ่มประสิทธิภาพการจำแนกข้อมูลในข้อมูลกลุ่มน้อยให้มีความแม่นยำมากขึ้น โดยนำเสนอแนวทางการปรับปรุงการจำแนกข้อมูลโดยใช้เทคนิคการปรับข้อมูลแบบผสมผสาน (Hybrid Data Level) บนชุดข้อมูลที่คลาสไม่สมดุลกันด้วยเทคนิคการจัดกลุ่มข้อมูลและการสุ่มข้อมูล (Sampling) ในชื่อ Clustering Switching Method for Sampling Imbalanced Data หรือ เรียกว่า ClusIM ซึ่งวิธีการที่นำเสนอเป็นการสร้างและสลับเปลี่ยนกลุ่มข้อมูลร่วมกับการผสมผสานเทคนิคการสุ่มเพิ่มข้อมูลกลุ่มน้อย โดยนำเทคนิคการจัดกลุ่มข้อมูลมาใช้ในการลดข้อมูลตัวอย่างแบบวนซ้ำ (Repetitive Under-sampling) เพื่อที่เพิ่มประสิทธิภาพในการจำแนกข้อมูลกลุ่มน้อยให้ดีขึ้นและไม่กระทบต่อประสิทธิภาพการจำแนกของข้อมูลกลุ่มมาก โดยในการทดลองผู้วิจัยอาศัยชุดข้อมูลจาก 2 แหล่งคือ แหล่งที่ 1 จาก UCI [1] จำนวน 6 ชุด และแหล่งที่ 2 คือข้อมูลของนักศึกษาในหลักสูตรเทคโนโลยีบัณฑิตที่เข้าศึกษาตั้งแต่ปีการศึกษา 2548-2553 สาขาวิชาเทคโนโลยีสารสนเทศ ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีและการจัดการ

อุตสาหกรรม มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ วิทยาเขตปราจีนบุรี จำนวน 1 ชุดมาประกอบการทดลอง ในส่วนของข้อมูลแหล่งที่ 2 ถูกนำมาใช้เพื่อแสดงให้เห็นว่าขั้นตอนวิธีที่พัฒนาขึ้นนี้สามารถนำมาประยุกต์ใช้ให้เกิดประโยชน์กับปัญหาที่มีอยู่จริง โดยสามารถสร้างเป็นโมเดลในการจำแนกข้อมูลดังกล่าวเพื่อนำผลที่ได้จากการวิจัยไปเป็นแนวทางในการจัดการเรียนการสอน หรือให้คำแนะนำในเรื่องการเรียนแก่นักศึกษาในอนาคต

2. วิธีดำเนินการวิจัย

2.1 สมมติฐานงานวิจัย

แนวคิดของงานวิจัยนี้เกิดจากการสังเกตจำนวนตัวอย่างของข้อมูลที่ผู้วิจัยนำมาใช้ในการวิจัยทางด้านการทำเหมืองข้อมูลโดยส่วนมากแล้วจะเป็นข้อมูลที่มีจำนวนตัวอย่างในแต่ละคลาสไม่สมดุลกัน ไม่ว่าจะเป็นข้อมูลนักศึกษา ข้อมูลประเภทของลูกค้า เป็นต้น ซึ่งบางชุดข้อมูลมีความไม่สมดุลน้อย บางชุดข้อมูลมีความไม่สมดุลมาก นอกจากนี้ยังพบว่าการกระจายตัวของข้อมูลนั้นมีความซ้อนทับกันในแต่ละคลาส ดังรูปที่1 เป็นชุดข้อมูลนักศึกษาที่แบ่งเป็นคลาส Yes และคลาส No ซึ่งจำนวนของนักศึกษาในคลาส Yes มีมากกว่าคลาส No จำนวนมาก นอกจากนี้ยังพบว่ามีการซ้อนทับกันระหว่างคลาส จึงทำให้ประสิทธิภาพในการจำแนกข้อมูลของตัวจำแนกต่ำลง ซึ่งคุณสมบัติของคลาสลบนั้นจะบดบังคุณสมบัติของคลาสบวกทำให้เกิดการจำแนกผิดพลาด



รูปที่ 1 แสดงตัวอย่างข้อมูลที่มีจำนวนตัวอย่างในแต่ละคลาสไม่สมดุลกันและมีคลาสที่ซ้อนทับกัน

ดังนั้น งานวิจัยนี้จึงมีแนวคิดในการพัฒนาขั้นตอนวิธีในการเพิ่มประสิทธิภาพการจำแนกข้อมูลของข้อมูลตัวอย่างในคลาสบวกโดยการลดการซ้อนทับกันระหว่างคลาสและสังเคราะห์ตัวอย่างในคลาสบวกให้มีจำนวนเพิ่มมากขึ้น ซึ่งน่าจะส่งผลให้ตัวจำแนกสามารถจำแนกข้อมูลได้มีประสิทธิภาพมากยิ่งขึ้น

2.2 ข้อมูลที่นำมาใช้ในการทดลอง

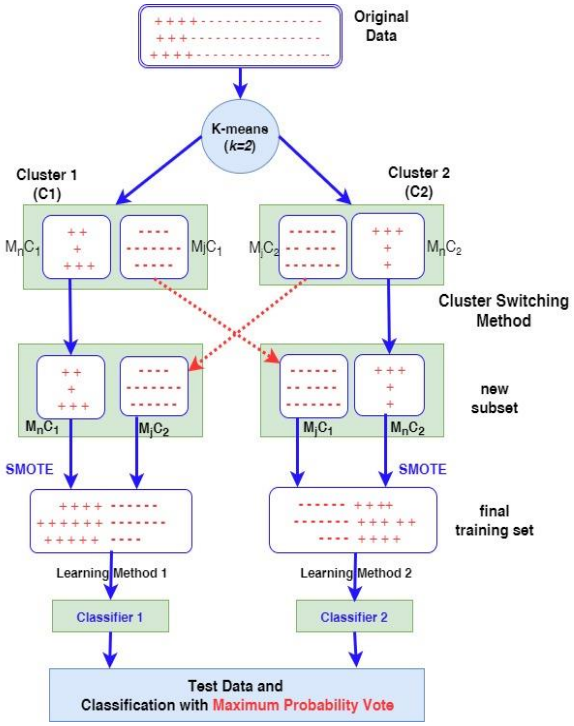
ในงานวิจัยนี้ได้แบ่งข้อมูลเป็น 2 ส่วนคือ ข้อมูลจาก UCI จำนวน 6 ชุดข้อมูล และข้อมูลนักศึกษาในหลักสูตรเทคโนโลยีบัณฑิต 1 ชุดข้อมูล คือ IT-FITM รายละเอียดของชุดข้อมูลแสดงในตารางที่ 1 โดย #Mj คือ จำนวนของคลาสลบ, #Mn คือจำนวนของคลาสบวก *IR* ย่อมาจาก Imbalanced Ratio หรืออัตราความไม่สมดุล โดยอัตราความไม่สมดุลคือ สัดส่วนของจำนวนตัวอย่างของคลาสลบหารด้วยจำนวนข้อมูลตัวอย่างของคลาสบวก

ตารางที่ 1 รายละเอียดชุดข้อมูลทั้ง 7 ชุดข้อมูล

Dataset	Instance	#Mj	# Mn	#Attr	IR
Hepatitis	155	32	123	20	3.84
Adult	32561	7841	24720	15	3.15
Pima	768	268	500	9	1.86
Monk2	169	64	105	6	1.64
Yeast	483	20	463	8	23.15
Ozone	2536	73	2463	72	3.73
IT-FITM	544	156	388	30	2.48

2.3 ขั้นตอนวิธี

งานวิจัยนี้ มีวัตถุประสงค์เพื่อพัฒนาขั้นตอนวิธีที่มีความสามารถในการจำแนกประเภทข้อมูลที่มีจำนวนตัวอย่างในแต่ละคลาสไม่สมดุลกันโดยข้อมูลที่ใช้ในการจำแนกจำเป็นต้องมีคลาสเป้าหมาย 2 คลาสเท่านั้น ผู้วิจัยได้พัฒนาขั้นตอนวิธีชื่อ Clustering Switching Method for Sampling Imbalanced Data หรือ ClusIM โดยขั้นตอนวิธีนี้ต้องสามารถเพิ่มประสิทธิภาพในการจำแนกคลาสบวกได้ดีขึ้น ซึ่งขั้นตอนวิธี ClusIM สามารถอธิบายในรูปแบบของกรอบการทำงานและรหัสเทียม (Pseudo Code) ดังรูปที่ 2 และ ตารางที่ 2 ตามลำดับ



รูปที่ 2 กรอบการทำงานของขั้นตอนวิธี ClusIM

ตารางที่ 2 รหัสเทียมของขั้นตอนวิธี ClusIM

Algorithm 1 : ClusIM

Input:
1) Given $S\{(x_1,y_1),\ldots,(x_m,y_m)\} x_i \in X$,with labels $y_i \in Y = 2$
2) M_j= the number of instances of the majority class
3) M_n= the number of instances of the minority class
4) $IR = M_j/M_n$
5) k is the number of clusters ($k=2$)

Begin:
1) $C_{1..2}$= Kmeans(S,k)
2) temp = M_nC_2
3) **for** i = 1 to **2** do
4) $T_i = M_jC_i \cup$ temp
5) if IR of $T_i > 1.5$ then $T_i = M_jC_i \cup$ SMOTE(temp) ;
6) temp = M_nC_1
7) h_i = baseclassifier(T_i)
8) **end for**

End

Output: H* = Maximum Probability Vote h_k

จากรูปที่ 2 สามารถสรุปขั้นตอนการทำงานได้ 3 ขั้นตอนดังนี้

1) การสลับเปลี่ยนกลุ่มข้อมูลเพื่อลดพื้นที่ซ้อนทับกัน (Clustering Switching Method for Reducing Overlapped Region)

ผู้วิจัยได้ประยุกต์ใช้เทคนิคการจัดกลุ่มข้อมูลมาทำการแยกข้อมูลที่ซ้อนทับ (Class overlapping) กันภายใต้แนวคิดการลดจำนวนข้อมูลตัวอย่างของคลาสลบ โดยใช้หลักการการจัดกลุ่มข้อมูลด้วยขั้นตอนวิธี k-means ซึ่งมีขั้นตอนการทำงานดังนี้

สำหรับทุกข้อมูลตัวอย่าง (Instances) จะถูกแบ่งกลุ่มด้วยขั้นตอนวิธี k-means โดยกำหนดค่า *k = 2 (เท่ากับจำนวนของคลาส)* ดังนั้น ผลลัพธ์ในขั้นตอนนี้จะได้จำนวนของชุดข้อมูลฝึก 2 ชุด และเลือกใช้ Euclidean distance เป็นมาตรวัดค่าความเหมือนระหว่างข้อมูลแต่ละเรคคอร์ด หลังจากผ่านขั้นตอนการจัดกลุ่มด้วย k-means แล้วจะได้กลุ่มของข้อมูล 2 กลุ่มคือ C_1, และ C_2 โดย C_1 แทนคลัสเตอร์ 1 และ C_2 แทนคลัสเตอร์ 2

พิจารณาข้อมูลตัวอย่างในแต่ละคลาสและแต่ละคลัสเตอร์ ขั้นตอนวิธี ClusIM จะสร้างชุดข้อมูลย่อยตามจำนวนคลัสเตอร์ที่กำหนด ภายใต้สมมติฐานที่ว่า **"ถ้าคุณสมบัติของข้อมูลตัวอย่างของแต่ละคลาสในชุดข้อมูลฝึกมีความแตกต่างกันและเมื่อนำไปจำแนกด้วยตัวจำแนก จะส่งผลให้ตัวจำแนกสามารถจำแนกข้อมูลได้มีประสิทธิภาพเพิ่มขึ้น"** ดังนั้น ข้อมูลตัวอย่างของคลาสลบ ในคลัสเตอร์หนึ่ง (C_1) จะถูกสลับกลุ่มโดยจะถูกนำไปจัดกลุ่มกับข้อมูลตัวอย่างของคลาสบวกในคลัสเตอร์สอง (C_2) เช่นเดียวกับข้อมูลตัวอย่างของคลาสลบในคลัสเตอร์สอง (C_2) จะถูกสลับกลุ่มโดยจะถูกนำไปจัดกลุ่มกับข้อมูลตัวอย่างของคลาสบวกในคลัสเตอร์หนึ่ง (C_1) ตัวอย่างรูปแบบการจับกลุ่มสำหรับข้อมูลย่อยทั้ง 2 ชุดมีดังต่อไปนี้

a) ข้อมูลย่อยชุดที่ 1 เกิดจากข้อมูลตัวอย่างของคลาสลบในคลัสเตอร์ที่ 1 (M_jC_1) จะถูกนำมารวมกับข้อมูลตัวอย่างของคลาสบวกในคลัสเตอร์ที่ 2 (M_nC_2).

b) ข้อมูลย่อยชุดที่ 2 เกิดจากข้อมูลตัวอย่างของคลาสลบในคลัสเตอร์ที่ 2 (M_jC_2) จะถูกนำมารวมกับข้อมูลตัวอย่างของคลาสบวกในคลัสเตอร์ที่ 1 (M_nC_1)

ดังนั้น ผลลัพธ์ที่ได้จากขั้นตอนนี้คือ การสร้างชุดข้อมูลย่อยทั้งหมด 2 ชุดเพื่อเป็นการลดข้อมูลตัวอย่างของคลาสลบให้มีจำนวนน้อยลงโดยสามารถเลี่ยงปัญหาการสูญเสียข้อมูลตัวอย่างของคลาสลบที่สำคัญ (Information loss problem) ที่เกิดจากการลบข้อมูลตัวอย่างของคลาสลบในระหว่างการเรียนรู้ อีกทั้งมีการสลับเปลี่ยนกลุ่มข้อมูลระหว่างคลาสบวกกับคลาสลบเพื่อลดพื้นที่ซ้อนทับกันของทั้งสองคลาส

ตารางที่ 3 แสดงตัวอย่างวิธีการทำงานของขั้นตอนนี้บนชุดข้อมูล Pima

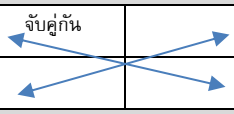
ข้อมูลตัวอย่าง Pima จากเดิมมีทั้งหมด 768 ตัวอย่าง มีคลาสบวก 268 ตัวอย่าง และคลาส ลบ 500 ตัวอย่าง เมื่อผ่านกระบวนการจัดกลุ่มข้อมูลแล้วจะได้ชุดข้อมูลตัวอย่างย่อย 2 ชุด

2) ขั้นตอนการสังเคราะห์ข้อมูลตัวอย่างในคลาสบวกด้วยเทคนิคการสุ่มเพิ่ม

สำหรับขั้นตอนนี้จะเป็นเทคนิคการเพิ่มข้อมูลตัวอย่างของคลาสบวกด้วยวิธีการสร้างตัวอย่างสังเคราะห์ของคลาสบวกด้วยขั้นตอนวิธี SMOTE ซึ่งข้อดีของขั้นตอนวิธีนี้คือจะสร้างตัวอย่างสังเคราะห์ของคลาสบวกโดยการอาศัยการ

ตารางที่ 3 แสดงตัวอย่างการสลับเปลี่ยนกลุ่มข้อมูลบนชุดข้อมูล Pima

ชุดข้อมูลย่อย	จำนวนตัวอย่างคลาสบวก	จำนวนตัวอย่างคลาสลบ
หลังการจัดกลุ่มด้วย k-means		
กลุ่มที่ 1	135	380
กลุ่มที่ 2	133	120
หลังดำเนินการสลับเปลี่ยนข้อมูลตัวอย่างระหว่างกลุ่ม		
ชุดข้อมูลที่ 1	135	120
ชุดข้อมูลที่ 2	133	380



ตารางที่ 4 แสดงตัวอย่างเพิ่มข้อมูลตัวอย่างของคลาสบวกบนชุดข้อมูล Pima

ชุดข้อมูลย่อย	จำนวนตัวอย่างคลาสบวก	จำนวนตัวอย่างคลาสลบ	หมายเหตุ
ชุดข้อมูลหลังดำเนินการสลับเปลี่ยนข้อมูลตัวอย่างระหว่างกลุ่ม			
ชุดข้อมูลที่ 1	135	120	IR = 0.88
ชุดข้อมูลที่ 2	133	380	**IR = 2.86**
พิจารณา IR ในแต่ละชุดข้อมูล ถ้า IR>1.5 ข้อมูลในคลาสบวกจะถูกสร้างด้วย SMOTE			
ชุดข้อมูลที่ 1	135	120	
ชุดข้อมูลที่ 2	**253**	380	คลาสบวกถูกเพิ่มด้วยวิธีการ SMOTE

วัดระยะห่างระหว่างข้อมูลตัวอย่างของคลาสบวกกับเพื่อนบ้านใกล้เคียงที่เป็นคลาสบวกจำนวน k ตัว แล้วสร้างตัวอย่างสังเคราะห์ของคลาสบวกที่อยู่ระหว่างระยะห่างนั้น จึงทำให้วิธีการนี้ไม่ได้สร้างตัวอย่างจากการทำซ้ำ แต่เป็นการสังเคราะห์ข้อมูลใหม่จากข้อมูลเดิม นอกจากนี้วิธีการนี้ยังบรรเทาปัญหา overfitting ที่อาจจะเกิดขึ้นจากการเพิ่มข้อมูลตัวอย่างของคลาสบวก ซึ่งรายละเอียดการทำงานของขั้นตอนนี้ มีดังนี้

นำขั้นตอนวิธี SMOTE มาสร้างข้อมูลตัวอย่างของคลาสบวก โดยพิจารณาจำนวนข้อมูลตัวอย่างที่สร้างจาก Imbalance ratio (IR) ซึ่ง IR หมายถึงสัดส่วนของจำนวนตัวอย่างของคลาสลบหารด้วยจำนวนข้อมูลตัวอย่างของคลาสบวก ดังนั้นในงานวิจัยนี้จำนวนตัวอย่างของคลาสบวกจะถูกสร้างจนกระทั่ง IR เท่ากับ 1.5

ตารางที่ 4 แสดงตัวอย่างวิธีการทำงานของขั้นตอนนี้บนชุดข้อมูล Pima

- หลังจากดำเนินการสลับเปลี่ยนข้อมูลตัวอย่างระหว่างกลุ่มข้อมูลแล้ว จะพิจารณาจำนวนข้อมูลตัวอย่างระหว่างคลาสของชุดข้อมูล Pima เพื่อเพิ่มข้อมูลตัวอย่างของคลาสบวกในแต่ละชุดข้อมูล

ดังนั้น จากวิธีการข้างต้นผลลัพธ์สุดท้ายของขั้นตอนนี้คือ ได้ชุดข้อมูลฝึกทั้งหมด 2 ชุด โดยในแต่ละชุดข้อมูลฝึกจะประกอบไปด้วยคลาสทั้ง

ตารางที่ 5 ขั้นตอนวิธีอื่น ๆ ที่นำมาใช้ในการเปรียบเทียบกับ ClusIM

ขั้นตอนวิธีที่นำมาเปรียบเทียบ	ชื่อย่อ
Support Vector Machine	SVM
Bagging with Support Vector Machine	BSVM
AdaboostM1 with Support Vector Machine	AbSVM
AdaCost with Support Vector Machine	AcSVM
SMOTE Bagging with Support Vector Machine	B_SM
SMOTE AdabootM1 with Support Vector Machine	Ab_SM
SMOTE Adacost with Support Vector Machine	Ac_SM

สองคลาส นอกจากนี้ในแต่ละชุดข้อมูลฝึกนั้นจะมีคุณสมบัติที่สำคัญคือ ข้อมูลตัวอย่างของคลาสทั้งสองจะมีคุณลักษณะที่แตกต่างกันอย่างชัดเจน ซึ่งสามารถบรรเทาปัญหาการซ้อนทับกันของข้อมูลตัวอย่างระหว่างคลาสได้

3) ขั้นตอนการจำแนกประเภท (Classification)

วัตถุประสงค์ของขั้นตอนนี้คือ สร้างตัวแบบเพื่อใช้ในการจำแนกประเภทข้อมูลที่มีจำนวนข้อมูลตัวอย่างในแต่ละคลาสไม่สมดุลกัน โดยนำชุดข้อมูลฝึกแต่ละชุดที่มีความซ้อนทับกันของข้อมูลน้อยจากการจัดกลุ่มข้อมูลมาเรียนรู้ด้วยขั้นตอนวิธีซัพพอร์ตเวกเตอร์แมชชีน หลังจากผ่านกระบวนการเรียนรู้แล้วจะทำให้ได้องค์ความรู้หรือตัวแบบ 2 ตัว (h_1 และ h_2) เพื่อนำไปทำนายผลภายใต้การเลือกค่าความน่าจะเป็นที่สูงที่สุด (Maximum Probability Vote)

2.4 การทดสอบความถูกต้อง และความน่าเชื่อถือของตัวแบบในการจำแนกประเภท

ในขั้นตอนนี้เป็นการนำผลลัพธ์ที่ได้จากการเรียนรู้มาทดสอบความถูกต้องและความน่าเชื่อถือของตัวแบบนั้น ๆ เพื่อให้ได้มาซึ่งขั้นตอนวิธีที่มีความสามารถในการเรียนรู้ข้อมูลในคลาสบวกได้ดีและไม่กระทบต่อประสิทธิภาพของการเรียนรู้ข้อมูลในคลาสลบ ซึ่งในขั้นตอนการทดลองจะใช้เครื่องมือ Weka 3.6.1 ควบคู่กับ Netbean IDE 6.7 โดยเลือกรูปแบบการทดสอบความถูกต้องแบบ 10-fold cross validation ซึ่งเป็นการแบ่งกลุ่มข้อมูลเพื่อนำมาใช้ในการทดสอบ โดยกำหนดจำนวนของกลุ่มข้อมูลทั้งสิ้น 10 รูปแบบ

สำหรับการวัดประสิทธิภาพจะเลือกใช้ตัววัดสำหรับปัญหาชุดข้อมูลที่มีจำนวนตัวอย่างในแต่ละคลาสไม่สมดุลกัน โดยใช้การวัดผลจากตาราง Confusion matrix สำหรับการจำแนกคลาส และวัดประสิทธิภาพด้วยค่า F-measure และ G-mean และเพื่อเป็นการวัดประสิทธิภาพของขั้นตอนวิธี ClusIM ผู้วิจัยได้เลือกขั้นตอนวิธีที่หลากหลายมาเปรียบเทียบกับขั้นตอนวิธีที่นำเสนอแสดงดังตารางที่ 5

ซึ่งขั้นตอนวิธี Bagging [2] เป็นขั้นตอนวิธีที่ถูกเลือกนำมาเปรียบเทียบเนื่องจากเป็นวิธีการที่สุ่มสร้างชุดข้อมูลฝึกหลายๆ ชุดและนำข้อมูลหลายๆ ชุดนั้นมาจำแนกด้วยขั้นตอนวิธีเดียวกันทั้งหมดตามจำนวนรอบที่ผู้ใช้กำหนด

สำหรับ AdaboostM1 [8] เป็นขั้นตอนวิธีที่ถูกพัฒนาขึ้นโดยสร้างชุดข้อมูลฝึกหลายๆ ชุด

เช่นกันกับ Bagging แต่มีการปรับค่าน้ำหนักของตัวอย่างในแต่ละรอบของการเรียนรู้ และเป็นขั้นตอนวิธีที่ถูกนำมาใช้กับการจำแนกประเภทข้อมูลที่มีจำนวนตัวอย่างในแต่ละคลาสไม่สมดุลกันในหลาย ๆ งานวิจัย

สำหรับ AdaCost [6] เป็นขั้นตอนวิธีที่พัฒนามาจาก Adaboost และอาศัยหลักการของ cost-sensitive boosting โดยมีการปรับค่า cost ให้กับตัวอย่างแต่ละตัวอย่างโดยเฉพาะอย่างยิ่งตัวอย่างที่มีการจำแนกผิด ซึ่งเมื่อนำมาจำแนกกับชุดข้อมูลตัวอย่างที่ไม่สมดุลกัน ข้อมูลตัวอย่างในคลาสบวกจะถูกจำแนกผิดเป็นจำนวนมาก ขั้นตอนวิธีนี้จะมีปรับค่าน้ำหนักกับข้อมูลตัวอย่างเหล่านี้เพื่อเพิ่มประสิทธิภาพการจำแนกให้ดีขึ้น

นอกจากนี้ผู้วิจัยได้ทำการทดลองโดยเปรียบเทียบประสิทธิภาพของขั้นตอนวิธี ClusIM กับ 3 ขั้นตอนวิธีข้างต้นมาใช้กับวิธีการ SMOTE เพื่อดำเนินการปรับชุดข้อมูลตัวอย่างของคลาสบวกให้มีความสมดุลกับคลาสลบแล้วนำมาจำแนกด้วยขั้นตอนวิธี Bagging, AdaboostM1 และ AdaCost โดยตั้งชื่อวิธีการใหม่นี้ว่า B_SM, Ab_SM, และ Ac_SM

3. ผลการวิจัยและวิจารณ์ผลการวิจัย

ในส่วนของผลการทดลอง ผู้วิจัยได้แบ่งเป็น 2 ส่วนคือ 1) การวิเคราะห์ค่าประสิทธิภาพของ ClusIM 2) การวิเคราะห์ผลของ ClusIM

3.1 การวิเคราะห์ค่าประสิทธิภาพของขั้นตอนวิธี ClusIM

จากชุดข้อมูลในตารางที่ 1 จะถูกนำมาทดสอบเพื่อดำเนินการวัดประสิทธิภาพในการจำแนกประเภทข้อมูลด้วยขั้นตอนวิธี ClusIM กับขั้นตอนวิธีอื่น ๆ ที่เปรียบเทียบตามตารางที่ 5 โดยจะใช้มาตรวัดที่มีความเหมาะสมกับการจำแนกชุดข้อมูลตัวอย่างที่มีจำนวนตัวอย่างในแต่ละคลาสไม่สมดุลกันได้แก่ F-measure และ G-mean สำหรับการเปรียบเทียบประสิทธิภาพในการจำแนกด้วยมาตรวัด F-measure จะแสดงดังตารางที่ 6 และการเปรียบเทียบประสิทธิภาพในการจำแนกด้วยมาตรวัด G-mean จะแสดงในตารางที่ 7 ตามลำดับ

จากตารางที่ 6 พบว่าทุกชุดข้อมูล ค่า F-measure ของวิธีการที่นำเสนอได้ผลดีกว่าวิธีการอื่นๆ ที่เปรียบเทียบ และเมื่อพิจารณาทีละชุดข้อมูลสามารถวิเคราะห์ผลการทดลองได้ดังต่อไปนี้

ชุดข้อมูลที่มีอัตราความไม่สมดุลต่ำ (Low Imbalanced Ratio) นั่นคือชุดข้อมูล Pima และ Monk2 พบว่า ขั้นตอนวิธี ClusIM มีค่า F-measure สูงกว่าขั้นตอนวิธีอื่นๆ อย่างมาก บนชุดข้อมูล Pima ขั้นตอนวิธี SVM มีค่า F-measure เท่ากับ 0.589 ในขณะที่ ClusIM มีค่า F-measure เท่ากับ 0.968 ซึ่งสูงกว่า SVM ถึง 0.382 ในขณะที่ชุดข้อมูล Monk2 นั้นขั้นตอนวิธีอื่นๆ ที่นำมาเปรียบเทียบมีค่าต่ำกว่า ClusIM อย่างมาก โดยเฉลี่ยแล้วขั้นตอนวิธีอื่นๆ มีค่าประสิทธิภาพเท่ากับ 0.55 ในขณะที่ ClusIM ให้ค่าประสิทธิภาพ F-measure เท่ากับ 0.962

ในส่วนของชุดข้อมูลที่มีอัตราความไม่สมดุลสูง (High Imbalanced Ratio) นั่นคือชุดข้อมูล Yeast และ Ozone พบว่าโดยรวมขั้นตอนวิธี ClusIM ให้ค่า F-measure ที่สูงกว่าขั้นตอนวิธีอื่น ๆ แต่อย่างไรก็ตามค่าที่สูงกว่านั้นเป็นค่าที่สูงกว่าเพียงเล็กน้อย ตัวอย่างเช่น บนชุดข้อมูล Yeast ขั้นตอนวิธี SVM ให้ค่า F-measure เท่ากับ 0.977 ส่วน ClusIM ให้ค่า F-measure เท่ากับ 0.979 สำหรับชุดข้อมูล Ozone ขั้นตอนวิธี SVM ให้ค่า F-measure เท่ากับ 0.937 ในขณะที่ ClusIM ให้ค่า F-measure เท่ากับ 0.991

ตารางที่ 6 แสดงประสิทธิภาพในการจำแนกด้วยมาตรวัด F-measure

ขั้นตอนวิธี / ชุดข้อมูล	SVM	BSVM	AbSVM	AcSVM	B_SM	Ab_SM	Ac_SM	ClusIM
Hepatitis	0.849	0.899	0.799	0.849	0.891	0.882	0.887	**0.967**
Adult	0.904	0.842	0.842	0.842	0.853	0.841	0.845	**0.991**
Pima	0.586	0.756	0.875	0.763	0.743	0.739	0.738	**0.968**
Monk2	0.763	0.527	0.529	0.493	0.510	0.533	0.513	**0.962**
Yeast	0.977	0.977	0.977	0.977	0.977	0.977	0.786	**0.979**
Ozone	0.937	0.906	0.875	0.906	0.872	0.875	0.874	**0.991**
IT-FITM	0.779	0.759	0.779	0.779	0.786	0.779	0.782	**0.970**

ตารางที่ 7 แสดงประสิทธิภาพในการจำแนกด้วยมาตรวัด G-mean

ขั้นตอนวิธี / ชุดข้อมูล	SVM	BSVM	AbSVM	AcSVM	B_SM	Ab_SM	Ac_SM	ClusIM
Hepatitis	0.624	0.899	0.900	0.849	0.892	0.882	0.887	**0.968**
Adult	0.731	0.846	0.845	0.845	0.851	0.832	0.842	**0.991**
Pima	0.771	0.763	0.876	0.771	0.743	0.740	0.739	**0.968**
Monk2	0.550	0.548	0.533	0.535	0.514	0.532	0.513	**0.963**
Yeast	0.978	0.978	0.978	0.978	0.978	0.978	0.824	**0.979**
Ozone	0.907	0.907	0.876	0.907	0.873	0.876	0.876	**0.991**
IT-FITM	0.794	0.762	0.794	0.794	0.824	0.794	0.781	**0.970**

จากตารางที่ 7 เป็นการแสดงค่าประสิทธิภาพในการจำแนกของขั้นตอนวิธี ClusIM กับขั้นตอนวิธีอื่น ๆ ที่เปรียบเทียบด้วยมาตรวัด G-mean ผลการทดลองพบว่าขั้นตอนวิธี ClusIM ให้ค่า G-mean สูงกว่าขั้นตอนวิธีอื่นๆ ที่เปรียบเทียบในทุก ๆ ชุดข้อมูล โดยสามารถวิเคราะห์การทดลองได้ดังต่อไปนี้

บนชุดข้อมูล Monk2 ขั้นตอนวิธี SVM และขั้นตอนวิธีอื่น ๆ ให้ค่า G-mean โดยเฉลี่ยประมาณ 0.53 ในขณะที่ขั้นตอนวิธี ClusIM ให้ค่า G-mean เท่ากับ 0.963 และเมื่อพิจารณาการนำเทคนิคการเพิ่มข้อมูลคลาสบวกด้วย SMOTE มาใช้ก่อนดำเนินการจำแนกด้วยขั้นตอนวิธี Bagging, AdaBoostM1, และ AdaCost ในทุกชุดข้อมูล พบว่าการเพิ่มข้อมูลด้วย SMOTE กลับไม่ช่วยเพิ่มประสิทธิภาพในการจำแนกในทางตรงกันข้ามกลับลดประสิทธิภาพในการจำแนกให้น้อยลงกว่าตัวจำแนกแบบมาตรฐาน

นอกจากนี้เมื่อนำเทคนิค Ensemble ประกอบไปด้วย Bagging, AdaboostM1, Adacost ซึ่งเป็นการรวมของโมเดลการเรียนรู้ที่หลากหลาย มีความแตกต่างและมีอิสระต่อกันเข้าด้วยกันเพื่อเพิ่ม

ตารางที่ 8 แสดงประสิทธิภาพในการจำแนกคลาสบวก (Minority Class) ด้วยมาตรวัด F-measure

ขั้นตอนวิธี / ชุดข้อมูล	SVM	BSVM	AbSVM	AcSVM	B_SM	Ab_SM	Ac_SM	ClusIM
Hepatitis	0.623	0.876	0.483	0.623	0.898	0.878	0.886	**0.984**
Adult	0.713	0.647	0.645	0.645	0.732	0.721	0.725	**0.991**
Pima	0.625	0.606	0.425	0.625	0.728	0.722	0.721	**0.968**
Monk2	0.103	0.234	0.319	0.103	0.533	0.520	0.510	**0.959**
Yeast	0.687	0.687	0.687	0.687	0.687	0.687	0.734	**0.974**
Ozone	0.000	0.000	0.878	0.000	0.875	0.878	0.877	**0.989**
IT-FITM	0.556	0.547	0.556	0.556	0.734	0.556	0.760	**0.967**

ตารางที่ 9 แสดงประสิทธิภาพในการจำแนกคลาสบวก (Minority Class) ด้วยมาตรวัด G-mean

ขั้นตอนวิธี / ชุดข้อมูล	SVM	BSVM	AbSVM	AcSVM	B_SM	Ab_SM	Ac_SM	ClusIM
Hepatitis	0.624	0.899	0.485	0.624	0.892	0.878	0.887	**0.971**
Adult	0.695	0.652	0.651	0.633	0.715	0.723	0.720	**0.991**
Pima	0.633	0.617	0.872	0.633	0.729	0.723	0.722	**0.968**
Monk2	0.134	0.251	0.321	0.134	0.535	0.520	0.510	**0.960**
Yeast	0.710	0.710	0.710	0.710	0.710	0.710	0.761	**0.974**
Ozone	0.000	0.000	0.878	0.000	0.875	0.878	0.878	**0.989**
IT-FITM	0.576	0.551	0.576	0.576	0.577	0.576	0.760	**0.966**

ประสิทธิภาพของโมเดลนั้น พบว่า Ensemble สามารถเพิ่มประสิทธิภาพในการจำแนกได้สูงขึ้นกว่าการใช้ขั้นตอนวิธี SVM พื้นฐานในบางชุดข้อมูลเท่านั้น สำหรับชุดข้อมูลที่เทคนิค Ensemble สามารถช่วยเพิ่มประสิทธิภาพจากขั้นตอนวิธีพื้นฐานได้แก่ Hepatitis, Adult และ IT-FITM ชุดข้อมูลที่เหลือเทคนิค Ensemble จะให้ค่าประสิทธิภาพได้เทียบเท่าหรือต่ำกว่าขั้นตอนวิธี SVM พื้นฐาน

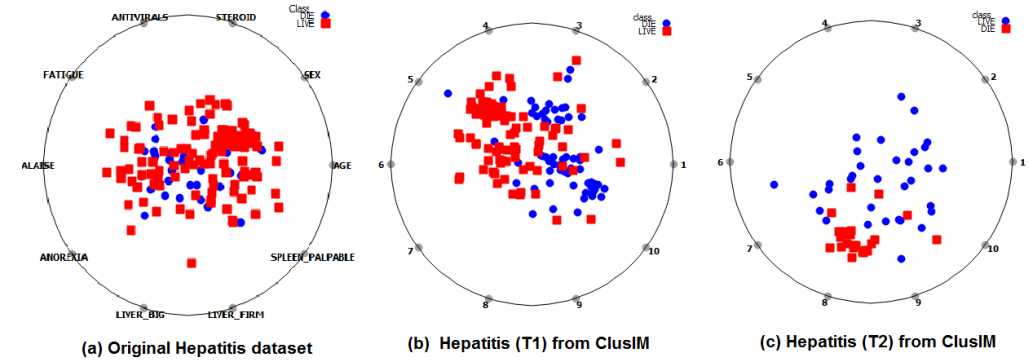
และเพื่อเป็นการแสดงให้เห็นถึงการเพิ่มประสิทธิภาพในการจำแนกคลาสบวกของขั้นตอนวิธี ClusIM ดังนั้น ผู้วิจัยจึงได้แสดงค่าประสิทธิภาพในการจำแนกเฉพาะคลาสบวก ทั้งในเชิงมาตรวัด F-measure และ G-mean โดยแสดงในตารางที่ 8 และ 9 ตามลำดับ จากตารางที่ 8 แสดงให้เห็นว่าขั้นตอนวิธี ClusIM สามารถเพิ่มประสิทธิภาพในการจำแนกของคลาสบวกด้วยค่า F-measure ได้ดีกว่าขั้นตอนวิธีอื่นๆ ที่เปรียบเทียบได้ในทุก ๆ ชุดข้อมูล โดยเฉพาะอย่างยิ่งในชุดข้อมูล Monk2 และ

ชุดข้อมูล Ozone ซึ่งจากผลลัพธ์ที่ได้แสดงให้เห็นว่า ขั้นตอนวิธี SVM BSVM และ AcSVM ไม่สามารถจำแนกข้อมูลในคลาสบวกได้ถูกต้องเลยทำให้ค่า F-measure มีเท่ากับ 0.000 ในขณะที่ AbSVM ซึ่งเป็นขั้นตอนวิธีที่นำมาใช้ในการแก้ไขปัญหาข้อมูลตัวอย่างในแต่ละคลาสไม่สมดุลกัน สามารถจำแนกข้อมูลคลาสบวกได้ถูกต้องที่ค่า F-measure เท่ากับ 0.878 แต่อย่างไรก็ตามขั้นตอนวิธี ClusIM ก็ให้ค่าประสิทธิภาพในการจำแนกคลาสบวกที่สูงที่สุดโดยมีค่า F-measure เท่ากับ 0.9689 สำหรับชุดข้อมูล Monk2 นั้น ขั้นตอนวิธี SVM, BSVM, AbSVM และ AcSVM ให้ค่า F-measure ของคลาสบวกเท่ากับ 0.103, 0.234, 0.319 และ 0.103 ตามลำดับ และหากนำวิธีการ SMOTE มาใช้กับชุดข้อมูล Monk2 แล้วนำไปจำแนกด้วย B_SM, Ab_SM, Ac_SM จะสามารถจำแนกได้ถูกต้องเพียงร้อยละ 50 เท่านั้น ในทางตรงกันข้ามขั้นตอนวิธี ClusIM สามารถจำแนกคลาสบวกในชุดข้อมูล Monk2 ได้ถูกต้องถึงร้อยละ 95

จากตารางที่ 9 เป็นการเปรียบเทียบประสิทธิภาพในการจำแนกคลาสบวกด้วยมาตรวัด G-mean ซึ่งในชุดข้อมูล Monk2 และ Ozone ยังเป็นชุดข้อมูลที่ขั้นตอนวิธีอื่นๆ ที่นำมาเปรียบเทียบจำแนกคลาสบวกได้ถูกต้องน้อยมากในชุดข้อมูล Ozone มี 3 ขั้นตอนวิธีที่ไม่สามารถจำแนกข้อมูลในคลาสบวกได้ถูกต้องเลย ในทางตรงกันข้ามขั้นตอนวิธี ClusIM สามารถจำแนกข้อมูลในคลาสบวกได้ถูกต้องมากที่สุด และในทุกชุดข้อมูลค่าประสิทธิภาพในการจำแนกข้อมูลในคลาสบวกของ ClusIM มีค่า G-mean มากกว่าร้อยละ 90 นั่นแสดงให้เห็นว่าขั้นตอนวิธี ClusIM สามารถเพิ่มประสิทธิภาพในการจำแนกข้อมูลในคลาสบวกในทุกชุดข้อมูลที่เปรียบเทียบได้ตรงตามวัตถุประสงค์ของงานวิจัย

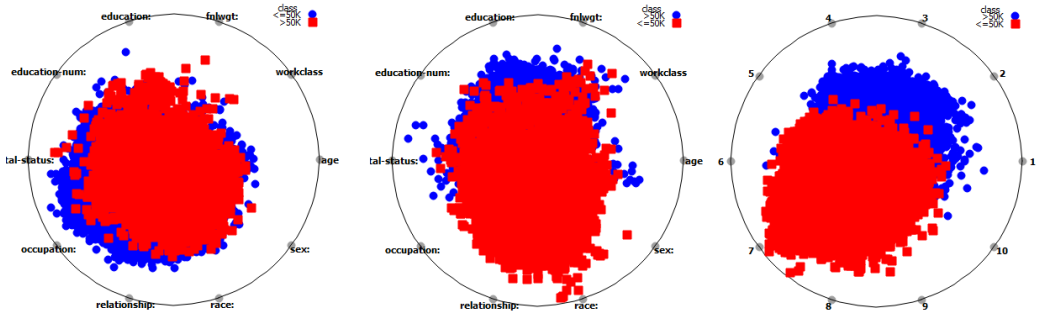
3.2 การวิเคราะห์ผลของ ClusIM

จากวัตถุประสงค์ในการวิจัยนี้คือต้องการพัฒนาวิธีการที่สามารถเพิ่มประสิทธิภาพในการจำแนกข้อมูลที่มีจำนวนตัวอย่างในแต่ละคลาสไม่สมดุลกันโดยวิธีการในระดับข้อมูลเพื่อปรับข้อมูลตัวอย่างโดยใช้เทคนิคการปรับตัวอย่างแบบผสมผสาน และเพื่อแสดงให้เห็นถึงประสิทธิภาพของขั้นตอนวิธี ClusIM ในการแก้ไขปัญหาระดับข้อมูลบนชุดข้อมูลตัวอย่างที่มีจำนวนแต่ละคลาสไม่สมดุลกัน ผู้วิจัยได้ดำเนินการแสดงลักษณะการกระจายตัวของข้อมูลในทุกชุดข้อมูลหลังผ่านกระบวนการสลับเปลี่ยนกลุ่ม (Clustering Switching Method) และการเพิ่มข้อมูลตัวอย่าง ซึ่งเมื่อผ่านขั้นตอนนี้แล้ว ClusIM จะสร้างชุดข้อมูลใหม่ขึ้นมา 2 ชุด ที่สามารถลดความซ้อนทับกันของข้อมูลระหว่างคลาสบวกและคลาสลบลงได้พร้อมสร้างตัวอย่างสังเคราะห์ของคลาสบวกในแต่ละชุดข้อมูลให้มีความสมดุลกับคลาสลบ ลักษณะการกระจายตัวของชุดข้อมูลทั้งหมดที่ใช้ในการวิจัยโดยเปรียบเทียบระหว่างชุดข้อมูลต้นฉบับกับชุดข้อมูลจาก ClusIM แสดงดังรูปที่ 3 ถึงรูปที่ 9



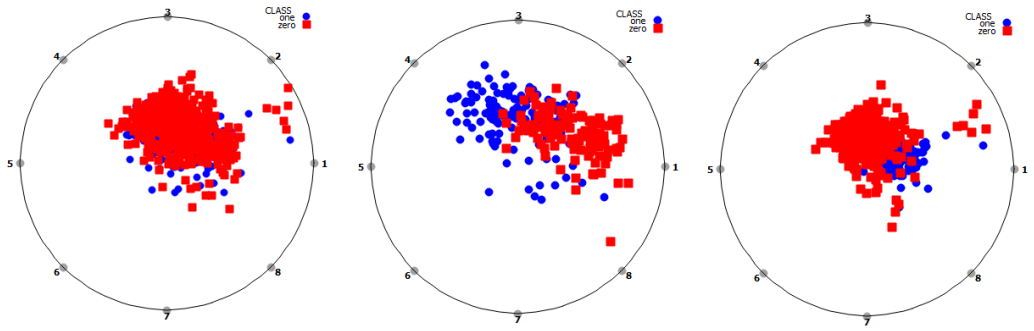
(a) Original Hepatitis dataset (b) Hepatitis (T1) from ClusIM (c) Hepatitis (T2) from ClusIM

รูปที่ 3 (a) ลักษณะการกระจายตัวของชุดข้อมูลต้นฉบับ Hepatitis (b) ลักษณะการกระจายตัวของชุดข้อมูล Hepatitis ในชุดที่ 1 หลังผ่านขั้นตอนวิธี ClusIM (c) ลักษณะการกระจายตัวของชุดข้อมูล Hepatitis ในชุดที่ 2 หลังผ่านขั้นตอนวิธี ClusIM



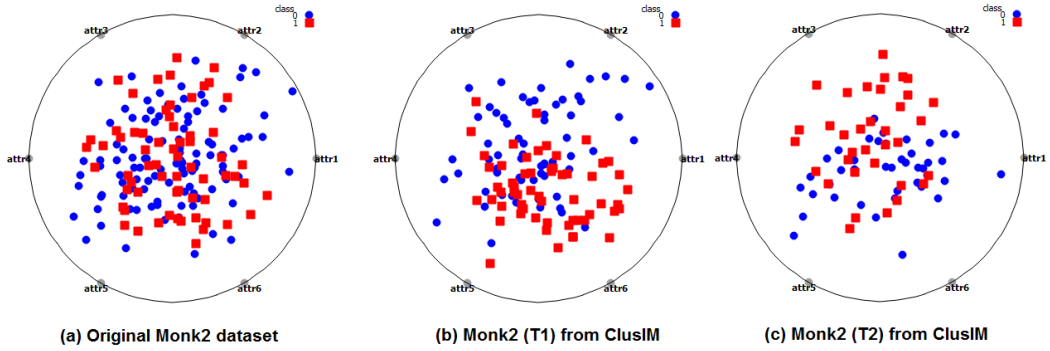
(a) Original Adult dataset (b) Adult (T1) from ClusIM (c) Adult (T2) from ClusIM

รูปที่ 4 (a) ลักษณะการกระจายตัวของชุดข้อมูลต้นฉบับ Adult (b) ลักษณะการกระจายตัวของชุดข้อมูล Adult ในชุดที่ 1 หลังผ่านขั้นตอนวิธี ClusIM (c) ลักษณะการกระจายตัวของชุดข้อมูล Adult ในชุดที่ 2 หลังผ่านขั้นตอนวิธี ClusIM

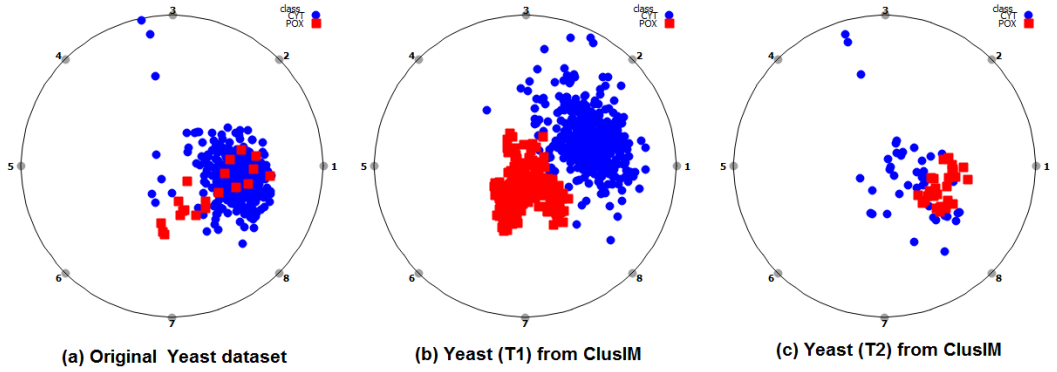


(a) Original Pima dataset (b) Pima (T1) from ClusIM (c) Pima (T2) from ClusIM

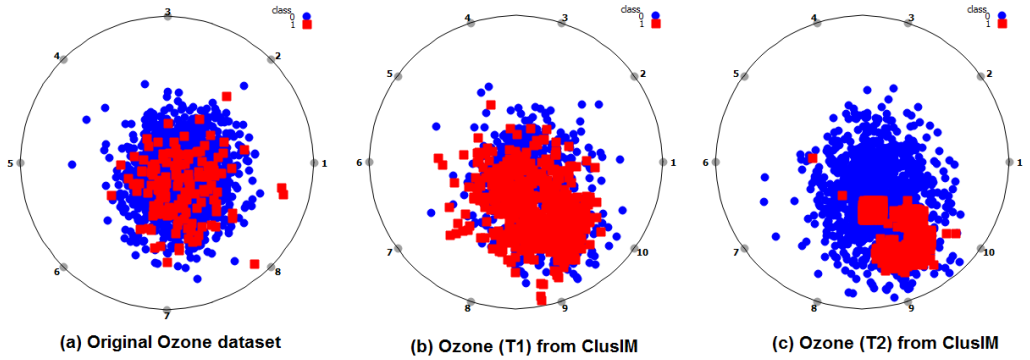
รูปที่ 5 (a) ลักษณะการกระจายตัวของชุดข้อมูลต้นฉบับ Pima (b) ลักษณะการกระจายตัวของชุดข้อมูล Pima ในชุดที่ 1 หลังผ่านขั้นตอนวิธี ClusIM (c) ลักษณะการกระจายตัวของชุดข้อมูล Pima ในชุดที่ 2 หลังผ่านขั้นตอนวิธี ClusIM



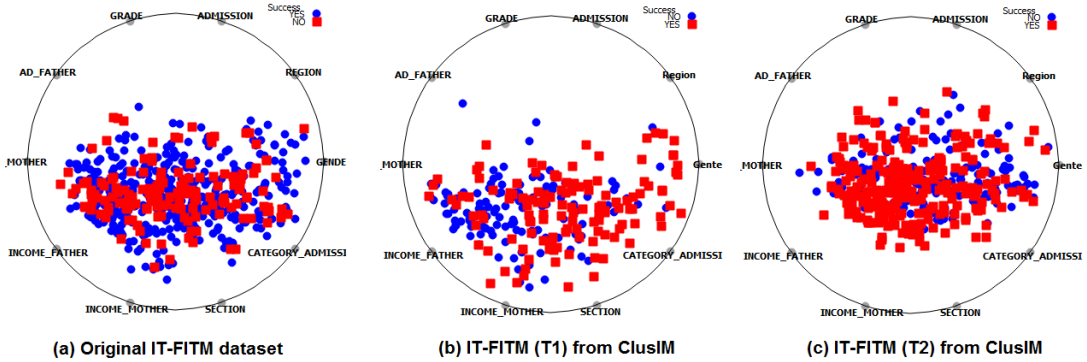
รูปที่ 6 (a) ลักษณะการกระจายตัวของชุดข้อมูลต้นฉบับ Monk2 (b) ลักษณะการกระจายตัวของชุดข้อมูล Monk2 ในชุดที่ 1 หลังผ่านขั้นตอนวิธี ClusIM (c) ลักษณะการกระจายตัวของชุดข้อมูล Monk2 ในชุดที่ 2 หลังผ่านขั้นตอนวิธี ClusIM



รูปที่ 7 (a) ลักษณะการกระจายตัวของชุดข้อมูลต้นฉบับ Yeast (b) ลักษณะการกระจายตัวของชุดข้อมูล Yeast ในชุดที่ 1 หลังผ่านขั้นตอนวิธี ClusIM (c) ลักษณะการกระจายตัวของชุดข้อมูล Yeast ในชุดที่ 2 หลังผ่านขั้นตอนวิธี ClusIM



รูปที่ 8 (a) ลักษณะการกระจายตัวของชุดข้อมูลต้นฉบับ Ozone (b) ลักษณะการกระจายตัวของชุดข้อมูล Ozone ในชุดที่ 1 หลังผ่านขั้นตอนวิธี ClusIM (c) ลักษณะการกระจายตัวของชุดข้อมูล Ozone ในชุดที่ 2 หลังผ่านขั้นตอนวิธี ClusIM



รูปที่ 9 (a) ลักษณะการกระจายตัวของชุดข้อมูลต้นฉบับ IT-FITM (b) ลักษณะการกระจายตัวของชุดข้อมูล IT-FITM ในชุดที่ 1 หลังผ่านขั้นตอนวิธี ClusIM (c) ลักษณะการกระจายตัวของชุดข้อมูล IT-FITM ในชุดที่ 2 หลังผ่านขั้นตอนวิธี ClusIM

จากรูปที่ 3 ถึง รูปที่ 9 แสดงการเปรียบเทียบชุดข้อมูลต้นฉบับกับชุดข้อมูลที่สร้างจากขั้นตอนวิธี ClusIM ผลที่ได้จากการทำงานของขั้นตอนวิธี ClusIM จะเห็นว่า ชุดข้อมูลที่สร้างขึ้นใหม่นี้มีการกระจายตัวของคลาสบวกและคลาสลบมีการซ้อนทับกันน้อยลงจากชุดข้อมูลต้นฉบับอย่างชัดเจน โดยเฉพาะอย่างยิ่งในชุดข้อมูล Pima , Monk2 และ Yeast ซึ่งจากการที่ชุดข้อมูลมีการซ้อนทับกันระหว่างคลาสที่ลดลงนี้ส่งผลให้ขั้นตอนวิธี ClusIM สามารถจำแนกข้อมูลในแต่ละคลาสได้ถูกต้องมากยิ่งขึ้น นอกจากการลดการซ้อนทับกันระหว่างคลาสแล้วชุดข้อมูลที่สร้างขึ้นมาใหม่ทั้งสองชุดนั้นยังมีจำนวนตัวอย่างในแต่ละคลาสที่สมดุลกันเพิ่มมากขึ้นจากชุดข้อมูลต้นฉบับ เนื่องจาก ClusIM ได้สร้างตัวอย่างสังเคราะห์ในคลาสบวกเพิ่มด้วยวิธีการ SMOTE โดยพิจารณาจากค่าอัตราความไม่สมดุลกัน *(IR)* ของทั้งสองคลาสจะต้องไม่เกิน 1.5

จากผลการทดลองทั้งหมดพบว่า ClusIM สามารถเพิ่มประสิทธิภาพในการจำแนกข้อมูลที่มีจำนวนตัวอย่างในแต่ละคลาสไม่สมดุลกันได้ดีมากในทุกชุดข้อมูลที่นำมาใช้ในงานวิจัยโดยเมื่อเทียบกับขั้นตอนวิธีทั้งหมดที่นำมาเปรียบเทียบโดยเฉพาะอย่างยิ่งกับคลาสบวกที่ ClusIM สามารถเพิ่มประสิทธิภาพในการจำแนกได้ดีกว่าขั้นตอนวิธีพื้นฐาน (SVM) ได้ถึงร้อยละ 90 บนชุดข้อมูลตัวอย่างที่มีความล้ำกันระหว่างคลาสอย่างมาก ตัวอย่างเช่น บนชุดข้อมูล Monk2 และ Ozone อีกทั้งขั้นตอนวิธี ClusIM สามารถเพิ่มประสิทธิภาพในการจำแนกได้ดีบนชุดข้อมูลที่มีอัตราความไม่สมดุลที่หลากหลาย นอกจากนี้ขั้นตอนวิธี ClusIM เป็นขั้นตอนวิธีที่ดำเนินการในระดับของข้อมูล ดังนั้นวิธีการนี้สามารถนำไปประยุกต์ใช้กับขั้นตอนวิธีในการจำแนกข้อมูลอื่น ๆ ที่หลากหลายตามวัตถุประสงค์ความต้องการของผู้ใช้

4. สรุปผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อออกแบบและพัฒนาขั้นตอนวิธีที่มีความสามารถในการจำแนกข้อมูลที่มีจำนวนตัวอย่างในแต่ละคลาสไม่สมดุลกันเพื่อเพิ่มประสิทธิภาพในการจำแนกข้อมูลด้วยเทคนิคผสมผสานระหว่างเทคนิคการจัดกลุ่มและเทคนิคการสุ่มตัวอย่างเพื่อปรับจำนวนตัวอย่างให้มีความสมดุลกันในชื่อ ClusIM ซึ่งหลักการ

ทำงานของ ClusIM ประกอบไปด้วยขั้นตอน 3 ขั้นตอนคือ 1) กระบวนสลับเปลี่ยนกลุ่มข้อมูลเพื่อลดความซ้อนทับของข้อมูลโดยใช้เทคนิคการจัดกลุ่มด้วยขั้นตอนวิธี K-means ซึ่งผลลัพธ์ที่ได้จากขั้นตอนนี้คือ ชุดข้อมูลตัวอย่าง 2 ชุดที่ถูกลดความซ้อนทับกันระหว่างคลาสบวกและคลาสลบ 2) การปรับข้อมูลตัวอย่างในชุดข้อมูลทั้ง 2 ชุด โดยปรับให้จำนวนตัวอย่างในคลาสบวกให้มีความสมดุลกับคลาสลบด้วยการใช้เทคนิค SMOTE โดยพิจารณาจากอัตราความไม่สมดุลต้องไม่มากกว่า 1.5 ซึ่งผลลัพธ์ที่ได้จากขั้นตอนนี้คือ ชุดข้อมูลตัวอย่าง 2 ชุดข้อมูลที่มีความสมดุลกันและคลาสบวกกับคลาสลบมีความซ้อนทับกันของระหว่างคลาสน้อยลง 3) การจำแนกข้อมูลด้วยตัวจำแนก ซึ่งงานวิจัยนี้เลือกใช้ขั้นตอนวิธี SVM และหาผลลัพธ์การจำแนกจากตัวแบบ 2 ตัวแบบด้วยวิธี Maximum Probability Vote

จากผลการทดลองทั้งหมดพบว่า ClusIM ให้ประสิทธิภาพในการจำแนกข้อมูลได้ดีกว่าทุกขั้นตอนวิธีที่เปรียบเทียบ เนื่องจาก ClusIM มีกระบวนการในการแยกความซ้อนทับกันของข้อมูลซึ่งส่งผลให้ตัวจำแนกสามารถจำแนกข้อมูลที่มีคุณลักษณะที่แตกต่างกันและส่งผลทำให้ตัวจำแนกให้ประสิทธิภาพในการจำแนกที่ได้ดีขึ้นกว่าการจำแนกบนข้อมูลต้นฉบับ นอกจากนี้ยังมีส่วนการเพิ่มข้อมูลตัวอย่างคลาสบวกโดยการใช้ SMOTE ซึ่งเป็นเทคนิคที่สร้างตัวอย่างสังเคราะห์ของคลาสบวกโดยไม่ทำให้เกิดปัญหา Overfitting ซึ่งจากทั้งสองส่วนนี้จึงส่งผลให้ประสิทธิภาพในการจำแนกของ ClusIM สามารถให้ค่า F-measure และ G-mean ได้ดีกว่าขั้นตอนวิธีอื่น ๆ ที่เปรียบเทียบ

อย่างไรก็ตาม งานวิจัยนี้นำเสนอขั้นตอนวิธี ClusIM ที่สามารถเพิ่มประสิทธิภาพในการจำแนกข้อมูลที่มีจำนวนตัวอย่างในแต่ละคลาสไม่สมดุลกันบนชุดข้อมูลที่มีคุณลักษณะ และจำนวนตัวอย่างที่ไม่มากนัก หากสามารถนำไปทดสอบกับชุดข้อมูลตัวอย่างที่มีจำนวนคุณลักษณะและจำนวนตัวอย่างที่มีขนาดใหญ่ จะสามารถแสดงให้เห็นถึงประสิทธิภาพในการจำแนกที่ชัดเจนมากขึ้น

5. กิตติกรรมประกาศ

ขอขอบคุณสำนักวิจัยวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ที่ให้ทุนสนับสนุนงานวิจัยนี้

6. เอกสารอ้างอิง

[1] Asuncion A. and Newman D., "*UCI machine learning repository.*". Available: http://archive.ics.uci.edu/ml/datasets.html, (2007).

[2] Breiman L., "*Bagging predictors*",Mach. Learn., 24(2): 123-140, (1996).

[3] Chawla N. V., Bowyer K. W., Hall L. O., and Kegelmeyer W. P., "*SMOTE: synthetic minority over-sampling technique*",Journal artificial intelligence research, 16(1): 321-357, (2002).

[4] Cieslak D. A., Chawla N. V., and Striegel A.,"*Combating imbalance in network intrusion datasets*", Granular Computing, 2006 IEEE International Conference on 732-737.(2006)

[5] Deng X., Zhong W., Ren J., Zeng D., and Zhang H.,"*An imbalanced data classification method based on automatic clustering under-sampling"*, 2016 IEEE 35th International Performance Computing and Communications Conference (IPCCC) 1-8.(2016)

[6] Fan W., Stolfo S. J., Zhang J., and Chan P. K., *"AdaCost: Misclassification Cost-Sensitive Boosting"*, presented at the Proceedings of the Sixteenth International Conference on Machine Learning, (1999).

[7] Fathi Ganji M., Abadeh M. S., Hedayati M., and Bakhtiari N.,"*Fuzzy classifcation of imbalanced data sets for medical diagnosis"*, Biomedical Engineering (ICBME), 2010 17th Iranian Conference of 1-5.(2010)

[8] Freund Y. and Schapire R.,"*Experiments with a New Boosting Algorithm"*, International Conference on Machine Learning 148-156.(1996)

[9] Gazzah S., Hechkel A., and Amara N. E. B.,"*A hybrid sampling method for imbalanced data"*, 2015 IEEE 12th International Multi-Conference on Systems, Signals & Devices (SSD15) 1-6.(2015)

[10] Maldonado S. and López J., "*Imbalanced data classification using second-order cone programming support vector machines"*,Pattern Recognition, 47(5): 2070-2079, (2014).

[11] Márquez-Vera C., Cano A., Romero C., and Ventura S., "*Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data"*,Applied Intelligence, 38(3): 315-330, (2013).

[12] Prachuabsupakij W. and Doungpaisan P.,"*Matching preprocessing methods for improving the prediction of student's graduation"*, 2016 2nd IEEE International Conference on Computer and Communications (ICCC) 33-37.(2016)

[13] Zhu M., Su B., and Ning G.,"*Research of Medical High-Dimensional Imbalanced Data Classification Ensemble Feature Selection Algorithm with Random Forest"*, 2017 International Conference on Smart Grid and Electrical Automation (ICSGEA) 273-277.(2017)