แนวทางการสุ่มข้อมูลโดยการผนวกเทคนิค K-Means และ SMOTE สำหรับการจำแนกประเภทข้อมูลที่ไม่สมดุล

อดิเทพ จิตพิทยาพร

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่
วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์
มหาวิทยาลัยธุรกิจบัณฑิตย์
ปีการศึกษา 2564

KMSM: TOWARDS MORE EFFICIENT SAMPLING TECHNIQUE FOR IMBALANCED CLASSIFICATION

ADITHEP CHITPITAYAPORN

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Engineering

Department of Big Data Engineering,

College of Innovative Technology and Engineering,

Dhurakij Pundit University

Academic Year 2021

ใบรับรองงานวิทยานิพนธ์

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์

ปริญญา วิศวกรรมศาสตรมหาบัณฑิต

หัวข้อวิทยานิพนธ์ แนวทางการสุ่มข้อมูลโดยการผนวกเทคนิค K-Means และ SMOTE สำหรับการจำแนกประเภทข้อมูลที่ไม่สมดุล

เสนอโดย อดิเทพ จิตพิทยาพร

สาขาวิชา วิศวกรรมข้อมูลขนาดใหญ่

อาจารย์ที่ปรึกษาวิทยานิพนธ์ ดร.เอกสิทธิ์ พัชรวงศ์ศักดา

ได้พิจารณาเห็นชอบโดยคณะกรรมการสอบวิทยานิพนธ์แล้ว

.....ประธานกรรมการ
(ดร.สรรพฤทธิ์ มฤคทัต)

.....กรรมการและอาจารย์ที่ปรึกษา
(ดร.เอกสิทธิ์ พัชรวงศ์ศักดา)

.....กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ดวงใจ จิตคงชื่น)

.....กรรมการ
(ดร.ธนภัทร ยังคะจิตร)

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์รับรองแล้ว

.....
(ดร.ชัยพร เขมะภาตะพันธ์)
คณบดีวิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์
วันที่ 3 เดือน พฤศจิกายน พ.ศ. 2564

หัวข้อวิทยานิพนธ์	แนวทางการสุ่มข้อมูลโดยการผนวกเทคนิค K-Means และ SMOTE สำหรับการจำแนกประเภทข้อมูลที่ไม่สมดุล
ชื่อผู้เขียน	อดิเทพ จิตพิทยาพร
อาจารย์ที่ปรึกษา	ดร. เอกสิทธิ์ พัชรวงศ์ศักดา
สาขาวิชา	วิศวกรรมข้อมูลขนาดใหญ่
ปีการศึกษา	2564

บทคัดย่อ

ปัจจุบันมีการนำข้อมูลต่างๆมาใช้ในการวิเคราะห์ด้วยวิธีการจำแนกหมวดหมู่ ปัญหาอย่างหนึ่งที่พบไม่ว่าจะใช้วิธีการจำแนกหมวดหมู่วิธีใดก็ตามคือ ถ้าข้อมูลที่ใช้สำหรับการเรียนรู้ (training data) มีความไม่สมดุลระหว่างแต่ละหมวดหมู่ มักจะทำให้ผลลัพธ์มีความเอนเอียงไปทางด้านของข้อมูลส่วนใหญ่ และยิ่งข้อมูลมีความไม่สมดุลระหว่างส่วนใหญ่กับส่วนน้อยมากเท่าไร ยิ่งทำให้มีความเอนเอียงสูงมากขึ้น

ในงานวิจัยนี้ได้นำเสนอแนวทางการปรับสมดุลของข้อมูลโดยการแบ่งกลุ่มของข้อมูล (Clustering) และใช้วิธีการเพิ่มจำนวนข้อมูลส่วนน้อยตามแต่ละกลุ่ม (Oversampling) รวมถึงการจำแนกประเภทข้อมูลตามแต่ละกลุ่มด้วย ทำให้ประสิทธิภาพของการจำแนกประเภทข้อมูลมีความแม่นยำ และถูกต้องมากขึ้น

จากผลการทดลองกับข้อมูลทั้งหมด 44 ชุดข้อมูล พบว่าวิธีการที่นำเสนอนี้มีความถูกต้องเฉลี่ย 0.913 เมื่อเทียบกับวิธี SMOTE ที่มีความถูกต้องเฉลี่ย 0.904

Thesis Title	KMSM: TOWARDS MORE EFFICIENT SAMPLING TECHNIQUES FOR IMBALANCED CLASSIFICATION
Author	Adithep Chitpitayaporn
Thesis Advisor	Dr. Eakasit Pacharawongsakda
Department	Big Data Engineering
Academic Year	2021

ABSTRACT

Nowadays, a wide range of data is analyzed through the use of classification methods. One of the problems faced, regardless of classification methods being adopted, is when class imbalance exits within the training data, the results tend to be biased towards classes which have number of instances. When the ratio between the majority class and the minority class is larger, it leads to more bias.

In this research, a data balancing method was proposed. Clustering, oversampling, and classification were combined to provide greater accuracy and precision of a classification method.

The results of experiments with 44 datasets revealed that the proposed method could achieve the average accuracy of 0.913 in comparison to SMOTE with the average accuracy of 0.904.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยกรุณาของ ดร.เอกสิทธิ์ พัชรวงศ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ได้ให้คำปรึกษาแนะนำ การปรับปรุงแก้ไขงานวิจัย และแก้ไขร่างวิทยานิพนธ์เป็นอย่างดีมาโดยตลอด

ผู้เขียนขอกราบขอบพระคุณ ดร.สรรพฤทธิ์ มฤคทัต ที่กรุณาให้เกียรติเป็นประธานโดยมี ดร.ธนภัทร ฆังคะจิตร และ ผศ.ดร. ดวงใจ จิตคงชื่น เป็นกรรมการในการสอบวิทยานิพนธ์ ซึ่งได้ให้คำแนะนำแนวทางที่เป็นประโยชน์ต่องานวิจัย และตรวจแก้ไขวิทยานิพนธ์ฉบับนี้ให้ถูกต้องสมบูรณ์ยิ่งขึ้น ตลอดจน นางสาวกุลธิดา รอดบุญ รวมถึงเจ้าหน้าที่บัณฑิตมหาวิทยาลัยธุรกิจบัณฑิตย์ ทุกท่านที่ช่วยอำนวยการ และประสานงาน ในการทำวิทยานิพนธ์ให้ผู้เขียนตลอดมา ส่งผลให้การจัดทำวิทยานิพนธ์ของผู้เขียนครั้งนี้สำเร็จลุล่วงไปด้วยดี

ท้ายนี้ ผู้วิจัยต้องกราบขออภัยเป็นอย่างสูงมา ณ โอกาสนี้ หากมีสิ่งใดที่ผู้วิจัยได้ทำผิดพลาดหรือบกพร่องประการใด และผู้วิจัยหวังเป็นอย่างยิ่งว่าวิทยานิพนธ์ฉบับนี้จะเป็นพื้นฐานในการต่อยอดองค์ความรู้ของผู้ที่สนใจศึกษาในงานด้านนี้ต่อไป

อดิเทพ จิตพิทยาพร

สารบัญ

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญภาพ

ภาพที่ **หน้า**

สารบัญภาพ (ต่อ)

ภาพที่ **หน้า**

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ปัจจุบันมีการนำข้อมูลต่างๆมาวิเคราะห์เพื่อวัตถุประสงค์ต่างๆมากมาย บางวัตถุประสงค์เพื่อต้องการแบ่งกลุ่มของข้อมูลที่มีความคล้ายคลึงกัน บางวัตถุประสงค์เพื่อต้องการจำแนกหมวดหมู่ของข้อมูลนั้นๆ สำหรับข้อมูลจริงที่นำมาใช้ในการเรียนรู้มักจะพบปัญหาความไม่สมดุลของข้อมูล ทำให้ผลลัพธ์ของการเรียนรู้ไม่มีประสิทธิภาพคือมีความเอนเอียงไปทางข้อมูลส่วนใหญ่ ในการแก้ปัญหาความเอนเอียงนี้ สามารถทำได้โดยการทำให้ข้อมูลระหว่างข้อมูลส่วนใหญ่ กับข้อมูลส่วนน้อยมีปริมาณสมดุลกันก่อน แล้วจึงนำข้อมูลไปใช้ในการเรียนรู้ต่อไป การแก้ปัญหาความไม่สมดุลมีอยู่หลายวิธี หนึ่งในเทคนิคการแก้ปัญหาของข้อมูลไม่สมดุลนั้น ทำโดยการลดจำนวนข้อมูลส่วนใหญ่ ด้วยการหาตัวแทนข้อมูลส่วนใหญ่ ทั้งที่เป็นข้อมูลที่สร้างขึ้นมาใหม่ หรือคัดเลือกจากข้อมูลจริง นอกจากนี้ ยังมีวิธีการเพิ่มจำนวนข้อมูลส่วนน้อยให้มีปริมาณมากขึ้นจนเทียบเท่ากับข้อมูลส่วนใหญ่ ทั้งนี้วิธีการดังกล่าว จะกระทำบนข้อมูลส่วนใหญ่ หรือข้อมูลส่วนน้อยเพียงอย่างเดียว หากพิจารณาข้อมูล จะพบว่าในบางครั้งจะมีข้อมูลส่วนใหญ่ที่อยู่บริเวณใกล้เคียงกับข้อมูลส่วนน้อย และคัดเลือกข้อมูลที่อยู่ใกล้เคียงกับศูนย์กลางของกลุ่ม จะทำให้ข้อมูลที่อยู่บริเวณใกล้เคียงกับข้อมูลส่วนน้อยนี้ไม่ได้นำมาใช้ในการเรียนรู้ ซึ่งข้อมูลบริเวณนี้อาจจะเป็นข้อมูลที่มีผลต่อการเรียนรู้ ทำให้การทำนายผลผิดพลาดได้

เช่นกัน เมื่อทำการแบ่งกลุ่มของข้อมูลที่จะใช้ในการเรียนรู้ทั้งหมดออกมาเป็นกลุ่มย่อยๆ แล้ว ในแต่ละกลุ่มย่อย อาจจะพบว่า ในบางกลุ่มนั้นจำนวนข้อมูลส่วนน้อย จะมีปริมาณมากกว่าจำนวนข้อมูลส่วนใหญ่ และการคัดเลือกข้อมูลส่วนใหญ่ของกลุ่ม จะมีผลทำให้ปริมาณข้อมูลส่วนน้อยของข้อมูลทั้งหมดมีปริมาณน้อยลง ทำให้ปริมาณข้อมูลที่จะใช้ในการเรียนรู้นั้นมีปริมาณน้อยลงจนมีผลต่อประสิทธิภาพของการเรียนรู้

ในงานวิจัยนี้จะศึกษาการแก้ปัญหาความไม่สมดุลของข้อมูลโดยวิธีการแบ่งกลุ่มของข้อมูลและการสร้างข้อมูลส่วนน้อยในแต่ละกลุ่มให้มีปริมาณใกล้เคียงกับจำนวนข้อมูลส่วนใหญ่ของกลุ่ม

นั้นๆ แล้วนำข้อมูลที่มีการปรับสมดุลแล้วไปใช้ในการเรียนรู้ เพื่อให้ได้โมเดลที่เหมาะสมสำหรับข้อมูลในแต่ละกลุ่มต่อไป

1.2 วัตถุประสงค์ของงานวิจัย

1.2.1 เพื่อศึกษาวิธีปรับปรุงข้อมูลที่มีความไม่สมดุล ให้มีเป็นข้อมูลที่มีความสมดุลสำหรับการนำไปใช้ในการเรียนรู้

1.2.2 เพื่อศึกษาการแบ่งกลุ่มของข้อมูล สำหรับการนำไปใช้ในการเรียนรู้

1.2.3 เพื่อศึกษาวิธีการเพิ่มจำนวนข้อมูลของข้อมูลส่วนน้อย ที่สามารถเพิ่มประสิทธิภาพของการเรียนรู้

1.2.4 เพื่อศึกษาวิธีการผสมผสานการแบ่งกลุ่มของข้อมูล และการเพิ่มข้อมูลส่วนน้อย สำหรับนำไปใช้ในการเรียนรู้และจำแนกหมวดหมู่ของข้อมูล

1.3 ขอบเขตงานวิจัย

1.3.1 ข้อมูลที่มีความไม่สมดุลจำนวน 44 ชุดข้อมูลจากฐานข้อมูล KEEL (Knowledge Extraction based on Evolutionary Learning)

1.3.2 การแบ่งกลุ่มของข้อมูลด้วย K-Mean

1.3.3 การเพิ่มจำนวนข้อมูลส่วนน้อยของแต่ละกลุ่มด้วยวิธีการ SMOTE

1.3.4 การจำแนกหมวดหมู่ข้อมูลด้วยต้นไม้ตัดสินใจ

1.4 สมมติฐานของงานวิจัย

1.4.1 ข้อมูลที่ไม่ได้เป็นตัวแทนของข้อมูลส่วนใหญ่ในวิธี UnderSampling อาจจะมีผลกระทบอย่างสูงต่อความถูกต้องของโมเดล

1.4.2 การแบ่งกลุ่มของข้อมูล และการจำแนกหมวดหมู่ข้อมูลสำหรับแต่ละกลุ่ม จะสามารถเพิ่มประสิทธิภาพของการเรียนรู้ได้

1.4.3 การแบ่งกลุ่มของข้อมูลและเพิ่มจำนวนข้อมูลส่วนน้อยที่เหมาะสมสำหรับแต่ละกลุ่ม สามารถเพิ่มประสิทธิภาพของการเรียนรู้ได้

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1.5.1 วิธีการผสมผสานของการแบ่งกลุ่มข้อมูล และการเพิ่มปริมาณข้อมูลส่วนน้อย จะช่วยให้พัฒนาโมเดลสำหรับการเรียนรู้มีประสิทธิภาพมากขึ้น

1.5.2 วิธีการผสมผสานจะช่วยให้การทำนายผลมีความถูกต้องมากขึ้น

1.5.3 เพื่อเป็นแนวทางในการพัฒนาการจัดการข้อมูลที่มีความไม่สมดุลให้มีความสอดคล้องกับข้อมูล

บทที่ 2
แนวคิด ทฤษฎี และผลงานวิจัยที่เกี่ยวข้อง

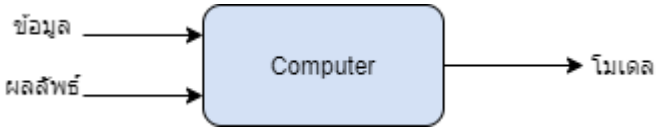
งานวิจัยเรื่อง วิธีการสุ่มตัวอย่างตามการแบ่งกลุ่มของข้อมูลที่ไม่สมดุล ผู้วิจัยได้ทำการศึกษาจากแหล่งความรู้ทางอินเตอร์เน็ต และจากงานวิจัยที่เกี่ยวข้อง โดยรายละเอียดเกี่ยวกับแนวความคิดและทฤษฎี รวมถึงผลงานวิจัยที่เกี่ยวข้อง มีรายละเอียดดังนี้

2.1 แนวคิดและทฤษฎีที่เกี่ยวข้อง

ในงานวิจัยนี้ มีแนวคิดและทฤษฎีที่ใช้ ไม่ว่าจะเป็น การเรียนรู้ของเครื่อง การแบ่งกลุ่มข้อมูล การจำแนกหมวดหมู่ รวมถึงการประเมินประสิทธิภาพการเรียนรู้ ดังรายละเอียดดังนี้

2.1.1. การเรียนรู้ของเครื่อง (Machine Learning)

การเรียนรู้ของเครื่อง เป็นการประยุกต์ใช้สถิติขั้นสูง โดยจะอาศัยการใส่ข้อมูลและผลลัพธ์เข้าไป แล้วใช้อัลกอริทึมเพื่อเรียนรู้ให้สามารถค้นพบรูปแบบหรือแบบแผนซ้ำๆ และคาดการณ์จากรูปแบบเหล่านั้น



ภาพที่ 2.1 การเรียนรู้ของเครื่อง (Machine Learning)

การเรียนรู้ของเครื่องสามารถแบ่งออกได้เป็น 2 ประเภทหลักๆคือ

2.1.1.1 การเรียนรู้แบบมีผู้สอน (Supervised Learning)

การเรียนรู้แบบมีผู้สอน จะเป็นนำชุดข้อมูลสำหรับเรียนรู้ โดยชุดข้อมูลนั้นจะมีคำตอบอยู่แล้ว และนำมาสร้างเป็นโมเดล เพื่อใช้คาดการณ์ผลของข้อมูลชุดใหม่ ซึ่งโมเดลสามารถเป็นได้ทั้งกฎ

(Rules) สมการทางคณิตศาสตร์ต่างๆ เช่น ระยะห่าง (Distance) หรือแม้กระทั่งโครงข่ายประสาท (Neural Network)

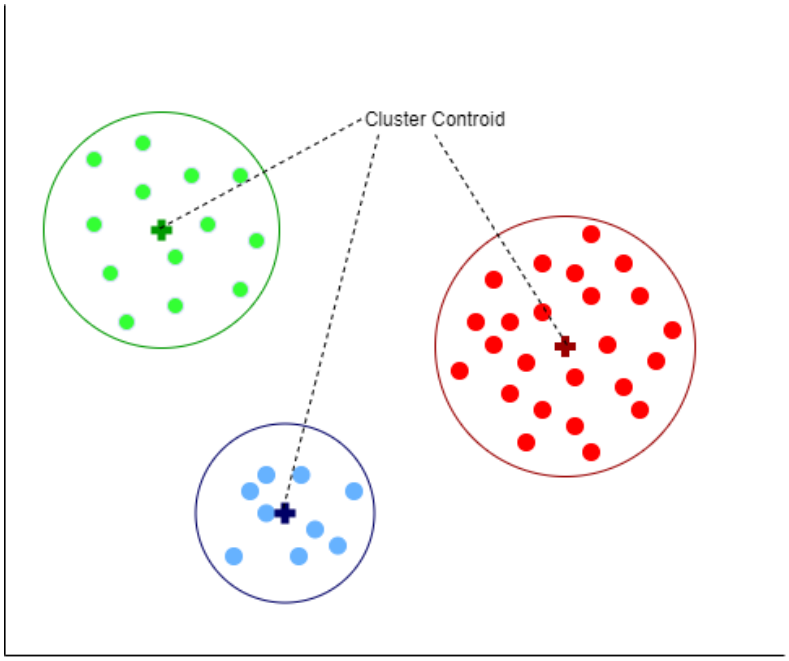
โดยที่การเรียนรู้แบบมีผู้สอนยังสามารถแบ่งออกได้เป็น 2 แบบคือ การจำแนกหมวดหมู่ (Classification) และการประมาณค่าข้อมูล (Regression) ซึ่งการเรียนรู้ทั้ง 2 ประเภทจะมีลักษณะคล้ายคลึงกันมาก แต่ต่างกันที่คำตอบของการทำนายที่ได้คือ การจำแนกหมวดหมู่จะเป็นคำตอบแบบเชิงคุณภาพ (Categorical) หรือแบบไม่ต่อเนื่อง (Discrete) ส่วนการประมาณค่าข้อมูล จะให้คำตอบแบบเชิงปริมาณ (Ordinal) หรือเป็นตัวเลขเท่านั้น

2.1.1.2 การเรียนรู้แบบไม่มีผู้สอน (Unsupervise Learning)

การเรียนรู้แบบไม่มีผู้สอน จะใช้ชุดข้อมูลที่ไม่มีคำตอบ และการเรียนรู้จะเป็นการพิจารณาความสัมพันธ์ของข้อมูลเป็นหลัก ซึ่งวิธีการเรียนรู้แบบไม่มีผู้สอนมีทั้งการค้นหาด้วยกฎความสัมพันธ์ (Association Rule) และการแบ่งกลุ่มข้อมูล (Clustering) ดังนั้น การเรียนรู้แบบไม่มีผู้สอน จะไม่สามารถนำไปใช้สำหรับคาดการณ์ได้แบบตรงไปตรงมา

2.1.2 การแบ่งกลุ่มข้อมูล (Clustering)

การแบ่งกลุ่มข้อมูล เป็นการจัดกลุ่มข้อมูลที่มีลักษณะเหมือนกันเอาไว้ในกลุ่มเดียวกัน (Cluster) ขั้นตอนที่ใช้ในการแบ่งกลุ่มจะอาศัยความเหมือน (Similarity) หรือความใกล้ชิด (Proximity) ซึ่งวิธีการแบ่งกลุ่มที่ใช้ความใกล้ชิดหรือระยะห่างที่นิยมใช้คือ การแบ่งกลุ่มด้วย k-means โดยจะกำหนดจำนวน k กลุ่มตามที่ต้องการ แล้ววัดระยะห่างระหว่างข้อมูลกับจุดศูนย์กลางของกลุ่ม ซึ่งจุดศูนย์กลางของกลุ่ม จะเป็นค่าเฉลี่ยระยะทางของแต่ละแอตทริบิวต์ (attribute) ของข้อมูลในกลุ่มนั้นๆ



ภาพที่ 2.2 การแบ่งกลุ่มข้อมูลด้วย k-means

การแบ่งกลุ่มข้อมูลด้วย k-mean นี้ ถึงแม้จะเป็นวิธีที่ง่าย สามารถกำหนดจำนวนกลุ่มได้ การคำนวณในการจัดกลุ่มสามารถทำได้เร็ว แต่มีข้อเสียคือ ยากในการเปรียบเทียบตอนสร้างกลุ่ม และไม่สามารถคำนวณจำนวนกลุ่มที่เหมาะสมได้

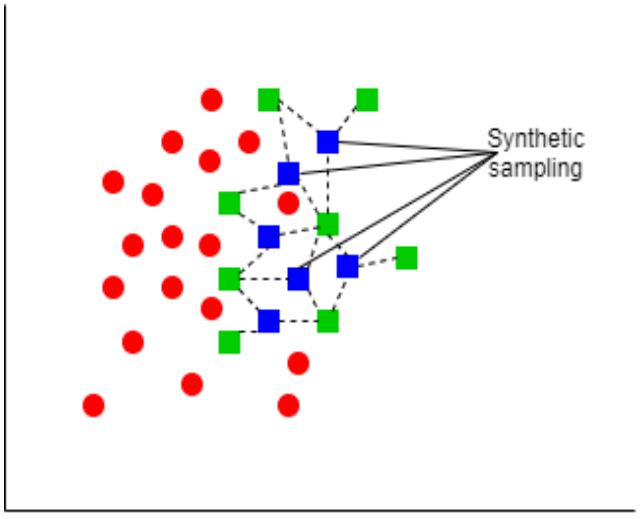
2.1.3 การปรับสมดุลข้อมูล

การนำข้อมูลมาสร้างโมเดลสำหรับการจำแนกหมวดหมู่ หากปริมาณข้อมูลส่วนใหญ่และข้อมูลส่วนน้อยมีไม่เท่ากัน หรือไม่สมดุลกัน จะทำให้การจำแนกหมวดหมู่มีความเอนเอียงไปทางข้อมูลส่วนใหญ่ เพื่อแก้ปัญหาความเอนเอียงในการจำแนกหมวดหมู่ สามารถทำได้โดยการปรับชุดข้อมูลให้มีความสมดุลระหว่างข้อมูลส่วนใหญ่และข้อมูลส่วนน้อยก่อนจะนำไปสร้างโมเดล ซึ่งวิธีการปรับสมดุลนี้ จะแบ่งได้ออกเป็น 2 วิธีคือ การลดจำนวนข้อมูลส่วนใหญ่ (Undersampling) และการเพิ่มจำนวนข้อมูลส่วนน้อย (Oversampling)

1. วิธีการลดจำนวนข้อมูลส่วนใหญ่ (Undersampling) ที่นิยมจะทำโดยการสุ่มข้อมูลส่วนใหญ่ ให้มีจำนวนเท่ากับจำนวนข้อมูลส่วนน้อย วิธีนี้สามารถทำได้ง่าย ไม่ซับซ้อน แต่ข้อเสียของวิธีนี้

คือ ไม่มีรูปแบบการคัดเลือกข้อมูลที่แน่นอน ซึ่งข้อมูลที่สุ่มได้ อาจจะไม่ได้เป็นตัวแทนของข้อมูลส่วนใหญ่ ทำให้ประสิทธิภาพในการจำแนกหมวดหมู่ไม่แน่นอน ส่วนอีกวิธีที่นิยมคือ การใช้ศูนย์กลางของกลุ่ม (ClusterControid) มาเป็นตัวแทนของกลุ่ม โดยการนำข้อมูลส่วนใหญ่มาแบ่งกลุ่มให้เท่ากับจำนวนข้อมูลส่วนน้อย แล้วใช้ศูนย์กลางของกลุ่มมาเป็นตัวแทน วิธีนี้จะเป็นการคัดเลือกข้อมูลส่วนใหญ่ได้ค่อนข้างแน่นอน และสามารถใช้เป็นตัวแทนของข้อมูลส่วนใหญ่ได้ แต่ข้อเสียคือ การที่ข้อมูลที่ได้นั้นไม่ได้เป็นข้อมูลจริงอาจจะทำให้ประสิทธิภาพการจำแนกหมวดหมู่ไม่ดีได้ในบางครั้ง จึงมีแนวคิดที่ใช้ข้อมูลจริงที่อยู่ใกล้ศูนย์กลางกลุ่มมากที่สุดมาเป็นตัวแทน และอีกปัญหาของวิธีนี้คือ ถ้าจำนวนข้อมูลส่วนน้อยมีปริมาณมาก การแบ่งกลุ่มที่มีจำนวนกลุ่มมากจะใช้เวลานาน

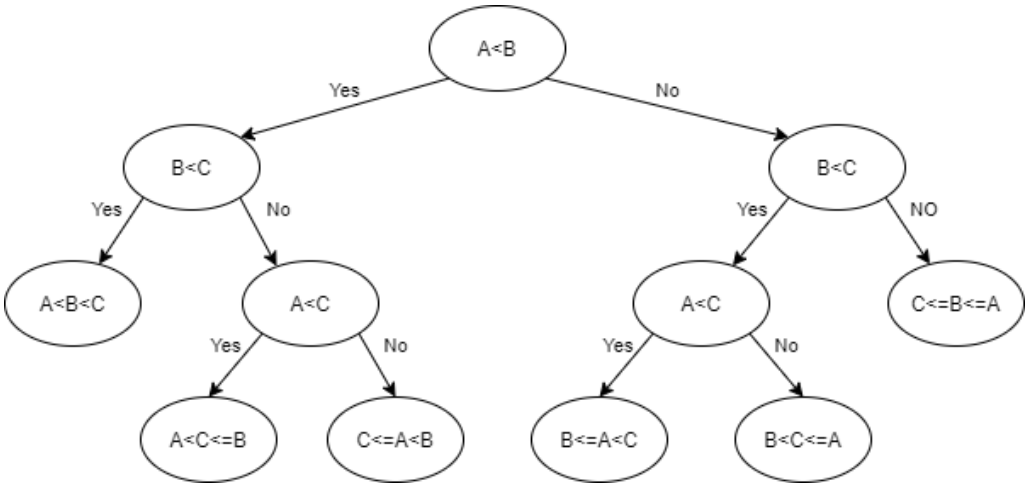
2. วิธีเพิ่มจำนวนข้อมูลส่วนน้อย (Oversampling) จะเป็นการสร้างข้อมูลในหมวดหมู่ของข้อมูลส่วนน้อยขึ้นมา โดยจะสร้างขึ้นมาให้มีปริมาณรวมใกล้เคียงหรือเท่ากับข้อมูลส่วนใหญ่ วิธีการของ Oversampling ที่นิยมใช้ส่วนใหญ่คือ วิธีสังเคราะห์ข้อมูลเพิ่ม (Synthetic Minority Oversampling TEchinque: SMOTE) วิธีนี้เป็นที่นิยมใช้กันมาก การสร้างข้อมูลขึ้นมาใหม่จากข้อมูลเดิม โดยใช้หลักการเพื่อนบ้านที่อยู่ใกล้ที่สุดในการขยายขอบเขต เริ่มต้นด้วยการสุ่มข้อมูลจากข้อมูลส่วนน้อย และกำหนดค่า k เพื่อนบ้านที่อยู่ใกล้ที่สุด จากนั้นจะสังเคราะห์ข้อมูลขึ้นระหว่างข้อมูลที่สุ่มและเพื่อนบ้านที่อยู่ใกล้



ภาพที่ 2.3 Oversampling ด้วย SMOTE

2.1.4. การจำแนกหมวดหมู่ (Classification)

การจำแนกหมวดหมู่ จะเป็นการนำข้อมูลมาเรียนรู้เพื่อให้รู้รูปแบบของข้อมูล แล้วจึงสร้างเป็นโมเดลขึ้นมา เพื่อนำมาใช้ในการหาคำตอบของข้อมูลชุดใหม่ ซึ่งข้อมูลที่ใช้ในการเรียนรู้จะเป็นข้อมูลที่มีเลเบลกำกับ เพื่อบอกว่าข้อมูลนั้นอยู่ในหมวดหมู่ใด วิธีการจำแนกหมวดหมู่ที่นิยมใช้คือต้นไม้ตัดสินใจ (Decision Tree) ซึ่งสามารถแปลความหมายและเข้าใจได้ง่าย



ภาพที่ 2.4 ต้นไม้ตัดสินใจ (Decision Tree)

2.1.5 การวัดประสิทธิภาพ

การที่จะรู้ว่าโมเดลใดดีหรือไม่ดี สามารถวัดได้จากการหาค่าประสิทธิภาพที่คำนวณได้จากตารางแจกแจงผลลัพท์ (Confusion Matrix) ซึ่งเป็นค่าที่เปรียบเทียบระหว่างค่าจริงและค่าที่ได้จากการทำนายโดยโมเดลดังตาราง

ตารางที่ 2.1 Confusion Matrix

	Predicted Negative	Predicted Positive
Actual Negative	True Negative	False Positive
Actual Positive	False Negative	True Positive

ค่าในตาราง True Negative จะเป็นค่าผลลัพท์ที่ทายถูกว่าเป็น class negative ส่วนค่า True Positive จะเป็นค่าผลลัพท์ที่ทายถูกว่าเป็น class positive

สำหรับค่า False Positive จะเป็นค่าที่ทำนายผลลัพท์ผิด โดยทำนายว่าเป็นค่า Positive แต่ค่าจริงเป็น Negative ส่วนค่า False Negative จะเป็นค่าที่ทำนายผลลัพท์ผิดโดยทำนายว่าเป็นค่า Negative แต่ค่าจริงเป็นค่า Positive

จากตารางแจกแจงผลลัพท์ สามารถนำมาคำนวณหาค่าความถูกต้อง (Accuracy) ความแม่นยำ (Precision) ค่าการตรวจสอบ (Recall) และค่า F-Measure

1. ค่าความถูกต้อง (Accuracy)

ค่าความถูกต้อง จะเป็นค่าที่อธิบายความถูกต้องโดยรวมของโมเดลเทียบกับข้อมูลทั้งหมด ซึ่งสามารถคำนวณได้จาก

$$Accuracy = \frac{(TruePositive + TrueNegative)}{(TruePositive + TrueNegative + FalsePositive + FNegative)}$$

2. ค่าความแม่นยำ (Precision)

ค่าความแม่นยำ จะเป็นค่าที่อธิบายความแม่นยำของโมเดลโดยพิจารณาแยกทีละคลาส ซึ่งสามารถคำนวณได้จาก

$$Precision = \frac{TruePositive}{(TruePositive + FalsePositive)}$$

3. ค่าการตรวจสอบ (Recall)

ค่าการตรวจสอบ จะเป็นค่าที่วัดความถูกต้องของโมเดล โดยพิจารณาแยกทีละคลาส ซึ่งสามารถคำนวณได้จาก

$$Recall = \frac{TruePositive}{(TruePositive + FalseNegative)}$$

4. F-Measure

ค่า F-Measure จะเป็นการวัดค่าโดยรวมของ Precision และ Recall โดยจะพิจารณาพร้อมกัน ซึ่งสามารถคำนวณได้จาก

$$F - measure = \frac{(2 * Precision * Recall)}{(Precision + Recall)}$$

2.2 งานวิจัยที่เกี่ยวข้อง

ในงานวิจัยของ Wei-Chao Lin และทีม ทำ undersampling ทำการแบ่งกลุ่มของข้อมูลส่วนใหญ่ให้มีจำนวนกลุ่มเท่ากับจำนวนข้อมูลในข้อมูลส่วนน้อย โดยวิธี k-means และใช้ศูนย์กลางของกลุ่มเป็นตัวแทนของข้อมูลส่วนใหญ่ ทำให้จำนวนข้อมูลที่คัดเลือกมาจากข้อมูลส่วนใหญ่มีจำนวนเท่ากับจำนวนข้อมูลส่วนน้อย วิธีนี้ไม่ได้นำข้อมูลจริงของข้อมูลส่วนใหญ่มาใช้สำหรับการเรียนรู้ เมื่อนำข้อมูลจริงมาทำการทดสอบ จะมีโอกาสเกิดความผิดพลาดได้ง่าย โดยเฉพาะถ้าข้อมูลนั้นเป็นข้อมูลส่วนใหญ่

และในงานวิจัยนี้ได้มีปรับปรุงการแก้ไขปัญหานี้เพิ่มเติม โดยใช้วิธีเดียวกับวิธีก่อนหน้านี้ แต่จะคัดเลือกข้อมูลในข้อมูลส่วนใหญ่ที่อยู่ใกล้ศูนย์กลางของกลุ่มมากที่สุดมาใช้เป็นตัวแทนของกลุ่ม นอกจากนี้การคัดเลือกข้อมูลจากวิธีการแบ่งกลุ่มของข้อมูลด้วยวิธี k-means ศูนย์กลางของกลุ่มอาจจะอยู่ใกล้กัน และข้อมูลที่อยู่ใกล้ศูนย์กลางของกลุ่มหนึ่ง อาจจะอยู่ใกล้ศูนย์กลางของอีกกลุ่มหนึ่งได้ ทำให้ข้อมูลจริงของข้อมูลส่วนใหญ่ที่คัดเลือกมานั้นเป็นตัวแทนของกลุ่มมากกว่า 1 กลุ่ม อย่างไรก็ตาม การทำ undersampling กับข้อมูลที่มีปริมาณของข้อมูลส่วนน้อยไม่มาก จะทำให้ได้ข้อมูลโดยรวมเพื่อใช้สำหรับการเรียนรู้มีปริมาณน้อย และทำให้โมเดลที่ได้จากการเรียนรู้ไม่มีประสิทธิภาพเท่าที่ควร

นอกจากนี้ Nitesh V. Chawla และทีม ได้เสนอวิธีการทำ oversampling ด้วยวิธี SMOTE แต่การสังเคราะห์ข้อมูล จะใช้วิธีสังเคราะห์จากข้อมูลจริงโดยดำเนินการในพื้นที่คุณลักษณะ แทนการดำเนินการในพื้นที่ข้อมูล โดยการสังเคราะห์ตัวอย่างบนแนวเส้นเชื่อมโยงของข้อมูลส่วนน้อย k ที่อยู่ใกล้กันมากที่สุด และในบางงานวิจัย (Ha &Bunke, 1997) จะทำการสังเคราะห์ข้อมูลโดยดำเนินการบนพื้นที่ข้อมูล เช่นการหมุนและปรับให้เอียง ทำให้ได้จำนวนข้อมูลส่วนน้อยเพิ่มขึ้น จากทั้ง 2 วิธีนี้ จะเป็นการสังเคราะห์ข้อมูลส่วนน้อยจากข้อมูลจริง การจำแนกหมวดหมู่จึงมีประสิทธิภาพ อย่างไรก็ตาม ข้อมูลส่วนน้อยอาจจะอยู่แบบกระจาย การสังเคราะห์ด้วยวิธีการนี้จากข้อมูลทั้งหมดในชุดข้อมูล จะทำให้ข้อมูลส่วนน้อยกระจายด้วยเช่นกัน และการจำแนกหมวดหมู่อาจจะไม่มีประสิทธิภาพมากนัก

ในงานวิจัยของ Elhassan AT จะมีการเตรียมข้อมูลโดยการนำข้อมูลที่เป็นข้อมูลรบกวนออกไปโดยใช้ T-Link algoritm และคัดเลือกข้อมูลส่วนใหญ่ด้วยวิธีการสุ่ม การนำข้อมูลที่เป็นข้อมูลรบกวนออกไปนั้น จะช่วยให้ข้อมูลส่วนที่เหลือนั้น มีผลต่อการจำแนกกลุ่มไม่เอนเอียงมากเกินไป และเมื่อมีการนำข้อมูลส่วนที่เหลือไปใช้งานต่อ ไม่ว่าจะนำไปทำการคัดเลือกข้อมูลส่วนใหญ่ด้วยวิธีการสุ่มหรือการสังเคราะห์ข้อมูลส่วนน้อยเพิ่มขึ้นก็ตาม มีผลทำให้การนำข้อมูลเหล่านั้นไปใช้ในการจำแนกกลุ่มมีประสิทธิภาพมากขึ้น อย่างไรก็ดี เมื่อมีการนำข้อมูลที่เป็นข้อมูลรบกวนออกไปแล้ว และทำการคัดเลือกข้อมูลส่วนใหญ่ด้วยวิธีการสุ่ม ถ้าปริมาณข้อมูลส่วนน้อยมีไม่มาก จะทำให้จำนวนข้อมูลทั้งหมดสำหรับการเรียนรู้มีปริมาณน้อย และมีผลทำให้โมเดลที่ได้จากการเรียนรู้ไม่มีประสิทธิภาพเท่าที่ควร

Chih-Fong Tsai, Wei-Chao Lin, Ya-Han Hu, and Guan-Ting Yao ได้มีการนำเสนอวิธี cluster-based instance selectin (CBIS) ซึ่งเป็นการลดจำนวนข้อมูลส่วนใหญ่โดยใช้วิธีการ Clustering analysis ร่วมกับ instance selection ในการทำ cluster analysis จะใช้ Affinity Propagation เพื่อแบ่งกลุ่มข้อมูล ซึ่งวิธีนี้จะมีข้อดีคือ สามารถแบ่งข้อมูลออกเป็นจำนวนกลุ่มที่เหมาะสม โดยไม่ต้องระบุว่าจะแบ่งข้อมูลเป็นกี่กลุ่ม ซึ่งแตกต่างจาก k-means ที่ต้องระบุจำนวนกลุ่ม จากนั้นจึงใช้วิธีการคัดเลือก instance ในแต่ละกลุ่ม ซึ่งในงานวิจัยนั้นมีการทดสอบโดยใช้ 3 วิธีคือ genetic algorithm (GA), IB3, และ DROP3 แล้วจึงนำไปรวมกับข้อมูลส่วนน้อย เพื่อนำไปใช้ในการเรียนรู้ ผลจากการคัดเลือกข้อมูลส่วนใหญ่ด้วยวิธีนี้ อาจจะได้ไม่เท่ากับจำนวนข้อมูลส่วนน้อย ซึ่งจะขึ้นอยู่กับการคัดเลือก instance ที่ใช้ ทำให้อัตราส่วนความไม่สมดุล (IR) จะยังคงมีอยู่ แต่จะขึ้นอยู่กับ algorithm การคัดเลือก instance ที่ใช้ โดยที่ IB3, DROP3, และ GA จะได้ IR ใหม่อยู่ระหว่าง 1.1 และ 98.92, 1.44 และ 123.1, และ 0.1 และ 55.72 นอกจากนี้ ระยะเวลาในการคัดเลือก instance จะใช้เวลาค่อนข้างนาน ทำให้วิธีการอาจจะไม่เหมาะกับชุดข้อมูลที่มีขนาดใหญ่ รวมถึงความเอนเอียงในการจำแนกหมวดหมู่จะยังคงมีอยู่เช่นเดิม

Behzad Mirzaei, Bahareh Nikpour, Hossein Nezamabadi-pour ได้มีการนำเสนอวิธี Clustering and Density-Based Hybrid (CHBH) เพื่อปรับสมดุลข้อมูล โดยใช้ข้อมูลการกระจายความหนาแน่นของ class เพื่อสร้างข้อมูลส่วนน้อยใหม่และลบข้อมูลส่วนใหญ่ที่ซ้ำซ้อน โดยจะเลือกข้อมูลที่อยู่ในบริเวณที่มีความหนาแน่นที่สุด เพราะมีข้อมูลมากกว่ากลุ่มอื่นๆ นอกจากนี้การจะเพิ่มข้อมูลส่วนน้อย จะทำกับชุดข้อมูลที่มีค่าอัตราความไม่สมดุล > 2 และเพิ่มข้อมูลส่วนน้อยจนมีอัตราความไม่สมดุลเท่ากับ 2 หลังจากนั้น จะทำการลดจำนวนข้อมูลส่วนใหญ่ลงจนทำให้ชุดข้อมูลนั้นมีความสมดุล แล้วจึง

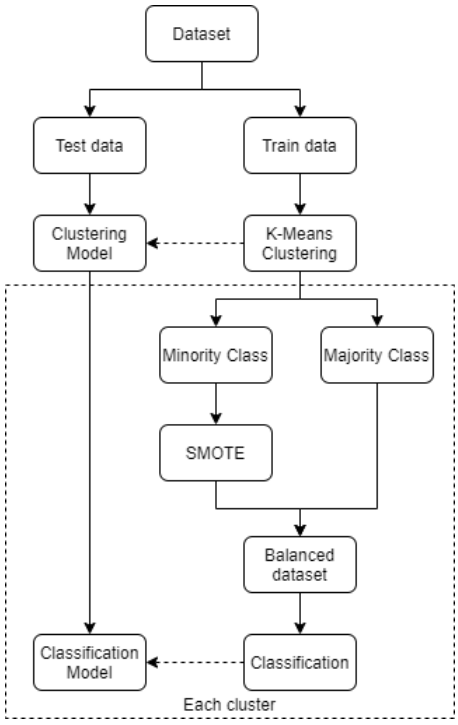
นำไปใช้ในการเรียนรู้ โดยที่ทั้งการเพิ่มจำนวนข้อมูลส่วนน้อยและการลดจำนวนข้อมูลส่วนใหญ่ จะมีการแบ่งกลุ่มข้อมูลด้วย k-means ซึ่งถ้าเป็นการเพิ่มข้อมูลส่วนน้อย จะทำการแบ่งกลุ่มข้อมูลของข้อมูลส่วนน้อย แล้วจึงทำการสุ่มเลือกกลุ่มเพื่อเพิ่มข้อมูลส่วนน้อยตามโดยการสุ่มเลือกจะใช้ความน่าจะเป็นจากความหนาแน่นของข้อมูล เมื่อสุ่มเลือกกลุ่มได้แล้ว จึงทำการเพิ่มข้อมูลส่วนน้อยด้วย SMOTE ส่วนการลดจำนวนข้อมูลส่วนใหญ่ จะทำการแบ่งกลุ่มข้อมูลส่วนใหญ่ด้วย k-means เช่นกัน แล้วจึงทำการลดข้อมูลส่วนใหญ่ลงโดยการสุ่มเลือกกลุ่มที่จะลดข้อมูลจากความน่าจะเป็นของความหนาแน่นข้อมูล ซึ่งกลุ่มข้อมูลที่มีความหนาแน่นมาก จะมีโอกาสถูกเลือกมาก จากขั้นตอนการเพิ่มข้อมูลส่วนน้อย และการลดจำนวนข้อมูลส่วนใหญ่ เมื่อข้อมูลมีความสมดุลแล้ว จึงนำชุดข้อมูลที่ได้ไปทำการเรียนรู้ การแบ่งกลุ่มข้อมูลของงานวิจัยนี้ จะเป็นการแบ่งกลุ่มเฉพาะข้อมูลส่วนใหญ่ หรือข้อมูลส่วนน้อย ซึ่งการแบ่งกลุ่มแบบนี้ไม่ได้เป็นการแบ่งกลุ่มทั้งชุดข้อมูล ทำให้การเพิ่ม หรือลดจำนวนข้อมูลไม่สอดคล้องกับข้อมูลอีกส่วนหนึ่ง ส่วนการจำแนกหมวดหมู่นั้น ข้อมูลในแต่ละกลุ่มจะมีลักษณะพฤติกรรมแตกต่างกัน เมื่อนำข้อมูลทั้งชุดไปเรียนรู้ จะทำให้ลักษณะบางอย่างของกลุ่มสูญเสียไปได้

บทที่ 3
ระเบียบวิธีวิจัย

การศึกษาวิจัยนี้ เป็นการศึกษาเพื่อหาแนวทางในการแก้ไขปัญหาความไม่สมดุลของข้อมูล โดยการเพิ่มจำนวนข้อมูลส่วนน้อยตามการแบ่งกลุ่มของข้อมูล เพื่อให้ข้อมูลมีความสมดุล และนำไปใช้ในการจำแนกหมวดหมู่ตามการแบ่งกลุ่มของข้อมูลได้มีประสิทธิภาพมากขึ้น

3.1 แนวทางการวิจัย

แนวทางการวิจัยในงานวิจัยนี้มีขั้นตอนดังภาพที่ 3.1



ภาพที่ 3.1 แผนภาพขั้นตอนในการวิจัย

3.1.1 ขั้นตอนการวิจัย

จากแผนภาพขั้นตอนในการวิจัย เริ่มจากนำข้อมูลมาแบ่งออกเป็น ข้อมูลสำหรับการเรียนรู้ (train data) และข้อมูลสำหรับการทดสอบ (test data) หลังจากนั้นจะนำข้อมูลมาปรับค่าให้อยู่ในมาตรฐานเดียวกัน เพื่อไม่ให้ค่าของข้อมูลมีผลต่อการปรับสมดุล และการจำแนกประเภท โดยการปรับค่านี้จะปรับให้มีค่าอยู่ระหว่าง 0-1 ทั้งหมด

หลังจากนั้น จะนำข้อมูลที่ได้มาใช้ในการทดลองปรับสมดุลของข้อมูล โดยใช้วิธีการ Random Undersampling, Cluster centroid undersampling, Cluster based undersampling, SMOTE, และวิธีการสุ่มตัวอย่างตามการแบ่งกลุ่มของข้อมูลที่ไม่สมดุลตามภาพที่ 3.1ซึ่งกระบวนการของวิธีการสุ่มตัวอย่างตามการแบ่งกลุ่มของข้อมูลที่ไม่สมดุลเป็นดังนี้

3.1.1.1 การแบ่งกลุ่มข้อมูลด้วย k-means

ในขั้นตอนนี้ จะนำข้อมูลสำหรับการเรียนรู้ทั้งหมดมาแบ่งกลุ่มด้วย k-means และเนื่องจากปริมาณข้อมูลส่วนน้อยมีไม่มาก ในการทดลองนี้จึงแบ่งกลุ่มออกเป็นกลุ่มย่อยเพียง 3-5 กลุ่มเท่านั้น เพื่อไม่ให้ข้อมูลในแต่ละกลุ่มมีปริมาณน้อยเกินไป เมื่อแบ่งกลุ่มเสร็จแล้ว จึงนำข้อมูลในแต่ละกลุ่มมาตรวจสอบเพื่อทำการเพิ่มจำนวนข้อมูลส่วนน้อยในแต่ละกลุ่ม โดยจะแบ่งออกมาได้ 2 กลุ่มหลักๆคือ

1. กลุ่มที่มีแต่ข้อมูลส่วนใหญ่ หรือข้อมูลส่วนน้อยเท่านั้น
2. กลุ่มที่มีทั้งข้อมูลส่วนใหญ่และข้อมูลส่วนน้อย

เมื่อแบ่งกลุ่มเสร็จแล้ว โมเดลสำหรับการแบ่งกลุ่มของข้อมูลที่ได้จะนำไปใช้ในการทดสอบกับชุดข้อมูลสำหรับทดสอบต่อไป

3.1.1.2 การเพิ่มปริมาณข้อมูลส่วนน้อยด้วย SMOTE

เมื่อนำข้อมูลสำหรับการเรียนรู้มาแบ่งกลุ่มแล้ว จึงนำข้อมูลในแต่ละกลุ่มมาทำการคัดแยกข้อมูลออกเป็นข้อมูลส่วนใหญ่ (Majority class) และข้อมูลส่วนน้อย (Minority class) และจัดการกับข้อมูลในแต่ละกลุ่มตามลักษณะของกลุ่มดังนี้

1. กลุ่มที่มีแต่ข้อมูลส่วนใหญ่ หรือข้อมูลส่วนน้อยเพียงอย่างเดียว ข้อมูลที่อยู่ในกลุ่มนี้จะไม่ทำการปรับสมดุลของปริมาณข้อมูล และจะไม่นำไปจำแนกหมวดหมู่ แต่จะถือว่าข้อมูลที่อยู่ในกลุ่มนี้เป็นข้อมูลส่วนใหญ่ หรือข้อมูลส่วนน้อยเลย

2. กลุ่มที่มีทั้งข้อมูลส่วนใหญ่และข้อมูลส่วนน้อยปนกัน จะนำมาแยกข้อมูลส่วนใหญ่และข้อมูลส่วนน้อย จากนั้นนำข้อมูลส่วนน้อยไปทำการเพิ่มปริมาณข้อมูลส่วนน้อยให้มีปริมาณเพิ่มขึ้นจนเท่ากับปริมาณข้อมูลส่วนใหญ่ในกลุ่มนั้นด้วยวิธี SMOTE ทั้งนี้ วิธีการเพิ่มจำนวนข้อมูลส่วน

น้อยด้วยวิธี SMOTE ยังมีข้อจำกัดคือ ถ้าปริมาณข้อมูลส่วนน้อยมีน้อยกว่าหรือเท่ากับค่าพารามิเตอร์จำนวนสมาชิกข้างเคียง (k_neighbors) จะไม่สามารถเพิ่มปริมาณข้อมูลได้ ในการทดลองนี้จะไม่แก้ปัญหานี้ แต่จะถือว่าข้อมูลที่อยู่ในกลุ่มนี้อยู่ในหมวดหมู่ของข้อมูลส่วนใหญ่

เนื่องจากปัญหาปริมาณข้อมูลส่วนน้อยในกลุ่มมีปริมาณน้อยมาก จึงมีการกำหนดพารามิเตอร์จำนวนสมาชิกข้างเคียงตั้งแต่ 2-3 สมาชิก เพื่อจะเพิ่มโอกาสสำหรับการเพิ่มปริมาณข้อมูลในกลุ่ม และสามารถนำไปจำแนกหมวดหมู่ต่อไป

3.1.1.3 การจำแนกหมวดหมู่ด้วยต้นไม้ตัดสินใจ (Decision Tree Classification)

เมื่อข้อมูลในแต่กลุ่มมีการปรับสมดุลแล้ว จึงนำข้อมูลในกลุ่มนั้นมาเรียนรู้เพื่อจำแนกหมวดหมู่ ซึ่งการจำแนกหมวดหมู่จะเป็นการจำแนกหมวดหมู่ที่เหมาะสมสำหรับกลุ่มนั้นๆ และโมเดลที่ได้ จะนำไปใช้สำหรับการทดสอบการจำแนกหมวดหมู่ด้วยชุดข้อมูลทดสอบต่อไป

3.1.1.4 การทดสอบด้วยชุดข้อมูลสำหรับทดสอบ

หลังจากที่ได้โมเดลการแบ่งกลุ่มและโมเดลการจำแนกหมวดหมู่แล้ว จะนำชุดข้อมูลทดสอบมาทดสอบกับโมเดลที่ได้ โดยการทดสอบนี้ จะเริ่มจากการนำชุดข้อมูลทดสอบมาทำการแบ่งกลุ่ม เพื่อให้รู้ว่าจะต้องใช้โมเดลจำแนกหมวดหมู่ใด

3.1.2 การตั้งค่าในการทดลอง

3.1.2.1 ชุดข้อมูล

ในงานวิจัยนี้ จะใช้ชุดข้อมูลจำนวน 44 ชุดข้อมูลที่ใช้โดย Galar et al. ซึ่งมีอัตราส่วนของความไม่สมดุลอยู่ระหว่าง 1.8 ถึง 129 และมีจำนวนข้อมูลระหว่าง 130 ถึง 5500 ข้อมูล ตามตารางที่ 1 โดยชุดข้อมูลจะมีการจำแนกประเภทได้ 2 ประเภท และในชุดข้อมูลแต่ละชุด มีการแบ่งเป็น 5-folds และใช้ชุดข้อมูลที่แบ่งแล้วในการทดลอง

ตารางที่ 3.1 ข้อมูลของชุดข้อมูลที่นำมาใช้

Datasets	No. of samples data	No. of features	Imbalance ratio
abalone19	4174.00	8.00	128.87
abalone9-18	731.00	8.00	16.68
ecoli-0_vs_1	220.00	7.00	1.86
ecoli-0-1-3-7_vs_2-6	281.00	7.00	39.15
ecoli1	336.00	7.00	3.36
ecoli2	336.00	7.00	5.46
ecoli3	336.00	7.00	8.19
ecoli4	336.00	7.00	13.84
glass-0-1-2-3_vs_4-5-6	192.00	9.00	10.29
glass-0-1-6_vs_2	184.00	9.00	19.44
glass-0-1-6_vs_5	214.00	9.00	1.82
glass0	214.00	9.00	3.19
glass1	214.00	9.00	10.39
glass2	214.00	9.00	15.47
glass4	214.00	9.00	22.81
glass5	214.00	9.00	22.81
glass6	214.00	9.00	6.38
haberman	306.00	3.00	2.68
iris0	150.00	4.00	2.00
new-thyroid1	215.00	5.00	5.14
new-thyroid2	215.00	5.00	4.92
page-blocks-1-3_vs_4	472.00	10.00	15.85
page-blocks0	5472.00	10.00	8.77
pima	768.00	8.00	1.90

ตารางที่ 3.1 (ต่อ)

Datasets	No. of samples data	No. of features	Imbalance ratio
segment0	2308.00	19.00	6.01
shuttle-c0-vs-c4	1829.00	9.00	13.87
shuttle-c2-vs-c4	129.00	9.00	20.50
vehicle0	846.00	18.00	3.23
vehicle1	846.00	18.00	2.52
vehicle2	846.00	18.00	2.52
vehicle3	846.00	18.00	2.52
vowel0	988.00	13.00	10.10
wisconsin	683.00	9.00	1.86
yeast-0-5-6-7-9_vs_4	528.00	8.00	9.35
yeast-1_vs_7	459.00	8.00	13.87
yeast-1-2-8-9_vs_7	947.00	8.00	30.56
yeast-1-4-5-8_vs_7	693.00	8.00	22.10
yeast-2_vs_4	514.00	8.00	9.08
yeast-2_vs_8	482.00	8.00	23.10
yeast1	1484.00	8.00	2.46
yeast3	1484.00	8.00	8.11
yeast4	1484.00	8.00	28.41
yeast5	1484.00	8.00	32.78
yeast6	1484.00	8.00	39.15

3.1.2.2 การแบ่งกลุ่มข้อมูล

ในงานวิจัยนี้ จะใช้การแบ่งกลุ่มด้วย k-means เป็นพื้นฐานสำหรับทุกๆขั้นตอนและวิธีการที่ต้องการแบ่งกลุ่มข้อมูล รวมถึงพารามิเตอร์ที่จะใช้สำหรับ k-means จะเป็นพารามิเตอร์เดียวกันทั้งหมดด้วยเช่นกัน โดยพารามิเตอร์ที่มีการปรับเปลี่ยนเพื่อทดสอบค่าที่เหมาะสมนั้น จะมีพารามิเตอร์จำนวนกลุ่มที่ต้องการแบ่ง (k_cluster) ซึ่งจะมีค่า 3-9

3.1.2.3 การเพิ่มปริมาณข้อมูลส่วนน้อยด้วย SMOTE

การเพิ่มปริมาณข้อมูลส่วนน้อยที่ใช้ในการทดลองนี้ จะใช้วิธี SMOTE เป็นพื้นฐานและพารามิเตอร์ที่ใช้ในการทดลองนี้ จะเป็นพารามิเตอร์เดียวกันทั้งหมดสำหรับทุกๆการทดสอบในแต่ละวิธีการ ส่วนพารามิเตอร์จำนวนสมาชิกข้างเคียง (k_neighbors) นั้น จะเป็นพารามิเตอร์ที่สามารถปรับเปลี่ยนเพื่อทดสอบหาค่าที่เหมาะสม ซึ่งในการทดลองนี้จะมีการกำหนดค่าตั้งแต่ 1-3 สมาชิก และยังคงควบคุมพารามิเตอร์นี้ให้มีค่าเดียวกันกับทุกๆการทดสอบในแต่ละวิธีการด้วยเช่นกัน

3.1.2.4 การจำแนกหมวดหมู่

ในกาทดลองนี้จะใช้ต้นไม้ตัดสินใจในการจำแนกหมวดหมู่เพียงอย่างเดียว และพารามิเตอร์ที่ใช้ในแต่ละวิธีการ จะเป็นพารามิเตอร์เดียวกันทั้งหมด รวมถึงต้นไม้ตัดสินใจที่ใช้ในแต่ละกลุ่มของข้อมูลของวิธีการสุ่มตัวอย่างตามการแบ่งกลุ่มของข้อมูลที่ไม่สมดุลด้วยเช่นกัน

3.1.3 การวัดประสิทธิภาพ

เมื่อทำการทดสอบโมเดลที่ได้กับชุดข้อมูลทดสอบ จะนำผลลัพธ์ที่ได้จากการจำแนกประเภทมาเปรียบเทียบกับค่าจริงของข้อมูลทดสอบ ทำให้สามารถวัดประสิทธิภาพของโมเดลได้ โดยค่าที่ใช้นี้ จะเป็นค่าต่างๆที่คำนวณได้มาจาก confusion matrix ที่แสดงตามตารางที่ 2

ตารางที่ 3.2 Confusion matrix

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

จากค่าใน confusion matrix สามารถนำมาคำนวณค่าความถูกต้อง)accuracy) ค่าความแม่นยำ)precision) ค่าการตรวจสอบ)recall) ค่า F-measure รวมถึงพื้นที่ใต้กราฟที่กำหนดได้ดังนี้

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

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

$$Recall = \frac{TP}{(TP + FN)}$$

$$F - measure = \frac{(2 * Precision * Recall)}{(Precision + Recall)}$$

3.2 เครื่องมือที่ใช้ในการวิจัย

เครื่องมือที่ใช้ในการวิจัยนี้ เป็นเครื่องมือที่เป็น OpenSource และสามารถดาวน์โหลดมาใช้ในเครื่องคอมพิวเตอร์ส่วนบุคคลทั่วไปได้ ไม่จำเป็นต้องใช้อุปกรณ์อื่นใดเพิ่มเติม เว้นแต่ต้องการให้การคำนวณทางคณิตศาสตร์ทำได้เร็วขึ้น สามารถใช้เครื่องคอมพิวเตอร์ที่มีหน่วยประมวลผลกราฟิกส์แทนได้ เครื่องมือต่างๆที่ใช้มีดังนี้

3.2.1 Anaconda3

Anaconda3 เป็นเครื่องมือสำหรับการพัฒนาแอปพลิเคชั่นของการทำเหมืองข้อมูล (data mining) ซึ่งใน Anaconda3 จะมีไลบรารี่ (library) ที่เกี่ยวกับการทำเหมืองข้อมูล ทำให้การพัฒนาสามารถทำได้ง่าย

3.2.2 Jupyter-Python

Jupyter-Python เป็นเครื่องมือสำหรับนักพัฒนาแอปพลิเคชั่นที่สามารถใช้งานผ่านทางบราวเซอร์ (browse) และสามารถประมวลผลทีละบรรทัด หรือทีละหลายๆบรรทัดได้ นอกจากนี้ยังสามารถแก้ไขบรรทัดที่ผิด หรือแก้ไขซอร์สโค้ด (source code) แล้วสามารถประมวลผลเฉพาะกลุ่ม (block) ที่แก้ไขได้

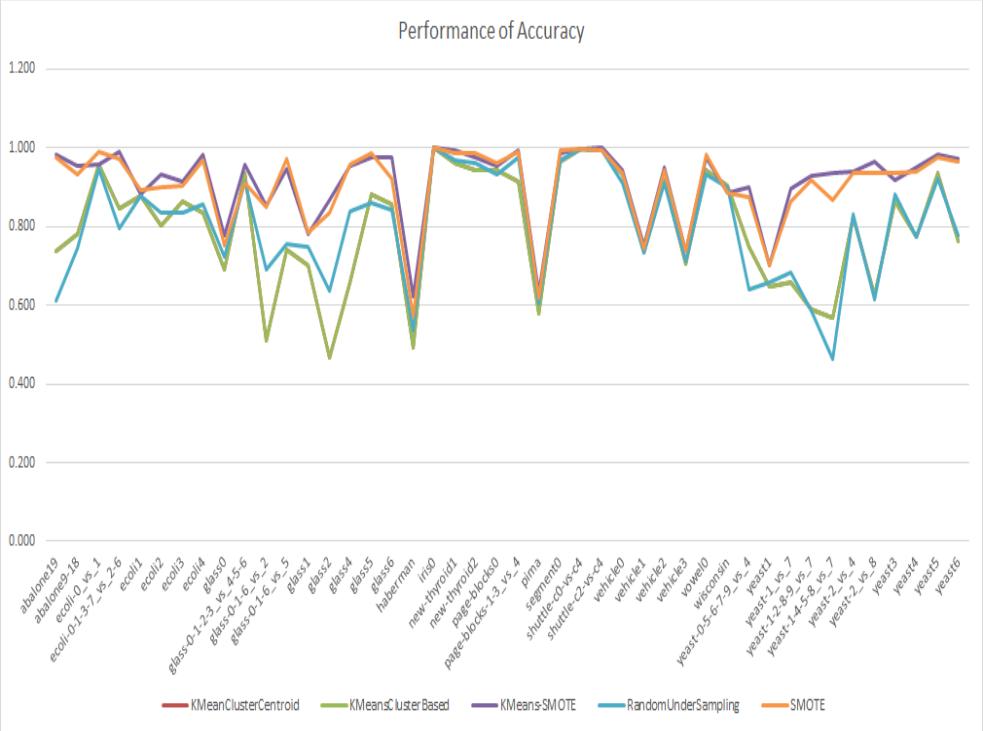
บทที่ 4
ผลการศึกษา

จากการวิจัยวิธีการสุ่มตัวอย่างตามการแบ่งกลุ่มของข้อมูลที่ไม่สมดุลนี้ จะเป็นแนวทางในการเพิ่มประสิทธิภาพการจำแนกหมวดหมู่ข้อมูลโดยการปรับสมดุลชุดข้อมูล ก่อนที่จะนำชุดข้อมูลที่มีการปรับปรุงสมดุลแล้ว ไปใช้การเรียนรู้ และในงานวิจัยนี้ ได้มีการทดลองกับชุดข้อมูลทั้งหมด 44 ชุดข้อมูล โดยแต่ละชุดข้อมูลจะมีอัตราส่วนของความไม่สมดุลแตกต่างกันไป ตั้งแต่ 1.8 จนถึง 129 และผลลัพธ์ของประสิทธิภาพการปรับสมดุลของข้อมูลจะไม่สามารถวัดได้โดยตรง แต่จะวัดจากประสิทธิภาพการจำแนกหมวดหมู่ด้วยต้นไม้ตัดสินใจ โดยนำชุดข้อมูลที่มีการปรับสมดุลด้วยวิธีต่างๆ มาทำการเรียนรู้การจำแนกหมวดหมู่ข้อมูลด้วยวิธีการเดียวกันและพารามิเตอร์เดียวกัน

ส่วนการวัดประสิทธิภาพการจำแนกหมวดหมู่ข้อมูล จะใช้การมาตราวัดทั้งหมด 3 แบบคือ ค่าความถูกต้อง (Accuracy), ค่าความแม่นยำ (F-measure), และพื้นที่ใต้กราฟ (AUC)

4.1 ผลการทดสอบประสิทธิภาพ

จากการทดลองการปรับสมดุลด้วยวิธีต่างๆ แล้วนำมาชุดข้อมูลนั้นมาเรียนรู้ และวัดประสิทธิภาพจากการเรียนรู้นั้น เมื่อวัดจากค่าความถูกต้อง (Accuracy) จะได้ดังภาพที่ 4.1 ได้



ภาพที่ 4.1 ค่าเฉลี่ยความถูกต้อง (Accuracy) จากประสิทธิภาพของการปรับสมดุล

ตารางที่ 4.1 ตารางเปรียบเทียบประสิทธิความถูกต้องของแต่ละวิธีเทียบกับชุดข้อมูล

Dataset	Kmean ClusterCentroid	Kmeans ClusterBased	KMSM	Random UnderSampling	SMOTE
abalone19	0.737	0.737	0.984	0.611	0.976
abalone9-18	0.780	0.780	0.954	0.746	0.932
ecoli-0_vs_1	0.959	0.959	0.959	0.945	0.991
ecoli-0-1-3-7_vs_2-6	0.847	0.847	0.989	0.797	0.971
ecoli1	0.878	0.878	0.884	0.878	0.893
ecoli2	0.804	0.804	0.931	0.837	0.902
ecoli3	0.864	0.864	0.914	0.834	0.905
ecoli4	0.834	0.834	0.985	0.858	0.967
glass0	0.692	0.692	0.776	0.724	0.753
glass-0-1-2-3_vs_4-5-6	0.935	0.935	0.958	0.916	0.911
glass-0-1-6_vs_2	0.511	0.511	0.854	0.692	0.849
glass-0-1-6_vs_5	0.740	0.740	0.946	0.756	0.973
glass1	0.700	0.700	0.780	0.748	0.785
glass2	0.467	0.467	0.869	0.635	0.836
glass4	0.659	0.659	0.953	0.840	0.958
glass5	0.883	0.883	0.977	0.860	0.986
glass6	0.855	0.855	0.977	0.841	0.920
haberman	0.494	0.494	0.621	0.536	0.572
iris0	1.000	1.000	1.000	1.000	1.000
new-thyroid1	0.963	0.963	0.995	0.968	0.986
new-thyroid2	0.944	0.944	0.977	0.963	0.986
page-blocks0	0.944	0.944	0.955	0.933	0.960
page-blocks-1-3_vs_4	0.913	0.913	0.994	0.975	0.991
pima	0.579	0.579	0.631	0.604	0.618
segment0	0.964	0.964	0.987	0.970	0.994
shuttle-c0-vs-c4	0.998	0.998	0.999	0.998	0.998
shuttle-c2-vs-c4	0.992	0.992	1.000	0.992	0.992
vehicle0	0.932	0.932	0.942	0.911	0.937
vehicle1	0.733	0.733	0.751	0.733	0.746
vehicle2	0.917	0.917	0.950	0.911	0.945
vehicle3	0.704	0.704	0.738	0.713	0.736
vowel0	0.944	0.944	0.977	0.933	0.984
wisconsin	0.903	0.903	0.884	0.899	0.887
yeast-0-5-6-7-9_vs_4	0.750	0.750	0.902	0.642	0.877
yeast1	0.646	0.646	0.702	0.659	0.701
yeast-1_vs_7	0.658	0.658	0.895	0.684	0.863
yeast-1-2-8-9_vs_7	0.592	0.592	0.930	0.585	0.917
yeast-1-4-5-8_vs_7	0.570	0.570	0.936	0.462	0.869
yeast-2_vs_4	0.825	0.825	0.941	0.832	0.938
yeast-2_vs_8	0.625	0.625	0.965	0.614	0.935
yeast3	0.866	0.866	0.917	0.884	0.937
yeast4	0.773	0.773	0.950	0.774	0.939
yeast5	0.937	0.937	0.982	0.921	0.977
yeast6	0.764	0.764	0.974	0.776	0.965
Average	**0.797**	**0.797**	**0.913**	**0.804**	**0.904**

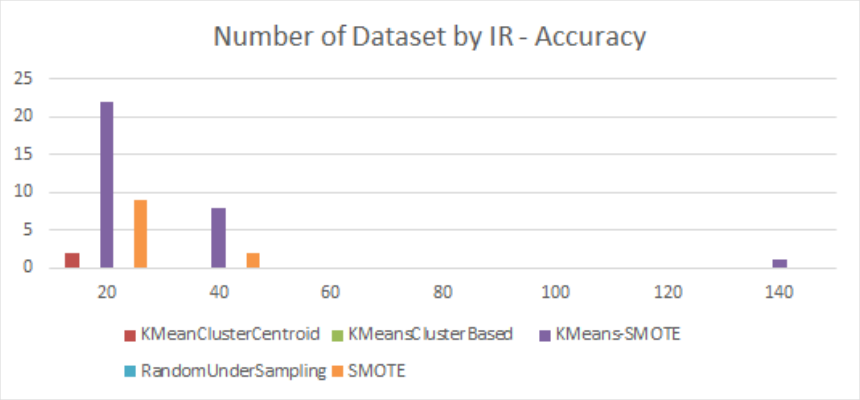
จากภาพที่ 4.1 และตารางที่ 4.1 ประสิทธิภาพความถูกต้องเฉลี่ยของการจำแนกหมวดหมู่ด้วยต้นไม้ตัดสินใจ ที่ใช้ชุดข้อมูลที่มีการปรับสมดุลด้วยวิธีการสุ่มตัวอย่างตามการแบ่งกลุ่มของข้อมูลจะมีค่าความถูกต้องสูงกว่าการปรับสมดุลโดยการเพิ่มจำนวนข้อมูลส่วนน้อยด้วยวิธี SMOTE ประมาณ 1.068% เพราะเนื่องจาก 3 ปัจจัยคือ

1. การแบ่งกลุ่มของข้อมูล มีโอกาสทำให้ข้อมูลบางกลุ่มมีแต่ข้อมูลส่วนใหญ่ ซึ่งทำให้ไม่จำเป็นต้องทำการจำแนกหมวดหมู่ และสามารถทำนายได้ว่า ข้อมูลที่อยู่ในกลุ่มนั้น จะเป็นอยู่ที่หมวดหมู่ใดหมวดหมู่หนึ่งได้ทันที ทำให้มีความถูกต้องสูงขึ้น

2. ข้อมูลที่มีการแบ่งกลุ่มแล้ว ในกรณีที่กลุ่มนั้นๆมีทั้งข้อมูลส่วนใหญ่และข้อมูลส่วนน้อย ทำให้ไม่จำเป็นต้องเพิ่มปริมาณข้อมูลส่วนน้อยในปริมาณมาก ซึ่งมีโอกาสที่ข้อมูลที่เพิ่มขึ้นไม่สอดคล้องกับความเป็นจริงได้

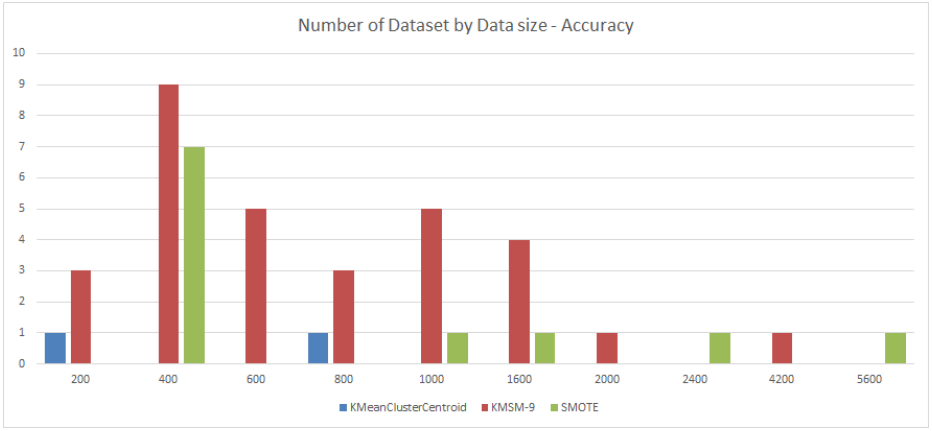
3. โมเดลต้นไม้ตัดสินใจที่ใช้ในแต่ละกลุ่มจะเหมาะสมกับกลุ่มนั้นๆมากกว่า เมื่อเทียบกับโมเดลต้นไม้ตัดสินใจที่ทำกับข้อมูลทั้งชุด

นอกจากนี้ เมื่อเทียบประสิทธิภาพความถูกต้องเฉลี่ยกับชุดข้อมูลที่มีการปรับสมดุลด้วยการลดจำนวนข้อมูลส่วนใหญ่ ของวิธี Cluster Centroid, Cluster Base, และ Random Undersampling ซึ่งประสิทธิภาพของการปรับสมดุลด้วยวิธีการสุ่มตัวอย่างตามการแบ่งกลุ่มของข้อมูลจะสูงกว่าประมาณ 14.565%, 14.565%, และ 13.534% ตามลำดับ ประสิทธิภาพที่เพิ่มขึ้นเพราะการเพิ่มจำนวนข้อมูลส่วนน้อย จะทำให้ปริมาณข้อมูลที่ใช้ในการเรียนรู้มีเพิ่มขึ้น แต่การปรับสมดุลด้วยการลดจำนวนข้อมูลส่วนใหญ่ อาจส่งผลให้สูญเสียข้อมูลที่สำคัญของข้อมูลส่วนใหญ่ไป จึงทำให้การปรับสมดุลด้วยวิธีการสุ่มตัวอย่างตามการแบ่งกลุ่มของข้อมูลมีประสิทธิภาพความถูกต้องสูงกว่าการปรับสมดุลโดยการลดจำนวนข้อมูลส่วนใหญ่มาก



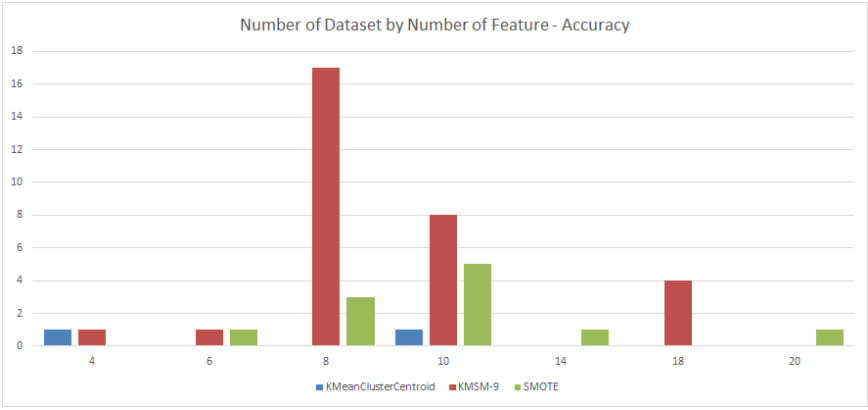
ภาพที่ 4.2 กราฟสรุปการจัดอันดับของจำนวนชุดข้อมูลค่าเฉลี่ยความถูกต้อง (Accuracy) จากประสิทธิภาพของการปรับสมดุล

จากภาพที่ 4.2 จะเห็นว่า การปรับสมดุลด้วยวิธีการสุ่มตัวอย่างตามการแบ่งกลุ่มของข้อมูลจะให้ประสิทธิภาพความถูกต้องดีกว่าการปรับสมดุลด้วยวิธีอื่นๆในทุกช่วงของอัตราความไม่สมดุลของข้อมูล แต่เนื่องจากชุดข้อมูลที่ใช้ในการวิจัยนี้ ส่วนใหญ่เป็นมีอัตราความไม่สมดุลอยู่ในช่วง 1-20 มากที่สุด แต่อัตราความไม่สมดุลในช่วง 40-120 ไม่มีเลย ทำให้ไม่มีวิธีใดที่ให้ประสิทธิภาพความถูกต้องในช่วงของอัตราความไม่สมดุลนี้

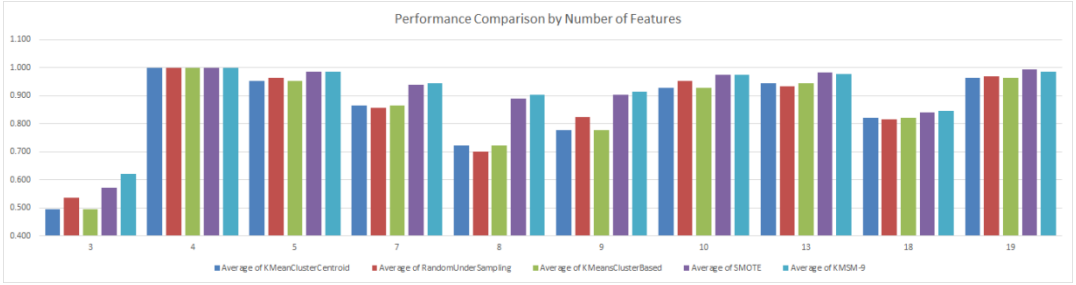


ภาพที่ 4.3 กราฟสรุปการจัดอันดับของจำนวนชุดข้อมูลค่าเฉลี่ยความถูกต้อง (Accuracy) จากประสิทธิภาพของการปรับสมดุลเทียบกับขนาดของชุดข้อมูล

จากภาพที่ 4.3 การปรับสมดุลด้วยวิธีการสุ่มตัวอย่างตามการแบ่งกลุ่มของข้อมูลจะให้ประสิทธิภาพความถูกต้องดีกว่าการปรับสมดุลด้วยวิธีอื่นๆในเกือบทุกช่วงของขนาดของข้อมูล โดยที่ชุดข้อมูลที่ใช้ในงานวิจัยนี้ มีค่าตั้งแต่ 129 – 5474 ข้อมูล แต่เนื่องจากชุดข้อมูลที่มีจำนวนข้อมูลตั้งแต่ 2000 ตัวอย่าง มีเพียง 3 ชุดข้อมูล จึงไม่สามารถระบุได้ชัดเจนว่า วิธี KMSM นี้สามารถให้ประสิทธิภาพการจำแนกข้อมูลได้ดี แต่หากเป็นชุดข้อมูลที่มีจำนวนข้อมูลไม่เกิน 2000 ตัวอย่างข้อมูล ซึ่งมีจำนวน 40 ชุดข้อมูล และในแต่ละช่วงของจำนวนข้อมูลนั้น KMSM สามารถให้ประสิทธิภาพได้สูงกว่าวิธีการอื่น

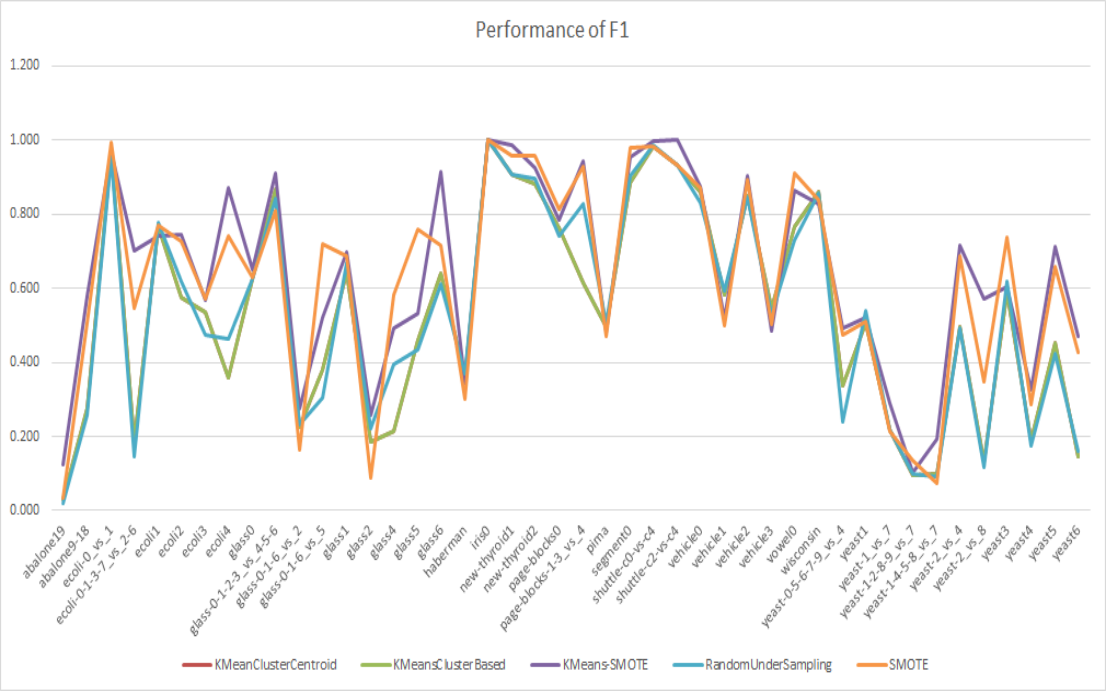


ภาพที่ 4.4 กราฟสรุปการจัดอันดับของจำนวนชุดข้อมูลค่าเฉลี่ยความถูกต้อง (Accuracy) จากประสิทธิภาพของการปรับสมดุลเทียบกับจำนวน feature



ภาพที่ 4.5 กราฟเปรียบเทียบค่าเฉลี่ยความถูกต้อง (Accuracy) เทียบกับจำนวน feature

จากภาพที่ 4.4 และภาพที่ 4.5 การปรับสมดุลด้วยวิธีการสุ่มตัวอย่างตามการแบ่งกลุ่มของข้อมูลจะให้ประสิทธิภาพความถูกต้องดีกว่าการปรับสมดุลด้วยวิธีอื่นๆในเกือบทุกช่วงของจำนวน feature ของข้อมูล โดยที่ชุดข้อมูลที่ใช้ในงานวิจัยนี้ มีจำนวน feature ตั้งแต่ 3-19 feature ทั้งนี้โดยเฉพาะในช่วงระหว่าง 7-18 features ประสิทธิภาพการจำแนกข้อมูล วิธี KMSM นั้น ให้ประสิทธิภาพในการจำแนกข้อมูลได้ดีกว่าวิธีอื่นๆ โดยที่ชุดข้อมูลที่มีจำนวน 8 feature จะให้ประสิทธิภาพความถูกต้องสูงกว่าวิธี SMOTE 1.76%

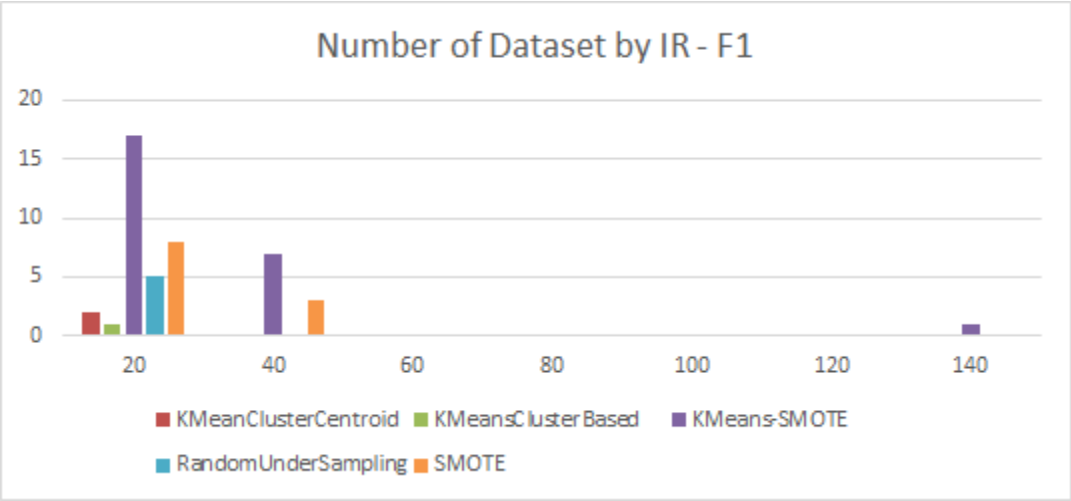


ภาพที่ 4.6 ค่าเฉลี่ยความแม่นยำ (F-measure) จากประสิทธิภาพของการปรับสมดุล

ตารางที่ 4.2 ตารางเปรียบเทียบประสิทธิความแม่นยำ (F-measure) ของแต่ละวิธีเทียบกับชุดข้อมูล

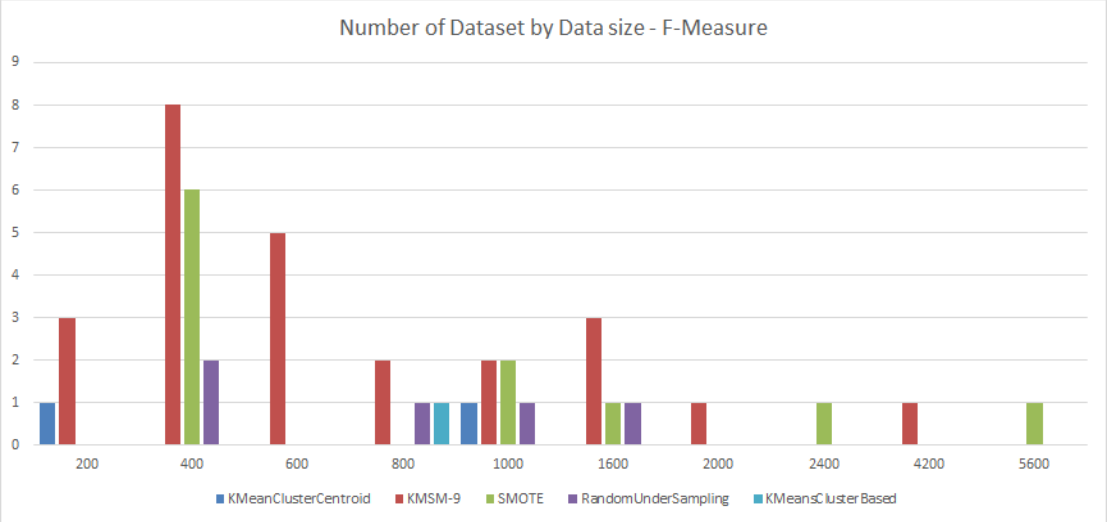
Dataset	Kmean ClusterCentroid	Kmeans ClusterBased	KMSM	Random UnderSampling	SMOTE
abalone19	0.029	0.029	0.126	0.021	0.036
abalone9-18	0.274	0.274	0.579	0.257	0.509
ecoli-0_vs_1	0.968	0.968	0.969	0.956	0.993
ecoli-0-1-3-7_vs_2-6	0.186	0.186	0.700	0.145	0.547
ecoli1	0.766	0.766	0.742	0.776	0.772
ecoli2	0.576	0.576	0.746	0.620	0.726
ecoli3	0.535	0.535	0.569	0.474	0.570
ecoli4	0.357	0.357	0.871	0.463	0.741
glass0	0.625	0.625	0.653	0.627	0.630
glass-0-1-2-3_vs_4-5-6	0.866	0.866	0.912	0.843	0.810
glass-0-1-6_vs_2	0.226	0.226	0.277	0.232	0.164
glass-0-1-6_vs_5	0.382	0.382	0.520	0.303	0.720
glass1	0.645	0.645	0.699	0.670	0.686
glass2	0.186	0.186	0.258	0.220	0.090
glass4	0.213	0.213	0.494	0.394	0.581
glass5	0.461	0.461	0.533	0.433	0.760
glass6	0.642	0.642	0.915	0.611	0.716
haberman	0.361	0.361	0.344	0.370	0.303
iris0	1.000	1.000	1.000	1.000	1.000
new-thyroid1	0.907	0.907	0.987	0.909	0.960
new-thyroid2	0.883	0.883	0.925	0.898	0.958
page-blocks0	0.762	0.762	0.786	0.741	0.813
page-blocks-1-3_vs_4	0.616	0.616	0.942	0.827	0.930
pima	0.495	0.495	0.483	0.508	0.472
segment0	0.886	0.886	0.955	0.902	0.979
shuttle-c0-vs-c4	0.984	0.984	0.996	0.988	0.984
shuttle-c2-vs-c4	0.933	0.933	1.000	0.933	0.933
vehicle0	0.862	0.862	0.875	0.833	0.870
vehicle1	0.583	0.583	0.518	0.594	0.501
vehicle2	0.851	0.851	0.902	0.846	0.893
vehicle3	0.548	0.548	0.487	0.530	0.502
vowel0	0.766	0.766	0.864	0.730	0.913
wisconsin	0.862	0.862	0.828	0.857	0.840
yeast-0-5-6-7-9_vs_4	0.339	0.339	0.494	0.241	0.474
yeast1	0.506	0.506	0.520	0.540	0.511
yeast-1_vs_7	0.219	0.219	0.291	0.219	0.214
yeast-1-2-8-9_vs_7	0.095	0.095	0.101	0.099	0.136
yeast-1-4-5-8_vs_7	0.098	0.098	0.192	0.093	0.073
yeast-2_vs_4	0.496	0.496	0.714	0.491	0.689
yeast-2_vs_8	0.130	0.130	0.571	0.118	0.350
yeast3	0.589	0.589	0.605	0.618	0.737
yeast4	0.191	0.191	0.326	0.176	0.288
yeast5	0.454	0.454	0.712	0.425	0.660
yeast6	0.146	0.146	0.470	0.160	0.429
Average	**0.534**	**0.534**	**0.647**	**0.539**	**0.624**

จากภาพที่ 4.6 และตารางที่ 4.2 ประสิทธิภาพความแม่นยำ (F-measure) เฉลี่ยของการจำแนกหมวดหมู่ด้วยต้นไม้ตัดสินใจ ที่ใช้ชุดข้อมูลที่มีการปรับสมดุลด้วยวิธีการสุ่มตัวอย่างตามการแบ่งกลุ่มของข้อมูล จะมีค่าความแม่นยำสูงกว่าการปรับสมดุลโดยการเพิ่มจำนวนข้อมูลส่วนน้อยด้วยวิธี SMOTE ประมาณ 3.609% เพราะเนื่องจาก 3 ปัจจัยเช่นเดียวกับประสิทธิภาพความถูกต้อง (Accuracy) เฉลี่ย



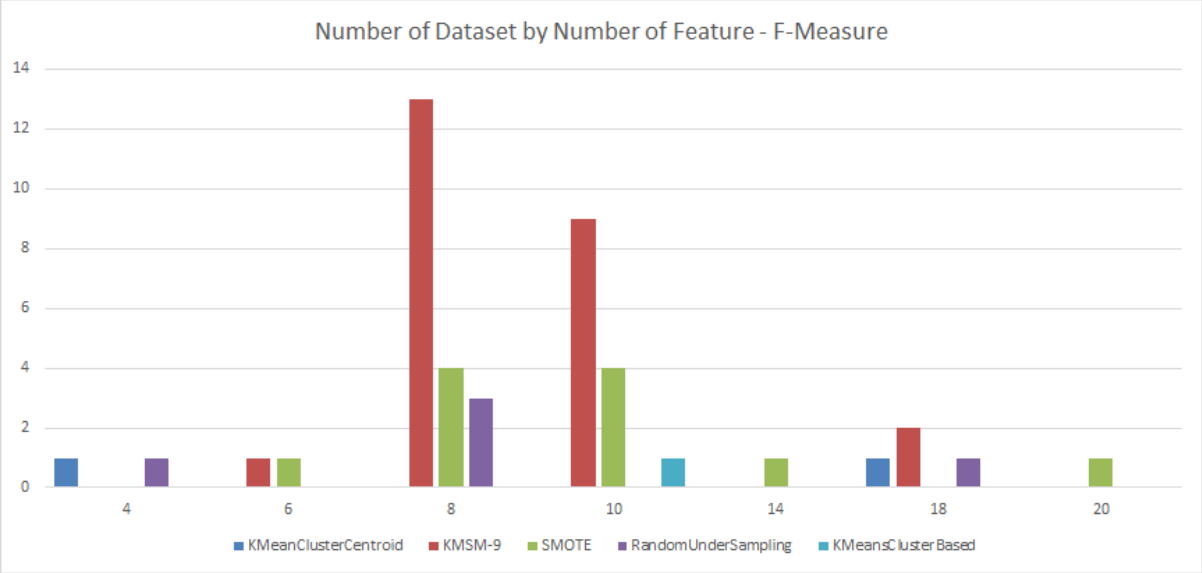
ภาพที่ 4.7 กราฟสรุปการจัดอันดับของจำนวนชุดข้อมูลค่าเฉลี่ยความแม่นยำ (F-measures) จากประสิทธิภาพของการปรับสมดุล

จากภาพที่ 4.7 จะเห็นว่า การปรับสมดุลด้วยวิธีการสุ่มตัวอย่างตามการแบ่งกลุ่มของข้อมูลจะให้ประสิทธิภาพความแม่นยำดีกว่าการปรับสมดุลด้วยวิธีอื่นๆในทุกช่วงของอัตราความไม่สมดุลของข้อมูลเช่นเดียวกับประสิทธิภาพความถูกต้อง และประสิทธิภาพความแม่นยำสำหรับชุดข้อมูลที่มีอัตราความไม่สมดุลสูง ยังคงให้ประสิทธิภาพดีกว่าวิธีอื่นเช่นเดียวกัน



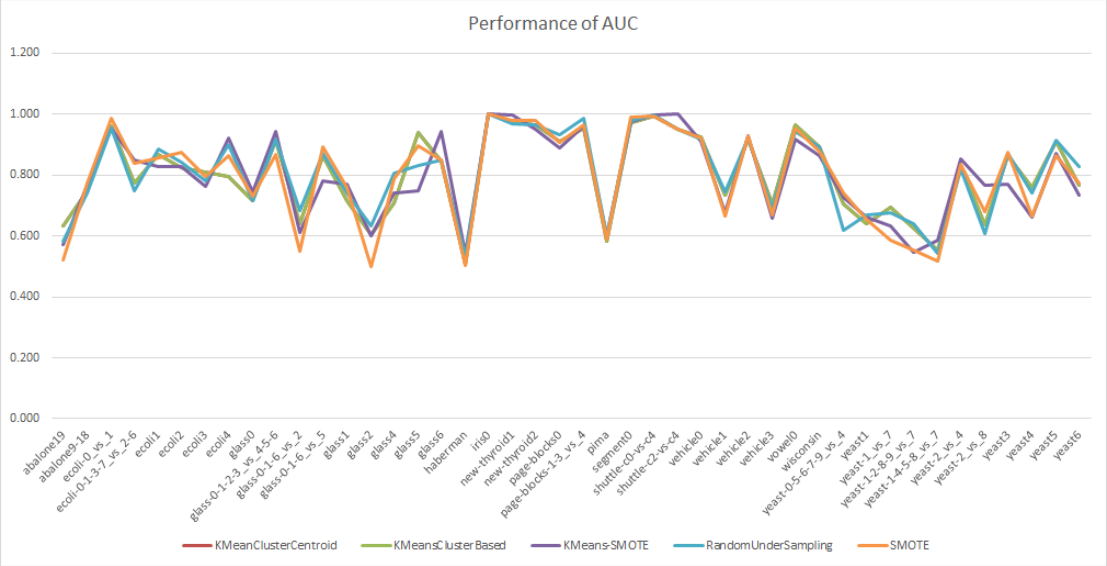
ภาพที่ 4.8 กราฟสรุปการจัดอันดับของจำนวนชุดข้อมูลค่าเฉลี่ยความแม่นยำ (F-Measure) จากประสิทธิภาพของการปรับสมดุลเทียบกับขนาดของชุดข้อมูล

จากภาพที่ 4.8 การปรับสมดุลด้วยวิธีการสุ่มตัวอย่างตามการแบ่งกลุ่มของข้อมูลจะให้ประสิทธิภาพความแม่นยำดีกว่าการปรับสมดุลด้วยวิธีอื่นๆในเกือบทุกช่วงของขนาดของข้อมูล แต่ชุดข้อมูลที่มีจำนวนข้อมูลตั้งแต่ 2000 ตัวอย่างนั้นไม่สามารถระบุได้ชัดเจนว่า วิธี KMSM นี้สามารถให้ประสิทธิภาพการจำแนกข้อมูลได้ดี แต่หากเป็นชุดข้อมูลที่มีจำนวนข้อมูลไม่เกิน 2000 ตัวอย่างข้อมูล ซึ่งมีจำนวน 40 ชุดข้อมูล และในแต่ละช่วงของจำนวนข้อมูลนั้น KMSM สามารถให้ประสิทธิภาพได้สูงกว่าวิธีการอื่นอย่างชัดเจน



ภาพที่ 4.9 กราฟสรุปการจัดอันดับของจำนวนชุดข้อมูลค่าเฉลี่ยความแม่นยำ (F-Measure) จากประสิทธิภาพของการปรับสมดุลเทียบกับจำนวน feature

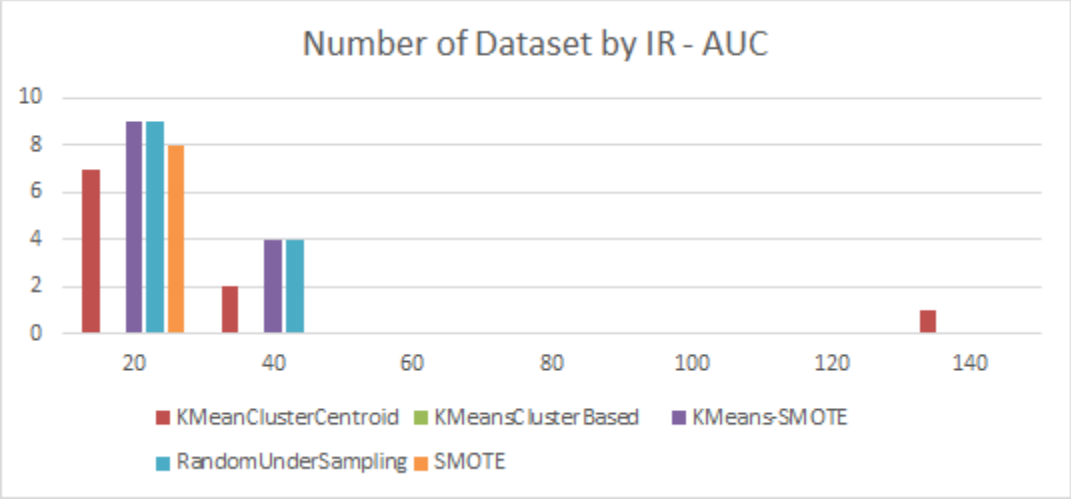
จากภาพที่ 4.9 การปรับสมดุลด้วยวิธีการสุ่มตัวอย่างตามการแบ่งกลุ่มของข้อมูลจะให้ประสิทธิภาพความถูกต้องดีกว่าการปรับสมดุลด้วยวิธีอื่นๆในเกือบทุกช่วงของจำนวน feature ของข้อมูล โดยเฉพาะในช่วงระหว่าง 5-18 features ประสิทธิภาพการจำแนกข้อมูล วิธี KMSM นั้น ให้ประสิทธิภาพในการจำแนกข้อมูลได้ดีกว่าวิธีอื่นๆอย่างชัดเจน



ภาพที่ 4.10 ค่าเฉลี่ยพื้นที่ใต้กราฟ (AUC) จากประสิทธิภาพของการปรับสมดุล

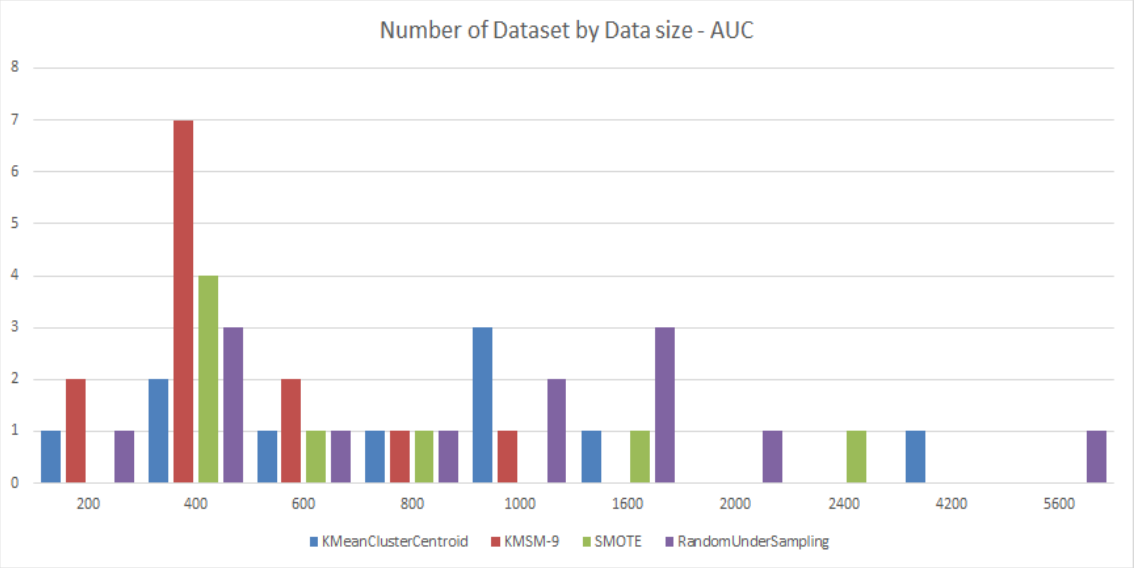
จากภาพที่ 4.10 ประสิทธิภาพพื้นที่ใต้กราฟ (AUC) เฉลี่ยของการจำแนกหมวดหมู่ด้วยต้นไม้ตัดสินใจ ที่ใช้ชุดข้อมูลที่มีการปรับสมดุลด้วยวิธีการสุ่มตัวอย่างตามการแบ่งกลุ่มของข้อมูล จะมีค่าพื้นที่ใต้กราฟสูงกว่าการปรับสมดุลโดยการเพิ่มจำนวนข้อมูลส่วนน้อยด้วยวิธี SMOTE ประมาณ 0.205% เพราะเนื่องจาก 3 ปัจจัยเช่นเดียวกับประสิทธิภาพความถูกต้องเฉลี่ย

นอกจากนี้ เมื่อเทียบประสิทธิภาพพื้นที่ใต้กราฟเฉลี่ยกับชุดข้อมูลที่มีการปรับสมดุลด้วยการลดจำนวนข้อมูลส่วนใหญ่ ของวิธี Cluster Centroid, Cluster Base, และ Random Undersampling ซึ่งประสิทธิภาพของการปรับสมดุลด้วยวิธีการสุ่มตัวอย่างตามการแบ่งกลุ่มของข้อมูลจะน้อยกว่าประมาณ 0.719%, 0.719%, และ 1.123% ตามลำดับ ประสิทธิภาพที่ลดลงเพราะการแบ่งกลุ่มของข้อมูลนั้น มีโอกาสทำให้ในบางกลุ่มมีปริมาณข้อมูลส่วนน้อยนั้น น้อยเกินไป ทำให้ไม่สามารถเพิ่มจำนวนข้อมูลด้วยวิธี SMOTE ได้ และเมื่อนำข้อมูลกลุ่มนี้ไปเรียนรู้โดยไม่มีการปรับสมดุล จะทำให้เกิดความเอนเอียงสูงมาก ในงานวิจัยนี้จึงปรับปรุงการจำแนกหมวดหมู่ของข้อมูลในกลุ่มนี้โดยการกำหนดให้ข้อมูลที่อยู่ในกลุ่มนี้ อยู่ในหมวดหมู่ของข้อมูลส่วนใหญ่



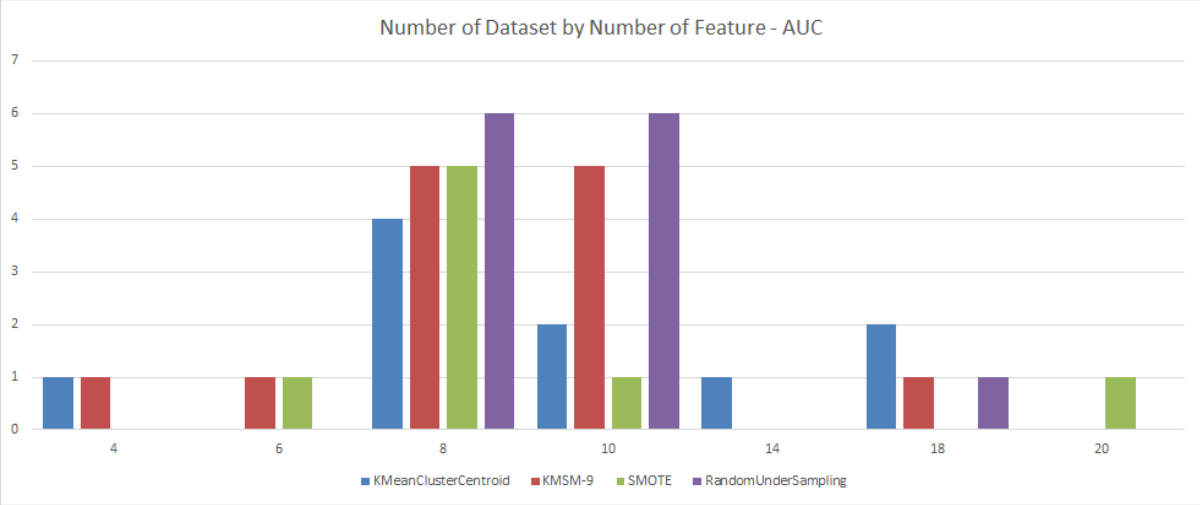
ภาพที่ 4.11 กราฟสรุปการจัดอันดับของจำนวนชุดข้อมูลค่าเฉลี่ยพื้นที่ใต้กราฟ (AUC) จากประสิทธิภาพของการปรับสมดุล

จากภาพที่ 4.11 จะเห็นว่า การปรับสมดุลด้วยวิธีการสุ่มตัวอย่างตามการแบ่งกลุ่มของข้อมูลจะให้ประสิทธิภาพพื้นที่ใต้กราฟใกล้เคียงกับการปรับสมดุลด้วยวิธีอื่นๆในช่วงของอัตราความไม่สมดุลของข้อมูลที่ 1-40 แต่ชุดข้อมูลที่มีอัตราความไม่สมดุลสูง (มากกว่า 120) มีเพียงชุดข้อมูลเดียวแต่ประสิทธิภาพพื้นที่ใต้กราฟสำหรับชุดข้อมูลที่มีอัตราความไม่สมดุลสูงนี้ วิธีการปรับสมดุลด้วยวิธี Cluster Centroid ให้ประสิทธิภาพดีกว่าวิธีอื่น



ภาพที่ 4.12 กราฟสรุปการจัดอันดับของจำนวนชุดข้อมูลค่าเฉลี่ยพื้นที่ใต้กราฟ (AUC) จากประสิทธิภาพของการปรับสมดุลเทียบกับขนาดของชุดข้อมูล

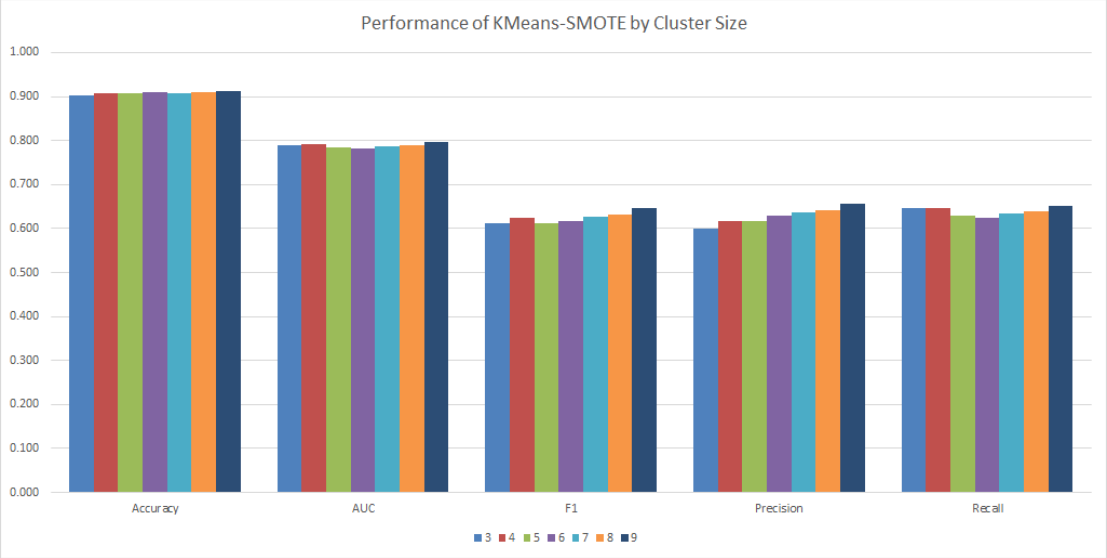
จากภาพที่ 4.12 การปรับสมดุลด้วยวิธีการสุ่มตัวอย่างตามการแบ่งกลุ่มของข้อมูลจะให้ประสิทธิภาพพื้นที่ใต้กราฟดีกว่าการปรับสมดุลด้วยวิธีอื่นๆในช่วงของขนาดของข้อมูลที่ไม่มาก แต่ชุดข้อมูลที่มีจำนวนข้อมูลตั้งแต่ 2000 ตัวอย่างนั้นไม่สามารถระบุได้ชัดเจนว่า วิธี KMSM นี้สามารถให้ประสิทธิภาพการจำแนกข้อมูลได้ดี แต่หากเป็นชุดข้อมูลที่มีจำนวนข้อมูลไม่เกิน 1000 ตัวอย่างข้อมูล และในแต่ละช่วงของจำนวนข้อมูลตั้งแต่ 300-600 ข้อมูลนั้น KMSM สามารถให้ประสิทธิภาพได้สูงกว่าวิธีการอื่นอย่างชัดเจน



ภาพที่ 4.13 กราฟสรุปการจัดอันดับของจำนวนชุดข้อมูลค่าเฉลี่ยพื้นที่ใต้กราฟ (AUC) จากประสิทธิภาพของการปรับสมดุลเทียบกับจำนวน feature

จากภาพที่ 4.12 การปรับสมดุลด้วยวิธีการสุ่มตัวอย่างตามการแบ่งกลุ่มของข้อมูลจะให้ประสิทธิภาพพื้นที่ใต้กราฟไม่ได้ดีกว่าการปรับสมดุลด้วยวิธีอื่นๆในทุกช่วงของจำนวน feature ของข้อมูล แต่ในช่วงข้อมูลที่มีจำนวน feature ตั้งแต่ 7-10 feature นั้น สามารถให้ประสิทธิภาพพื้นที่ใต้กราฟอยู่ในระดับที่สูงเมื่อเทียบกับวิธีอื่นๆ

นอกจากการเปรียบเทียบประสิทธิภาพการปรับสมดุลด้วยวิธีต่างๆแล้ว ในการวิจัยนี้ ยังได้ทำการทดลองหาความสัมพันธ์ระหว่างจำนวนกลุ่มที่ใช้การปรับสมดุลด้วยการสุ่มตัวอย่างตามการแบ่งกลุ่มของข้อมูล โดยในการทดลองนี้ จะมีการแบ่งกลุ่มตั้งแต่ 3-9 กลุ่ม ประสิทธิภาพเฉลี่ยในด้านต่างๆเป็นตามตาราง



ภาพที่ 4.14 กราฟเปรียบเทียบประสิทธิภาพเฉลี่ยกับจำนวนการแบ่งกลุ่มข้อมูล

ตารางที่ 4.3 ตารางเปรียบเทียบประสิทธิภาพเฉลี่ยกับจำนวนการแบ่งกลุ่มข้อมูล

Number of Cluster	Accuracy	AUC	F1	Precision	Recall
3	0.904	0.787	0.610	0.601	0.641
4	0.907	0.788	0.616	0.612	0.640
5	0.908	0.782	0.608	0.610	0.626
6	0.910	0.785	0.622	0.631	0.629
7	0.910	0.785	0.625	0.638	0.629
8	0.912	0.790	0.634	0.648	0.639
9	0.913	0.792	0.638	0.648	0.643

ตารางที่ 4.4 ตารางความถี่ของชุดข้อมูลที่ประสิทธิภาพสูงสุด

Method	Accuracy	AUC	F1	Precision	Recall
KMeanClusterCentroid	2	10	2	2	22
KMeansClusterBased	0	0	1	0	1
KMeans-SMOTE	31	13	25	31	2
RandomUnderSampling	0	13	5	1	18
SMOTE	11	8	11	10	1

จากผลการทดลองพบว่า การปรับสมดุลของข้อมูลโดยการสุ่มตัวอย่างตามการแบ่งกลุ่มของข้อมูลนี้ เมื่อเพิ่มจำนวนกลุ่มของข้อมูลที่แบ่ง ประสิทธิภาพของการจำแนกหมวดหมู่จะเพิ่มขึ้นตามจำนวนกลุ่มของข้อมูล เพราะเนื่องจาก 3 ปัจจัยเช่นเดียวกับประสิทธิภาพความถูกต้องเฉลี่ย กล่าวคือ

1. การแบ่งกลุ่มของข้อมูล มีโอกาสทำให้ข้อมูลบางกลุ่มมีแต่ข้อมูลส่วนใหญ่ และถ้าจำนวนกลุ่มที่ต้องการแบ่งมาก จะทำให้มีโอกาสที่แต่ละกลุ่มมีแต่ข้อมูลส่วนใหญ่มากขึ้น

2. ข้อมูลที่มีการแบ่งกลุ่มแล้ว ในกรณีที่กลุ่มนั้นๆมีทั้งข้อมูลส่วนใหญ่และข้อมูลส่วนน้อย มีโอกาสที่จะไม่ต้องเพิ่มปริมาณข้อมูลส่วนน้อย หรือถ้าต้องเพิ่ม ก็จะเพิ่มในปริมาณที่ไม่มาก

3. โมเดลต้นไม้ตัดสินใจที่ใช้จะมีจำนวนโมเดลมากขึ้นตามจำนวนกลุ่มที่แบ่ง ซึ่งจะทำให้โมเดลที่ได้เหมาะสมในกลุ่มนั้นๆขึ้น

การเพิ่มจำนวนกลุ่มของข้อมูลนี้ อาจจะทำให้ประสิทธิภาพลดลงได้ เพราะถ้าแบ่งกลุ่มยิ่งมาก จะมีโอกาสทำให้แต่ละกลุ่มมีปริมาณของข้อมูลส่วนน้อยมีจำนวนน้อยจนไม่สามารถใช้วิธี SMOTE เพื่อเพิ่มจำนวนข้อมูลส่วนน้อยได้ จึงทำให้การจำแนกหมวดหมู่ของข้อมูลส่วนน้อยมีโอกาสผิดพลาดได้สูง

บทที่ 5
บทสรุปและข้อเสนอแนะ

งานวิจัยในวิทยานิพนธ์นี้ เป็นการทดสอบประสิทธิภาพของการปรับสมดุลข้อมูลด้วยวิธีการสุ่มตัวอย่างตามการแบ่งกลุ่มของข้อมูล บทสรุปของการทดลอง และข้อจำกัดของวิธีการที่พบจากการทดสอบ ตลอดจนข้อเสนอแนะแนวทางในการพัฒนางานวิจัยนี้ต่อไป เพื่อนำไปปรับปรุงและแก้ไขข้อบกพร่องต่างๆที่พบในงานวิจัยนี้

5.1 สรุปผลการศึกษา

งานวิจัยนี้ ได้นำเสนอวิธีปรับสมดุลข้อมูลเพื่อลดความเอนเอียงรวมถึงเพิ่มประสิทธิภาพของการจำแนกหมวดหมู่ เมื่อนำชุดข้อมูลไปใช้ในการเรียนรู้ โดยนำหลักการแบ่งกลุ่มข้อมูล เพื่อลดอัตราความไม่สมดุลของข้อมูลในแต่ละกลุ่ม ผนวกกับการเพิ่มปริมาณข้อมูลส่วนน้อยให้มีจำนวนเพิ่มขึ้น เพื่อให้ปริมาณข้อมูลที่จะใช้มีปริมาณเพียงพอสำหรับการเรียนรู้ ซึ่งจะให้ประสิทธิภาพความถูกต้อง (Accuracy) และความแม่นยำ (F-Measure) สูงขึ้น

จากการทดลองพบว่า การแบ่งกลุ่มข้อมูลนั้น ช่วยให้แยกหมวดหมู่ข้อมูลได้ โดยเฉพาะกลุ่มของข้อมูลที่มีแต่ข้อมูลส่วนใหญ่ หรือข้อมูลส่วนน้อยเพียงอย่างเดียว แต่เนื่องจากข้อมูลส่วนน้อยมีปริมาณไม่มาก จึงมีโอกาสน้อยมากหรืออาจจะไม่มีความเป็นไปได้ที่การแบ่งกลุ่มข้อมูลจะมีเพียงข้อมูลส่วนน้อยเพียงอย่างเดียว

สำหรับกลุ่มข้อมูลที่แบ่งแล้วที่มีทั้งข้อมูลส่วนใหญ่และข้อมูลส่วนน้อย อัตราความไม่สมดุลลดลง ทำให้การเพิ่มจำนวนข้อมูลส่วนน้อย จะเพิ่มในปริมาณที่ไม่มากนัก ซึ่งจะใกล้เคียงกับความเป็นจริงของข้อมูลมากกว่า เมื่อเทียบกับการเพิ่มปริมาณข้อมูลส่วนน้อยสำหรับชุดข้อมูลทั้งชุด เมื่อแบ่งกลุ่มข้อมูลแล้ว การจำแนกหมวดหมู่ตามการแบ่งกลุ่มของข้อมูล จะมีความเหมาะสมสำหรับข้อมูลแต่ละกลุ่มมากกว่า

ผลที่ได้จากการทดลองเมื่อเพิ่มปริมาณข้อมูลส่วนน้อยตามการแบ่งกลุ่มข้อมูล รวมถึงการจำแนกหมวดหมู่ตามการแบ่งกลุ่มข้อมูล ตามวิธีการที่นำเสนอนั้นสามารถเพิ่มประสิทธิภาพการจำแนก

หมวดหมู่ข้อมูลได้ โดยที่สามารถเพิ่มประสิทธิภาพความแม่นยำ (F-Measure)ได้สูงกว่าการเพิ่มปริมาณข้อมูลส่วนน้อยด้วย SMOTE กับทั้งชุดข้อมูลถึง 3.609% และยังสูงกว่าการลดปริมาณข้อมูลส่วนใหญ่ได้สูงถึง 21.061%

นอกจากนี้ วิธีการที่นำเสนอ ยังสามารถใช้ได้กับชุดข้อมูลที่มีอัตราความไม่สมดุลได้ทุกช่วงอัตรา รวมถึงชุดข้อมูลที่มีอัตราความไม่สมดุลสูงถึง 128.870 ได้อย่างมีประสิทธิภาพ

ในด้านของจำนวน feature นั้น วิธีที่นำเสนอ สามารถให้ประสิทธิภาพความถูกต้องของการจำแนกข้อมูลได้ดีสำหรับข้อมูลที่มีจำนวน feature ระหว่าง 7-18 feature โดยเฉพาะอย่างยิ่งกับชุดข้อมูลที่มี 8 feature สามารถทำให้ประสิทธิภาพความถูกต้องมากกว่าวิธี SMOTE ประมาณ 1.76%

5.2 ข้อจำกัดและแนวทางพัฒนาของงานวิจัย

งานวิจัยนี้มีข้อจำกัดที่เกิดขึ้นดังนี้

5.2.1 การเพิ่มจำนวนการแบ่งกลุ่มมีผลทำให้บางกลุ่มจะมีปริมาณข้อมูลส่วนน้อยที่ไม่เพียงพอสำหรับการเพิ่มจำนวนด้วยวิธี SMOTE จึงต้องปรับพารามิเตอร์จำนวนสมาชิกข้างเคียง (k_neighbors) ลงโดยใช้ค่า 2 ซึ่งเป็นจำนวนน้อยที่สุด แต่ยังคงพบกลุ่มของข้อมูลที่มีปริมาณข้อมูลส่วนน้อยเพียงแค่ 1 ตัวเท่านั้น เมื่อนำข้อมูลกลุ่มนี้ไปเรียนรู้เพื่อจำแนกหมวดหมู่ จะทำให้เกิดความเอนเอียงมา ในงานวิจัยจึงกำหนดให้ข้อมูลในกลุ่มนี้อยู่ในหมวดหมู่ของข้อมูลส่วนใหญ่ และเป็นผลทำให้ความถูกต้องลดลง แนวทางแก้ไขของข้อจำกัดนี้ สามารถทำได้โดยการหาสมาชิกของข้อมูลส่วนน้อยที่อยู่ใกล้ที่สุดจากชุดข้อมูลทั้งหมด และนำมาเพียงจำนวนน้อยที่สุดเท่าที่วิธี SMOTE ต้องการ และทำการเพิ่มจำนวนข้อมูลส่วนน้อยเพื่อมาเป็นตัวแทนในกลุ่มที่พบปัญหานี้

5.2.2 การแบ่งกลุ่มข้อมูลด้วย k-means เป็นวิธีที่ง่าย แต่มีความอ่อนไหวต่อค่าที่ผิดปกติ ซึ่งจะมีผลทำให้การแบ่งอาจจะแบ่งได้ไม่เหมาะสม ถ้าค่าผิดปกตินั้นเป็นข้อมูลส่วนน้อย อาจจะมีผลต่อการแบ่งกลุ่มและการเพิ่มปริมาณข้อมูลส่วนน้อย แนวทางแก้ไขนั้น สามารถทำได้โดยการเลือกใช้วิธีการแบ่งกลุ่มวิธีอื่นที่ไม่อ่อนไหวต่อค่าผิดปกติ เช่น DBSCAN และถ้าค่าที่ผิดปกตินั้นเป็นข้อมูลส่วนน้อย DBSCAN ยังคงสามารถแบ่งแยกข้อมูลนี้ออกมาได้ ทำให้สามารถจำแนกหมวดหมู่ของข้อมูลกลุ่มนี้ออกมาได้ดีขึ้น

บรรณานุกรม

บรรณานุกรม

Behzad Mirzaei, Bahareh Nikpour, Hossein Nezamabadi-pour, "CDBH: A clustering and density-based hybrid approach for imbalanced data classification," Expert Systems With Applications Volume 164, 2021

Chawla N.V., Lazarevic A., Hall L.O., Bowyer K.W. "SMOTEBoost: Improving Prediction of the Minority Class in Boosting," In: Lavrač N., Gamberger D., Todorovski L., Blockeel H. (eds) Knowledge Discovery in Databases: PKDD 2003. PKDD 2003. Lecture Notes in Computer Science, vol 2838. Springer, Berlin, Heidelberg

Chih-Fong Tsai, Wei-Chao Lin, Ya-Han Hu, Guan-Ting Yao, "Under-sampling class imbalanced datasets by combining clustering analysis and instance selection," Information Sciences Volume 477, 2019, Pages 47-54.

Jerzy Błaszczyński n, Jerzy Stefanowski, "Neighbourhood sampling in bagging for imbalanced data, " Neurocomputing vol. 150, Part B, 20 February 2015, pp 529-542.

M. Galar, A. Fernandez, E. Barrenechea, H. Bustince and F. Herrera, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches," in IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 42, no. 4, pp. 463-484

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research vol. 16, (2002) 321–357

S. Na, L. Xumin and G. Yong, "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm," 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, 2010, pp. 63-67.

Wei-Chao Lina, Chih-Fong Tsai, Ya-Han Huc, Jing-Shang Jhang, "Clustering-based undersampling in class-imbalanced data, " Information Sciences vol. 410–409, October 2017, pp 26–17

X. Guo, Y. Yin, C. Dong, G. Yang and G. Zhou, "On the Class Imbalance Problem," 2008 Fourth International Conference on Natural Computation, 2008, pp. 192-201.

YanminSunaMohamed S.KamelaAndrew K.C.WongbYangWang, "Cost-sensitive boosting for classification of imbalanced da, " Pattern Recognition Volume 40, Issue 12, December 2007, Pages 3358-3378.

Y. Zhang, L. Zhang and Y. Wang, "Cluster-based majority under-sampling approaches for class imbalance learning," 2010 2nd IEEE International Conference on Information and Financial Engineering, 2010, pp. 400-404

ภาคผนวก

ภาคผนวก ก
ผลการทดสอบประสิทธิภาพเทียบกับจำนวนการแบ่งกลุ่มข้อมูลของแต่ละชุดข้อมูล

ตารางที่ 7.1 ตารางเปรียบเทียบประสิทธิภาพความถูกต้อง (Accuracy) กับจำนวนการแบ่งกลุ่มข้อมูล

No. Cluster	3	4	5	6	7	8	9
abalone19	0.977	0.976	0.978	0.977	0.979	0.981	0.984
abalone9-18	0.930	0.931	0.940	0.949	0.951	0.951	0.954
ecoli-0_vs_1	0.950	0.955	0.968	0.977	0.946	0.964	0.959
ecoli-0-1-3-7_vs_2-6	0.978	0.978	0.982	0.982	0.989	0.979	0.989
ecoli1	0.863	0.878	0.861	0.875	0.881	0.890	0.884
ecoli2	0.931	0.940	0.937	0.934	0.931	0.928	0.931
ecoli3	0.908	0.916	0.914	0.914	0.902	0.902	0.914
ecoli4	0.970	0.967	0.973	0.988	0.988	0.985	0.985
glass0	0.752	0.733	0.743	0.775	0.738	0.790	0.776
glass-0-1-2-3_vs_4-5-6	0.925	0.930	0.934	0.930	0.925	0.939	0.958
glass-0-1-6_vs_2	0.828	0.875	0.875	0.859	0.875	0.839	0.854
glass-0-1-6_vs_5	0.978	0.973	0.968	0.968	0.951	0.951	0.946
glass1	0.785	0.780	0.761	0.766	0.794	0.743	0.780
glass2	0.869	0.855	0.836	0.822	0.827	0.822	0.869
glass4	0.935	0.948	0.953	0.977	0.986	0.958	0.953
glass5	0.958	0.967	0.953	0.962	0.962	0.962	0.977
glass6	0.953	0.962	0.972	0.953	0.962	0.967	0.977
haberman	0.572	0.614	0.604	0.634	0.608	0.621	0.621
iris0	1.000	1.000	1.000	1.000	1.000	1.000	1.000
new-thyroid1	0.986	0.986	0.986	0.991	0.977	0.986	0.995
new-thyroid2	0.986	0.986	0.991	1.000	0.991	0.981	0.977
page-blocks0	0.957	0.955	0.959	0.958	0.957	0.956	0.955
page-blocks-1-3_vs_4	0.991	0.998	0.991	0.994	0.991	0.991	0.994
pima	0.608	0.620	0.638	0.628	0.633	0.648	0.631
segment0	0.993	0.990	0.990	0.989	0.993	0.988	0.987
shuttle-c0-vs-c4	0.999	0.999	0.997	0.998	0.999	0.999	0.999
shuttle-c2-vs-c4	0.977	0.969	0.961	0.961	0.977	1.000	1.000
vehicle0	0.942	0.949	0.941	0.936	0.947	0.949	0.942
vehicle1	0.751	0.735	0.736	0.746	0.741	0.739	0.751
vehicle2	0.949	0.953	0.962	0.956	0.952	0.954	0.950
vehicle3	0.741	0.756	0.729	0.726	0.727	0.738	0.738
vowel0	0.980	0.975	0.977	0.972	0.971	0.978	0.977
wisconsin	0.890	0.890	0.892	0.895	0.880	0.880	0.884

ตารางที่ 7.1 (ต่อ)

No. Cluster	3	4	5	6	7	8	9
yeast-0-5-6-7-9_vs_4	0.883	0.892	0.901	0.896	0.888	0.903	0.902
yeast1	0.687	0.694	0.675	0.696	0.678	0.677	0.702
yeast-1_vs_7	0.867	0.911	0.922	0.889	0.895	0.900	0.895
yeast-1-2-8-9_vs_7	0.918	0.934	0.929	0.935	0.926	0.932	0.930
yeast-1-4-5-8_vs_7	0.878	0.896	0.905	0.912	0.919	0.928	0.936
yeast-2_vs_4	0.943	0.942	0.944	0.953	0.945	0.936	0.941
yeast-2_vs_8	0.952	0.948	0.952	0.950	0.955	0.963	0.965
yeast3	0.929	0.932	0.927	0.922	0.922	0.919	0.917
yeast4	0.942	0.947	0.942	0.956	0.954	0.955	0.950
yeast5	0.980	0.977	0.978	0.983	0.980	0.983	0.982
yeast6	0.968	0.968	0.975	0.972	0.973	0.972	0.974
Average	**0.904**	**0.909**	**0.908**	**0.910**	**0.908**	**0.910**	**0.913**

ตารางที่ 7.2 ตารางเปรียบเทียบประสิทธิภาพความแม่นยำ (F-measure) กับจำนวนการแบ่งกลุ่มข้อมูล

No. Cluster	3	4	5	6	7	8	9
abalone19	0.039	0.036	0.043	0.045	0.019	0.068	0.126
abalone9-18	0.534	0.524	0.522	0.543	0.576	0.541	0.579
ecoli-0_vs_1	0.960	0.965	0.976	0.982	0.958	0.972	0.969
ecoli-0-1-3-7_vs_2-6	0.460	0.527	0.567	0.500	0.700	0.500	0.700
ecoli1	0.704	0.734	0.671	0.708	0.731	0.754	0.742
ecoli2	0.779	0.799	0.805	0.774	0.779	0.732	0.746
ecoli3	0.561	0.610	0.590	0.583	0.558	0.524	0.569
ecoli4	0.765	0.752	0.787	0.893	0.893	0.871	0.871
glass0	0.640	0.622	0.576	0.671	0.629	0.680	0.653
glass-0-1-2-3_vs_4-5-6	0.844	0.849	0.864	0.844	0.831	0.864	0.912
glass-0-1-6_vs_2	0.144	0.276	0.247	0.235	0.238	0.177	0.277
glass-0-1-6_vs_5	0.720	0.633	0.567	0.667	0.600	0.560	0.520
glass1	0.686	0.682	0.660	0.653	0.689	0.628	0.699
glass2	0.235	0.205	0.237	0.146	0.238	0.191	0.258
glass4	0.446	0.477	0.537	0.770	0.865	0.663	0.494
glass5	0.293	0.500	0.267	0.200	0.200	0.267	0.533
glass6	0.808	0.860	0.893	0.826	0.863	0.875	0.915
haberman	0.320	0.353	0.341	0.368	0.341	0.358	0.344
iris0	1.000	1.000	1.000	1.000	1.000	1.000	1.000
new-thyroid1	0.960	0.960	0.960	0.975	0.929	0.960	0.987
new-thyroid2	0.958	0.958	0.971	1.000	0.975	0.942	0.925
page-blocks0	0.800	0.791	0.807	0.797	0.795	0.795	0.786
page-blocks-1-3_vs_4	0.938	0.982	0.927	0.942	0.924	0.926	0.942
pima	0.458	0.464	0.488	0.486	0.484	0.496	0.483
segment0	0.976	0.965	0.965	0.962	0.974	0.959	0.955
shuttle-c0-vs-c4	0.996	0.996	0.980	0.988	0.996	0.996	0.996
shuttle-c2-vs-c4	0.600	0.400	0.200	0.200	0.533	1.000	1.000
vehicle0	0.878	0.893	0.875	0.862	0.885	0.892	0.875
vehicle1	0.508	0.483	0.501	0.506	0.506	0.500	0.518
vehicle2	0.903	0.908	0.928	0.912	0.905	0.910	0.902
vehicle3	0.512	0.521	0.461	0.453	0.452	0.473	0.487
vowel0	0.887	0.852	0.872	0.839	0.834	0.872	0.864
wisconsin	0.846	0.842	0.847	0.849	0.824	0.827	0.828

ตารางที่ 7.2 (ต่อ)

No. Cluster	3	4	5	6	7	8	9
yeast-0-5-6-7-9_vs_4	0.456	0.507	0.506	0.452	0.436	0.488	0.494
yeast1	0.475	0.492	0.468	0.507	0.459	0.467	0.520
yeast-1_vs_7	0.236	0.378	0.366	0.254	0.291	0.290	0.291
yeast-1-2-8-9_vs_7	0.206	0.258	0.224	0.175	0.147	0.191	0.101
yeast-1-4-5-8_vs_7	0.117	0.084	0.169	0.155	0.152	0.295	0.192
yeast-2_vs_4	0.703	0.704	0.717	0.769	0.734	0.687	0.714
yeast-2_vs_8	0.502	0.496	0.524	0.511	0.526	0.556	0.571
yeast3	0.687	0.706	0.686	0.660	0.654	0.637	0.605
yeast4	0.278	0.303	0.210	0.362	0.336	0.310	0.326
yeast5	0.698	0.630	0.653	0.731	0.686	0.724	0.712
yeast6	0.430	0.461	0.522	0.437	0.471	0.394	0.470
Average	**0.612**	**0.624**	**0.613**	**0.618**	**0.628**	**0.632**	**0.647**

ตารางที่ 7.3 ตารางเปรียบเทียบประสิทธิภาพพื้นที่ใต้กราฟ (AUC) กับจำนวนการแบ่งกลุ่มข้อมูล

No. Cluster	3	4	5	6	7	8	9
abalone19	0.526	0.523	0.524	0.523	0.510	0.539	0.571
abalone9-18	0.795	0.761	0.745	0.740	0.764	0.742	0.766
ecoli-0_vs_1	0.953	0.949	0.963	0.976	0.942	0.963	0.953
ecoli-0-1-3-7_vs_2-6	0.745	0.843	0.845	0.746	0.848	0.746	0.848
ecoli1	0.812	0.825	0.782	0.799	0.813	0.838	0.829
ecoli2	0.876	0.863	0.894	0.851	0.866	0.817	0.827
ecoli3	0.759	0.802	0.775	0.775	0.768	0.743	0.762
ecoli4	0.890	0.889	0.892	0.923	0.923	0.922	0.922
glass0	0.731	0.717	0.699	0.752	0.721	0.763	0.745
glass-0-1-2-3_vs_4-5-6	0.897	0.899	0.910	0.881	0.871	0.894	0.945
glass-0-1-6_vs_2	0.537	0.615	0.585	0.590	0.585	0.574	0.612
glass-0-1-6_vs_5	0.894	0.844	0.747	0.794	0.786	0.786	0.783
glass1	0.761	0.757	0.740	0.740	0.769	0.720	0.769
glass2	0.571	0.563	0.583	0.531	0.578	0.553	0.601
glass4	0.747	0.737	0.756	0.861	0.945	0.868	0.740
glass5	0.693	0.745	0.643	0.600	0.600	0.648	0.750
glass6	0.875	0.908	0.928	0.903	0.922	0.911	0.945
haberman	0.511	0.543	0.533	0.562	0.536	0.548	0.544
iris0	1.000	1.000	1.000	1.000	1.000	1.000	1.000
new-thyroid1	0.980	0.980	0.980	0.994	0.952	0.980	0.997

ตารางที่ 7.3 (ต่อ)

No. Cluster	3	4	5	6	7	8	9
new-thyroid2	0.980	0.980	0.983	1.000	0.994	0.966	0.952
page-blocks0	0.904	0.900	0.904	0.893	0.889	0.898	0.890
page-blocks-1-3_vs_4	0.995	0.999	0.961	0.962	0.946	0.961	0.962
pima	0.577	0.586	0.604	0.599	0.600	0.613	0.599
segment0	0.986	0.980	0.978	0.976	0.983	0.977	0.975
shuttle-c0-vs-c4	1.000	1.000	0.991	0.995	0.996	0.996	0.996
shuttle-c2-vs-c4	0.800	0.700	0.600	0.600	0.750	1.000	1.000
vehicle0	0.922	0.932	0.920	0.909	0.923	0.930	0.912
vehicle1	0.670	0.651	0.667	0.669	0.668	0.666	0.677
vehicle2	0.937	0.935	0.955	0.935	0.933	0.937	0.931
vehicle3	0.675	0.683	0.640	0.635	0.634	0.649	0.657
vowel0	0.934	0.911	0.927	0.904	0.899	0.923	0.917
wisconsin	0.884	0.878	0.885	0.882	0.862	0.866	0.865
yeast-0-5-6-7-9_vs_4	0.707	0.749	0.736	0.699	0.694	0.722	0.728
yeast1	0.629	0.641	0.622	0.650	0.618	0.621	0.660
yeast-1_vs_7	0.619	0.673	0.663	0.615	0.634	0.621	0.634
yeast-1-2-8-9_vs_7	0.635	0.644	0.625	0.595	0.575	0.578	0.545
yeast-1-4-5-8_vs_7	0.554	0.516	0.584	0.572	0.576	0.644	0.585
yeast-2_vs_4	0.838	0.838	0.847	0.878	0.865	0.833	0.853
yeast-2_vs_8	0.784	0.781	0.784	0.783	0.761	0.765	0.766
yeast3	0.831	0.849	0.836	0.816	0.813	0.801	0.771
yeast4	0.648	0.651	0.602	0.675	0.664	0.637	0.662
yeast5	0.867	0.821	0.843	0.891	0.846	0.879	0.870
yeast6	0.761	0.775	0.792	0.735	0.749	0.693	0.736
Average	**0.788**	**0.792**	**0.784**	**0.782**	**0.786**	**0.789**	**0.797**

ตารางที่ 7.4 ตารางเปรียบเทียบประสิทธิภาพ Precision กับจำนวนการแบ่งกลุ่มข้อมูล

No. Cluster	3	4	5	6	7	8	9
abalone19	0.028	0.026	0.033	0.035	0.013	0.060	0.109
abalone9-18	0.477	0.528	0.551	0.607	0.622	0.584	0.623
ecoli-0_vs_1	0.978	0.967	0.973	0.986	0.966	0.979	0.966
ecoli-0-1-3-7_vs_2-6	0.433	0.433	0.500	0.500	0.700	0.500	0.700
ecoli1	0.706	0.749	0.724	0.781	0.790	0.784	0.768
ecoli2	0.783	0.864	0.787	0.841	0.816	0.861	0.851
ecoli3	0.580	0.611	0.587	0.667	0.530	0.536	0.658
ecoli4	0.753	0.779	0.860	0.950	0.950	0.900	0.900
glass0	0.614	0.585	0.600	0.665	0.595	0.679	0.665
glass-0-1-2-3_vs_4-5-6	0.849	0.867	0.874	0.916	0.911	0.936	0.911
glass-0-1-6_vs_2	0.123	0.283	0.290	0.210	0.267	0.152	0.275
glass-0-1-6_vs_5	0.667	0.600	0.700	0.800	0.650	0.583	0.467
glass1	0.708	0.700	0.665	0.679	0.737	0.643	0.680
glass2	0.347	0.197	0.239	0.123	0.229	0.169	0.267
glass4	0.430	0.487	0.583	0.900	0.883	0.620	0.500
glass5	0.233	0.500	0.300	0.200	0.200	0.300	0.600
glass6	0.900	0.900	0.938	0.843	0.866	0.933	0.933
haberman	0.277	0.322	0.310	0.342	0.309	0.329	0.320
iris0	1.000	1.000	1.000	1.000	1.000	1.000	1.000
new-thyroid1	0.956	0.956	0.956	0.956	0.956	0.950	0.975
new-thyroid2	0.946	0.946	0.971	1.000	0.956	0.956	0.946
page-blocks0	0.765	0.756	0.783	0.787	0.787	0.768	0.767
page-blocks-1-3_vs_4	0.896	0.967	0.933	0.960	0.960	0.943	0.960
pima	0.446	0.457	0.486	0.470	0.477	0.496	0.474
segment0	0.976	0.964	0.970	0.968	0.980	0.959	0.954
shuttle-c0-vs-c4	0.992	0.992	0.977	0.985	1.000	1.000	1.000
shuttle-c2-vs-c4	0.600	0.400	0.200	0.200	0.600	1.000	1.000
vehicle0	0.872	0.887	0.871	0.869	0.894	0.890	0.897
vehicle1	0.514	0.488	0.484	0.501	0.496	0.487	0.513
vehicle2	0.893	0.919	0.918	0.937	0.916	0.916	0.916
vehicle3	0.485	0.511	0.459	0.456	0.459	0.477	0.482
vowel0	0.901	0.887	0.882	0.865	0.873	0.902	0.904
wisconsin	0.832	0.849	0.835	0.861	0.847	0.837	0.864

ตารางที่ 7.4 (ต่อ)

No. Cluster	3	4	5	6	7	8	9
yeast-0-5-6-7-9_vs_4	0.434	0.488	0.512	0.464	0.438	0.490	0.486
yeast1	0.461	0.471	0.445	0.477	0.445	0.447	0.484
yeast-1_vs_7	0.187	0.386	0.490	0.227	0.261	0.293	0.268
yeast-1-2-8-9_vs_7	0.153	0.240	0.206	0.153	0.124	0.279	0.082
yeast-1-4-5-8_vs_7	0.087	0.077	0.136	0.127	0.131	0.272	0.229
yeast-2_vs_4	0.730	0.715	0.717	0.759	0.718	0.672	0.699
yeast-2_vs_8	0.453	0.437	0.493	0.462	0.542	0.577	0.675
yeast3	0.677	0.682	0.663	0.641	0.636	0.629	0.633
yeast4	0.239	0.294	0.197	0.399	0.364	0.337	0.305
yeast5	0.674	0.626	0.614	0.686	0.680	0.713	0.679
yeast6	0.362	0.397	0.465	0.423	0.458	0.413	0.486
Average	**0.600**	**0.618**	**0.618**	**0.629**	**0.637**	**0.642**	**0.656**

ตารางที่ 7.5 ตารางเปรียบเทียบประสิทธิภาพ Recall กับจำนวนการแบ่งกลุ่มข้อมูล

No. Cluster	3	4	5	6	7	8	9
abalone19	0.067	0.062	0.062	0.062	0.033	0.091	0.153
abalone9-18	0.642	0.570	0.525	0.503	0.553	0.506	0.553
ecoli-0_vs_1	0.945	0.965	0.979	0.979	0.951	0.966	0.972
ecoli-0-1-3-7_vs_2-6	0.500	0.700	0.700	0.500	0.700	0.500	0.700
ecoli1	0.717	0.727	0.638	0.661	0.688	0.742	0.728
ecoli2	0.794	0.751	0.831	0.731	0.771	0.656	0.676
ecoli3	0.571	0.657	0.600	0.600	0.600	0.543	0.571
ecoli4	0.800	0.800	0.800	0.850	0.850	0.850	0.850
glass0	0.671	0.672	0.571	0.686	0.671	0.686	0.657
glass-0-1-2-3_vs_4-5-6	0.844	0.840	0.864	0.785	0.765	0.805	0.920
glass-0-1-6_vs_2	0.183	0.300	0.233	0.267	0.233	0.250	0.317
glass-0-1-6_vs_5	0.800	0.700	0.500	0.600	0.600	0.600	0.600
glass1	0.672	0.672	0.661	0.648	0.675	0.635	0.727
glass2	0.217	0.217	0.283	0.183	0.283	0.233	0.283
glass4	0.533	0.500	0.533	0.733	0.900	0.767	0.500
glass5	0.400	0.500	0.300	0.200	0.200	0.300	0.500
glass6	0.767	0.833	0.867	0.833	0.866	0.833	0.900
haberman	0.382	0.393	0.382	0.409	0.382	0.395	0.381
iris0	1.000	1.000	1.000	1.000	1.000	1.000	1.000
new-thyroid1	0.971	0.971	0.971	1.000	0.914	0.971	1.000

ตารางที่ 7.5 (ต่อ)

No. Cluster	3	4	5	6	7	8	9
new-thyroid2	0.971	0.971	0.971	1.000	1.000	0.943	0.914
page-blocks0	0.837	0.830	0.835	0.812	0.803	0.825	0.807
page-blocks-1-3_vs_4	1.000	1.000	0.927	0.927	0.893	0.927	0.927
pima	0.474	0.474	0.493	0.507	0.492	0.496	0.493
segment0	0.976	0.967	0.961	0.958	0.970	0.961	0.958
shuttle-c0-vs-c4	1.000	1.000	0.984	0.992	0.992	0.992	0.992
shuttle-c2-vs-c4	0.600	0.400	0.200	0.200	0.500	1.000	1.000
vehicle0	0.885	0.900	0.879	0.859	0.879	0.894	0.855
vehicle1	0.507	0.480	0.525	0.512	0.516	0.516	0.525
vehicle2	0.913	0.899	0.941	0.890	0.895	0.904	0.890
vehicle3	0.542	0.537	0.462	0.453	0.448	0.472	0.495
vowel0	0.878	0.833	0.866	0.822	0.811	0.855	0.844
wisconsin	0.862	0.837	0.862	0.841	0.804	0.820	0.799
yeast-0-5-6-7-9_vs_4	0.489	0.571	0.531	0.455	0.453	0.496	0.513
yeast1	0.492	0.515	0.496	0.543	0.478	0.489	0.564
yeast-1_vs_7	0.333	0.400	0.367	0.300	0.333	0.300	0.333
yeast-1-2-8-9_vs_7	0.333	0.333	0.300	0.234	0.200	0.200	0.133
yeast-1-4-5-8_vs_7	0.200	0.100	0.233	0.200	0.200	0.333	0.200
yeast-2_vs_4	0.707	0.709	0.727	0.784	0.765	0.705	0.744
yeast-2_vs_8	0.600	0.600	0.600	0.600	0.550	0.550	0.550
yeast3	0.706	0.742	0.718	0.681	0.674	0.650	0.582
yeast4	0.333	0.333	0.236	0.373	0.353	0.296	0.353
yeast5	0.747	0.656	0.700	0.795	0.703	0.770	0.750
yeast6	0.543	0.571	0.600	0.486	0.514	0.400	0.486
Average	**0.646**	**0.647**	**0.630**	**0.624**	**0.633**	**0.639**	**0.652**

ภาคผนวก ข

ผลงานตีพิมพ์

IBDAP and BDAP 2021, August 26-27, Bangkok, Thailand, 2021

วิธีการสุ่มตัวอย่างตามการแบ่งกลุ่มของข้อมูลที่ไม่สมดุล
Cluster-based ensemble sampling approaches for Imbalanced Data

อดิเทพ จิตพิทยาพร *(Adithep Chitpityaporn)*, เอกสิทธิ์ พัชรวงศ์ศักดา *(Eakasit Pacharawongsakda)*
สาขาวิศวกรรมข้อมูลขนาดใหญ่ วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์
60516202004@dpu.ac.th, eakasit.pac@dpu.ac.th

บทคัดย่อ

ปัจจุบันมีการนำข้อมูลต่างๆมาใช้ในการวิเคราะห์ด้วยวิธีการจำแนกหมวดหมู่ ปัญหาอย่างหนึ่งที่พบไม่ว่าจะใช้วิธีการจำแนกหมวดหมู่วิธีใดก็ตามคือ ถ้าข้อมูลที่ใช้สำหรับการเรียนรู้ (training data) มีความไม่สมดุลระหว่างแต่ละหมวดหมู่ มักจะทำให้ผลลัพธ์มีความเอนเอียงไปทางด้านของข้อมูลส่วนใหญ่ และยิ่งข้อมูลมีความไม่สมดุลระหว่างส่วนใหญ่กับส่วนน้อยมากเท่าไร ยิ่งทำให้มีความเอนเอียงสูงมากขึ้น

ในงานวิจัยนี้ได้นำเสนอแนวทางการปรับสมดุลของข้อมูลโดยการแบ่งกลุ่มของข้อมูล (clustering) และใช้วิธีการเพิ่มจำนวนข้อมูลส่วนน้อยตามแต่ละกลุ่ม (oversampling) รวมถึงการจำแนกประเภทข้อมูลตามแต่ละกลุ่มด้วย ทำให้ประสิทธิภาพของการจำแนกประเภทข้อมูลมีความแม่นยำ และถูกต้องมากขึ้น

จากผลการทดลองพบว่าวิธีการที่นำเสนอมีประสิทธิภาพโดยรวมมากขึ้นกว่าวิธีการก่อนหน้านี้เป็นจำนวน 0.64 %

คำสำคัญ: ข้อมูลไม่สมดุล, การแบ่งกลุ่มข้อมูล, การเพิ่มปริมาณข้อมูล

Abstract

At present, various data are used for analysis by classification method. One of the problems when using any classification method is if the training data is imbalanced between each category, the results are biased towards the bigger class. If the ratio between the majority class and minority class is higher, it has more bias.

In this research, the data balancing method is proposed by clustering followed by oversampling for each group and also classification based on each cluster. The precise and correctness of classification is higher.

According the results of the experiment, the performance of the proposed method is greater than the previous method 0.44%

Keyword: Imbalance Data, Clustering, Oversampling

1. บทนำ

ปัจจุบันมีการนำข้อมูลต่างๆมาวิเคราะห์เพื่อวัตถุประสงค์ต่างๆมากมาย บางวัตถุประสงค์เพื่อต้องการแบ่งกลุ่มของข้อมูลที่มีความแตกต่างกัน บางวัตถุประสงค์เพื่อต้องการจำแนกหมวดหมู่ของข้อมูลนั้นๆ สำหรับข้อมูลจริงที่นำมาใช้ในการเรียนรู้มักจะพบปัญหาความไม่สมดุลของข้อมูล ทำให้ผลลัพธ์ของการเรียนรู้ไม่มีประสิทธิภาพโดยมีความเอนเอียงไปทางข้อมูลส่วนใหญ่ ในการเพิ่มประสิทธิภาพของการเรียนรู้ สามารถทำได้โดยทำให้ข้อมูลที่จะนำมาใช้เพื่อการเรียนรู้นั้นมีความสมดุล ในการแก้ปัญหาของข้อมูลไม่สมดุลนั้นมีอยู่หลายวิธี หากจำแนกวิธีการนั้นสามารถแบ่งได้เป็น 2 ประเภทใหญ่ๆคือ การลดจำนวนของข้อมูลส่วนใหญ่ให้มีปริมาณใกล้เคียงหรือเท่ากับจำนวนของข้อมูลส่วนน้อย (Under Sampling) และการเพิ่มจำนวนข้อมูลส่วนน้อยให้มีจำนวนมากขึ้นจนใกล้เคียงหรือเท่ากับข้อมูลส่วนใหญ่ (Over Sampling) การลดจำนวนของข้อมูลส่วนใหญ่นั้น ปัญหาอย่างหนึ่งที่มักจะพบได้คือ ถ้าข้อมูลส่วนน้อยมีปริมาณน้อยมาก เมื่อ

ลดจำนวนข้อมูลส่วนใหญ่แล้ว จะทำให้ปริมาณข้อมูลที่ใช้ในการเรียนรู้มีปริมาณน้อยลง อาจจะมีผลทำให้โมเดลของการจำแนกหมวดหมู่ไม่มีประสิทธิภาพเท่าที่ควร

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ปัญหาความไม่สมดุลของข้อมูล

ความไม่สมดุลของข้อมูลเป็นปัญหาอย่างหนึ่งของชุดข้อมูลที่มีข้อมูลประเภทหนึ่ง[4] มีจำนวนน้อยกว่าข้อมูลอีกประเภทหนึ่ง ข้อมูลในลักษณะนี้ มีความน่าสนใจมาก เพราะข้อมูลจริงที่เกิดขึ้นนั้นจะมีความไม่สมดุล ซึ่งประเภทข้อมูลที่มีจำนวนน้อย มักจะเป็นข้อมูลที่สำคัญในการจำแนกประเภทข้อมูล ตัวอย่างของข้อมูลในโลกแห่งความเป็นจริง เช่น การตรวจจับการทุจริต เซ็นเซอร์ของเครื่องจักร เป็นต้น จะเห็นได้ว่า ข้อมูลที่เกิดขึ้นจริง ข้อมูลส่วนน้อย จะมีจำนวนน้อยกว่าข้อมูลส่วนใหญ่มาก การจำแนกประเภทข้อมูลด้วยอัลกอริทึมการเรียนรู้ที่นิยม จะมีความโน้มเอียงไปทางข้อมูลส่วนใหญ่ และปัญหาความไม่สมดุลของข้อมูลจะมีคุณลักษณะดังนี้

ตัวอย่างข้อมูลมีน้อย ข้อมูลส่วนน้อยนี้มีปริมาณน้อยมาก ซึ่งไม่เพียงพอต่อการนำมาใช้ในการสร้างโมเดล วิธีการหนึ่งที่นิยมใช้กันคือ การทำให้ข้อมูลมีความสมดุลก่อนนำไปในการสร้างโมเดล

การซ้อนทับหรือความสามารถในการแยกประเภท เมื่อข้อมูลส่วนน้อยมีการกระจายอยู่ในข้อมูลส่วนใหญ่ จะมีปัญหาในการจำแนกประเภทอย่างมีประสิทธิภาพ ในกรณีนี้ การจำแนกประเภทจะทำให้ข้อมูลส่วนน้อยถูกจัดอยู่ในประเภทของข้อมูลส่วนใหญ่

ข้อมูลแบ่งเป็นกลุ่มเล็ก ในกรณีที่ข้อมูลส่วนน้อยมีการแบ่งแยกออกมาชัดเจน และแบ่งเป็นกลุ่มย่อยๆ จะเพิ่มความซับซ้อนของปัญหา เนื่องจากจำนวนข้อมูลในแต่ละกลุ่มมีความไม่สมดุลกัน

2.2 ประเภทของการสุ่มตัวอย่าง

การสุ่มตัวอย่างเป็นเทคนิคหนึ่งที่ใช้ในการเตรียมข้อมูลเพื่อปรับชุดข้อมูลให้มีความสมดุลก่อนจะนำไปใช้ในการสร้างโมเดล โดยเทคนิคที่ใช้จะแบ่งเป็น 3 ประเภท คือ

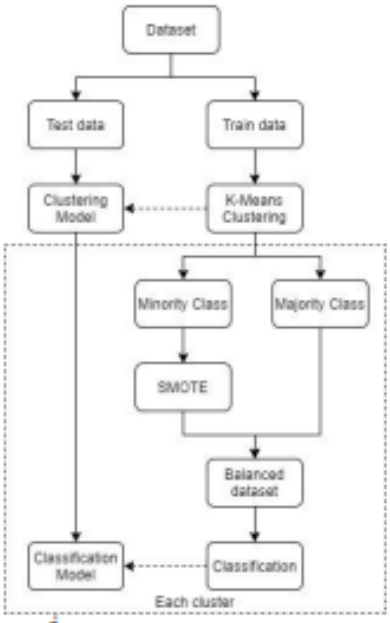
1. การลดจำนวนข้อมูล ซึ่งจะเป็นการลดจำนวนข้อมูลส่วนใหญ่ให้เหลือเท่ากับจำนวนข้อมูลส่วนน้อย แนวทางการทำ undersampling [1] จะการแบ่งกลุ่มของข้อมูลส่วนใหญ่ให้มีจำนวนกลุ่มเท่ากับจำนวนข้อมูลในข้อมูลส่วนน้อย โดยวิธี k-means และใช้ศูนย์กลางของกลุ่มเป็นตัวแทนของข้อมูลส่วนใหญ่ ทำให้จำนวนข้อมูลที่คัดเลือกมาจากข้อมูลส่วนใหญ่มีจำนวนเท่ากับจำนวนข้อมูลส่วนน้อย วิธีนี้ไม่ได้ใช้ข้อมูลจริงของข้อมูลส่วนใหญ่มาใช้สำหรับการเรียนรู้ เมื่อนำข้อมูลจริงมาทำการทดสอบ จะมีโอกาสเกิดความผิดพลาดได้ง่าย โดยเฉพาะถ้าข้อมูลนั้นเป็นข้อมูลส่วนใหญ่และในงานวิจัยนี้ได้มีการแก้ไขปัญหานี้โดยใช้วิธีเดียวกับวิธีก่อนหน้านี้ แต่จะคัดเลือกข้อมูลในข้อมูลส่วนใหญ่ที่อยู่ใกล้ศูนย์กลางของกลุ่มมากที่สุดมาใช้เป็นตัวแทนของกลุ่ม

ถ้าข้อมูลที่นำมาใช้ในการจำแนกนั้น มีจำนวนของข้อมูลส่วนน้อยปริมาณมาก จะทำให้การแบ่งกลุ่มของข้อมูลส่วนใหญ่ให้มีขนาดเท่ากับจำนวนข้อมูลส่วนน้อยนั้นใช้เวลานาน จึงมีการเสนออีกวิธีการคือ จะแบ่งกลุ่มของข้อมูลส่วนใหญ่ให้เป็นกลุ่มย่อยๆจำนวนไม่มาก[2] และทำการคัดเลือกจำนวนข้อมูลส่วนใหญ่ของแต่ละกลุ่มย่อยออกมาจำนวนหนึ่งตามอัตราส่วนระหว่างจำนวนข้อมูลส่วนน้อยต่อจำนวนข้อมูลส่วนใหญ่ แต่วิธีการคัดเลือกข้อมูลนี้จะใช้ 2 วิธีมีคือใช้วิธีสุ่มเลือกข้อมูลในกลุ่มนั้นๆ และใช้วิธีหาจากข้อมูลที่อยู่ใกล้ศูนย์กลางของกลุ่มที่สุดจากนั้นจึงนำข้อมูลที่ได้จากทั้ง 2 กลุ่มมารวมกัน เพื่อเป็นตัวแทนของข้อมูลส่วนใหญ่ แล้วนำไปรวมกับข้อมูลส่วนย่อยทั้งหมดเพื่อใช้ในสำหรับการเรียนรู้ การแบ่งกลุ่มข้อมูลให้เป็นกลุ่มย่อยจำนวนไม่มากขนาดนั้นไปนั้น จะสามารถทำได้รวดเร็วกว่าการแบ่งกลุ่มของข้อมูลให้มีจำนวนกลุ่มของข้อมูลส่วนใหญ่เท่ากับจำนวนข้อมูลส่วนย่อย และการเลือกตัวแทนของกลุ่มนี้จะได้ข้อมูลจริง รวมทั้งกระจายข้อมูลที่เลือกให้มีลักษณะการกระจายตัวของข้อมูลใกล้เคียงกับการกระจายตัวของข้อมูลส่วนใหญ่ ทำให้ประสิทธิภาพในการจำแนกหมวดหมู่ดีขึ้น ทั้งนี้ การคัดเลือกตัวแทนข้อมูลยังมีโอกาสที่ทำให้ข้อมูล

เกาะกลุ่มบริเวณจุดศูนย์กลางของกลุ่ม มีผลทำให้เกิดความเอนเอียงไปทางศูนย์กลางของกลุ่มได้

2.การเพิ่มจำนวนข้อมูล จะเป็นการเพิ่มจำนวนข้อมูลส่วนน้อย ให้มีปริมาณเพิ่มขึ้นเท่ากับจำนวนข้อมูลส่วนใหญ่ทำ oversampling ด้วยวิธี SMOTE [3,5] เป็นการเพิ่มข้อมูลของข้อมูลส่วนน้อยโดยจะใช้วิธีสังเคราะห์จากข้อมูลจริง โดยการหาข้อมูลส่วนน้อยที่อยู่ใกล้ แล้วสุ่มหาจุดที่อยู่ระหว่างทาง 2 จุดนั้นและทำการเพิ่มข้อมูลส่วนน้อยแบบนี้จนมีจำนวนเท่ากับจำนวนข้อมูลส่วนใหญ่ วิธีการนี้จะเป็นการสร้างข้อมูลขึ้นมา

3.การสุ่มเลือกแบบขั้นสูง วิธีการนี้แตกต่างจาก 2 วิธีดังกล่าวข้างต้น แต่จะเป็นการสุ่มตัวอย่างจากผลลัพธ์ที่ได้จากการจำแนกประเภท การส่งเสริม (boosting) [9] เป็นวิธีการในการทำวนซ้ำด้วยวิธีการปรับน้ำหนักที่แตกต่างกันในแต่ละรอบของการเรียนรู้ โดยจะปรับเพิ่มน้ำหนักที่เกี่ยวข้องกับตัวอย่างที่จำแนกประเภทไม่ถูกต้อง และปรับลดน้ำหนักที่เกี่ยวข้องกับตัวอย่างที่จำแนกประเภทถูกต้อง ซึ่งวิธีการนี้จะมุ่งเน้นที่ตัวอย่างที่จำแนกประเภทไม่ถูกต้องเพื่อให้มีความถูกต้องในการจำแนกประเภทครั้งต่อไป



ภาพที่ 1: กระบวนการวิธีการสุ่มตัวอย่างตามการแบ่งกลุ่มของข้อมูลที่ไม่สมดุล

3. วิธีการดำเนินการวิจัย

3.1 ขั้นตอนการทดลอง

ในการทดลอง จะนำข้อมูลมาปรับค่าให้อยู่ในมาตรฐานเดียวกัน เพื่อไม่ให้ค่าของข้อมูลมีผลต่อการปรับสมดุล และการจำแนกประเภท โดยการปรับค่านี้จะปรับให้มีค่าอยู่ระหว่าง 0-1 ทั้งหมด

หลังจากนั้น จะนำข้อมูลที่ได้มาใช้ในการทดลอง โดยจะทดสอบด้วยวิธีการ Random Undersampling, Cluster centroid undersampling, Cluster based undersampling, SMOTE, และวิธีการสุ่มตัวอย่างตามการแบ่งกลุ่มของข้อมูลที่ไม่สมดุลตามภาพที่ 1. ซึ่งกระบวนการของวิธีการสุ่มตัวอย่างตามการแบ่งกลุ่มของข้อมูลที่ไม่สมดุลเป็นดังนี้

นำข้อมูลทั้งหมดมาแบ่งกลุ่มข้อมูลด้วย k-means โดยจะแบ่งออกเป็นกลุ่มย่อย 4 กลุ่ม เพื่อไม่ให้ข้อมูลในแต่ละกลุ่มมีปริมาณน้อยเกินไป

เมื่อแบ่งกลุ่มเสร็จแล้ว จึงนำข้อมูลในแต่ละกลุ่มมาตรวจสอบเพื่อทำการเพิ่มจำนวนข้อมูลส่วนน้อยในแต่ละกลุ่ม โดยจะแบ่งออกมาได้ 2 กลุ่มหลักๆคือ

1. กลุ่มที่มีแต่ข้อมูลส่วนใหญ่ หรือข้อมูลส่วนน้อยเท่านั้น

2. กลุ่มที่มีทั้งข้อมูลส่วนใหญ่และข้อมูลส่วนน้อย

เมื่อมีการแบ่งกลุ่มของข้อมูลแล้ว จึงนำข้อมูลแต่ละกลุ่มมาจำแนกประเภท โดยที่ ถ้าข้อมูลอยู่ในกลุ่มที่มีเพียงข้อมูลประเภท ก็สามารถสรุปได้ว่า ข้อมูลนั้นเป็นข้อมูลประเภทเดียวกับข้อมูลส่วนใหญ่ แต่ถ้าในกลุ่มนั้นมีทั้งข้อมูลส่วนใหญ่และข้อมูลส่วนน้อย จะเอาข้อมูลในกลุ่มนี้

มาปรับสมดุลโดยเพิ่มจำนวนข้อมูลส่วนน้อยให้มีปริมาณเท่ากับข้อมูลส่วนใหญ่ด้วยวิธีการ SMOTE

เมื่อได้ชุดข้อมูลที่ปรับสมดุลแล้ว จึงนำไปจำแนกประเภทข้อมูลโดยใช้ต้นไม้ตัดสินใจ โมเดลต้นไม้ตัดสินใจที่ใช้บั้น จะแยกตามแต่ละกลุ่มของข้อมูลที่ได้จากแบ่งกลุ่มด้วย k-means

3.2 การตั้งค่าในการทดลอง

1. ชุดข้อมูล

ในงานวิจัยนี้ จะใช้ชุดข้อมูล 44 ชุดข้อมูลที่ใช้โดย Ginlar et al.[4] ซึ่งมีค่าอัตราส่วนของความไม่สมดุลอยู่ระหว่าง 1.8 ถึง 129 และมีจำนวนข้อมูลระหว่าง 130 ถึง 5500 โดยชุดข้อมูลจะมีการจำแนกประเภทได้ 2 ประเภท และในชุดข้อมูลแต่ละชุด มีการแบ่งเป็น 5-folds และใช้ชุดข้อมูลที่แบ่งแล้วในการทดลอง

ตารางที่ 1: ข้อมูลของชุดข้อมูลที่นำมาใช้

Datasets	No. of data	No. of features	IR
abalone19	4174.00	8.00	128.87
abalone9-18	731.00	8.00	16.68
ecoli-0_vs_1	220.00	7.00	1.86
ecoli-0137vs26	281.00	7.00	39.15
ecoli1	336.00	7.00	3.36
ecoli2	336.00	7.00	5.46
ecoli3	336.00	7.00	8.19
ecoli4	336.00	7.00	13.84
glass0123vs456	192.00	9.00	20.29
glass-0-1-6vs2	205.00	9.00	39.44
glass-0-1-6vs5	214.00	9.00	1.82
glass0	214.00	9.00	3.19
glass1	214.00	9.00	10.39
glass2	214.00	9.00	11.47
glass4	214.00	9.00	22.81
glass5	214.00	9.00	22.81
glass6	214.00	9.00	6.38
haberman	306.00	3.00	2.68
iris0	150.00	4.00	2.00
new-thyroid1	215.00	5.00	5.14
new-thyroid2	215.00	5.00	4.92
page-blocks-1-3-vs-4	472.00	10.00	15.85
page-blocks0	5472.00	10.00	8.77
pima	768.00	8.00	1.90
segment0	2308.00	19.00	6.01
shuttle-c0-vs-c4	1829.00	9.00	13.87
shuttle-c2-vs-c4	129.00	9.00	20.50
vehicle0	846.00	18.00	3.23
vehicle1	846.00	18.00	2.92
vehicle2	846.00	18.00	2.92
vehicle3	846.00	18.00	2.92
vowel0	988.00	13.00	10.10
wisconsin	683.00	9.00	1.86
yeast-0-5-6-7-9vs4	528.00	8.00	9.35
yeast-1_vs_7	459.00	8.00	14.87
yeast-1-2-8-9_vs_7	947.00	8.00	30.56
yeast-1-4-5-8_vs_7	693.00	8.00	22.10
yeast-2_vs_4	514.00	8.00	9.08
yeast-2_vs_8	482.00	8.00	23.10
yeast1	1484.00	8.00	2.46
yeast3	1484.00	8.00	8.10
yeast4	1484.00	8.00	28.10
yeast5	1484.00	8.00	32.78
yeast6	1484.00	8.00	39.15

2. การจำแนกประเภท

ในงานวิจัยนี้จะใช้ต้นไม้ตัดสินใจเป็นพื้นฐานในการจำแนกประเภท โดยพารามิเตอร์ชุดเดียวกันทั้งหมดเพื่อใช้ในการจำแนกประเภทของแต่ละวิธีการสุ่มเลือกตัวอย่างข้อมูล

3. การแบ่งกลุ่มข้อมูล

สำหรับวิธีต่างๆที่มีการแบ่งกลุ่มข้อมูล จะมีการใช้ k-means[10] เป็นวิธีการพื้นฐาน และพารามิเตอร์ที่ใช้ จะเป็นพารามิเตอร์เดียวกันทั้งหมด

4. การวัดประสิทธิภาพ

เพื่อทำการทดสอบโมเดลที่ได้กับชุดข้อมูลทดสอบ จะนำผลลัพธ์ที่ได้จากการจำแนกประเภทมาเปรียบเทียบกับค่าจริงของข้อมูลทดสอบ ทำให้สามารถวัดประสิทธิภาพของโมเดลได้ โดยค่าที่ใช้นี้ จะเป็นค่าต่างๆที่คำนวณได้มาจาก confusion matrix ซึ่งสามารถนำมาคำนวณค่าความถูกต้อง (accuracy) ค่าความแม่นยำ (precision) ค่าการตรวจสอบ (recall) ค่า F-measure รวมถึงพื้นที่ใต้กราฟ

3.3 ผลการดำเนินงาน

จากการทดลอง ได้มีการนำผลลัพธ์ที่ได้มาหาค่าความถูกต้อง (accuracy) ค่าความแม่นยำ (precision) ค่าการตรวจสอบ (recall) ค่า F-measure และพื้นที่ใต้กราฟ (AUC) โดยได้ผลเปรียบเทียบค่าความถูกต้องและ F-measure ของแต่ละชุดข้อมูลตามตารางที่ 2 และตารางที่ 3 ตามลำดับ

ตารางที่ 2: ตารางเปรียบเทียบค่าความถูกต้อง (accuracy) ของการจำแนกประเภทของแต่ละชุดข้อมูล

Datasets	Random Under Sampling	KMeans Cluster Centroid	KMeans Cluster Based	SMOTE	KMeans-SMOTE
abalone19	0.615	0.737	0.735	0.956	**0.977**
abalone9-18	0.756	0.780	0.780	**0.932**	**0.932**
ecoli-0_vs_1	0.985	0.989	0.990	**0.991**	0.988
ecoli-0137vs26	0.797	0.847	0.847	0.971	**0.978**
ecoli1	0.878	0.878	0.878	**0.893**	0.881
ecoli2	0.837	0.864	0.864	0.902	**0.921**
ecoli3	0.830	0.864	0.864	**0.909**	0.902
ecoli4	0.850	0.850	0.834	0.967	**0.979**
glass0123vs456	0.916	**0.985**	**0.985**	0.911	0.930
glass-0-1-6vs2	0.692	0.711	0.711	0.849	**0.899**
glass-0-1-6vs5	0.750	0.750	0.750	**0.972**	**0.972**
glass0	0.721	0.692	0.692	**0.785**	0.721
glass1	0.740	0.760	0.780	0.785	**0.790**
glass2	0.833	0.867	0.867	0.836	**0.869**
glass4	0.868	0.859	0.859	**0.988**	0.933
glass5	0.868	0.885	0.885	**0.986**	0.962
glass6	0.842	0.895	0.895	0.920	**0.967**
haberman	0.556	0.494	0.494	0.572	**0.637**
iris0	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
new-thyroid1	0.968	0.963	0.963	**0.986**	**0.986**
new-thyroid2	0.963	0.968	0.968	**0.986**	**0.986**
page-blocks-1-3-vs-4	0.975	0.913	0.913	0.991	**0.998**
page-blocks0	0.935	0.966	0.966	**0.969**	0.952

pima	0.681	0.679	0.679	**0.688**	0.686
segment0	0.970	0.964	0.966	**0.994**	0.990
shuttle-c0-vs-c4	0.998	0.998	0.998	0.998	**0.999**
shuttle-c2-vs-c4	**0.992**	**0.992**	**0.992**	**0.992**	0.969
vehicle0	0.911	0.922	0.932	0.937	**0.948**
vehicle1	0.735	0.735	0.735	**0.746**	0.739
vehicle2	0.931	0.917	0.917	0.943	**0.955**
vehicle3	0.713	0.704	0.708	0.736	**0.760**
vowel0	0.933	0.944	0.944	**0.984**	0.976
wisconsin	0.959	**0.963**	**0.963**	0.957	0.957
yeast-050-70-vs4	0.642	0.750	0.750	0.877	**0.898**
yeast-1_vs_7	0.688	0.658	0.658	0.863	**0.906**
yeast-1-2-8-9_vs_7	0.595	0.582	0.592	0.917	**0.934**
yeast-1-4-5-8_vs_7	0.862	0.870	0.870	0.869	**0.902**
yeast-2_vs_4	0.852	0.823	0.823	0.958	**0.968**
yeast-2_vs_8	0.634	0.623	0.623	0.933	**0.948**
yeast1	0.659	0.646	0.646	**0.781**	0.695
yeast3	0.884	0.866	0.866	**0.927**	0.926
yeast4	0.731	0.773	0.773	0.939	**0.949**
yeast5	0.921	0.937	0.937	**0.977**	0.976
yeast6	0.776	0.764	0.766	0.963	**0.968**

ตารางที่ 3: ตารางเปรียบเทียบประสิทธิภาพ (F-measure) ของการจำแนกประเภทของแต่ละชุดข้อมูล

Dataset	Random Under Sampling	KMean Cluster Centroid	KMeans Cluster Based	SMOTE	KMeans-SMOTE
abalone19	0.221	0.229	0.229	0.036	**0.428**
abalone9-18	0.257	0.274	0.276	0.309	**0.521**
ecoli-0_vs_1	0.956	0.964	0.966	**0.993**	0.966
ecoli-0-1-3-7_vs_2-6	0.143	0.186	0.186	**0.667**	0.393
ecoli1	**0.776**	0.766	0.766	0.772	0.747
ecoli2	0.620	0.576	0.576	0.726	**0.777**
ecoli3	0.474	0.333	0.333	**0.578**	0.549
ecoli4	0.861	0.357	0.357	0.731	**0.814**
glass0123vs456	0.843	**0.866**	**0.866**	0.810	0.849
glass-0-6-vs-2	0.232	0.226	0.226	0.164	**0.276**
glass-0-6-vs-5	0.393	0.382	0.382	**0.720**	0.633
glass0	0.627	0.623	0.625	**0.630**	0.589
glass1	0.670	0.603	0.605	0.686	**0.701**
glass2	0.220	0.186	0.186	0.090	**0.230**
glass4	0.594	0.213	0.213	0.581	**0.618**
glass5	0.833	0.461	0.461	**0.760**	0.653
glass6	0.631	0.642	0.642	0.716	**0.879**
haberman	**0.370**	0.361	0.361	0.303	0.360
iris0	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
new-thyroid1	0.969	0.927	0.927	**0.968**	**0.968**
new-thyroid2	0.959	0.883	0.883	**0.958**	**0.958**
page-blocks-1-3-vs-4	0.827	0.810	0.810	0.930	**0.982**
page-blocks0	0.743	0.762	0.762	**0.853**	0.750
pima	**0.588**	0.491	0.498	0.472	0.464
segment0	0.962	0.886	0.886	**0.979**	0.967
shuttle-c0-vs-c4	0.988	0.984	0.988	0.984	**0.996**
shuttle-c2-vs-c4	**0.913**	**0.913**	**0.913**	**0.913**	0.880
vehicle0	0.833	0.862	0.862	0.870	**0.899**
vehicle1	**0.594**	0.583	0.583	0.583	0.579
vehicle2	0.846	0851	0.851	0.893	**0.912**
vehicle3	0.530	**0.548**	**0.548**	0.502	0.532
vowel0	0.780	0.766	0.766	**0.953**	0.859
wisconsin	0.937	**0.961**	**0.961**	0.880	0.928
yeast-050-70-vs4	0.241	0.359	0.359	0.874	**0.877**
yeast-1_vs_7	0.219	0.219	0.219	0.214	**0.289**
yeast-1-2-8-9_vs_7	0.099	0.093	0.095	0.136	**0.253**
yeast-1-4-5-8_vs_7	0.093	0.098	0.098	0.073	**0.176**
yeast-2_vs_4	0.491	0.496	0.496	0.689	**0.758**
yeast-2_vs_8	0.130	0.130	0.130	0.350	**0.589**
yeast1	**0.540**	0.506	0.506	0.513	0.497
yeast3	0.630	0.589	0.589	**0.727**	0.686
yeast4	0.176	0.191	0.191	0.288	**0.366**
yeast5	0.822	0.454	0.458	**0.868**	0.643
yeast6	0.160	0.146	0.146	0.529	**0.661**

ผลเปรียบเทียบค่าความถูกต้องของการจำแนกหมวดหมู่เทียบกับจำนวน feature ตามตารางที่ 4 พบว่าวิธีการที่นำเสนอ มีประสิทธิภาพดีกับเกือบทุกขนาดของจำนวน feature แต่เนื่องจากจำนวนชุดข้อมูลที่มีปริมาณ feature ที่มากกว่า 10 feature ขึ้นไป มีเพียง 1 ชุดข้อมูลต่อช่วงของจำนวน feature จึงไม่สามารถบอกได้ชัดเจนว่าวิธีการที่นำเสนอสามารถใช้ได้กับทุกช่วงของจำนวน feature

ตารางที่ 4: ตารางเปรียบเทียบค่าความถูกต้อง (accuracy) ของการจำแนกประเภทของแต่ละชุดข้อมูลกับจำนวน feature

Feature	Random Under Sampling	KMean Cluster Centroid	KMeans Cluster Based	SMOTE	KMeans-SMOTE
4	0.760	0.747	0.747	0.786	**0.818**
6	0.963	0.954	0.954	**0.986**	**0.986**
8	0.747	0.764	0.764	0.904	**0.916**
10	0.841	0.888	0.888	0.918	**0.923**
14	0.933	0.944	0.944	**0.984**	0.977
18	0.837	0.822	0.822	0.841	**0.889**
20	0.870	0.864	0.864	**0.954**	0.947

ผลเปรียบเทียบจำนวนชุดข้อมูลที่มีประสิทธิภาพของการจำแนกสูงสุดตามตารางที่ 5 และ ผลเปรียบเทียบของค่าเฉลี่ยประสิทธิภาพจากทุกชุดข้อมูลตามตารางที่ 6

ตารางที่ 5: ตารางเปรียบเทียบจำนวนชุดข้อมูลที่มีประสิทธิภาพของการจำแนกประเภทสูงสุด

Measure	Random Under Sampling	KMean Cluster Centroid	KMeans Cluster Based	SMOTE	KMeans-SMOTE
Accuracy	2	1	1	20	**26**
Precision	3	3	3	18	**27**
Recall	**29**	24	24	4	4
F1	7	5	5	19	**24**
AUC	**14**	10	10	13	**14**

ตารางที่ 6: ตารางเปรียบเทียบประสิทธิภาพของการจำแนกประเภท

Measure	Random Under Sampling	KMean Cluster Centroid	KMeans Cluster Based	SMOTE	KMeans-SMOTE
Accuracy	0.804	0.797	0.797	0.900	**0.908**
Precision	0.463	0.457	0.457	0.605	**0.620**
Recall	0.806	**0.830**	**0.830**	0.660	0.653
F1	0.559	0.554	0.554	0.620	**0.626**
AUC	**0.806**	0.802	0.802	0.795	0.794

3.4 สรุป

ความไม่สมดุลของข้อมูลมีผลต่อการประสิทธิภาพของการจำแนกประเภทข้อมูล ทำให้การทำนายผลการจำแนกประเภทมีความลำเอียงไปทางข้อมูลส่วนใหญ่ ในการเพิ่มประสิทธิภาพการจำแนกประเภทข้อมูลนั้นสามารถทำได้โดยการปรับสมดุลของข้อมูลก่อนนำไปใช้ในการเรียนรู้ โดยวิธีการปรับสมดุลของข้อมูลสามารถทำได้โดยการลดจำนวนข้อมูลส่วนใหญ่ หรือเพิ่มจำนวนข้อมูลส่วนน้อย ในงานวิจัยนี้จึงได้นำเสนอวิธีการสุ่มตัวอย่างตามการแบ่งกลุ่มของข้อมูล โดยการแบ่งข้อมูล

ทั้งหมดเป็นกลุ่มย่อยก่อน แล้วจึงนำข้อมูลในแต่ละกลุ่มไปปรับสมดุล และนำข้อมูลในแต่ละกลุ่มไปใช้ในการเรียนรู้ตามแต่ละกลุ่มต่อไป ซึ่งวิธีการนี้ทำให้ค่าความถูกต้องและประสิทธิภาพโดยรวมของการจำแนกประเภทข้อมูลสูงกว่าการปรับสมดุลด้วยการเพิ่มจำนวนข้อมูลส่วนน้อยที่ทำกับข้อมูลทั้งชุด 0.44% และ 0.64% ตามลำดับ ทั้งนี้เพราะกลุ่มของข้อมูลที่ได้จากการแบ่งกลุ่มนั้น มีลักษณะไม่เหมือนกัน การเพิ่มจำนวนข้อมูลส่วนน้อย และการเรียนรู้แยกตามกลุ่ม จะทำให้ได้โมเดลที่ตรงกับลักษณะของข้อมูลในแต่ละกลุ่มมากกว่า

ทั้งนี้ ตัวอย่างชุดข้อมูลที่นำมาใช้ในการวิจัยเป็นชุดข้อมูลขนาดเล็ก ทำให้กลุ่มที่ได้จากการแบ่งกลุ่ม มีโอกาสที่จะมีจำนวนข้อมูลส่วนน้อยซึ่งมีจำนวนน้อยอยู่แล้วนั้นน้อยลงไปอีก ทำให้มีผลต่อการเพิ่มจำนวนข้อมูลส่วนน้อยในกลุ่มนั้น โดยเฉพาะกลุ่มที่มีจำนวนข้อมูลส่วนน้อยเพียงตัวเดียว ในกรณีเช่นนี้ อาจจะปรับวิธีการปรับสมดุล โดยใช้วิธีการสุ่มเพื่อเพิ่มจำนวนข้อมูลส่วนน้อยแทน SMOTE รวมถึงการทดสอบกับข้อมูลขนาดใหญ่ เพื่อหาอัตราส่วนระหว่างจำนวนกลุ่ม กับขนาดของข้อมูลที่เหมาะสมต่อไป

เอกสารอ้างอิง

[1] Wei-Chao Lina, Chih-Fong Tsai, Ya-Han Hue, Jing-Shang Jhang, "Clustering-based undersampling in class-imbalanced data, " *Information Sciences* vol. 409–410, October 2017, pp 17–26

[2] Y. Zhang, L. Zhang and Y. Wang, "Cluster-based majority under-sampling approaches for class imbalance learning," *2010 2nd IEEE International Conference on Information and Financial Engineering*, 2010, pp. 400-404

[3] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research* vol. 16, (2002) 321–357

[4] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince and F. Herrera, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches," *in IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 4, pp. 463-484

[5] Chawla N.V., Lazarevic A., Hall L.O., Bowyer K.W. "SMOTEBoost: Improving Prediction of the Minority Class in Boosting," *In: Lavrač N., Gamberger D., Todorovski L., Blockeel H. (eds) Knowledge Discovery in Databases: PKDD 2003. PKDD 2003. Lecture Notes in Computer Science*, vol 2838. Springer, Berlin, Heidelberg.

[6] X. Guo, Y. Yin, C. Dong, G. Yang and G. Zhou, "On the Class Imbalance Problem," *2008 Fourth International Conference on Natural Computation*, 2008, pp. 192-201.

[7] S. Na, L. Xumin and G. Yong, "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm," *2010 Third International Symposium on Intelligent Information Technology and Security Informatics*, 2010, pp. 63-67.

[8] Jerzy Blaszczyński n, Jerzy Stefanowski, "Neighbourhood sampling in bagging for imbalanced data, " *Neurocomputing* vol. 150, Part B, 20 February 2015, pp 529-542.

[9] YanminSunaMohamed S.KamelaAndrew K.C.WongbYangWang, "Cost-sensitive boosting for classification of imbalanced da, " *Pattern Recognition* Volume 40, Issue 12, December 2007, Pages 3358-3378.

[10] S. Na, L. Xumin and G. Yong, "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm," *2010 Third International Symposium on Intelligent Information Technology and Security Informatics*, 2010, pp. 63-67.

ประวัติผู้เขียน

ชื่อ-นามสกุล	นายอดิเทพ จิตพิทยาพร
ประวัติการศึกษา	วิทยาศาสตรบัณฑิต สาขาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยเชียงใหม่ ปีการศึกษา 2541
ตำแหน่งและสถานที่ทำงานปัจจุบัน	Head of Architect , บริษัท ทรูมันนี่ จำกัด