**ชื่อโครงการ (ภาษาไทย)** วิธีใหม่ในการสุ่มตัวอย่างแบบผสมสำหรับการจำแนกประเภทของชุดข้อมูลที่ไม่สมดุล

**ชื่อโครงการ (ภาษาอังกฤษ)** A New Hybrid Sampling Method for Imbalanced Datasets Classification

**แหล่งเงิน** เงินรายได้คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

**ประจำปีงบประมาณ** 2560 จำนวนเงินที่ได้รับการสนับสนุน 45,000 บาท

**ระยะเวลาทำการวิจัย** 1 ปี ตั้งแต่ วันที่ 1 ตุลาคม พ.ศ. 2559 ถึง วันที่ 30 กันยายน พ.ศ. 2560

**ชื่อ-สกุล หัวหน้าโครงการ**

**หัวหน้าโครงการวิจัย** ผศ.ดร.อนันตพร หรรษคุณาฒัย สัดส่วน 100%

**หน่วยงานต้นสังกัด** ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

## บทคัดย่อ

ความไม่สมดุลของประเภทข้อมูลเป็นปัญหาหนึ่งที่สำคัญในกระบวนการเรียนรู้ของเครื่อง ซึ่งปัญหาดังกล่าวส่งผลกระทบต่อประสิทธิภาพการทำนายของโมเดล หนึ่งในวิธีการแก้ปัญหาความไม่สมดุลของประเภทข้อมูลที่ได้รับความนิยม คือ เทคนิคการสุ่มตัวอย่างซึ่งเป็นการแก้ปัญหาในระดับข้อมูล งานวิจัยนี้จึงทำการพัฒนาเทคนิคการสุ่มตัวอย่างแบบใหม่ขึ้นมาที่มีชื่อว่า “DBSM” ซึ่งเป็นเทคนิคผสมระหว่างการเพิ่มจำนวนข้อมูลร่วมกับการลดจำนวนข้อมูล นอกจากนี้ได้นำขั้นตอนวิธีเชิงพันธุกรรมมาประยุกต์ใช้กับอัลกอริทึม DBSM ในการหาคำตอบที่เหมาะสมในการแก้ปัญหาความไม่สมดุลของข้อมูล (GADBSM) จากผลการทดลอง เมื่อเปรียบเทียบเทคนิค DBSM กับเทคนิคการสุ่มตัวอย่างทั้ง 3 เทคนิค ได้แก่ SMOTE Tomek Links และ SMOTE+Tomek Links ในอัลกอริทึมการเรียนรู้ต้นไม้ตัดสินใจ ตัวจำแนกแบบเบย์อย่างง่าย และเพื่อนบ้านใกล้เคียงที่สุด k ตัว พบว่าเทคนิค GADBSM ให้ค่าเฉลี่ยของ F-measure และ AUC สูงที่สุดเมื่อเทียบกับเทคนิคการสุ่มตัวอย่างแบบอื่นในอัลกอริทึมการเรียนรู้ตัวจำแนกแบบเบย์อย่างง่ายและต้นไม้ตัดสินใจ ตามลำดับ นอกจากนี้เทคนิค GADBSM ยังสามารถเพิ่มประสิทธิภาพในการจำแนกประเภทข้อมูลในอัลกอริทึมการเรียนรู้ตัวจำแนกแบบเบย์อย่างง่าย ได้ถึง 7.18%

**คำสำคัญ :** ความไม่สมดุลของประเภทข้อมูล เทคนิคการสุ่มตัวอย่างแบบผสม เทคนิค SMOTE

**Research Title:** A New Hybrid Sampling Method for Imbalanced Datasets Classification

**Researcher:** Asst.Prof.Dr. Anantaporn Hanskunatai

**Faculty:** Science **Department:** Computer Science

## ABSTRACT

The class imbalance is a major problem in machine learning. This problem affects the performance of a model prediction. A popular technique to handle the class imbalance problem is a sampling technique that is solving in data level. Thus, this research proposes a new hybrid-sampling algorithm, called DBSM. This technique combines over-sampling and under-sampling techniques to deal with the class imbalance for two-classes classification problem. In addition, genetic algorithm is applied for parameters tuning in the DBSM algorithm and called GADBSM. The experimental results of DBSM are compared with three sampling techniques which are SMOTE, Tomek Links, and SMOTE+Tomek Links based on three learning algorithms which are decision tree, naivebayes, and k-nearest neighbors. The results show that the GADBSM algorithm yields the best in averages of F-measure and AUC when compared with other sampling techniques on naivebayes and decision tree learning algorithms. Moreover, GADBSM can improve the classification performance on naivebayes algorithm upto 7.18%.

**Keywords :** imbalanced dataset, hybrid-sampling, SMOTE