กรอบงานการสุ่มเพิ่มตัวอย่างข้างน้อยสำหรับปัญหาความไม่ดุลระหว่างกลุ่ม

นายวัชรศักดิ์ ศิริเสรีวรรณ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต
สาขาวิชาวิทยาการคณนา ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2556

MINORITY OVERSAMPLING FRAMEWORK FOR CLASS IMBALANCE PROBLEM

Mr. Wacharasak Siriseriwan

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy Program in Computational Science
Department of Mathematics and Computer Science
Faculty of Science
Chulalongkorn University
Academic Year 2013

Thesis Title	MINORITY OVERSAMPLING FRAMEWORK FOR CLASS IMBALANCE PROBLEM
By	Mr. Wacharasak Siriseriwan
Field of Study	Computational Science
Thesis Advisor	Assistant Professor Krung Sinapiromsaran, Ph.D.

Accepted by the Faculty of Science, Chulalongkorn University in Partial Fulfillment of the Requirements for the Doctoral Degree

....Dean of the Faculty of Science

(Professor Supot Hannongbua, Dr.rer.nat.)

THESIS COMMITTEE

....Chairman

(Assistant Professor Khamron Mekchay, Ph.D.)

....Thesis Advisor

(Assistant Professor Krung Sinapiromsaran, Ph.D.)

....Examiner

(Assistant Professor Jaruloj Chongstitvatana, Ph.D.)

....Examiner

(Boonyarit Intiyot, Ph.D.)

....Examiner

(Phantipa Thipwiwatpotjana, Ph.D.)

....External Examiner

(Kamol Keatruengkammala, Ph.D.)

วัชรศักดิ์ ศิริเสรีวรรณ : กรอบงานการสุ่มเพิ่มตัวอย่างข้างน้อยสำหรับปัญหาความไม่ดุลระหว่างกลุ่ม. (MINORITY OVERSAMPLING FRAMEWORK FOR CLASS IMBALANCE PROBLEM) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ. ดร.กรุง สินอภิรมย์สราญ, 129 หน้า.

วิทยานิพนธ์นี้ได้ปรับปรุงแก้ไขวิธีการสุ่มเพิ่มตัวอย่างที่ใช้ในปัญหาความไม่ดุลระหว่างกลุ่ม จุดด้อยของวิธีการสุ่มเพิ่มตัวอย่างที่มีอยู่ได้ถูกวิเคราะห์และกรอบงานสุ่มตัวอย่างข้างน้อยได้ถูกเสนอเพื่อแก้ไขจุดด้อยเหล่านี้พร้อมการเพิ่มประสิทธิภาพในการแบ่งกลุ่ม งานวิจัยสามชิ้นในกรอบงานนี้ได้จัดการกับแง่มุมที่เป็นจุดด้อยของวิธีการสุ่มตัวอย่างที่มีอยู่ งานชิ้นแรกคือ Relocating Safe-level SMOTE ที่หลีกเลี่ยงการสังเคราะห์ข้อมูลใกล้กับจุดข้อมูลกลุ่มข้างมาก งานชิ้นที่สองคือ Adaptive Neighbor SMOTE (ANS) ที่ให้จำนวนเพื่อนบ้านแบบพลวัต ที่เป็นกระบวนการหนึ่งในวิธีการ SMOTE งานชิ้นสุดท้ายคือ ขั้นตอนการจัดการจุดข้อมูลข้างน้อยนอกคอกด้วยเพื่อนบ้านที่ใกล้ที่สุด สำหรับจุดข้อมูลส่วนเกินของกลุ่มข้างน้อย เพื่อพัฒนาผลลัพธ์ในการแบ่งกลุ่ม โดยที่ minority outcast handling นี้จะเป็นส่วนเพิ่มเติมของ RSLS และ ANS เพื่อเพิ่มความแม่นยำของทั้งสองวิธี ผลการทดลองบนชุดข้อมูลมาตรฐาน 14 ชุดและตัวแบบจำแนกประเภท 5 แบบ แสดงว่าวิธีการสุ่มเพิ่มตัวอย่างทั้งสองและขั้นตอนการจัดการจุดข้อมูลข้างน้อยนอกคอก สามารถเอาชนะวิธีการสุ่มเพิ่มตัวอย่างข้างน้อยอื่นๆ ในชุดข้อมูลส่วนใหญ่ ภายใต้ตัววัด F-measure, geometric mean และ adjusted geometric mean นอกจากนี้การทดสอบวิลคอกซันถูกใช้เพื่อแสดงให้เห็นว่าการพัฒนาขึ้นโดยรวมที่เกิดจากวิธีการทั้งสองมีนัยสำคัญทางสถิติ

ภาควิชา	คณิตศาสตร์และวิทยาการคอมพิวเตอร์	ลายมือชื่อนิสิต วัชรศักดิ์ ศิริเสรี
		ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก
สาขาวิชา	วิทยาการคณนา	
ปีการศึกษา	2556	

5273886323 : MAJOR COMPUTATIONAL SCIENCE
KEYWORDS: CLASS IMBALANCE PROBLEM / DATA MINING / CLASSIFICATION

WACHARASAK SIRISERIWAN: MINORITY OVERSAMPLING FRAMEWORK FOR CLASS IMBALANCE PROBLEM. ADVISOR: ASST. PROF. KRUNG SINAPIROMSARAN, Ph.D., 129 pp.

This dissertation enhances oversampling techniques which are used in a class imbalance problem. Several weaknesses of existing oversampling techniques are investigated and the minority oversampling framework is suggested to overcome these weaknesses and improves the classification performances. This dissertation provides the framework which contains three research works that deal with different aspects of existing oversampling techniques. The first work is Relocating Safe-level SMOTE (RSLS) to avoid conflicted synthetic instances near majority instances. The second work is Adaptive Neighbor SMOTE (ANS) which provides the dynamic number of nearest neighbors in SMOTE algorithm. The final work is the minority outcast handling process with 1-nearest neighbor to handle noises of positive instances in the dataset for improving the classification performance. This minority outcast handling process is augmented into RSLS and ANS to boost their accuracies. The experimental results on 14 benchmark datasets and 5 classifiers confirm that both oversampling techniques with minority outcast handling outperform other oversampling techniques in most datasets under three performance measures; F-measure, geometric mean and adjusted geometric mean. Wilcoxon sign ranked test is conducted to verify that the improvements caused by these two oversampling techniques are statistically significant.

Department:	Mathematics and Computer Science	Student's Signature	....
Field of Study:	Computational Science	Advisor's Signature	....
Academic Year:	2013		

ACKNOWLEDGEMENTS

Firstly, I would like to express my deepest gratitude to Assistant Professor Dr. Krung Sinapiromsaran, who is my advisor for the time throughout my master and doctorate degrees. Without his knowledge, suggestion and guidance in this dissertation, none of this work would become reality. I would like to thank all of my dissertation committees who provide suggestions and advices to complete this dissertation. Moreover, I am grateful to every academic staff in the department of Mathematics and Computer Science, Chulalongkorn University for their knowledge, suggestion and support. I want to reserve my appreciation to Applied Mathematics and Computational Science program for supporting in resources.

Also, I would like to thank my financial sponsor, the Development and Promotion of Science and Technology (DPST), Institute of the Promotion of Teaching Science and Technology (IPST), for the scholarship.

Finally, I would like to thank my family and friends for believing in me on pursuing doctoral degree and giving all kinds of encouragement. I also want to thank all colleagues, friends, seniors and juniors of AMCS and CU who stayed with me and provided their supports in many ways during my hard time in this doctorate course. Without any of them, I could not stay in this difficult and steep path until reaching the finish line.

CONTENTS

Page

Page

LIST OF FIGURES

Page

Page

LIST OF TABLES

Page

Page

Page