

กรอบงานการสุ่มเพิ่มเติมตัวอย่างข้างน้อยสำหรับปัญหาความไม่ดุลระหว่างกลุ่ม



นายวัชรศักดิ์ ศรีเสวีวรรณ



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต

สาขาวิชาวิทยาการคณนา ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์

คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2556

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

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


Copyright of Chulalongkorn University



5273886323

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.

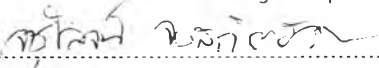
Accepted by the Faculty of Science, Chulalongkorn University in Partial Fulfillment of the Requirements for the Doctoral Degree



..... Dean of the Faculty of Science
(Professor Supot Hannongbua, Dr.rer.nat.)

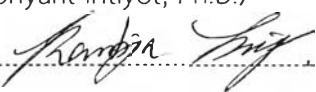
THESIS COMMITTEE



..... 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 Thipwivatpotjana, Ph.D.)


..... External Examiner
(Kamol Keatruengkammala, Ph.D.)

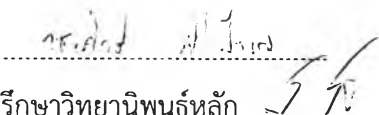
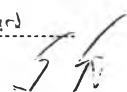
วัชรศักดิ์ ศิริเสวีวรรณ : กรอบงานการสุ่มเพิ่มตัวอย่างข้างน้อยสำหรับปัญหาความไม่ดุลระหว่างกลุ่ม. (MINORITY OVERSAMPLING FRAMEWORK FOR CLASS IMBALANCE PROBLEM) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ. ดร.กรุง สินอภิรมย์สราน, 129 หน้า.

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 คณิตศาสตร์และวิทยาการ
คอมพิวเตอร์

สาขาวิชา วิทยาการคณนา

ปีการศึกษา 2556

ลายมือชื่อนิสิต 
ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก 

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.

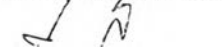
This dissertation enhances oversampling techniques which are used in a class imbalance problem. Several weaknesses of existing oversampling techniques are investigated and the minority oversampling framework is suggested to overcome these weaknesses and improves the classification performances. This 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

Field of Study: Computational Science

Academic Year: 2013

Student's Signature 

Advisor's Signature 

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
THAI ABSTRACT	iv
ENGLISH ABSTRACT	v
ACKNOWLEDGEMENTS	vi
CONTENTS	vii
LIST OF FIGURES	ix
LIST OF TABLES	xi
CHAPTER 1 INTRODUCTION	1
CHAPTER 2 BACKGROUND KNOWLEDGE.....	6
2.1 Classification models	6
2.1.1 Decision Tree	6
2.1.2 Naïve Bayes Model for classification.....	8
2.1.3 Support Vector Machine.....	11
2.1.4 Neural network.....	14
2.1.5 K -nearest neighbor (K -NN).....	16
2.2 Class imbalance problem	18
2.2.1 Data preprocessing techniques for class imbalance problem.	19
Synthetic oversampling techniques.....	21
2.2.1.1 Synthetic Minority Oversampling TEchnique (SMOTE).....	21
2.2.1.2 Adaptive Synthetic Sampling (ADASYN)	22
2.2.1.3 Safe-level SMOTE : Safe-level Synthetic Minority Oversampling TEchnique	24
2.2.1.4 Density-based Synthetic Minority Oversampling TEchnique (DBSMOTE).....	27
2.2.2 Cost-sensitive learning techniques.....	30
2.2.3 Algorithmic techniques for class imbalance problem	33
2.3 Performance measures.....	35

CHAPTER 3 MINORITY OVERSAMPLING FRAMEWORK FOR CLASS IMBALANCE PROBLEM	39
3.1 Minority outcast handling	39
3.2 Triangular minority oversampling technique	42
3.3 Relocating framework for safe-level SMOTE	44
3.4 Adaptive neighbors Synthetic Minority Oversampling TEchnique under 1-NN outcast handling	49
3.5 The time complexity analysis	52
CHAPTER 4 EXPERIMENTAL RESULTS AND ANALYSIS	55
4.1 Datasets and experimental settings	55
4.1.1 The description of benchmark datasets	55
4.1.2 Experimental settings	56
4.1.3 Wilcoxon signed-rank test	59
4.2 The result analysis	62
4.2.1 Triangular minority oversampling technique	62
4.2.2 Relocating safe-level SMOTE	66
4.2.3 Adaptive neighbors SMOTE	70
CHAPTER 5 DISCUSSION AND CONCLUSION	83
Future works	84
REFERENCES	86
VITA	129

LIST OF FIGURES

	Page
Figure 1: Knowledge discovery in databases process (KDD process).....	1
Figure 2: An example of decision tree	7
Figure 3: A visualization of simple support vector machine	12
Figure 4: The multilayer in neural network.....	16
Figure 5: The visualization of k -nearest neighbor	18
Figure 6: The scatter plots of generated datasets; a) an original imbalanced dataset and b) a balanced dataset with SMOTE	22
Figure 7: The scatter plot of a generated dataset after balancing with ADASYN.....	24
Figure 8: The visualization of SLS on the adjusted range due to the safe-level ratio.	26
Figure 9: The scatter plots of generated dataset a) an original imbalanced dataset, b) a balanced dataset by SMOTE and c) a balanced dataset by safe-level SMOTE.....	26
Figure 10: The scatter plots of generated dataset; a) an original imbalanced dataset and b) positive instances clustering with DBSCAN.....	28
Figure 11: The synthetic generation process of DBSMOTE	28
Figure 12: The scatter plot of a generated dataset after balanced by DBSMOTE	29
Figure 13: An example of a minority outcast in a dataset.....	40
Figure 14: A visualization showing that synthetic instances are not generated outside the convex hull of the original positive region	43
Figure 15: The comparison of synthetic generation of SMOTE and TMOT.....	44
Figure 16: An example of relocating a synthetic instance	46
Figure 17: A diagram of relocating safe-level SMOTE with 1-NN minority outcast handling.....	49
Figure 18: A visualization of assigning the number of K process.....	50
Figure 19: The flowchart of Adaptive neighbors Synthetic Minority Oversampling TEchnique under 1-NN outcast handling.....	52
Figure 20: The graph showing the percentage of outcast instances in each dataset when the value of c is varied.	57
Figure 21: The diagram of the experimental process in each round of train-test sampling	58
Figure 22: The table of the critical upper and lower bound values of W when n is no more than 20.....	61
Figure 23: The comparison of the average F-measure from ORIG, SMOTE and TMOT using a decision tree as a classifier	62

Figure 24: The comparison of the average F-measure from ORIG, SMOTE and TMOT using a naïve Bayes classifier as a classifier.....	63
Figure 25: The comparison of the average F-measure from ORIG, SMOTE and TMOT using a multilayer perceptron as a classifier.....	63
Figure 26: The comparison of the average F-measure from ORIG, SMOTE and TMOT using a support vector machine as a classifier.....	64
Figure 27: The comparison of the average F-measure from ORIG, SMOTE and TMOT using a k -nearest neighbor as a classifier.....	64
Figure 28: The bar chart of the number of datasets each oversampling technique achieves the best F-measure	67
Figure 29: The bar chart of the number of datasets which ANS1 and each oversampling technique achieves the best F-measure.....	71
Figure 30: The bar chart of the number of datasets which ANS1 and each oversampling technique achieves the top three F-measure	72
Figure 31: The bar chart of the number of datasets which ANS2 and each oversampling technique achieves the best F-measure.....	74
Figure 32: The bar chart of the number of datasets which ANS2 and each oversampling technique achieves the top three F-measure	74

LIST OF TABLES

	Page
Table 1: The condition of safe-level and safe-level ratio and their corresponding ranges	25
Table 2: A cost matrix of binary classification	30
Table 3: A simpler cost matrix with an equivalent optimal classification.....	31
Table 4: A confusion matrix of binary classification	36
Table 5: The summary of time complexities of SMOTE, safe-level SMOTE and suggested oversampling techniques in the framework	54
Table 6: The description of datasets used in the experiments.	56
Table 7: The null and alternative hypotheses in each type of Wilcoxon signed-rank test.....	59
Table 8: The Wilcoxon signed-rank test on F-measures from TMOT against ones from ORIG and SMOTE.....	65
Table 9: The list of dataset names which RSLs achieves the best, second best and third best F-measure in each classifier.	68
Table 10: The number of cases each technique achieves the average F-measure in the ranking 1 st -3 rd	68
Table 11: The Wilcoxon signed-rank of the difference of F-measure from RSLs against other sampling techniques.....	69
Table 12: The Wilcoxon signed-rank of the difference of F-measure from RSLs against other sampling techniques in each classifier.....	70
Table 13: The list of dataset names which ANS1 achieves the best, second best and third best F-measure in each classifier.	73
Table 14: The number of cases each oversampling technique achieves the F-measure in the ranking 1 st -3 rd	73
Table 15: The list of dataset names which ANS2 achieves the best, second best and third best F-measure in each classifier.	76
Table 16: The number of cases each oversampling technique achieves the F-measure in the ranking 1 st -3 rd	76
Table 17: The Wilcoxon signed-rank of the difference of F-measure from ANS1 and ANS2 against other oversampling techniques	77
Table 18: The Wilcoxon signed-rank of the difference of F-measure from ANS1 and ANS2 against other sampling techniques in each classifier	78
Table 19: The number of cases which averaged F-measure of ANS1 or ANS2 is	

	Page
higher/lower than one of SMOTEO-1 or SMOTEO-2.....	80
Table 20: The ANOVA table between F-measure values from SMOTE with the fixed $k = 5$ and the ones from ANS.....	81
Table 21: The ANOVA table between F-measure values from oversampling techniques without applying minority outcast handling and the ones with minority outcast handling	81
Table 22: The average percentage of minority outcasts in positive instances in each dataset when the number of c is varied	95
Table 23: The comparison of Triangular minority oversampling technique with using original imbalanced dataset (ORIG) and SMOTE under F-measure	97
Table 24: The comparison with relocating safe-level SMOTE under F-measure	99
Table 25: The comparison with adaptive neighbors SMOTE without minority outcast handling under F-measure.....	102
Table 26: The comparison with adaptive neighbors SMOTE with minority outcast handling under F-measure.....	105
Table 27: The comparison of F-measure from SMOTE with the default setting k as 5 and ANS.	108
Table 28: The comparison with relocating safe-level SMOTE under geometric mean	111
Table 29: The number of cases each technique achieves the average geometric mean in the ranking 1st -3rd.....	113
Table 30: The comparison with adaptive neighbors SMOTE without minority outcast handling under geometric mean.....	114
Table 31: The number of cases each technique achieves the average geometric mean in the ranking 1st -3rd.....	116
Table 32: The comparison with adaptive neighbors SMOTE with minority outcast handling under geometric mean.....	117
Table 33: The number of cases each technique achieves the average geometric mean in the ranking 1st -3rd.....	119
Table 34: The comparison with relocating safe-level SMOTE under adjusted geometric mean.....	120
Table 35: The number of cases each technique achieves the average adjusted geometric mean in the ranking 1st -3rd.....	122
Table 36: The comparison with adaptive neighbors SMOTE without minority outcast handling under adjusted geometric mean.....	123

	Page
Table 37: The number of cases each technique achieves the average adjusted geometric mean in the ranking 1st -3rd.....	125
Table 38: The comparison with adaptive neighbors SMOTE with minority outcast handling under adjusted geometric mean.....	126
Table 39: The number of cases each technique achieves the average adjusted geometric mean in the ranking 1st -3rd.....	128