การปรับการกระจายของข้อมูลที่ไม่สมดุลโดยใช้วิธีการหาเพื่อนบ้านที่ใกล้ที่สุด แบบย้อนกลับ K ตัว

BALANCING IMBALANCED DATA BY REVERSE K NEAREST NEIGHBOR

กัลยา พันธุนะผล

KALAYA PHUNTUMAPOL

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาวิทยาการคอมพิวเตอร์

คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2556

KMITL-2013-SC-M-022-021

การปรับการกระจายของข้อมูลที่ไม่สมดุลโดยใช้วิธีการหาเพื่อนบ้านที่ใกล้ที่สุดแบบย้อนกลับ K ตัว

BALANCING IMBALANCED DATA BY REVERSE K NEAREST NEIGHBOR

กัลยา พันธุมะผล

KALAYA PHUNTUMAPOL

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาวิทยาการคอมพิวเตอร์

คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2556

KMITL-2013-SC-M-002-021

BALANCING IMBALANCED DATA BY REVERSE K NEAREST NEIGHBOR

KALAYA PHUNTUMAPOL

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF SCIENCE IN COMPUTER SCIENCE

FACULTY OF SCIENCE

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2013

KMITL-2013-SC-M-002-021

คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ใบรับรองวิทยานิพนธ์

หัวข้อวิทยานิพนธ์	การปรับการกระจายของข้อมูลที่ไม่สมดุลโดยใช้วิธีการหาเพื่อนบ้านที่ใกล้ที่สุดแบบย้อนกลับ K ตัว Balancing Imbalanced Data by Reverse K Nearest Neighbor
นักศึกษา	นางสาวกัลยา พันธุมะผล
รหัสประจำตัว	51067507
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	วิทยาการคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์	ผศ.ดร.นวลสวาท หิรัญสกลวงศ์

คณะกรรมการสอบวิทยานิพนธ์		ลายมือชื่อ
ดร.สายชล	ใจเย็น	
รศ.ดร.จีรพร	วีระพันธุ์	
ดร.ชาคริต	วัชโรภาส	
ผศ.ดร.นวลสวาท	หิรัญสกลวงศ์	

วัน / เดือน/ ปี ที่สอบ 15 พฤษภาคม พ.ศ. 2556 เวลา 10.00 – 12.00 น.
สถานที่สอบ ณ ห้อง 304 ชั้น 3 อาคารจุฬาภรณวลัยลักษณ์ 1

คณะวิทยาศาสตร์รับรองแล้ว

(รองศาสตราจารย์ ดร.ดุษณี ธนะบริพัฒน์)
คณบดีคณะวิทยาศาสตร์
วันที่ 27 **เดือน** เมษายน **พ.ศ.** 56

หัวข้อวิทยานิพนธ์	การปรับการกระจายของข้อมูลที่ไม่สมดุลโดยใช้วิธีการหาเพื่อนบ้านที่ใกล้ที่สุดแบบย้อนกลับ K ตัว
นักศึกษา	นางสาวกัลยา พันธุมะผล
รหัสประจำตัว	51067507
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	วิทยาการคอมพิวเตอร์
พ.ศ.	2556
อาจารย์ที่ปรึกษา	ผศ.ดร.นวลสวาท หิรัญสกลวงศ์

บทคัดย่อ

ข้อมูลไม่สมดุล (imbalanced data set) คือข้อมูลที่มีการกระจายตัวที่ไม่เท่ากันในแต่ละประเภทข้อมูล (class) กล่าวคืออัตราส่วนของประเภทข้อมูลโน้มเอียงไปยังประเภทข้อมูลประเภทใดประเภทหนึ่ง ปัจจุบันเราสามารถพบข้อมูลไม่สมดุลในงานจริงจำนวนมาก ในงานวิจัยชิ้นนี้มุ่งเน้นเฉพาะการคัดแยกข้อมูลเพียง 2 ประเภท (binary classification) ซึ่งประกอบด้วยประเภทข้อมูลเล็ก (minority class) และประเภทข้อมูลใหญ่ (majority class) โดยกำหนดให้ประเภทข้อมูลเล็กคือประเภทข้อมูลที่มีอัตราส่วนข้อมูลจำนวนน้อยๆ และประเภทข้อมูลใหญ่คือ ประเภทข้อมูลที่มีอัตราส่วนข้อมูลจำนวนมากๆ อาทิ เช่น ในข้อมูลใดๆ อาจประกอบด้วย ประเภทข้อมูลเล็ก 0.5% และประเภทข้อมูลใหญ่ 95.5% เป็นต้น ปัญหาหลักของข้อมูลไม่สมดุลคือโมเดลการคัดแยกข้อมูลจะโน้มเอียงไปยังประเภทข้อมูลใหญ่ ส่งผลให้คัดแยกข้อมูลทั้งหมดเป็นประเภทข้อมูลใหญ่ ทั้งที่ในความเป็นจริงแล้วประกอบด้วยประเภทข้อมูลเล็กด้วย วิธีที่ใช้แก้ปัญหาข้อมูลไม่สมดุลก็คือการปรับการกระจายตัวของประเภทข้อมูล โดยการสร้างข้อมูลสังเคราะห์ประเภทเล็ก (synthetic minority data) ให้อัตราส่วนข้อมูลประเภทเล็กมีจำนวนใกล้เคียงกับข้อมูลประเภทใหญ่ ในการวิจัยชิ้นนี้นำเสนอวิธีการเพิ่มจำนวนข้อมูลโดยใช้วิธีย้อนกลับไปดูเพื่อนบ้านในการเลือกข้อมูลประเภทเล็กที่จะเพิ่มเข้าไปตามค่าความรู้ (knowledge value) ในการทดลองได้นำวิธีการที่นำเสนอวัดผลกับวิธีการสร้างข้อมูลสังเคราะห์ประเภทเล็กอื่นๆ อาทิ เช่น สโมท (SMOTE) และ การสร้างข้อมูลสังเคราะห์บริเวณขอบ (Borderline-SMOTE) จากผลการทดลองแสดงให้เห็นว่าวิธีการที่นำเสนอ

สามารถเพิ่มประสิทธิภาพการคัดแยกข้อมูลได้ดีกับชุดข้อมูลส่วนใหญ่ เมื่อใช้มาตรวัดค่าเอฟ (F-Measure) และพื้นที่ใต้เส้นโค้ง (AUC)

คำสำคัญ : ปัญหาจำนวนข้อมูลไม่เท่ากัน, การเพิ่มจำนวนข้อมูล, SMOTE

Thesis Title	Balancing Imbalanced Data by Reverse K Nearest Neighbor
Student	Miss. Kalaya Phuntumapol
Student ID	51067507
Degree	Master of Science
Program	Computer Science
Year	2013
Thesis Advisor	Asst. Prof. Dr. Nualsawat Hiransakolwong

Abstract

Imbalanced data set is a data set that has the skewed class distribution. The imbalanced data set is found in many practical works. This paper proposes only a binary classification. Each data set consists of two classes, i.e., minority and majority classes. The minority class is a class that has less proportion of instances while the majority class has higher proportion of instances. There is a problem with many traditional classifiers biased to the majority class. The biased classification model always misclassifies the minority class into the majority class. One of the previous methods to solve imbalanced class problem is synthetic over-sampling. The synthetic over-sampling generates synthetic minority data to balance class distribution. The proposed method selects the synthetic minority instances based on knowledge value of the Reverse K-Nearest Neighbors. Then the proposed method is compared with SMOTE and Borderline-SMOTE. The experimental results show that the proposed method can improve performance in term of F-Measure and AUC.

Keywords : Class imbalanced problem, Over-sampling data, SMOTE

กิตติกรรมประกาศ

วิทยานิพนธ์เล่มนี้สำเร็จลุลวงไปด้วยดี หากมิได้รับคำแนะนำ คำชี้แจ้ง ความรู้ และความเอาใจใส่จาก ผศ.ดร.นวลสวาท หิรัญสกลวงศ์ ผู้เป็นอาจารย์ที่ปรึกษา ซึ่งได้สละเวลาให้อย่างเต็มที่ ให้คำแนะนำให้คำปรึกษาอย่างใกล้ชิด และเสนอแนะแนวทางแก้ปัญหา รวมทั้งตรวจแก้วิทยานิพนธ์ฉบับนี้หลายรอบมาก ให้มีความสมบูรณ์เพิ่มขึ้น จึงใคร่ขอขอบพระคุณเป็นอย่างสูง

ขอขอบพระคุณ รศ.ดร. จีรพร วีระพันธุ์ ดร. สายชล ใจเย็น และดร. ชาคริต วัชโรภาส คณะกรรมการสอบวิทยานิพนธ์ ที่ให้คำแนะนำ ให้คำปรึกษา และเสนอแนะแนวทางแก้ปัญหา

ขอขอบพระคุณทุนสนับสนุนการทำวิทยานิพนธ์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ที่ให้ทุนสนับสนุนการทำวิทยานิพนธ์ และให้ทุนสนับสนุนในการนำเสนอการประชุมวิชาการ

ขอขอบคุณเจ้าหน้าที่ประจำสาขาวิทยาการคอมพิวเตอร์ รวมทั้งเจ้าหน้าที่ประจำภาคบัณฑิต คณะวิทยาศาสตร์ ที่ให้ความร่วมมือ และอำนวยความสะดวก ในการทำวิทยานิพนธ์ให้สำเร็จลุลวงไปด้วยดี

สุดท้ายนี้ผู้จัดทำ ขอขอบพระคุณ บิดา มารดา และบุคคลในครอบครัว ที่ได้ให้ความช่วยเหลือทุกๆด้าน รวมทั้งเพื่อนๆ พี่ๆ ที่ให้กำลังใจตลอดระยะเวลาในการทำวิทยานิพนธ์

กัลยา พันธุมะผล

สารบัญ

หน้า

สารบัญ(ต่อ)

หน้า

สารบัญ(ต่อ)

หน้า

สารบัญตาราง

สารบัญตาราง(ต่อ)

ตารางที่ หน้า

สารบัญรูป

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ปัญหาข้อมูลที่ไม่สมดุลกัน (imbalanced data set) หรือมีอัตราส่วนของจำนวนข้อมูลที่แตกต่างกันมากในแต่ละประเภทของข้อมูล (class) ซึ่งปัญหานี้สามารถเกิดขึ้นได้กับข้อมูลโดยทั่วไป ในงานวิจัยนี้พิจารณาข้อมูลเพียงสองประเภท (binary classification) โดยแบ่งประเภทข้อมูลเป็น ข้อมูลประเภทเล็ก (minority instance) กับข้อมูลประเภทใหญ่ (majority instance) ข้อมูลประเภทเล็กเป็นข้อมูลที่มีปริมาณข้อมูลเป็นอัตราส่วนน้อยของชุดข้อมูล ในขณะที่ข้อมูลประเภทใหญ่นั้นกลับมีอัตราส่วนข้อมูลเป็นปริมาณมาก อาทิ เช่น ในชุดข้อมูลใดๆ อาจประกอบด้วย ข้อมูลประเภทเล็ก 0.5% และ ข้อมูลประเภทใหญ่ 95.5% เป็นต้น แต่ชุดข้อมูลโดยทั่วไปแล้ว มักมีประเภทข้อมูลมากกว่าสองประเภท (multiclass) จึงต้องแปลงประเภทข้อมูลให้เหลือเพียงสองประเภท การแปลงประเภทข้อมูลเริ่มจากเลือกประเภทข้อมูลว่าข้อมูลประเภทใดที่จะกำหนดเป็นข้อมูลประเภทเล็กและประเภทข้อมูลนั้นต้องมีปริมาณข้อมูลเป็นอัตราส่วนน้อยของชุดข้อมูล ส่วนประเภทข้อมูลที่เหลือในชุดข้อมูลจะถูกรวมและแปลงเป็นข้อมูลประเภทใหญ่ สาเหตุที่ต้องทำการปรับสมดุลในข้อมูลที่ไม่สมดุลนั้นเนื่องมาจาก หากสร้างโมเดลจำแนกประเภทข้อมูลจากชุดข้อมูลที่ไม่สมดุล โมเดลที่สร้างขึ้นจะโน้มเอียงไปยังประเภทข้อมูลใหญ่ อีกทั้งในการวัดประสิทธิภาพการจำแนกประเภทข้อมูลนั้นโดยทั่วไปแล้วจะใช้ค่าความแม่นยำ (accuracy) และค่าอัตราความผิดพลาด (error rate) ซึ่งทั้งค่าความแม่นยำและอัตราความผิดพลาดนั้นใช้ค่าความถี่ในการคำนวณทำให้ทั้งค่าความแม่นยำและอัตราความผิดพลาดโน้มเอียงไปสู่ประเภทข้อมูลใหญ่ มีผลมาจากประเภทข้อมูลใหญ่นั้นจะมีปริมาณข้อมูลเป็นจำนวนมากหรือมีความถี่ของข้อมูลสูง ยังผลให้โมเดลจำแนกประเภทข้อมูลที่สร้างขึ้นจากข้อมูลที่ไม่สมดุลนั้นจะจำแนกข้อมูลทั้งหมดเป็นประเภทข้อมูลใหญ่ทั้งหมด ทั้งที่ในความเป็นจริงแล้วประกอบด้วยประเภทข้อมูลเล็กด้วย

จากการศึกษานั้นมีวิธีหลักอยู่ 2 วิธีที่ใช้ในการแก้ปัญหาข้อมูลที่ไม่สมดุล วิธีที่หนึ่งคือการเพิ่มตัวอย่างข้อมูล (over-sampling) และวิธีที่สองคือการลดตัวอย่างข้อมูล (under-sampling) การเพิ่มตัวอย่างข้อมูลเป็นการเพิ่มจำนวนข้อมูลประเภทเล็กให้มีปริมาณใกล้เคียงกับปริมาณข้อมูลใหญ่ โดยสุ่มข้อมูลประเภทเล็กมาสร้างซ้ำ แต่วิธีนี้อาจจะสร้างข้อมูลที่ซ้ำซ้อนกันกับข้อมูลเดิม ทำให้เกิดปัญหาการเข้ากันเกินไป (over-fitting) เป็นปรากฏการที่แบบจำลองหรือตัวจำแนกประเภทที่ได้มีความพอดีเกินไปกับชุดข้อมูลสอนโดยจะให้ผลการจำแนกประเภทข้อมูลที่ถูกต้องสำหรับตัวอย่างในชุดข้อมูลสอน (training data set) แต่ใช้ได้ไม่ดีกับกรณีตัวอย่างข้อมูลอื่น ส่วนการลดตัวอย่างข้อมูลเป็นการลดจำนวนข้อมูลประเภทใหญ่ให้มีปริมาณใกล้เคียงกับข้อมูลประเภทเล็ก ข้อเสียของ

การลดตัวอย่างข้อมูล ก็คือข้อมูลที่มีประโยชน์ต่อการแยกประเภทข้อมูลนั้นอาจถูกลบออกได้ ซึ่งในงานวิจัยนี้จะศึกษาเพียงวิธีการเพิ่มตัวอย่างข้อมูลโดยการสังเคราะห์ข้อมูลประเภทเล็กเท่านั้น

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

เพื่อแก้ปัญหาแบบจำลองหรือตัวจำแนกประเภทที่สร้างจากข้อมูลที่มีอัตราส่วนของจำนวนข้อมูลที่แตกต่างกันมากในแต่ละประเภทของข้อมูล ให้สามารถจำแนกประเภทข้อมูลเล็กได้ถูกต้องมากขึ้น ไม่ให้จำแนกข้อมูลทั้งหมดเป็นประเภทข้อมูลใหญ่เพียงอย่างเดียว ทั้งที่ในความเป็นจริงแล้วชุดข้อมูลที่นำมาเรียนรู้และทดสอบประกอบด้วยข้อมูลประเภทเล็กด้วย ซึ่งในงานวิจัยนี้ใช้วิธีการเพิ่มตัวอย่างข้อมูลประเภทเล็กโดยการสังเคราะห์ข้อมูลประเภทเล็กจากวิธีสโมท (SMOTE) และนำข้อมูลที่ถูกสังเคราะห์มาเลือก เพื่อให้ได้ข้อมูลประเภทเล็กที่ดีที่สุดมาเพิ่มให้กับชุดข้อมูลที่ไม่สมดุล แล้วนำชุดข้อมูลที่ถูกปรับสมดุลหรือเพิ่มตัวอย่างข้อมูลเสร็จมาสร้างแบบจำลองเพื่อใช้จำแนกประเภทข้อมูล รวมทั้งศึกษามาตรที่ใช้วัดในกรณีข้อมูลไม่สมดุลในแต่ละประเภทของข้อมูลเพื่อใช้ตรวจสอบว่าข้อมูลที่ถูกปรับสมดุลแล้ว สามารถจำแนกประเภทข้อมูลได้ถูกต้องมากขึ้นหรือไม่

1.3 ขอบเขตของงานวิจัย

งานวิจัยนี้วิเคราะห์ข้อมูลเพียงสองประเภท ข้อมูลที่สนใจซึ่งเป็นข้อมูลที่มีอัตราส่วนน้อยเมื่อเทียบกับปริมาณชุดข้อมูลจัดเป็นประเภทข้อมูลเล็ก และข้อมูลที่มีอัตราส่วนมากเมื่อเทียบปริมาณชุดข้อมูลจัดเป็นประเภทข้อมูลใหญ่ ในกรณีที่มีข้อมูลมากกว่าสองประเภทจะจัดข้อมูลส่วนที่เหลือนอกจากข้อมูลประเภทเล็ก นำมารวมแล้วจัดเป็นประเภทข้อมูลใหญ่ การแก้ปัญหาข้อมูลไม่สมดุลกันนั้น จะใช้เพียงวิธีการเพิ่มปริมาณข้อมูลประเภทเล็ก โดยสังเคราะห์ข้อมูลจากข้อมูลประเภทเล็กสองตัว เพื่อให้แบบจำลองที่สร้างจากชุดข้อมูลที่ถูกปรับสมดุลแล้วสามารถจำแนกประเภทข้อมูลได้ถูกต้อง ไม่จำแนกเป็นประเภทข้อมูลใหญ่เพียงอย่างเดียว รวมถึงศึกษาการสร้างแบบจำลองเพื่อจำแนกประเภทข้อมูล โดยใช้ต้นไม้การตัดสินใจ C4.5 (Decision tree, C4.5)

การหาเพื่อนบ้านใกล้ที่สุด (k-Nearest Neighbors Algorithm, k-NN) ในงานวิจัยใช้ k = 5 เนื่องมาจากวิธีที่เกี่ยวข้อง เช่น สโมท(SMOTE) และ การสร้างข้อมูลสังเคราะห์บริเวณขอบ (Borderline-SMOTE) ใช้ค่านี้ในการคำนวณหาเพื่อนบ้านที่ใกล้ที่สุด เพื่อให้สามารถนำวิธีที่เสนอไปเปรียบเทียบกับวิธีที่เกี่ยวข้องได้

ชุดข้อมูลที่นำมาใช้ในการทดลองนั้นต้องมีค่าครบทุกข้อมูล ไม่มีข้อมูลที่สูญหาย (Missing value) ต้องเป็นข้อมูลที่เป็นตัวเลขหรือจำนวนจริง (Integer data and Real data) ทั้งหมด แต่หากข้อมูลเป็นข้อมูลแบบประเภท (Category data) จะต้องแปลงให้อยู่ในรูปจำนวนเต็ม ยกเว้นข้อมูลที่

ใช้ระบุเป็นคลาสเท่านั้นที่ต้องเป็นข้อมูลแบบประเภท และชุดข้อมูลที่นำมาใช้นั้นต้องเป็นชุดข้อมูลที่ใช้ในการจำแนกประเภท หรือจัดกลุ่มข้อมูลเท่านั้น

1.4 ส่วนประกอบวิทยานิพนธ์

ส่วนประกอบที่เหลือของวิทยานิพนธ์ฉบับนี้ประกอบด้วย

บทที่ 2 กล่าวถึงความรู้ทั่วไปของปัญหาข้อมูลไม่สมดุล วิธีที่ใช้แก้ปัญหาข้อมูลไม่สมดุลกัน อาทิเช่น การสุ่มข้อมูลซ้ำ (Resampling) และ การสังเคราะห์ข้อมูลประเภทเล็ก (Synthetic minority instances) เป็นต้น รวมถึงการสร้างแบบจำลองโดยใช้ต้นไม้การตัดสินใจเพื่อสร้างโมเดลจำแนกประเภทข้อมูล

บทที่ 3 อธิบายเกี่ยวกับวิธีการเลือกข้อมูลสังเคราะห์ที่ทำการพัฒนาและนำมาประยุกต์ใช้กับการสังเคราะห์ข้อมูลประเภทเล็กของวิธีสโมท

บทที่ 4 เป็นการศึกษาเกี่ยวกับวิธีการวัดคุณภาพของการจำแนกประเภทข้อมูลที่ได้และประสิทธิภาพของชุดข้อมูลที่ถูกเพิ่มตัวอย่างข้อมูลของวิธีการที่ทำการพัฒนา เทียบกับวิธีการเพิ่มตัวอย่างข้อมูลแบบสโมท, วิธีการเพิ่มตัวอย่างข้อมูลโดยการสร้างข้อมูลสังเคราะห์บริเวณขอบ และวิธีการที่ไม่มีการเพิ่มตัวอย่างแก่ข้อมูลต้นฉบับ เพื่อใช้เปรียบเทียบกับวิธีการต่างๆว่าสามารถเพิ่มประสิทธิภาพได้ปริมาณเท่าใด

บทที่ 5 สรุปผลการทดลองและข้อเสนอแนะเกี่ยวกับวิธีการเพิ่มตัวอย่างข้อมูลโดยใช้การปรับการกระจายของข้อมูลที่ไม่สมดุลโดยใช้วิธีการหาเพื่อนบ้านที่ใกล้ที่สุดแบบย้อนกลับ K ตัว

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงความรู้ทั่วไปของปัญหาข้อมูลที่ไม่สมดุลกัน วิธีที่ใช้แก้ปัญหาข้อมูลที่ไม่สมดุลกัน อาทิเช่น การสุ่มข้อมูลซ้ำ (resampling) และการสังเคราะห์ข้อมูลประเภทเล็ก (synthetic minority data) เป็นต้น รวมถึงการสร้างแบบจำลองหรือโมเดลโดยใช้ต้นไม้การตัดสินใจเพื่อสร้างโมเดลจำแนกประเภทข้อมูล

2.1 ปัญหาข้อมูลไม่สมดุลกัน

ปัญหาข้อมูลไม่สมดุลกัน [7] เป็นปัญหาที่เกิดจากชุดข้อมูลมีปริมาณข้อมูลประเภทหนึ่งน้อยมากหรือมีค่าแตกต่างมากจากประเภทข้อมูลที่เหลือ ในงานวิจัยนี้จะแบ่งข้อมูลเพียงสองประเภท คือ ข้อมูลประเภทเล็ก (minority instance) กับข้อมูลประเภทใหญ่ (majority instance) โดยข้อมูลประเภทเล็กเป็นข้อมูลที่มีอัตราส่วนน้อยเมื่อเทียบกับข้อมูลใหญ่ ส่วนข้อมูลใหญ่เป็นข้อมูลที่มีอัตราส่วนมากเมื่อเทียบกับจำนวนข้อมูลทั้งหมด สามารถพบปัญหานี้ได้ในข้อมูลจริง อาทิ เช่น การตรวจเนื้อเยื่อมะเร็ง, การตรวจสอบการรั่วไหลของน้ำมันจากภาพเรดาร์ของดาวเทียม [6], การคัดแยกประเภทข้อความ (text classification) [11], การสืบค้นสารสนเทศ (information retrieval), การคัดกรองงาน (data filtering) [4] และอื่นๆ เหตุผลที่ต้องปรับการกระจายตัวของข้อมูลที่ไม่สมดุลเพราะว่าแบบจำลองการคัดแยกประเภทข้อมูลโดยทั่วไปแล้วจะใช้ค่าความแม่นยำและอัตราส่วนความผิดพลาด (error rate) เพื่อวัดประสิทธิภาพของแบบจำลองที่สร้างขึ้นมา ซึ่งมีค่าโน้มเอียงไปตามข้อมูลจำนวนมากหรือข้อมูลประเภทใหญ่ ทำให้แบบจำลองที่ใช้คัดแยกประเภทข้อมูลจะทำนายข้อมูลเป็นประเภทข้อมูลใหญ่เพียงอย่างเดียว ทั้งที่จริงแล้วชุดข้อมูลที่ใช้ทดสอบนั้นประกอบด้วยข้อมูลเล็กและข้อมูลใหญ่ จากที่กล่าวมาจะเห็นได้ว่าปัญหาข้อมูลที่ไม่สมดุลกันนั้นเป็นปัญหาที่สำคัญอย่างหนึ่งของการทำเหมืองข้อมูล (data mining) [10]

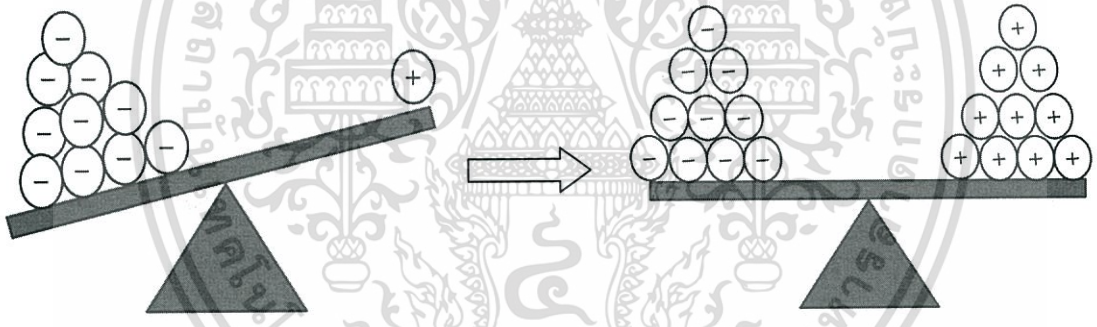
2.2 วิธีการแก้ปัญหาข้อมูลไม่สมดุลกัน

2.2.1 การสุ่มข้อมูลซ้ำ

การสุ่มข้อมูลซ้ำแบ่งเป็น การเพิ่มตัวอย่างข้อมูล (Over-sampling) กับการลดตัวอย่างข้อมูล (Under-sampling) [8]

2.2.1.1 การเพิ่มตัวอย่างข้อมูล

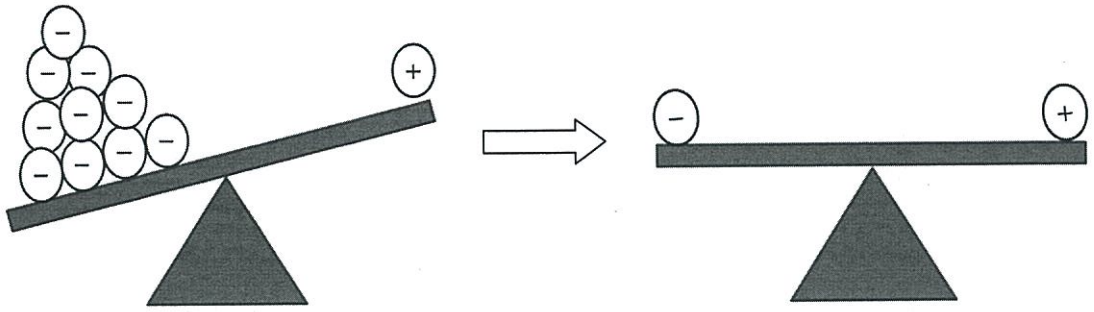
การเพิ่มตัวอย่างข้อมูล (Over-sampling) เป็นการสุ่มข้อมูลประเภทเล็กมาสร้างเพิ่มขึ้นจากข้อมูลเดิม มาสร้างซ้ำจนมีจำนวนใกล้เคียงกับข้อมูลประเภทใหญ่ ทำให้ข้อมูลเกิดความซ้ำซ้อนและทำให้ขอบเขตของข้อมูลประเภทเล็กมีความเฉพาะเจาะจง ยังให้เกิดปัญหาการเข้ากันเกินไปของข้อมูล ทำให้แบบจำลองหรือตัวจำแนกประเภทที่ได้มีความพอดีเกินไปกับชุดข้อมูลเรียนรู้ ทำให้ผลการจำแนกประเภทข้อมูลนั้นถูกต้องสำหรับตัวอย่างในชุดข้อมูลสอนเท่านั้น แต่ใช้ไม่ได้ดีกับกรณีชุดข้อมูลอื่น



รูปที่ 2.1 แสดงการเพิ่มตัวอย่างข้อมูล

2.2.1.2 การลดตัวอย่างข้อมูล

การลดตัวอย่างข้อมูล (Under-sampling) เป็นการสุ่มข้อมูลประเภทใหญ่มาลดจำนวนให้มีปริมาณใกล้เคียงกับข้อมูลประเภทเล็ก ข้อเสียของวิธีนี้ ก็คือข้อมูลที่มีประโยชน์ต่อการคัดแยกประเภทข้อมูลนั้นอาจถูกลบออกไป ส่งผลให้เกิดปัญหาไม่สามารถเรียนรู้ (under-fitting) เป็นปัญหาที่เกิดจากการมีตัวอย่างข้อมูลที่ใช้เรียนรู้น้อยเกินไปทำให้ โมเดลที่สร้างขึ้นเมื่อผ่านการลดตัวอย่างข้อมูลแล้วไม่สามารถคัดแยกประเภทข้อมูลได้ถูกต้อง

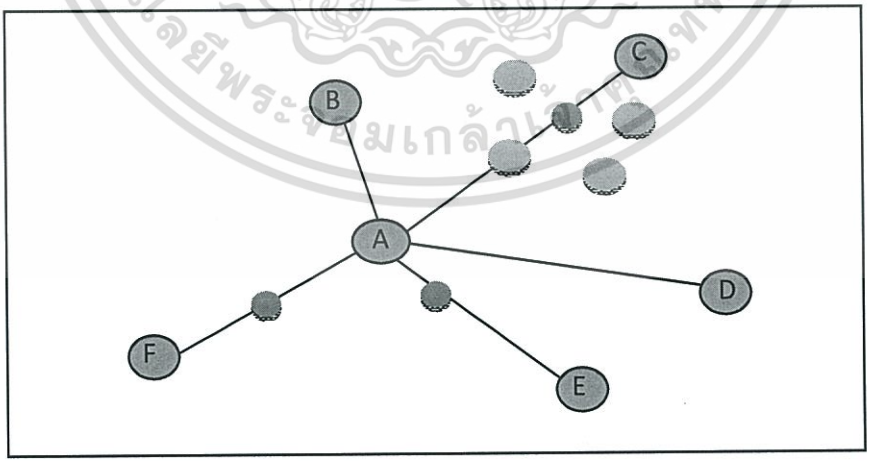


รูปที่ 2.2 แสดงการลดตัวอย่างข้อมูล

2.2.2 การสังเคราะห์ข้อมูลประเภทเล็ก

2.2.2.1 สโมท

สโมท (Synthetic Minority Oversampling Technique, SMOTE) [3] เป็นการเพิ่มข้อมูลประเภทเล็กโดยสร้างข้อมูลใหม่อย่างสุ่มระหว่างเส้นตรงที่ลากผ่านข้อมูลสองตัวที่อยู่ใกล้กัน โดยเลือกข้อมูลสองตัวนั้นจากวิธีการหาเพื่อนบ้าน k ตัวอย่างที่ใกล้ที่สุด (k-NN : k =5) ที่เป็นประเภทข้อมูลเล็ก ข้อดีของสโมทคือเพิ่มขอบเขตของข้อมูลประเภทเล็ก ลดความซ้ำซ้อนของข้อมูล อีกทั้งยังช่วยปรับปรุงการทำนายประเภทข้อมูลประเภทเล็กให้ดีขึ้น และไม่ทำให้ค่าความแม่นยำของข้อมูลลดลงอีกด้วย ส่วนข้อเสียของวิธีสโมท ก็คือข้อมูลประเภทเล็กที่สร้างขึ้นอาจจะสร้างอยู่ในเขตของข้อมูลประเภทใหญ่ได้



รูปที่ 2.3 การสังเคราะห์ตัวอย่างข้อมูลโดยวิธีสโมท

จากรูปที่ 2.3 วงกลมขนาดใหญ่แทนตัวอย่างข้อมูลประเภทเล็ก, วงกลมขนาดเล็กแทนข้อมูลที่ถูกสังเคราะห์ขึ้น, วงกลมขนาดกลางแทนตัวอย่างข้อมูลประเภทใหญ่ และเส้นตรงเชื่อมระหว่างวงกลมใหญ่สองอันแทนความสัมพันธ์ k-NN (k = 5) จะเห็นได้ว่าการสังเคราะห์ข้อมูลโดยวิธีสโมทจะทำการสุ่มข้อมูลเริ่มต้นมา (A) แล้วจึงหาเพื่อนบ้านที่ใกล้ที่สุด k ตัว (B, C, D, E, F) และเป็นข้อมูลประเภทเล็กมาทำการสังเคราะห์ข้อมูล ข้อมูลสังเคราะห์ถูกสร้างขึ้นระหว่างเส้นเชื่อมของข้อมูลเริ่มต้นกับหนึ่งในเพื่อนบ้านที่ใกล้ที่สุด 5 ตัว ด้วยการสุ่ม แต่ปัญหาของวิธีนี้ ก็คือข้อมูลที่ถูกสังเคราะห์ขึ้นอาจถูกสร้างในขอบเขตของประเภทข้อมูลใหญ่ก็ได้ ดังรูปตัวอย่างข้อมูลสังเคราะห์ที่สร้างระหว่างเส้นเชื่อมระหว่างวงกลม A ไป C ข้อมูลที่ถูกสังเคราะห์ขึ้นนั้นสร้างอยู่ในขอบเขตของข้อมูลประเภทใหญ่

Algorithm SMOTE(T; N; k)

Input : Number of minority class samples T

Amount of SMOTE N%

Number of nearest neighbors k

Output: (N=100) x T synthetic minority class samples

1. (# If N is less than 100%, randomize the minority class samples as only a random percent of them will be SMOTEd. #)
2. if N < 100
3. then Randomize the T minority class samples
4. T = (N=100) x T
5. N = 100
6. end if
7. N = (int)(N=100) (# The amount of SMOTE is assumed to be in integral multiples of 100. #)
8. k = Number of nearest neighbors
9. numattrs = Number of attributes
10. Sample[ ][ ]: array for original minority class samples
11. newindex: keeps a count of number of synthetic samples generated, initialized to 0
12. Synthetic[ ][ ]: array for synthetic samples

(# Compute k nearest neighbors for each minority class sample only. #)

```
13. for i  1 to T
14.     Compute k nearest neighbors for i, and save the indices in the nnarray
15.     Populate(N, i, nnarray)
16. end for
    Populate(N; i; nnarray) (# Function to generate the synthetic samples. #)
17. while N ≠ 0
18.     Choose a random number between 1 and k, call it nn. This step chooses one of the k
    nearest neighbors of i.
19.     for attr 1 to numattrs
20.             Compute: dif = Sample[nnarray[nn]][attr] - Sample[i][attr]
21.             Compute: gap = random number between 0 and 1
22.             Synthetic[newindex][attr] = Sample[i][attr] + gap x dif
23.     end for
24.      newindex++
25.     N = N - 1
26. end while
27. return (# End of Populate. #)
End of Pseudo-Code.
```

ตัวอย่าง: ถ้ามีตัวอย่างข้อมูลเป็น (6,4) แล้ว (4,3) เป็นเพื่อนบ้านที่ใกล้ที่สุด

เมื่อ (6,4) เป็นตัวอย่างข้อมูลเล็กที่สุ่มขึ้นมา เพื่อนำมาสังเคราะห์ข้อมูล

(4,3) เป็นหนึ่งในเพื่อนบ้านที่ใกล้ที่สุด k ตัว ที่สุ่มมา

เริ่มจาก หาค่า dif ของแต่ล่ะ attribute

dif(attr1) = 4-6 = -2

dif(attr2) = 3-4 = -1

Synthetic(x,y) = (6,4) + rand(0-1) x (-2,-1)

(# rand(0-1) สุ่มค่าระหว่าง 0-1 #)

2.2.2.2 การสร้างข้อมูลสังเคราะห์บริเวณขอบ

การสร้างข้อมูลสังเคราะห์บริเวณขอบ (Borderline-SMOTE) [5] เป็นการเลือกเฉพาะข้อมูลที่อยู่ใกล้ขอบของข้อมูลประเภทเล็ก (borderline region) มาสังเคราะห์ข้อมูลโดยใช้วิธีสโมท ส่วนข้อมูลที่อยู่ไกลจากขอบ (noise region) หรืออยู่บนแก่นข้อมูลประเภทเล็ก (safe region) จะถูกกรองทิ้ง ไม่นำมาสังเคราะห์ข้อมูล ข้อเสียของการสร้างข้อมูลสังเคราะห์บริเวณขอบ ถ้าชุดข้อมูลมีข้อมูลประเภทเล็กที่กระจายตัวห่างกันมาก อาจทำให้ไม่สามารถหาข้อมูลที่อยู่ใกล้ขอบของข้อมูลประเภทเล็กได้ ทำให้ไม่สามารถสร้างข้อมูลสังเคราะห์มาเพิ่มให้กับชุดข้อมูลตัวอย่างได้ (กำหนดให้ n เป็นจำนวนข้อมูลใหญ่ของตัวอย่างข้อมูลประเภทเล็กเมื่อหาเพื่อนบ้านที่ใกล้ที่สุด k ตัว : k =5)

ตารางที่ 2.1 คำนิยามตำแหน่งข้อมูล ของวิธีการสร้างข้อมูลสังเคราะห์บริเวณขอบ

Region	Definition
Noise	$n = k$
Borderline	$1/2k \leq n < k$
Safe	$0 \leq n < 1/2k$

แนวความคิดเบื้องต้นของวิธีนี้ ก็คือตัวอย่างข้อมูลที่อยู่บนขอบและใกล้ขอบข้อมูลประเภทเล็กนั้นมักจะคัดแยกประเภทไม่ถูก มากกว่าตัวอย่างข้อมูลที่อยู่บนแก่นข้อมูล ทำให้ตัวอย่างข้อมูลที่อยู่บนขอบและใกล้ขอบข้อมูลประเภทเล็กเป็นข้อมูลสำคัญที่ใช้เพื่อสร้างข้อมูลสังเคราะห์ประเภทเล็กแก่ชุดข้อมูลที่ไม่สมดุล

การสังเคราะห์ข้อมูลประเภทเล็กของวิธีนี้ เริ่มจากเลือกเฉพาะข้อมูลที่อยู่ใกล้ขอบของข้อมูลประเภทเล็ก แล้วสุ่มเลือกตัวอย่างข้อมูลประเภทเล็กที่อยู่ใกล้ขอบมาสร้างข้อมูลใหม่อย่างสุ่มระหว่างเส้นตรงที่ลากผ่านข้อมูลสองตัวที่อยู่ใกล้กัน โดยข้อมูลตัวแรกคือ ตัวอย่างข้อมูลประเภทเล็กที่อยู่ใกล้ขอบที่สุ่มมา ข้อมูลตัวที่สองคือหนึ่งในเพื่อนบ้าน k ตัวอย่างที่ใกล้ที่สุด (k-NN : k =5) ของตัวอย่างข้อมูลประเภทเล็กที่อยู่ใกล้ขอบที่สุ่มมา จากข้อมูลประเภทเล็กทั้งหมดบนชุดเรียนรู้

2.3 วิธีการสร้างแบบจำลองโดยใช้ต้นไม้การตัดสินใจ

ต้นไม้ตัดสินใจ (Decision Tree) คือ โครงสร้างต้นไม้ที่ประกอบไปด้วยโหนดตัดสินใจ (Decision Nodes) แล้วเชื่อมต่อกันด้วยสาขา (Branches) โดยขยายจากโหนดราก (Root Node) มายังโหนดใบ (Leaf Nodes) ค่าคุณสมบัติ (Attribute) ที่ใช้สำหรับเปรียบเทียบเงื่อนไขจะอยู่ในโหนดตัดสินใจ ซึ่งผลลัพธ์การเปรียบเทียบจะอยู่ในสาขา (Branches) และในแต่ละสาขาจะนำไปสู่ โหนดตัดสินใจอื่น หรือโหนดใบ ซึ่งโหนดใบจะเก็บผลลัพธ์ของการจำแนกไว้

ในงานวิจัยฉบับนี้เลือกใช้วิธีการสร้างต้นไม้ตัดสินใจ (Decision Tree) ด้วยขั้นตอนวิธีแบบ C4.5 โดยนิยามให้

$$\text{Entropy(s)} = \sum_{i=1}^{n} p_i \log p_i \tag{2.1}$$

โดยที่ p_i จำนวนความถี่ของประเภท i ใน s เพื่อใช้ในการหาค่าความน่าจะเป็น ซึ่งจะบอกเป็นหนึ่งประเภทเท่านั้นโดยที่ค่า Entropy จะมีค่าเป็น 0 และถ้ามีค่าเป็น 1 นั้นหมายถึงทุกประเภทมีความน่าจะเป็นที่เท่ากันซึ่งจะมีโอกาสเกิดขึ้นได้ โดยนิยาม

$$\mathrm{p_i} = (r_i \mid \mathrm{N}) \tag{2.2}$$

โดย N เท่ากับจำนวนประเภทข้อมูล (Class Labeled) โดย r_i จะเท่ากับเหตุการณ์ที่เกิดขึ้นใน N

กระบวนการสร้างต้นไม้ตัดสินใจ จะเริ่มจากการที่ไม่มีต้นไม้อยู่ แล้วนำ ข้อมูลกลุ่มการเรียนรู้ (Training set) เข้ามาคำนวณ และวนทำจนกระทั่งไม่สามารถแตกข้อมูลออกไปได้อีก

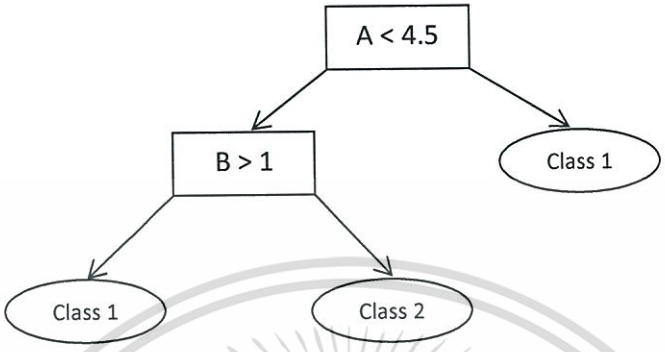
1. คำนวณค่า Entropy

$$\text{Entropy} = \sum_{i=1}^{n} -(r_i|N) \log_2 (r_i|N) \tag{2.3}$$

2. เลือกคุณลักษณะที่มีค่า สารสนเทศเกน (Information Gain) ที่มีค่ามากที่สุดเพื่อนำมาสร้างเป็น root nodeของต้นไม้ตัดสินใจ (Decision Tree)

$$\text{Information Gain} = \text{Entropy(Before)} - \text{Entropy(After)} \tag{2.4}$$

3. สร้างต้นไม้ในการตัดสินใจในระดับถัดไปด้วยการทำซ้ำในขั้นตอนที่ 1 และ 2 จนกระทั่งข้อมูลย่อยมีเหลือเพียงประเภทเดียว (one label) และ ค่าของ Entropy มีค่าเป็นศูนย์



รูปที่ 2.4 แสดงต้นไม้ตัดสินใจ

ในส่วนสุดท้ายที่เป็นโหนดใบ (Leaf Node) จะเป็นส่วนที่ตัดสินใจว่าค่าของประเภทข้อมูล ควรจะเป็นประเภทใด (รูปที่ 2.4) อย่างไรก็ตามต้นไม้ในการตัดสินใจก็ไม่สามารถตัดสินใจได้ถูกต้องเสมอไป

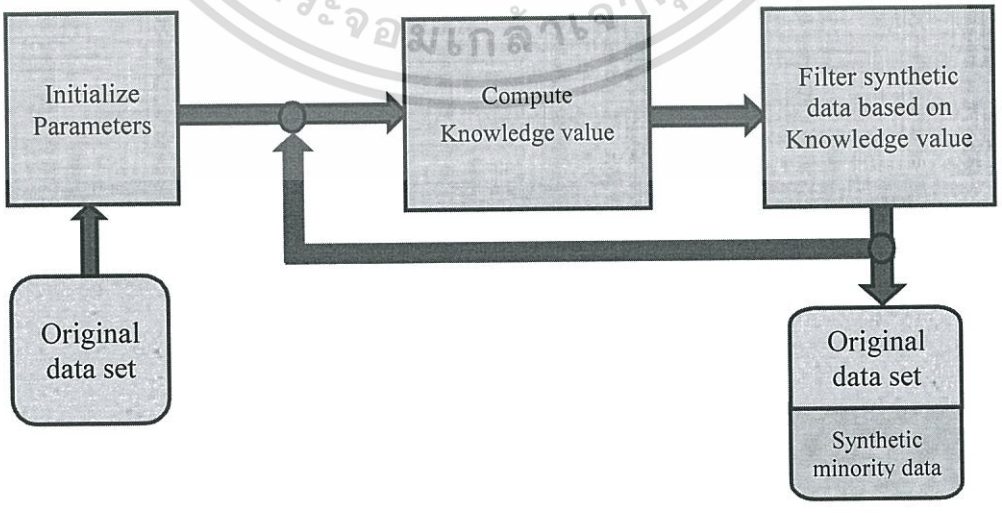
สำหรับตัวจำแนก C4.5 ได้ทำการขยายต่อส่วนของการจำแนกข้อมูลที่เป็นตัวเลข ด้วยการ แบ่งช่วงของข้อมูล เพื่อใช้ในการสร้างต้นไม้ในการตัดสินใจ หรือเรียกอีกอย่างว่าการทำการแบ่งค่าต่อเนื่องออกเป็นช่วงย่อยๆ (Discretization)

บทที่ 3

การปรับการกระจายของข้อมูลที่ไม่สมดุลโดยใช้วิธีการหาเพื่อนบ้านที่ใกล้ที่สุดแบบย้อนกลับ K ตัว

จากบทที่ 2 ได้บอกถึงลักษณะข้อมูลที่ไม่สมดุล รวมถึงปัญหาเมื่อนำชุดข้อมูลที่ไม่สมดุลกันไปสร้างแบบจำลองเพื่อจำแนกประเภทข้อมูล ปัญหาข้อมูลไม่สมดุลที่เกิดได้ในข้อมูลจริง รวมทั้งวิธีการแก้ปัญหาข้อมูลไม่สมดุลทั้งแบบการสุ่มตัวอย่างข้อมูลมาทำซ้ำ และการสังเคราะห์ข้อมูลประเภทเล็ก ซึ่งได้บอกทั้งข้อดีและข้อเสียที่เกิดขึ้นเมื่อใช้วิธีการแก้ปัญหาดังกล่าว ส่วนในบทนี้จะกล่าวถึงการปรับการกระจายของข้อมูลที่ไม่สมดุลโดยใช้วิธีการหาเพื่อนบ้านที่ใกล้ที่สุดแบบย้อนกลับ K ตัว

แนวความคิดเบื้องต้นของวิธีการที่นำเสนอคือ เพิ่มตัวอย่างเฉพาะข้อมูลประเภทข้อมูลเล็ก และคงตัวอย่างข้อมูลส่วนมากที่คัดแยกไว้ดีแล้ว สำหรับข้อมูลใดๆการที่จะดูว่าข้อมูลนั้นได้ถูกคัดแยกไว้ดีแล้วหรือไม่ ให้หาเพื่อนบ้านที่ใกล้ที่สุด k ตัวอย่าง และทำการโหวตประเภทข้อมูลจากประเภทข้อมูลของเพื่อนบ้านทั้งหมด ถ้าประเภทข้อมูลที่ได้จากการโหวตตรงกับประเภทของข้อมูลเริ่มต้นจะสรุปว่าข้อมูลนั้นได้ถูกคัดแยกไว้ดีแล้ว ในทางกลับกันถ้าประเภทข้อมูลที่ได้จากการโหวตไม่ตรงกับประเภทของข้อมูลเริ่มต้นจะสรุปว่าข้อมูลนั้นได้ถูกคัดแยกไว้ไม่ดี



รูปที่ 3.1 แสดงถึงภาพรวมของวิธีนำเสนอ

วิธีที่นำเสนอนั้นใช้ค่าความรู้ (knowledge value : Info) ในการคัดเลือกข้อมูลสังเคราะห์ประเภทเล็ก ที่ถูกสังเคราะห์จากวิธีสโมท ตามรูปที่ 3.1 เริ่มจากนำข้อมูลที่ไม่สมดุล(Original data set) แบ่งข้อมูลเป็นชุดสำหรับทดสอบ (testing data set) และชุดสำหรับเรียนรู้ (training data set) โดยใช้ 10-Fold Cross-validation เพื่อแก้ปัญหาจากการเลือกข้อมูลที่ดีและง่ายมาเป็นข้อมูลชุดทดสอบทำให้ผลการจำแนกประเภทข้อมูลนั้นดีเกินไป แล้วจึงกำหนดค่าเริ่มต้นของตัวแปร อาทิ เช่น เปอร์เซ็นต์การเพิ่มข้อมูลประเภทเล็ก โดยวิธีทีนำเสนอจะสังเคราะห์ข้อมูลประเภทเล็กที่ละ 25 เปอร์เซ็นต์ของเปอร์เซ็นต์การเพิ่มข้อมูลประเภทเล็กที่กำหนด นำมาคำนวณหาจำนวนรอบและจำนวนข้อมูลที่ต้องสังเคราะห์ในแต่ล่ะรอบตามหัวข้อ 3.2 เมื่อได้ข้อมูลสังเคราะห์ประเภทเล็กนำมาคำนวณหาค่าความรู้ (Compute knowledge value ตามหัวข้อ 3.3-3.5) เพื่อเลือกเฉพาะข้อมูลที่สามารถช่วยเหลือข้อมูลที่คัดแยกประเภทผิดได้เท่านั้นไปเพิ่มในชุดข้อมูล ตามหัวข้อ 3.6 แล้วจึงเริ่มการสังเคราะห์ข้อมูลใหม่ จนกว่าจะครบจำนวนรอบที่คำนวณไว้

วิธีการปรับการกระจายของข้อมูลที่ไม่สมดุลโดยใช้วิธีการหาเพื่อนบ้านที่ใกล้ที่สุดแบบย้อนกลับ K ตัว มีขั้นตอนหลักอยู่ 6 ขั้นตอน

1. การตรวจสอบว่าชุดข้อมูลที่นำมาใช้นั้นเหมาะกับวิธีที่นำเสนอหรือไม่
2. การคำนวณจำนวนรอบที่ต้องทำและจำนวนข้อมูลที่ต้องสังเคราะห์ในแต่ละรอบ
3. การหาเพื่อนบ้าน K ตัวอย่างที่ใกล้ที่สุดแบบกลับด้าน
4. การคำนวณค่าความช่วยเหลือของข้อมูล
5. การคำนวณหาค่าความรู้แก่ตัวอย่างสังเคราะห์
6. การเลือกข้อมูลสังเคราะห์จากค่าความรู้

3.1 การตรวจสอบว่าชุดข้อมูลที่นำมาใช้นั้นเหมาะกับวิธีที่นำเสนอหรือไม่

ในงานวิจัยนี้ใช้ระยะที่ใกล้ที่สุด 5 ตัว ของตัวอย่างข้อมูลทุกตัวบนชุดข้อมูลมาคำนวณหาค่าความแปรปรวน (Variance) เพื่อดูว่าชุดข้อมูลมีการกระจายตัวอย่างไร ถ้าข้อมูลนั้นมีการกระจายตัวที่มากชุดข้อมูลนี้จะไม่เหมาะสมที่จะมาใช้วิธีที่นำเสนอ (กำหนดให้ N แทนจำนวนข้อมูล และ k แทนจำนวนข้อมูลที่ระยะทางใกล้ที่สุด : ในงานวิจัยนี้ใช้ k = 5)

$$\text{Variance} = \sum_{i=1}^{N}\sum_{j=1}^{k}\frac{(x_{ij}-\bar{x})^2}{N-1} \qquad (3.1)$$

ซึ่งจากผลการทดลองข้อมูลทั้ง 10 ชุดแล้ว ชุดข้อมูลที่มีค่าความแปรปรวนมากเกินกว่า 1000 แล้ว จะให้ข้อมูลสังเคราะห์ที่ไม่ดีเมื่อเทียบกับวิธีการสร้างข้อมูลสังเคราะห์บริเวณขอบ

3.2 การคำนวณจำนวนรอบที่ต้องทำและจำนวนข้อมูลที่ต้องสังเคราะห์ในแต่ละรอบ

การสร้างข้อมูลสังเคราะห์ ในงานวิจัยนี้ได้ใช้วิธีการสโมทในการสร้างข้อมูลประเภทเล็ก โดยกำหนดให้ทุกรอบ สร้างข้อมูลสังเคราะห์ เพียง 25% ของเปอร์เซ็นต์การเพิ่มข้อมูลประเภทน้อยซึ่งกำหนดโดยผู้ใช้ ทำให้ต้องคำนวณจำนวนรอบทั้งหมด (numIteration) และจำนวนข้อมูลที่ต้องสังเคราะห์ในแต่ละรอบ (numSampling) ตามสมการที่ 3.2 และ 3.3 ตามลำดับ (กำหนดให้ Percentage แทนเปอร์เซ็นต์การเพิ่มข้อมูลประเภทเล็กซึ่งกำหนดโดยผู้ใช้, |Min| แทนจำนวนข้อมูลประเภทเล็กทั้งหมดในชุดข้อมูลที่ใช้สำหรับเรียนรู้ และ |Maj| แทนจำนวนข้อมูลประเภทใหญ่ทั้งหมดในชุดข้อมูลที่ใช้สำหรับเรียนรู้)

$$numIteration = Percentage/25 \tag{3.2}$$

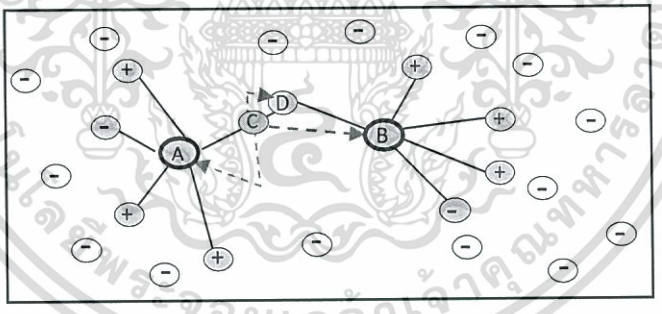
$$numSampling = \lceil |Min| * 0.25 \rceil \tag{3.3}$$

เมื่อได้จำนวนข้อมูลประเภทเล็กที่ต้องสร้างแล้ว ให้นำมาสังเคราะห์ข้อมูลด้วยวิธีการสโมทเป็นจำนวน 10 เท่าของจำนวนข้อมูลประเภทเล็กนั้น เพื่อให้ข้อมูลที่สังเคราะห์ขึ้นมีจำนวนเพียงพอให้สามารถเลือกตัวอย่างข้อมูลสังเคราะห์ที่ดีที่สุดมาเพิ่มแก่ข้อมูลต้นฉบับ โดยส่วนที่เหลือจะถูกกรองทิ้ง

สาเหตุที่ต้องแบ่งสร้างข้อมูลสังเคราะห์ที่ละ 25% ของเปอร์เซ็นต์การเพิ่มข้อมูลประเภทเล็ก เนื่องจากเมื่อสังเคราะห์ข้อมูลประเภทเล็กเสร็จในแต่ละรอบนั้น จะนำข้อมูลประเภทเล็กที่ดีที่สุดมาเพิ่มแก่ข้อมูลเรียนรู้ ทำให้รอบถัดมามีข้อมูลประเภทเล็กที่ถูกสังเคราะห์ไว้ดีแล้วมาเพิ่มเป็นตัวตั้งต้นให้ใช้สังเคราะห์ข้อมูลในรอบถัดไปรวมกับข้อมูลประเภทเล็กเดิม ส่วนการกำหนดเปอร์เซ็นต์ที่ใช้สังเคราะห์นั้น เป็นการปรับความละเอียดของการสังเคราะห์ข้อมูล สามารถปรับเปลี่ยนได้ โดยเปอร์เซ็นต์ยิ่งน้อยความละเอียดของการสังเคราะห์ข้อมูลยิ่งมากทำให้เวลาที่ใช้ในการคำนวณมากขึ้นตามไปด้วย

3.3 การหาเพื่อนบ้าน K ตัวอย่างที่ใกล้ที่สุดแบบกลับด้าน

ในงานวิจัยชิ้นนี้ ใช้วิธีการหาเพื่อนบ้าน K ตัวอย่างที่ใกล้ที่สุดแบบกลับด้าน (reverse K-NN, rKNN) สำหรับข้อมูลสังเคราะห์ใดๆ จะค้นกลับหาว่า ตัวอย่างสังเคราะห์นั้นเป็นเพื่อนบ้าน k ตัวอย่างที่ใกล้ที่สุด (k-NN) แก่ข้อมูลใดบ้าง ข้อมูลนั้นก็จะเป็นเพื่อนบ้าน K ตัวอย่างที่ใกล้ที่สุดแบบกลับด้านของตัวอย่างสังเคราะห์ที่กล่าวมา ดังแสดงในรูปที่ 3.2 ข้อมูลประเภทเล็กที่ถูกสังเคราะห์ C ซึ่งสร้างขึ้นจากวิธีการสโมท เมื่อคำนวณหาเพื่อนบ้านที่ใกล้ที่สุด k ตัวอย่าง บนข้อมูล A, B และ D พบว่ามีข้อมูล C ประกอบอยู่ด้วย จึงสามารถระบุได้ว่า A, B และ D เป็นเพื่อนบ้าน K ตัวอย่างที่ใกล้ที่สุดแบบกลับด้านของ C เช่นกัน เพื่อคำนวณหาค่าความรู้จากตัวอย่างสังเคราะห์ C จะต้องคำนวณค่าความช่วยเหลือจาก A, B และ D เนื่องจากค่าความรู้ของตัวอย่างสังเคราะห์คิดจากค่าความช่วยเหลือเฉพาะที่ตัวอย่างสังเคราะห์นั้นไปปรากฏเป็นเพื่อนบ้านทั้งหมดมารวมกัน (จากรูปที่ 3.2 กำหนดให้ วงกลมเครื่องหมายบวกแสดงถึงข้อมูลประเภทเล็ก, วงกลมเครื่องหมายลบแสดงถึงข้อมูลประเภทใหญ่, เส้นทึบแสดงถึงความสัมพันธ์แบบ k-NN และเส้นปะแสดงถึงความสัมพันธ์แบบ rKNN)

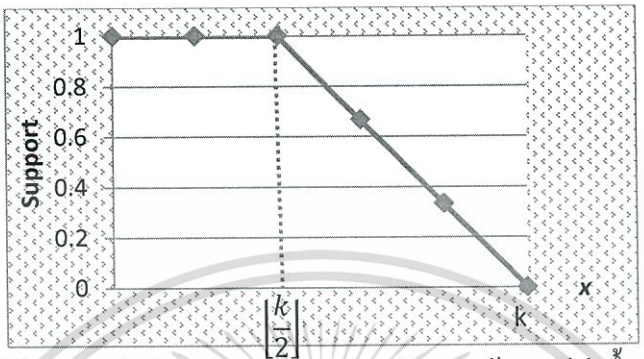


รูปที่ 3.2 ตัวอย่างการทำ reverse K-NN ของ node C

3.4 การคำนวณค่าความช่วยเหลือของข้อมูล

เพื่อคำนวณค่าความรู้นั้นเบื้องต้นต้องคำนวณค่าความช่วยเหลือ (support, *S*) ของทุกข้อมูลที่มีข้อมูลสังเคราะห์นั้นว่าเป็นหนึ่งในเพื่อนบ้าน k ตัวอย่างที่ใกล้ที่สุด (k-NN) หรือกล่าวคือการหาเพื่อนบ้าน K ตัวอย่างที่ใกล้ที่สุดแบบกลับด้านของข้อมูลสังเคราะห์ โดยค่าความช่วยเหลือแสดงได้ดังรูปที่ 3.3 โดยแกนนอนเป็นค่า x ซึ่งเป็นจำนวน minority instance ใน k-NN ของข้อมูลที่ตัวอย่างสังเคราะห์ไปปรากฏเป็นเพื่อนบ้าน และ แกนตั้งเป็นค่าความช่วยเหลือ ซึ่งจะมีค่ามากก็ต่อเมื่อ

ข้อมูลประเภทเล็กที่ตัวอย่างสังเคราะห์ไปปรากฏเป็นเพื่อนบ้านนั้นคัดแยกประเภทผิด และจะมีค่าน้อยก็ต่อเมื่อข้อมูลประเภทเล็กที่ตัวอย่างสังเคราะห์ไปปรากฏเป็นเพื่อนบ้านนั้นคัดแยกประเภทข้อมูลถูกอยู่แล้ว



รูปที่ 3.3 กราฟเส้นแสดงถึงค่า Support ของแต่ละตัวอย่างข้อมูล มีค่าขึ้นอยู่กับจำนวน x

ค่าความช่วยเหลือจะเป็น 1 ต่อเมื่อจำนวนเพื่อนบ้าน k ตัวอย่างที่ใกล้ที่สุด (k-NN) เกินกว่าครึ่งเป็นข้อมูลประเภทใหญ่ ดังแสดงในสมการที่ 3.4 บรรทัดบน และค่าความช่วยเหลือจะมีค่าลดลงเป็นกราฟเส้นตรงตามลำดับก็ต่อเมื่อจำนวนเพื่อนบ้าน k ตัวอย่างที่ใกล้ที่สุด (k-NN) น้อยกว่าครึ่งเป็นข้อมูลประเภทใหญ่ดังแสดงในสมการที่ 3.4 บรรทัดล่าง

$$S(x) = \begin{cases} 1 & , x \leq \left\lfloor \frac{k}{2} \right\rfloor \\ v_1(x) + v_2 & , x > \left\lfloor \frac{k}{2} \right\rfloor \end{cases} \quad (3.4)$$

$$v_1 = \frac{1}{\left\lfloor \frac{k}{2} \right\rfloor - k} \quad , \quad v_2 = \frac{-k}{\left\lfloor \frac{k}{2} \right\rfloor - k} \quad (3.5)$$

กำหนดให้ x เป็นจำนวน minority instanceใน k-NN ของข้อมูลที่ตัวอย่างสังเคราะห์ไปปรากฏเป็นเพื่อนบ้าน, v_1 แทนความชันของเส้นตรง และ v_2 จุดตัดบนแกน x ของเส้นตรงในรูปที่ 3.3

เนื่องจากงานวิจัยนี้ใช้ค่า k = 5 ในการหาเพื่อนบ้านที่ใกล้ที่สุด ดังนั้น ค่าความช่วยเหลือจะเป็น 1 ต่อเมื่อจำนวนเพื่อนบ้าน 5 ตัวอย่างที่ใกล้ที่สุด (5-NN) เกินกว่าครึ่งเป็นข้อมูลประเภทใหญ่ ดังแสดงในสมการที่ 3.6 บรรทัดบน และค่าความช่วยเหลือจะมีค่าลดลงเป็นกราฟเส้นตรงตามลำดับก็ต่อเมื่อจำนวนเพื่อนบ้าน 5 ตัวอย่างที่ใกล้ที่สุด (5-NN) น้อยกว่าครึ่งเป็นข้อมูลประเภทใหญ่ดังแสดงในสมการที่ 3.6 บรรทัดล่าง

$$S(x) = \begin{cases} 1 & , x \leq 2 \\ -\frac{1}{3}(x) + \frac{5}{3} & , x > 2 \end{cases} \quad (3.6)$$

3.5 การคำนวณหาค่าความรู้แก่ตัวอย่างสังเคราะห์

ค่าความรู้(knowledge value : Info) เป็นค่าที่ใช้บอกว่าตัวอย่างที่สังเคราะห์ขึ้นมานั้นสามารถช่วยข้อมูลประเภทเล็กที่คัดแยกประเภทข้อมูลผิดได้มากน้อยเพียงใด ถ้าค่าความรู้มีค่ามากแสดงว่าตัวอย่างสังเคราะห์นี้สามารถช่วยข้อมูลประเภทเล็กที่คัดแยกประเภทข้อมูลผิดได้เป็นจำนวนมาก หากเพิ่มตัวอย่างสังเคราะห์นี้เข้าชุดข้อมูลเดิม ซึ่งถ้าค่าความรู้มีค่าเป็น 0 หรือน้อยกว่า 0 แสดงว่าตัวอย่างสังเคราะห์นี้ไม่สามารถช่วยข้อมูลประเภทเล็กที่คัดแยกประเภทข้อมูลผิดได้ ซ้ำยังอาจทำให้ข้อมูลที่คัดแยกประเภทไว้ดีอยู่แล้วคัดแยกประเภทข้อมูลผิด หากเพิ่มตัวอย่างสังเคราะห์นี้เข้าชุดข้อมูลเดิม

ค่าความรู้เป็นค่าที่คิดขึ้นจากการหาเพื่อนบ้าน K ตัวอย่างที่ใกล้ที่สุดแบบกลับด้านดังหัวข้อ 3.3 แล้วนำเพื่อนบ้านแต่ละตัวมาคิดค่าความช่วยเหลือดังหัวข้อ 3.4 เมื่อได้ค่าความช่วยเหลือของแต่ละเพื่อนบ้านให้ปรับค่าความช่วยเหลือตามประเภทข้อมูลตามสมการที่ 3.7 และ 3.8 หลังจากปรับค่าความช่วยเหลือเสร็จ ให้นำค่าความช่วยเหลือมารวมกันดังสมการที่ 3.9 จะได้เป็นค่าความรู้ของแต่ละตัวของข้อมูลที่ถูกสังเคราะห์ กำหนดให้ |Min| แทนจำนวนข้อมูลประเภทเล็กของชุดข้อมูลที่นำมาเรียนรู้ และ |Maj| แทนจำนวนข้อมูลประเภทใหญ่ของชุดข้อมูลที่นำมาเรียนรู้

ถ้าตัวอย่างข้อมูลที่ตัวอย่างสังเคราะห์ไปปรากฏว่าเป็นเพื่อนบ้านเป็นข้อมูลประเภทเล็ก ให้นำค่าความช่วยเหลือที่คิดจากหัวข้อ 3.6 มาคิดค่าความช่วยเหลือดังสมการที่ 3.7 ที่กำหนดให้มีค่าเป็นบวกเนื่องมาจากถ้าเพิ่มตัวอย่างสังเคราะห์ประเภทเล็กนี้แล้ว จะทำให้ตัวอย่างข้อมูลประเภทเล็กสามารถจำแนกประเภทข้อมูลได้ถูกต้องมากขึ้น ส่วนการคูณด้วย $\frac{|Min| + |Maj|}{|Min| \times 2}$ เป็นการปรับให้ข้อมูลประเภทเล็กมีความสำคัญเทียบเท่ากับข้อมูลประเภทใหญ่

$$S_i = (+1) * (S * \frac{|Min| + |Maj|}{|Min| \times 2}) \quad (3.7)$$

ถ้าตัวอย่างข้อมูลที่ตัวอย่างสังเคราะห์ไปปรากฏว่าเป็นเพื่อนบ้านเป็นข้อมูลประเภทใหญ่แล้ว ให้นำค่าความช่วยเหลือที่คิดจากหัวข้อ 3.6 มาปรับค่าความช่วยเหลือดังสมการที่ 3.8 ที่กำหนดให้มีค่าเป็นลบเนื่องมาจากถ้าเพิ่มตัวอย่างสังเคราะห์ประเภทเล็กนี้เข้าไป อาจทำให้ข้อมูลประเภทใหญ่

ที่จำแนกประเภทข้อมูลดีอยู่แล้วนั้นจำแนกข้อมูลผิดได้ ส่วนการคูณด้วย $\frac{|Min|+|Maj|}{|Maj| \times 2}$ เป็นการปรับให้ข้อมูลประเภทใหญ่มีความสำคัญเทียบเท่ากับข้อมูลประเภทเล็ก

$$S_i = (-1) * (S * \frac{|Min|+|Maj|}{|Maj| \times 2}) \quad (3.8)$$

แล้วจำนำค่าความช่วยเหลือที่ถูกปรับตามสมการที่ 3.7 กับ 3.8 มาคำนวณค่าความรู้ของตัวอย่างสังเคราะห์ เฉพาะที่ตัวอย่างสังเคราะห์นั้นไปปรากฏเป็นเพื่อนบ้านตามสมการที่ 3.9

$$\text{Info} = \sum_{i=1}^{K} S_i \quad (3.9)$$

3.6 การเลือกข้อมูลสังเคราะห์จากค่าความรู้

เมื่อคำนวณค่าความรู้แก่ตัวอย่างสังเคราะห์ทั้งหมด และเลือกเฉพาะตัวอย่างสังเคราะห์ที่มีค่าสูงกว่า ค่าน้อยที่สุดที่ยอมรับได้ (threshold) โดยจะไม่เลือกตัวอย่างสังเคราะห์ที่มีค่าความรู้ต่ำกว่า 0 แล้วจึงทำการเรียงตัวอย่างสังเคราะห์จากมากไปน้อยตามค่าความรู้ เพื่อเลือกข้อมูลจำนวนอย่างมากเท่ากับ จำนวนข้อมูลที่ต้องสังเคราะห์ในแต่ละรอบ (กำหนดให้ N เป็นจำนวนข้อมูลที่สังเคราะห์ทั้งหมดในแต่ล่ะรอบ, วิธี RSMOTE1 ใช้ค่าน้อยที่สุดที่ยอมรับได้เป็นค่าเฉลี่ยตามสมการที่ 3.10 ในการเลือกตัวอย่างสังเคราะห์ และ วิธี RSMOTE2 ใช้ค่าน้อยที่สุดที่ยอมรับได้เป็นค่ามัธยฐานตามสมการที่ 3.11 ในการเลือกตัวอย่างสังเคราะห์)

$$\text{threshold} = \text{Mean}(Info) = \frac{\sum_{j=1}^{N} Info_j}{N} \quad (3.10)$$

$$\text{threshold} = \text{Median}(Info) \quad (3.11)$$

นำข้อมูลสังเคราะห์ที่ถูกเลือกมาแล้วเพิ่มไปในชุดข้อมูลแล้วจึงเริ่มการสังเคราะห์ข้อมูลโดยสโมท จนกว่าจะครบจำนวนรอบที่คำนวณ

สำหรับทุกๆ ชุดข้อมูล ถ้าค่าความแปรปรวน (variance) นั้นมีค่ามากกกว่า 1000 แสดงว่าชุดข้อมูลที่นำมาสังเคราะห์ข้อมูลนั้นมีการกระจายตัวของข้อมูลที่สูงมาก หลังจากการเตรียมข้อมูลแล้วนั้นจะสร้างข้อมูลสังเคราะห์ประเภทเล็กเป็นจำนวน 10 เท่าของ numSampling มาเลือกตัวอย่างที่ดีที่สุดจำนวนหนึ่ง โดยนำข้อมูลสังเคราะห์ทั้งหมดจะนำมาคำนวณค่าความรู้ ดังแสดงในสมการ 7 และ 8 โดยที่ค่าความรู้จะเพิ่มขึ้นสูงเมื่อตัวอย่างเพื่อนบ้านเป็นข้อมูลประเภทเล็กที่ไม่สามารถคัดแยกประเภทข้อมูลได้ถูกต้อง แต่จะโดนลดค่าเมื่อเพื่อนบ้านเป็นข้อมูลประเภทใหญ่

เพื่อคำนวณค่าความรู้นั้นเบื้องต้นต้องคำนวณค่าความช่วยเหลือ ของทุกข้อมูลที่มีข้อมูลสังเคราะห์นั้นว่าเป็นหนึ่งในเพื่อนบ้าน k ตัวอย่างที่ใกล้ที่สุด (k-NN) หรือกล่าวคือการหาเพื่อนบ้าน k ตัวอย่างที่ใกล้ที่สุดแบบกลับด้านของข้อมูลสังเคราะห์ โดยค่าความช่วยเหลือแสดงได้ดังสมการที่ 4, 5, และ 6 ตามลำดับ ค่าความช่วยเหลือจะมีค่าเท่ากับ 1 เมื่อจำนวนเพื่อนบ้าน k ตัวอย่างที่ใกล้ที่สุด (k-NN) เกินกว่าครึ่งเป็นข้อมูลประเภทใหญ่ดังแสดงในสมการที่ 4 และมีค่าลดลงเป็นกราฟเส้นตรงตามลำดับดังแสดงในสมการที่ 5 และ 6 เมื่อคำนวณค่าความรู้แก่ตัวอย่างสังเคราะห์ทั้งหมดแล้ว จึงทำการเรียงตัวอย่างสังเคราะห์จากมากไปน้อยตามค่าความรู้ และเลือกเฉพาะตัวอย่างสังเคราะห์ที่มีค่าสูงกว่า ค่าน้อยที่สุดที่ยอมรับได้ (threshold) ดังแสดงในสมการที่ 9 ซึ่งวิธีการที่กล่าวมาจะสร้างข้อมูลสังเคราะห์และคัดเลือกข้อมูลไปเรื่อยๆจนกว่าครบรอบที่กำหนด

บทที่ 4

การทดลองและผลการทดลอง

จากบทที่ 3 ได้กล่าวถึงวิธีการปรับการกระจายของข้อมูลที่ไม่สมดุลโดยใช้วิธีการหาเพื่อนบ้านที่ใกล้ที่สุดแบบย้อนกลับ K ตัว ซึ่งประยุกต์มาจากการสังเคราะห์ตัวอย่างข้อมูลโดยวิธีสโมท ในบทนี้จะแสดงรายละเอียดในการทดลอง, ลักษณะของชุดข้อมูลที่ใช้ในการทดลอง และวิธีการทดลอง โดยเริ่มจากการเตรียมข้อมูลเพื่อใช้สำหรับการสังเคราะห์ข้อมูล การวัดประสิทธิภาพของข้อมูลที่ปรับสมดุลและข้อมูลที่ไม่ได้ปรับสมดุล รวมถึงการวิเคราะห์ผลการทดลอง

4.1 การทดลอง

4.1.1 ข้อมูลที่ใช้ในการทดลอง

ข้อมูลที่ใช้ในการทดลองเป็นชุดข้อมูลในกลุ่มปัญหาการจำแนกประเภท โดยนำมาจาก UCI Machine Learning Repository ประกอบด้วย

ตารางที่ 4.1 แสดงรายละเอียดของชุดข้อมูล

#	Data set	number of Examples	number of Attributes	Class label (minority : majority)	Percentage of minority class
1.	Ionosphere	351	35	bad:good	35.8%
2.	Pima	768	8	1:0	34.77%
3.	Seeds	210	8	1:2	33.33%
4.	Vehicle	946	18	van:other	30.75%
5.	Contraceptive	1473	10	Long term:other	29.21%
6.	Haberman	306	3	2:1	26.47%
7.	Breast Tissue	106	10	car:other	24.71%
8.	Satimage	6435	37	4:other	10.75%
9.	Glass_3	214	10	3:other	8.63%
10.	Ecoli_om	336	8	om:other	6.32%

ตาราง 4.1 แสดงถึงรายละเอียดชุดข้อมูลโดยสรุป ซึ่งประกอบไปด้วย ชนิดของชุดข้อมูล (Data set) จำนวนตัวอย่างข้อมูล (number of Examples) จำนวนลักษณะเฉพาะของข้อมูล (number of

วิธีการปรับการกระจายของข้อมูลที่ไม่สมดุลโดยใช้วิธีการหาเพื่อนบ้านที่ใกล้ที่สุดแบบย้อนกลับ K ตัว โดยใช้ค่าน้อยที่สุดที่ยอมรับได้เป็นค่าเฉลี่ยของค่าความรู้ สามารถดูวิธีอย่างละเอียดได้จากซูโดโค้ดข้างล่าง

Algorithm: Reverse K-NN SMOTE (k=5)

Input: a set of all training set instances: D, |D| = N

a set of all training set minority instances: Min

a set of all training set majority instances: Maj

where |Maj| + |Min| = |D|

Percentage of over-sampling: Percentage,

i.e., 100, 200, 300, so on

Output: set of synthetic minority instances L

1. Define L = Ø
2. Calculate distance (x_{ij}) for all i in D and for all j in its k nearest neighbor, and then find their average distances ($\bar{x}$)

   Variance = $\sum_{i=1}^{N}\sum_{j=1}^{k}\frac{(x_{ij}-\bar{x})^2}{N-1}$ (1)
3. Do sampling 25% per iteration until reach the Percentage

   3.1 Calculate number of iteration

   numIteration = Percentage/25 (2)

   3.2 Calculate number of sampling

   numSampling = $\lceil |Min| * 0.25 \rceil$ (3)
4. Generate synthetic minority data 10 times of numSampling with SMOTE

   Candidate = set of synthetic data that was generated by SMOTE
5. D = D Union L
6. for each instance D

   Find set of k-nearest neighbors of D_i ($k\text{-}NN_i$)
7. Find set of reverse K-NN ($rKNN_j$) of each $Candidate_j$

   for $Candidate_j$ in Candidate {

   for each D_i in D {

   if $Candidate_j$ is in $k\text{-}NN_i$ {

   Add D_i to $rKNN_j$ }

   }

}

8. Calculate support (Support)

for each D_i in D {

Count number of minority instance in k-NN_i ($|Min_{knn\_i}|$)

Count number of majority instance in k-NN_i ($|Maj_{knn\_i}|$)

if ($|Min_{knn\_i}| \leq \lfloor (|Min_{knn\_i}| + |Maj_{knn\_i}|)/2 \rfloor$)

{$support_i = 1$} (4)

else

{ $support_i = v_1 |minor| + v_2$ } (5)

where $v_1 = \frac{1}{\lfloor \frac{k}{2} \rfloor - k}$, $v_2 = \frac{-k}{\lfloor \frac{k}{2} \rfloor - k}$ (6)

}

9. Calculate knowledge value ($Info_i$)

for $Candidate_j$ in Candidate {

$Info_i = 0$

for each $D_n \in rKNN_j$ {

if (D_n is minority class) { $Info_i = Info_i + (S_n * \frac{|Min| + |Maj|}{|Min| \times 2})$} (7)

if (D_n is majority class) { $Info_i = Info_i - (S_n * \frac{|Min| + |Maj|}{|Maj| \times 2})$} (8)

}

}

10. Sort Candidate with knowledge value by descending order

11. Calculate threshold

if (average(Info) < 0){ threshold = 0 } else { threshold = average(Info) } (9)

12. Prune all candidate that knowledge value lower than the threshold

13. Add candidate to L

14. go to the 4rd step until reach to numIteration

15. return L

Attributes) แบ่งข้อมูลเล็กและข้อมูลใหญ่เป็นข้อมูลประเภทใด (Class label) เช่น ข้อมูล Ionosphere จากตารางที่ 4.1ข้อมูลประเภทเล็ก คือ bad และข้อมูลประเภทใหญ่ คือ good โดยชุดข้อมูลเรียงตามเปอร์เซ็นต์ของข้อมูลประเภทเล็ก (Percentage of minority class) จากมากไปน้อย โดยแต่ละชุดข้อมูลมีรายละเอียดดังนี้

1. ปัญหา Ionosphere เป็นปัญหาที่เกี่ยวการจำแนกประเภทชั้นบรรยากาศ ว่าเป็นชั้นบรรยากาศดี (good) หรือชั้นบรรยากาศไม่ดี (bad) โดยพิจารณาจากจากข้อมูลเรดาห์ 34 ข้อมูล ซึ่งเป็นข้อมูลแบบต่อเนื่อง (continuous data)

2. ปัญหา Pima Indians Diabetes เป็นปัญหาที่เกี่ยวกับการจำแนกประเภทของการเกิดโรคเบาหวานของเพศหญิง โดยพิจารณาจากข้อมูลพื้นฐาน เช่น อายุ จำนวนครั้งในการตั้งครรภ์ และผลการทดสอบทางด้านการแพทย์

3. ปัญหา Seeds เป็นปัญหาการจำแนกประเภทของเมล็ด โดยพิจารณาจากขนาดเมล็ด เส้นรอบวง ความหนาแน่น ความยาวและความกว้างของแก่นกลาง เป็นต้น

4. ปัญหา Vehicle เป็นปัญหาการจำแนกประเภทของรถยนต์ โดยพิจารณาจากข้อมูลของรถยนต์ ทั้งความหนาแน่น ความยาวสูงสุด รัศมี เป็นต้น

5. ปัญหา Contraceptive เป็นปัญหาการจำแนกประเภทยาคุม ว่าผู้หญิงที่แต่งงานแล้วใช้ยาคุมแบบระยะยาว ระยะสั้นในการคุมกำเนิด หรือไม่ใช้เลย โดยพิจารณาจากข้อมูลพื้นฐาน เช่น อายุ ระดับการศึกษาทั้งตัวเองและสามี ศาสนา จำนวนบุตร ทำงานหรือไม่ อาชีพของสามี ระดับการครองชีพ และเปิดรับสื่อหรือไม่

6. ปัญหา Haberman เป็นปัญหาการจำแนกประเภทผู้ป่วยที่ได้รับการผ่าตัดมะเร็งที่หน้าอก ว่าผู้ป่วยเสียชีวิตภายใน 5 ปี หรือสามารถมีชีวิตได้มากกว่า 5 ปี โดยพิจารณาจาก อายุของผู้ป่วย ณ เวลาที่ได้รับการผ่าตัด ปีที่ได้รับการผ่าตัด และจำนวนจุดที่มีปัญหา

7. ปัญหา Breast Tissue เป็นปัญหาการจำแนกประเภทเนื้อเยื่อที่หน้าอก โดยพิจารณาจากข้อมูลทางการแพทย์ ว่าเป็นเนื้อเยื่อประเภทมะเร็ง เนื้องอก โรคเต้านม คล้ายต่อม และเนื้อเยื่อยึดต่อ ซึ่งในงานวิจัยนี้พิจารณาเพียง เป็นเนื้อเยื่อมะเร็งหรือไม่ เท่านั้น

8. ปัญหา Satimage เป็นปัญหาการจำแนกประเภทของดิน โดยพิจาณาจากภาพถ่ายดาวเทียม ขนาด 3x3 พิกเซล

9. ปัญหา Glass_3 เป็นปัญหาการจำแนกประเภทกระจก โดยพิจาณาจากส่วนประกอบ เช่น โซเดียม, เหล็ก, โพแทสเซียม, แคลเซียม และอื่นๆ ว่าเป็นกระจกส่วนไหน กระจกอาคาร กระจกรถยนต์ หรือกระจกบรรจุภัณฑ์

10. ปัญหา Ecoli_om เป็นปัญหาการจำแนกตำแหน่งของโปรตีนว่าอยู่ตำแหน่งของร่างกาย โดยพิจารณาจากข้อมูลทางการแพทย์

ชุดข้อมูลที่นำมาใช้ในการทดลองนั้นต้องมีค่าครบทุกข้อมูล ไม่มีข้อมูลที่สูญหาย (Missing value) ต้องเป็นข้อมูลที่เป็นตัวเลขหรือจำนวนจริงทั้งหมดยกเว้นข้อมูลที่ใช้ระบุเป็นคลาส และเป็นชุดข้อมูลที่นำมาใช้ในการจำแนกประเภทหรือจัดกลุ่มข้อมูลเท่านั้น

ข้อมูลของปัญหาข้างต้นจะแบ่งออกเป็น 3 ชนิด คือ ข้อมูลจำนวนเต็มหรือจำนวนจริง (Integer data or Real data) ข้อมูลแบบต่อเนื่อง (Continues data) และข้อมูลแบบประเภท (Category data)

4.1.2 การเตรียมข้อมูล

ในขั้นตอนแรกของการทดลองจะต้องทำการเตรียมข้อมูล (Data preparation) ก่อนเพราะวิธีที่เกี่ยวข้อง (วิธีสโมทและวิธีสร้างข้อมูลสังเคราะห์บริเวณขอบ) และวิธีที่นำเสนอ ใช้การหาเพื่อนบ้านที่ใกล้ที่สุด k ตัว (k-nearest neighbors : k = 5) มาใช้สังเคราะห์ข้อมูลประเภทเล็กเพื่อปรับการสมดุลของชุดข้อมูล ซึ่งวิธีการเพื่อนบ้านที่ใกล้ที่สุดนั้น จะใช้ค่าของข้อมูล (attribute) ที่เป็นตัวเลขมาคำนวณหาระยะทาง เพื่อหาเพื่อนบ้านที่ใกล้ที่สุด k ตัวของตัวอย่างข้อมูล ดังนั้นชุดข้อมูลที่สามารถนำมาใช้ได้ต้องเป็นข้อมูลที่เป็นตัวเลขทั้งหมด ยกเว้นข้อมูลที่ใช้ระบุเป็นคลาสเท่านั้น มีขั้นตอนเตรียมข้อมูลดังนี้

1. ข้อมูลแบบประเภทจะถูกแปลงให้เป็นข้อมูลจำนวนเต็ม
2. ข้อมูลแบบประเภทที่ใช้ระบุเป็นคลาส จะถูกแปลงให้เหลือเพียงสองคลาสเท่านั้น (minority : majority)

ตัวอย่างที่ 4.1 แสดงการเตรียมข้อมูลของปัญหา Contraceptive

จากตารางที่ 4.2 เป็นตัวอย่างของข้อมูลของปัญหา Contraceptive โดยประกอบด้วย 10 คอลัมน์ C1 และ C4 เป็นข้อมูลแบบจำนวนเต็ม คอลัมน์อื่นเป็นข้อมูลแบบประเภท

จากตารางที่ 4.3 จะแสดงข้อมูลที่ผ่านการเตรียมข้อมูลแล้ว โดยปรับข้อมูลแบบประเภทแปลงให้เป็นข้อมูลจำนวนเต็ม และข้อมูลแบบประเภทที่ใช้ระบุเป็นคลาสนั้นจะถูกแปลงให้เหลือเพียงสองคลาสเท่านั้น แต่จะไม่แปลงถ้าข้อมูลแบบประเภทนั้นเป็นข้อมูลที่เป็นจำนวนเต็มอยู่แล้ว

ตารางที่ 4.2 ตัวอย่างข้อมูลที่มีปัญหาของ Contraceptive

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
24	low	1	3	1	Yes	2	low	No	No-use
26	medium	2	5	0	No	2	high	Good	Long term
30	high	4	2	0	Yes	4	high	Good	Short term

ตารางที่ 4.3 ตัวอย่างการเตรียมข้อมูลของปัญหา Contraceptive

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
24	1	1	3	1	0	2	1	1	other
26	3	2	5	0	1	2	4	0	Long term
30	4	4	2	0	0	4	4	0	other

4.1.3 ขั้นตอนในการทดลอง

1. การหาเพื่อนบ้านใกล้ที่สุด (k-Nearest Neighbors Algorithm) ในงานวิจัยใช้ k=5 เพื่อให้สามารถนำวิธีที่เสนอไปเปรียบเทียบกับวิธีที่เกี่ยวข้องได้ เนื่องมาจากวิธีที่เกี่ยวข้อง เช่น สโมท (SMOTE) และการสร้างข้อมูลสังเคราะห์บริเวณขอบ(Borderline-SMOTE) ใช้ค่า k=5 ในการคำนวณหาเพื่อนบ้านที่ใกล้ที่สุด เพื่อสังเคราะห์ข้อมูล แล้วจึงใช้ต้นไม้ตัดสินใจ C4.5 ในการสร้างโมเดลหรือแบบจำลองเพื่อจำแนกประเภทข้อมูล

2. ในการทดลองจะทำการแบ่งข้อมูลออกเป็น 10 ส่วนด้วยกันโดยใช้ 10-fold cross validation ในการแบ่งข้อมูล เพื่อนำไปใช้เป็นข้อมูลชุดเรียนรู้ และข้อมูลชุดทดสอบ โดยแบ่งอัตราส่วนเป็น 90% และ 10% ตามลำดับ

วิธีการ *K*-Fold Cross-validation

เป็นวิธีการที่แบ่งข้อมูลออกเป็นกลุ่มจำนวน *K* กลุ่ม (*K*-Fold) ในตอนแรกเลือกข้อมูลกลุ่มที่ 1 เป็นข้อมูลชุดทดสอบ และข้อมูลชุดที่เหลือจะเป็นข้อมูลชุดสอน นำข้อมูลไป classify จากนั้นจะสลับข้อมูลกลุ่มที่ 2 มาเป็นชุดทดสอบและข้อมูลกลุ่มอื่นๆที่เหลือเป็นชุดทดสอบ สลับอย่างนี้ไปเรื่อยๆจนครบ *K* กลุ่ม ในขั้นตอนสุดท้ายจะหาค่าเฉลี่ยของค่าความถูกต้องในแต่ละกลุ่ม วิธีการนี้ข้อมูลทุกตัวอย่างจะได้เป็นทั้งชุดทดสอบและชุดสอน

3. ทำการเปรียบเทียบประสิทธิภาพการจำแนกประเภทข้อมูลทั้ง 5 วิธี โดย 3 วิธีแรกเป็นวิธีที่ใช้สำหรับเปรียบเทียบหรือวิธีที่เกี่ยวข้อง ส่วนอีก 2 วิธีที่เหลือเป็นวิธีที่ทำการพัฒนา ซึ่งจะมีรายละเอียดดังนี้

3.1 ชุดข้อมูลดั้งเดิมที่ไม่มีการปรับสมดุล (Original data set, C4.5)

3.2 การสังเคราะห์ข้อมูลประเภทเล็กโดยวิธีสโมท (Synthetic Minority Oversampling Technique, SMOTE)

3.3 การสังเคราะห์ข้อมูลประเภทเล็กโดยวิธีการสร้างข้อมูลสังเคราะห์บริเวณขอบ (Borderline-SMOTE, BSMOTE)

3.4 การเลือกข้อมูลสังเคราะห์โดยวิธีการปรับการกระจายของข้อมูลที่ไม่สมดุลโดยใช้วิธีการหาเพื่อนบ้านที่ใกล้ที่สุดแบบย้อนกลับ K ตัว กำหนดค่าน้อยสุดที่ยอมรับได้ (Threshold) เป็นค่าเฉลี่ยของค่าความรู้ (RSMOTE1)

3.5 การเลือกข้อมูลสังเคราะห์โดยวิธีการปรับการกระจายของข้อมูลที่ไม่สมดุลโดยใช้วิธีการหาเพื่อนบ้านที่ใกล้ที่สุดแบบย้อนกลับ K ตัว กำหนดค่าน้อยสุดที่ยอมรับได้ (Threshold) เป็นค่ามัธยฐานของค่าความรู้ (RSMOTE2)

4.1.4 มาตรที่ใช้วัดประสิทธิภาพการคัดแยกประเภทข้อมูล

4.1.4.1 มาตรวัดเอฟ

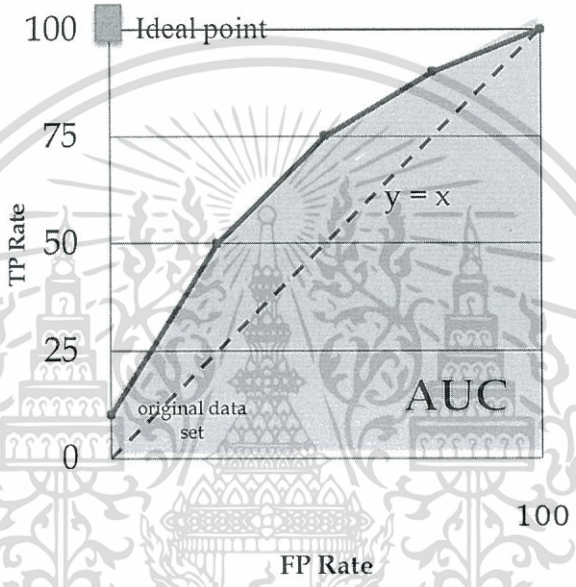
ค่าเอฟ (F-Measure) [2] เป็นผลการเฉลี่ยของค่าความแม่นยำ (precision) และค่าค้นคืน (recall) สาเหตุที่ไม่ใช้ค่าความแม่นยำหรือค่าค้นคืนเป็นมาตรที่ใช้วัดเพียงอย่างเดียวเท่านั้น เนื่องมาจากถ้าวัดเพียงค่าความแม่นยำอย่างเดียว ข้อมูลอาจค้นคืนมาเพียงนิดเดียวแต่เป็นข้อมูลที่จำแนกประเภทถูก จึงให้ค่าความแม่นยำสูง ในขณะที่ค่าค้นคืนมีค่าน้อย และถ้าวัดเพียงค่าค้นอย่างเดียว ข้อมูลอาจค้นคืนมาได้ทั้งหมด จึงให้ค่าค้นคืนที่สูง แต่อาจจะจำแนกประเภทผิดหมดทำให้ค่าความแม่นยำน้อย (กำหนดให้ positive example คือ ข้อมูลประเภทเล็ก และ negative example คือ ข้อมูลประเภทใหญ่)

ตารางที่ 4.4 แสดงรายละเอียดการคำนวณมาตรวัดเอฟ

Confusion matrix		
	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN
TP is the number of positive examples correctly classified. TN is the number of negative examples correctly classified. FP is the number of negative examples incorrectly classified as positive FN is the number of positive examples incorrectly classified as negative		
FP rate = FP/(TN+FP) TP rate = Recall = TP/(TP+FN) Precision = TP/(TP+FP) F-Measure = $((1+\beta^2)$ Recall x Precision) $/(\beta^2$ Recall + Precision) β : adjust recall and precision ($\beta = 1$)		

4.1.4.2 มาตรวัดพื้นที่ใต้เส้นโค้ง

พื้นที่ใต้เส้นโค้ง(Area Under Curve: AUC) ของกราฟรูปแบบของตัวรับสัญญาณ (Receiver Operator Characteristic, ROC) [1] ซึ่งรูปแบบของตัวรับสัญญาณเป็นกราฟที่ใช้ในการแสดงผลระหว่าง อัตราความถูกต้องที่เป็นบวก (true positive rate, TP rate) และอัตราส่วนความผิดพลาดที่เป็นบวก (false positive rate, FP rate) ในกรณีที่พื้นที่ใต้เส้นโค้งที่มีค่ามากกว่าแสดงถึงการจำแนกที่มีประสิทธิภาพสูงกว่า มีค่า TP Rate สูง ในทางกลับกันพื้นที่ใต้โค้งที่มีค่าน้อยกว่าแสดงถึงการจำแนกที่มีประสิทธิภาพต่ำกว่า มีค่า TP Rate น้อย



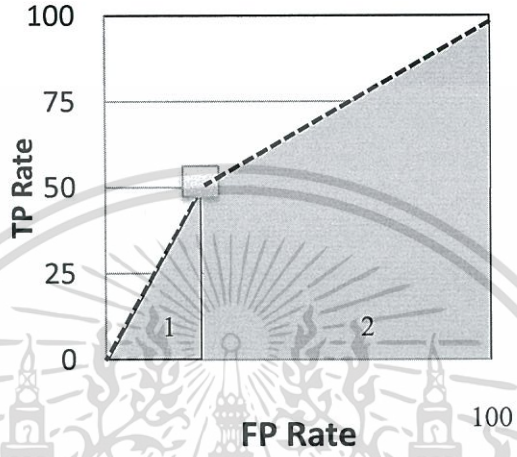
รูปที่ 4.1 กราฟของมาตรวัดพื้นที่ใต้เส้นโค้ง

รูปที่ 4.1 แสดงกราฟของมาตรวัดพื้นที่ใต้เส้นโค้ง(AUC) แกนตั้งเป็นค่าเปอร์เซ็นต์ TP Rate แกนนอนเป็นค่าเปอร์เซ็นต์ FP Rate โดยเส้นกราฟแทนด้วยเส้นประที่ y = x เป็นเส้นกราฟเดาจำแนกประเภทข้อมูลได้ถูกและผิดอย่างละครึ่ง ซึ่งข้อมูลที่ถูกปรับสมดุลหรือสังเคราะห์ข้อมูลเพิ่มแล้วควรจะมีค่าเหนือเส้นกราฟเดานี้ โดยสี่เหลี่ยมสีส้มคือ จุดในจินตนาการ เป็นจุดที่ข้อมูลสามารถจำแนกข้อมูลได้ถูกหมดไม่มีผิดเลย TP Rate = 100 เปอร์เซ็นต์ และเส้นกราฟแทนด้วยเส้นโค้ง จุดแรกด้านซ้ายแทนข้อมูลที่ไม่มีการปรับสมดุล จุดที่เหลือแทนข้อมูลที่ปรับสมดุลแล้ว แต่ใช้เปอร์เซ็นต์ในการสังเคราะห์ข้อมูลที่ไม่เท่ากัน

การคำนวณพื้นที่ใต้เส้นโค้ง ขั้นแรกต้องหาค่าอัตราความถูกต้องที่เป็นบวก (TP Rate) และอัตราส่วนความผิดพลาดที่เป็นบวก (FP Rate) มาพล็อตจุดบนกราฟดังรูปสี่เหลี่ยมในรูปที่ 4.2 แล้วจึงลากเส้นประจากจุด (0,0) กับจุด (100,100) มาลงจุดสี่เหลี่ยมดังรูป จากรูปจะประกอบด้วยสี่เหลี่ยมคางหมูสองอัน อันหนึ่งเป็นสามเหลี่ยมและอันสองเป็นสี่เหลี่ยมคางหมู หาผลรวมของ

พื้นที่ทั้งสองอัน จะได้เป็นค่าพื้นที่ใต้เส้นโค้ง(AUC) สูตรการหาพื้นที่ของสี่เหลี่ยมคางหมูตามสมการที่ 4.1

$$\text{พื้นที่สี่เหลี่ยมคางหมู} = \tfrac{1}{2} \times (\text{ผลบวกด้านคู่ขนาน}) \times \text{สูง} \qquad (4.1)$$



รูปที่ 4.2 แสดงการคำนวณพื้นที่ใต้เส้นโค้ง

4.1.4.3 สถิติทดสอบที

สถิติทดสอบที (t- test Statistic) เป็นการทดสอบสมมติฐานชนิดหนึ่งที่ผู้วิจัยใช้กลุ่มตัวอย่างขนาดเล็ก ($n < 30$) การทดสอบผู้วิจัยจะต้องทราบค่าความแปรปรวนของประชากร หรือในกรณีไม่ทราบค่าความแปรปรวนของประชากรเพราะในงานวิจัยผู้วิจัยจะไม่มีโอกาสทราบค่าความแปรปรวนของประชากรผู้วิจัยก็อาจจะใช้ค่าความแปรปรวนของกลุ่มตัวอย่าง (S^2) แทน

เป็นการทดสอบสมมติฐานเพื่อเปรียบเทียบค่าเฉลี่ยของกลุ่มตัวอย่างสองกลุ่ม ในกรณีที่ไม่ทราบความแปรปรวนของประชากร และกลุ่มตัวอย่างทั้งสองกลุ่มที่มีขนาดเล็ก กล่าวคือ $n_1 < 30$ และ $n_2 < 30$ ซึ่งก่อนที่จะทำการทดสอบโดยใช้สถิติทดสอบที จะต้องนำค่าความแปรปรวนของกลุ่มตัวอย่างทั้งสองกลุ่มไปทดสอบเพื่อสรุปว่า ประชากรที่ศึกษานั้นมีความแปรปรวนเท่ากันหรือไม่ มีขั้นตอนในการทดสอบมีดังนี้

1. ตรวจสอบข้อตกลงเบื้องต้นของสถิติทดสอบ มีดังนี้
   1.1 กลุ่มตัวอย่างทั้งสองกลุ่มได้มาโดยการสุ่มอย่างเป็นอิสระจากกัน
   1.2 ประชากรทั้งสองกลุ่มมีการแจกแจงแบบปกติ
   1.3 ข้อมูลอยู่ในมาตราอันตรภาคหรืออัตราส่วน
   1.4 ไม่ทราบค่าความแปรปรวนของประชากร

2. กำหนดสมมติฐานทางสถิติ

สำหรับการทดสอบแบบสองทิศทาง

H_o: $\mu_1 = \mu_2$

H_1: $\mu_1 \neq \mu_2$

สำหรับการทดสอบแบบทิศทางเดียว

H_o: $\mu_1 = \mu_2$ (Equal Significant)

H_1: $\mu_1 > \mu_2$ (Win Significant)

$\mu_1 < \mu_2$ (Lose Significant)

3. กำหนด α หรือระดับนัยสำคัญ

4. คำนวณค่าสถิติ t จากสูตรใดสูตรหนึ่งใน 2 สูตร ดังนี้

4.1 เมื่อทดสอบได้ว่า $\alpha_1^2 = \alpha_2^2$ เรียกสูตรนี้ว่า t – test ชนิด Pooled Variance

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{S_p^2\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} \tag{4.2}$$

มี $df = n_1 + n_2 - 2$

S_p^2 แทน ความแปรปรวนร่วม (Pooled Variance)

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \tag{4.3}$$

4.2 เมื่อทดสอบได้ว่า $\alpha_1^2 \neq \alpha_2^2$ เรียกสูตรนี้ว่า t – test ชนิด Separated Variance

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{4.4}$$

โดยมี

$$df = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{(n_1 - 1)} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{(n_2 - 1)}} \tag{4.5}$$

5. กำหนดขอบเขตวิกฤตโดยหาค่า t วิกฤต

6. สรุปผลการทดลอง

พิจารณาตัวเลขเท่านั้นไม่คิดเครื่องหมาย

$t \geq t$ วิกฤต จะปฏิเสธ H_0

$t < t$ วิกฤต จะยอมรับ H_0

4.2 ผลการทดลอง

4.2.1 มาตรวัดเอฟ และสถิติทดสอบที

เป็นผลการเฉลี่ยของค่าความแม่นยำ (precision) กับค่าค้นคืน (recall) ค่ามากแสดงว่าข้อมูลสามารถค้นคืนได้มากและจำแนกประเภทได้อย่างแม่นยำด้วย โดยจะทำการทดลองทั้งสิ้น 10 ครั้ง (10 fold cross validation) ในแต่ละเปอร์เซ็นต์การสังเคราะห์ข้อมูลประเภทเล็ก (%Sampling) แล้วนำค่าที่ได้มาเฉลี่ยผลการทดลองแสดงดังตารางที่ 4.5 - 4.14 ซึ่งแสดงผลการเปรียบเทียบคุณภาพของการจำแนกประเภทข้อมูลบนมาตรวัดค่าเอฟทั้งวิธีที่เกี่ยวข้องและวิธีที่พัฒนา รวมทั้งหมด 5 วิธี สามารถอ่านค่าตาราง ได้ดังนี้

1. ค่าที่แสดงตัวหนาในตาราง แทนผลลัพธ์ที่ดีที่สุดในแต่ละ%Sampling
2. %Minority แทนเปอร์เซ็นต์ข้อมูลประเภทเล็กในชุดข้อมูล และ %Sampling แทนเปอร์เซ็นต์ในการสังเคราะห์ข้อมูลประเภทเล็ก เช่น ที่ %Samplingที่ 100 ในกรณีมีข้อมูลประเภทเล็ก 10 ตัวอย่าง จะสังเคราะห์ข้อมูลประเภทเล็กเพิ่มขึ้นอีก 10 ตัวอย่าง เป็นต้น
3. Variance ความแปรปรวนของชุดข้อมูล เพื่อดูว่าชุดข้อมูลมีการกระจายตัวแบบใด สามารถคำนวณได้ตามสมการที่ 3.1 ในบทที่ 3 ถ้าค่าความแปรปรวนมีค่ามากแสดงว่าชุดข้อมูลมีการกระจายตัวที่มาก ไม่เหมาะสมกับวิธีที่นำเสนอ
4. C4.5 แทนวิธีการที่ไม่มีการสร้างตัวอย่างสังเคราะห์เพิ่มแก่ข้อมูลต้นฉบับ, SMOTE แทนวิธีสโมท, BSMOTE แทนวิธีการสร้างข้อมูลสังเคราะห์บริเวณขอบ, RSMOTE1 แทนวิธีที่นำเสนอแต่ใช้ Threshold เป็นค่าเฉลี่ยของค่าความรู้ และ RSMOTE2 แทนวิธีที่นำเสนอแต่ใช้ Threshold เป็นค่ามัธยฐานของค่าความรู้
5. Frequency of 1st Rank แทนผลลัพธ์รวมที่เป็นคำตอบดีที่สุด เช่น ตารางที่ 4.5 ที่คอลัมน์วิธี C4.5 มีค่า Frequency of 1st Rank เป็น 5 แสดงว่าที่ชุดข้อมูล Ionosphere ในวิธี C 4.5 ให้ผลลัพธ์ในการจำแนกประเภทข้อมูลที่ดีที่สุด 5 การทดลองจากทั้งหมด 5 การทดลอง บนมาตรวัดค่าเอฟ
6. W/E/L (Win/Equal/Lose Significant [9]) เป็นค่าที่บอกว่าวิธีที่เปรียบเทียบให้ผล มากกว่า/เท่ากัน/น้อยกว่า อย่างมีนัยสำคัญทางสถิติที่ความเชื่อมั่น 95% หรือไม่ โดยใช้วิธีการสถิติทดสอบที (t-test)

Win Significant (Win: W) ค่าเปรียบเทียบระหว่างสองวิธีให้ผลลัพธ์ที่มากกว่าและแตกต่างอย่างมีนัยสำคัญ

Equal Significant (Equal: E) ค่าเปรียบเทียบระหว่างสองวิธีให้ผลลัพธ์ที่แตกต่างอย่างไม่มีนัยสำคัญ

Lose Significant (Lose: L) ค่าเปรียบเทียบระหว่างสองวิธีให้ผลลัพธ์ที่น้อยและแตกต่างอย่างมีนัยสำคัญ

เช่น ตารางที่ 4.5 ที่แถว RSMOTE1 กับ คอลัมน์ SMOTE มีค่า W/E/L เท่ากับ 1/4/0 แสดงว่าวิธีที่นำเสนอ RSMOTE1 มีผลลัพธ์ที่ดีกว่าและแตกต่างอย่างมีนัยสำคัญอยู่ 1 การทดลอง, ผลลัพธ์ที่แตกต่างอย่างไม่มีนัยสำคัญอยู่ 4 การทดลอง และผลลัพธ์ที่น้อยกว่าและแตกต่างอย่างมีนัยสำคัญอยู่ 0 เมื่อเปรียบเทียบกับวิธี SMOTE และที่แถว RSMOTE1 กับ คอลัมน์ RSMOTE2 มีค่า W/E/L เท่ากับ 1/4/0 แสดงว่าวิธีที่นำเสนอ RSMOTE1 มีผลลัพธ์ที่ดีกว่าและแตกต่างอย่างมีนัยสำคัญอยู่ 1 การทดลอง, ผลลัพธ์ที่แตกต่างอย่างไม่มีนัยสำคัญอยู่ 4 การทดลอง และผลลัพธ์ที่น้อยกว่าและแตกต่างอย่างมีนัยสำคัญอยู่ 0 เมื่อเปรียบเทียบกับวิธี RSMOTE2 เป็นต้น

ตารางที่ 4.5 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Ionosphere บนมาตรวัดเอฟ

%Minority	Variance	%Sampling	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
35.8	35.63	100	**83.3%**	82.2%	82.6%	**83.3%**	82.1%
		200	**83.3%**	82.6%	82.9%	82.6%	81.5%
		300	**83.3%**	82.3%	82.5%	82.1%	80.7%
		400	**3.3%**	81.7%	81.5%	82.3%	80.3%
		500	**83.3%**	80.4%	81.2%	82.1%	80.9%
Frequency of 1st Rank			5	0	0	1	0

	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
RSMOTE1 (W/E/L)	0/4/1	1/4/0	0/5/0	0/5/0	1/4/0
RSMOTE2 (W/E/L)	0/1/4	0/3/2	0/2/3	0/4/1	0/5/0

จากตารางที่ 4.5 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Ionosphere บนมาตรวัดเอฟ ที่เปอร์เซ็นต์สังเคราะห์ข้อมูลที่ 100 วิธี C4.5 และวิธีที่นำเสนอ RSMOTE1 ให้ผลลัพธ์ในการคัดแยกประเภทข้อมูลที่ดีที่สุด ส่วนเปอร์เซ็นต์สังเคราะห์ข้อมูลที่ 200 - 500 วิธี C4.5 ให้ผลลัพธ์ที่ดีที่สุด ในขณะที่วิธีที่นำเสนอ RSMOTE1 ให้ผลลัพธ์รองลงมา ส่วนวิธีที่นำเสนอ RSMOTE2 ให้ผลลัพธ์ในการคัดแยกประเภทข้อมูลด้อยที่สุดจาก 5 วิธี

การวัดผลเปรียบเทียบด้วยวิธีการสถิติทดสอบที (t-test) สาเหตุที่ต้องเปรียบเทียบว่าผลลัพธ์ที่ได้แตกต่างอย่างมีนัยสำคัญหรือไม่นั้น เนื่องมาจากการดูเพียงค่าผลรวมผลลัพธ์ที่ดีที่สุดไม่เพียง

พอที่จะบอกว่าได้ว่าวิธีที่นำเสนอนั้นสามารถเพิ่มประสิทธิภาพในการคัดแยกประเภทข้อมูลได้จริง อาจจะให้ผลลัพธ์ที่เท่ากันก็ได้ จึงต้องนำมาเปรียบเทียบว่าผลลัพธ์ที่ได้นั้นต้องแตกต่างอย่างมีนัยสำคัญจริง ซึ่งจากตารางที่ 4.5 ส่วนล่าง วิธีที่นำเสนอ RSMOTE1 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญเพียง 1 การทดลอง บนวิธี SMOTE เท่านั้น และให้ประสิทธิภาพต่ำกว่าอย่างมีนัยสำคัญเพียง 1 การทดลอง บนวิธี C4.5 ส่วนวิธีที่นำเสนอ RSMOTE2 ไม่ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญบนการทดลองใดเลย และให้ประสิทธิภาพต่ำกว่าอย่างมีนัยสำคัญ 4, 2 และ 3 การทดลอง เมื่อเปรียบเทียบกับวิธีการ C4.5, SMOTE และ BSMOTE ตามลำดับ จึงสรุปได้ว่าวิธีที่นำเสนอ RSMOTE1 สามารถคัดแยกประเภทข้อมูลได้ดีกว่าวิธี SMOTE เท่านั้น แต่คัดแยกได้ด้อยกว่าในวิธี C4.5 ส่วนวิธีที่นำเสนอ RSMOTE2 คัดแยกประเภทข้อมูลได้ด้อยที่สุด

ตารางที่ 4.6 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Pima Indians Diabetes บนมาตรวัดเอฟ

%Minority	Variance	%Sampling	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
34.77	741,156	100	62.8%	63.3%	65.3%	**65.4%**	64.7%
		200	62.8%	63.8%	**65.1%**	64.5%	64.3%
		300	62.8%	63.4%	**65.2%**	64.7%	64.7%
		400	62.8%	63.2%	**65.4%**	64.7%	64.4%
		500	62.8%	63.8%	**65.8%**	64.7%	64.0%
Frequency of 1st Rank			0	0	4	1	0

	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
RSMOTE1 (W/E/L)	5/0/0	4/1/0	0/4/1	0/5/0	0/5/0
RSMOTE2 (W/E/L)	4/1/0	3/2/0	0/4/1	0/5/0	0/5/0

จากตารางที่ 4.6 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Pima Indians Diabetes บนมาตรวัดเอฟ วิธี BSMOTE ให้ผลรวมผลลัพธ์ในการคัดแยกประเภทข้อมูลที่ดีที่สุด ส่วนวิธีที่นำเสนอ RSMOTE1 ให้ผลลัพธ์รองลงมา โดยที่ปัญหา Pima Indians Diabetes มีค่าความแปรปรวน (Variance) ของชุดข้อมูลมากกกว่า 1000 แสดงว่าชุดข้อมูลที่นำมาสังเคราะห์ข้อมูลนั้นมีการกระจายตัวของข้อมูลที่สูงมากไม่เหมาะกับวิธีที่นำเสนอทั้ง RSMOTE1 และ RSMOTE2

การวัดผลเปรียบเทียบด้วยวิธีการสถิติทดสอบที (t-test) จากตารางที่ 4.6 ส่วนล่าง วิธีที่นำเสนอ RSMOTE1 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 5 และ4 การทดลอง บนวิธี C4.5 และ SMOTE เท่านั้น แต่ให้ประสิทธิภาพต่ำกว่าอย่างมีนัยสำคัญเพียง 1 การทดลอง บนวิธี BSMOTE ส่วนวิธีที่

นำเสนอ RSMOTE2 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 4 และ 3 การทดลอง บนวิธี C4.5 และ SMOTE แต่ให้ประสิทธิภาพต่ำกว่าอย่างมีนัยสำคัญเพียง 1 การทดลอง บนวิธี BSMOTE จึงสรุปได้ว่าวิธีที่นำเสนอทั้งสองวิธีสามารถคัดแยกประเภทข้อมูลได้ดีกว่าวิธี C4.5 และ SMOTE แต่ด้อยกว่าวิธี BSMOTE

ตารางที่ 4.7 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Seeds บนมาตรวัดเอฟ

%Minority	Variance	%Sampling	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
33.33	0.55	100	77.5%	**89.3%**	84.9%	85.5%	84.9%
		200	77.5%	84.7%	84.7%	**85.5%**	84.8%
		300	77.5%	84.7%	84.2%	**86.0%**	84.2%
		400	77.5%	85.3%	84.2%	**86.0%**	83.5%
		500	77.5%	85.3%	84.3%	**86.0%**	84.2%
Frequency of 1st Rank			0	1	0	4	0

	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
RSMOTE1 (W/E/L)	5/0/0	1/3/1	4/1/0	0/5/0	2/3/0
RSMOTE2 (W/E/L)	5/0/0	0/3/2	0/5/0	0/3/2	0/5/0

จากตารางที่ 4.7 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Seeds บนมาตรวัดค่าเอฟ วิธีที่นำเสนอ RSMOTE1 ให้ผลลัพธ์ที่ดีที่สุดมากกว่าวิธีอื่น คือ ให้ผลลัพธ์ที่ดีที่สุด 4 การทดลอง จากทั้งหมด 5 การทดลอง ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 200-500 โดยวิธี SMOTE ให้ผลลัพธ์ที่ดีที่สุดเพียง 1 การทดลอง ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 100

การวัดผลเปรียบเทียบด้วยวิธีการสถิติทดสอบที (t-test) จากตารางที่ 4.7 ส่วนล่าง วิธีที่นำเสนอ RSMOTE1 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 5, 1 และ 4 การทดลอง บนวิธี C4.5, SMOTE และ BSMOTE ตามลำดับ แต่ให้ประสิทธิภาพต่ำกว่าอย่างมีนัยสำคัญเพียง 1 การทดลอง บนวิธี SMOTE ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 100 ส่วนวิธีที่นำเสนอ RSMOTE2 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 5 การทดลอง บนวิธี C4.5 เท่านั้น แต่ให้ประสิทธิภาพต่ำกว่าอย่างมีนัยสำคัญ 2 การทดลอง บนวิธี SMOTE จึงสรุปได้ว่าวิธีที่นำเสนอ RSMOTE1 สามารถคัดแยกประเภทข้อมูลได้ดีกว่าวิธี C4.5 และ BSMOTE แต่คัดแยกได้พอๆกันในวิธี SMOTE ส่วนวิธีที่นำเสนอ RSMOTE2 สามารถคัดแยกประเภทข้อมูลได้ดีกว่าวิธี C4.5 เท่านั้น แต่คัดแยกได้ด้อยกว่าในวิธี SMOTE

ตารางที่ 4.8 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Vehicle บนมาตรวัดเอฟ

%Minority	Variance	%Sampling	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
30.75	637	100	86.6%	87.4%	87.4%	**88.1%**	**88.1%**
		200	86.6%	86.7%	87.3%	**88.1%**	87.3%
		300	86.6%	87.2%	86.8%	**88.0%**	86.3%
		400	86.6%	86.3%	87.0%	**88.3%**	87.3%
		500	86.6%	86.7%	86.9%	**88.4%**	87.0%
Frequency of 1st Rank			0	0	0	5	1

	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
RSMOTE1 (W/E/L)	5/0/0	3/2/0	3/2/0	0/5/0	2/3/0
RSMOTE2 (W/E/L)	1/4/0	1/3/1	0/5/0	0/3/2	0/5/0

จากตารางที่ 4.8 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Vehicle บนมาตรวัดค่าเอฟ วิธีที่นำเสนอ RSMOTE1 ให้ผลรวมผลลัพธ์ที่ดีที่สุดมากกว่าวิธีอื่น คือ ให้ผลลัพธ์ที่ดีที่สุด 5 การทดลอง จากทั้งหมด 5 การทดลอง ส่วนวิธีที่นำเสนอ RSMOTE2 ให้ผลลัพธ์ที่ดีที่สุดเพียง 1 การทดลอง ที่เปอร์เซ็นต์สังเคราะห์ข้อมูลที่ 100

การวัดผลเปรียบเทียบด้วยวิธีการสถิติทดสอบที (t-test) จากตารางที่ 4.8 ส่วนล่าง วิธีที่นำเสนอ RSMOTE1 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 5, 3 และ 3 การทดลอง บนวิธี C4.5, SMOTE และ BSMOTE ตามลำดับ ส่วนวิธีที่นำเสนอ RSMOTE2 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญเพียง 1 การทดลอง บนวิธี C4.5 และ BSMOTE แต่ให้ประสิทธิภาพต่ำกว่าอย่างมีนัยสำคัญเพียง 1 การทดลอง บนวิธี SMOTE จึงสรุปได้ว่าวิธีที่นำเสนอ RSMOTE1 สามารถคัดแยกประเภทข้อมูลได้ดีกว่าวิธีที่เกี่ยวข้องทั้งหมด ส่วนวิธีที่นำเสนอ RSMOTE2 สามารถคัดแยกประเภทข้อมูลได้ดีกว่าวิธี C4.5 เท่านั้น แต่คัดแยกได้พอๆกันในวิธี SMOTE และ BSMOTE

ตารางที่ 4.9 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Contraceptive บนมาตรวัดเอฟ

%Minority	Variance	%Sampling	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
29.21	10.18	100	42.9%	44.1%	**44.5%**	44.4%	44.4%
		200	42.9%	44.7%	44.3%	**45.0%**	44.3%
		300	42.9%	42.7%	44.0%	**44.9%**	44.1%
		400	42.9%	42.6%	43.3%	**44.8%**	44.1%
		500	42.9%	42.5%	42.9%	**44.7%**	43.9%
Frequency of 1st Rank			0	0	1	4	0

	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
RSMOTE1 (W/E/L)	5/0/0	3/2/0	2/3/0	0/5/0	0/5/0
RSMOTE2 (W/E/L)	5/0/0	3/2/0	1/4/0	0/5/0	0/5/0

จากตารางที่ 4.9 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Contraceptive บนมาตรวัดค่าเอฟ วิธีที่นำเสนอ RSMOTE1 ให้ผลรวมผลลัพธ์ที่ดีที่สุดมากกว่าวิธีอื่น คือ ให้ผลลัพธ์ที่ดีที่สุด 4 การทดลอง จากทั้งหมด 5 การทดลอง ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 200-500 โดยวิธี BSMOTE ให้ผลลัพธ์ที่ดีที่สุดเพียง 1 การทดลอง ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 100

การวัดผลเปรียบเทียบด้วยวิธีการสถิติทดสอบที (t-test) จากตารางที่ 4.9 วิธีที่นำเสนอทั้ง RSMOTE1 และ RSMOTE2 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 5, 3 และ 2 การทดลอง บนวิธี C4.5, SMOTE และ BSMOTE ตามลำดับ จึงสรุปได้ว่าวิธีที่นำเสนอทั้งสองวิธีสามารถคัดแยกประเภทข้อมูลได้ดีกว่าวิธีที่เกี่ยวข้องทั้งหมด

ตารางที่ 4.10 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Haberman บนมาตรวัดเอฟ

%Minority	Variance	%Sampling	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
26.47	300	100	36.3%	50.0 %	**50.2%**	47.4%	49.5%
		200	36.3%	49.5%	49.2%	**49.7%**	49.2%
		300	36.3%	47.1%	48.5%	**48.9%**	**48.9%**
		400	36.3%	48.0%	48.2%	**49.2%**	48.7%
		500	36.3%	47.9%	48.0%	**49.4%**	48.3%
Frequency of 1st Rank			0	0	1	4	1

	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
RSMOTE1 (W/E/L)	5/0/0	2/2/1	1/3/1	0/5/0	0/5/0
RSMOTE2 (W/E/L)	5/0/0	1/4/0	0/5/0	0/5/0	0/5/0

จากตารางที่ 4.10 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Haberman บนมาตรวัดค่าเอฟ วิธีที่นำเสนอ RSMOTE1 ให้ผลรวมผลลัพธ์ที่ดีที่สุดมากกว่าวิธีอื่น คือ ให้ผลลัพธ์ที่ดีที่สุด 4 การทดลอง จากทั้งหมด 5 การทดลอง ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 200-500 ส่วนวิธีที่นำเสนอ RSMOTE2 ให้ผลลัพธ์ที่ดีที่สุดเพียง 1 การทดลอง ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 300 และ วิธี BSMOTE ให้ผลลัพธ์ที่ดีที่สุดเพียง 1 การทดลอง ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 100

การวัดผลเปรียบเทียบด้วยวิธีการสถิติทดสอบที (t-test) จากตารางที่ 4.10 ส่วนล่าง วิธีที่นำเสนอ RSMOTE1 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 5, 2 และ 1 การทดลอง บนวิธี C4.5, SMOTE และ BSMOTE ตามลำดับ แต่ให้ประสิทธิภาพต่ำกว่าอย่างมีนัยสำคัญเพียง 1 การทดลอง บนวิธี SMOTE และ BSMOTE ส่วนวิธีที่นำเสนอ RSMOTE2 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 5 และ 1 การทดลอง บนวิธี C4.5 และ SMOTE จึงสรุปได้ว่าวิธีที่นำเสนอทั้งสองวิธีสามารถคัดแยกประเภทข้อมูลได้ดีกว่าวิธี C4.5 และ SMOTE แต่คัดแยกประเภทข้อมูลได้พอๆกันกับวิธีBSMOTE

ตารางที่ 4.11 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Breast Tissue บนมาตรวัดเอฟ

%Minority	Variance	%Sampling	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
24.71	1.95	100	82.6%	83.6%	**84.6%**	84.3%	83.0%
		200	82.6%	85.2%	85.7%	**85.7%**	**85.7%**
		300	82.6%	85.5%	85.7%	**86.5%**	**86.5%**
		400	82.6%	84.0%	83.4%	86.5%	**87.0%**
		500	82.6%	84.9%	84.7%	86.5%	**87.0%**
Frequency of 1st Rank			0	0	1	2	4

	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
RSMOTE1 (W/E/L)	5/0/0	3/2/0	3/2/0	0/5/0	0/5/0
RSMOTE2 (W/E/L)	4/1/0	3/2/0	2/3/0	0/5/0	0/5/0

จากตารางที่ 4.11 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Breast Tissue บนมาตรวัดค่าเอฟ วิธีที่นำเสนอ RSMOTE2 ให้ผลรวมผลลัพธ์ที่ดีที่สุดมากกว่าวิธีอื่น คือ ให้ผลลัพธ์ที่ดีที่สุด 4 การทดลอง จากทั้งหมด 5 การทดลอง ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 200-500 ส่วนวิธีที่นำเสนอ RSMOTE1 ให้ผลลัพธ์ที่ดีที่สุดเพียง 2 การทดลอง ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 300-400 และ วิธี BSMOTE ให้ผลลัพธ์ที่ดีที่สุดเพียง 1 การทดลอง ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 100

การวัดผลเปรียบเทียบด้วยวิธีการสถิติทดสอบที (t-test) จากตารางที่ 4.11 ส่วนล่าง วิธีที่นำเสนอ RSMOTE1 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 5, 3 และ 3 การทดลอง บนวิธี C4.5, SMOTE และ BSMOTE ตามลำดับ ส่วนวิธีที่นำเสนอ RSMOTE2 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 4, 3 และ 2 การทดลอง บนวิธี C4.5, SMOTE และ BSMOTE ตามลำดับ จึงสรุปได้ว่าวิธีที่นำเสนอทั้งสองวิธีสามารถคัดแยกประเภทข้อมูลได้ดีกว่าวิธีที่เกี่ยวข้องทั้งหมด

ตารางที่ 4.12 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Satimage บนมาตรวัดเอฟ

%Minority	Variance	%Sampling	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
10.75	640,827	100	55.7%	**57.6%**	57.4%	56.5%	57.4%
		200	55.7%	57.3%	**58.0%**	57.4%	57.0%
		300	55.7%	58.3%	**58.7%**	57.7%	58.0%
		400	55.7%	57.7%	57.8%	57.3%	**58.0%**
		500	55.7%	57.4%	**58.2%**	57.6%	57.5%
Frequency of 1st Rank			0	1	3	0	1

	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
RSMOTE1 (W/E/L)	5/0/0	0/4/1	0/3/2	0/5/0	0/5/0
RSMOTE2 (W/E/L)	5/0/0	0/5/0	0/5/0	0/5/0	0/5/0

จากตารางที่ 4.12 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Satimage บนมาตรวัดค่าเอฟ วิธี BSMOTE ให้ผลลัพธ์ในการคัดแยกประเภทข้อมูลที่ดีที่สุด ให้ผลลัพธ์ที่ดีที่สุด 3 การทดลอง จากทั้งหมด 5 การทดลอง ส่วนวิธีที่นำเสนอ RSMOTE1 ให้ผลลัพธ์รองลงมา โดยที่ปัญหา Satimage มีค่าความแปรปรวน (Variance) ของชุดข้อมูลมากกกว่า 1000 เช่นเดียวกับปัญหา Pima Indians Diabetes แสดงว่าชุดข้อมูลที่นำมาสังเคราะห์ข้อมูลนั้นมีการกระจายตัวของข้อมูลที่สูงมาก ไม่เหมาะกับวิธีที่นำเสนอทั้ง RSMOTE1 และ RSMOTE2

การวัดผลเปรียบเทียบด้วยวิธีการสถิติทดสอบที (t-test) จากตารางที่ 4.12 ส่วนล่าง วิธีที่นำเสนอ RSMOTE1 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 5 การทดลอง บนวิธี C4.5 เท่านั้น แต่ให้ประสิทธิภาพต่ำกว่าอย่างมีนัยสำคัญเพียง 1 และ 2การทดลอง บนวิธี SMOTE และ BSMOTE ส่วนวิธีที่นำเสนอ RSMOTE2 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 5 การทดลอง บนวิธี C4.5 จึงสรุปได้ว่าวิธีที่นำเสนอ RSMOTE1 สามารถคัดแยกประเภทข้อมูลได้ดีกว่าวิธี C4.5 แต่คัดแยกประเภทข้อมูลได้ด้อยกว่าวิธี SMOTE และ BSMOTE ส่วนวิธีที่นำเสนอ RSMOTE2 สามารถคัดแยกประเภทข้อมูลได้ดีกว่าวิธี C4.5 และคัดแยกประเภทข้อมูลได้พอๆกัน กับวิธี SMOTE และ BSMOTE

ตารางที่ 4.13 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Glass_3 บนมาตรวัดเอฟ

%Minority	Variance	%Sampling	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
8.63	11.29	100	**52.1%**	50.4%	48.5%	44.2%	49.2%
		200	52.1%	52.9%	50.6%	55.4%	**58.1%**
		300	52.1%	52.1%	46.9%	**57.4%**	**57.4%**
		400	52.1%	52.3%	44.4%	56.4%	**57.0%**
		500	52.1%	50.2%	42.1%	55.4%	**55.6%**
Frequency of 1st Rank			1	0	0	1	4

	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
RSMOTE1 (W/E/L)	4/0/1	3/1/1	4/1/0	0/5/0	0/4/1
RSMOTE2 (W/E/L)	4/0/1	4/1/0	4/1/0	1/4/0	0/5/0

จากตารางที่ 4.13 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Glass_3 บนมาตรวัดค่าเอฟ วิธีที่นำเสนอ RSMOTE2 ให้ผลรวมผลลัพธ์ที่ดีที่สุดมากกว่าวิธีอื่น คือ ให้ผลลัพธ์ที่ดีที่สุด 4 การทดลอง จากทั้งหมด 5 การทดลอง ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 200-500 ส่วนวิธีที่นำเสนอ RSMOTE1 ให้ผลลัพธ์ที่ดีที่สุดเพียง 1 การทดลอง ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 300 และ วิธี C4.5 ให้ผลลัพธ์ที่ดีที่สุดเพียง 1 การทดลอง ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 100

การวัดผลเปรียบเทียบด้วยวิธีการสถิติทดสอบที (t-test) จากตารางที่ 4.13 ส่วนล่าง วิธีที่นำเสนอ RSMOTE1 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 4, 3 และ 4 การทดลอง บนวิธี C4.5, SMOTE และ BSMOTE ตามลำดับ ส่วนวิธีที่นำเสนอ RSMOTE2 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 4, 4 และ 4 การทดลอง บนวิธี C4.5, SMOTE และ BSMOTE ตามลำดับ จึงสรุปได้ว่าวิธีที่นำเสนอทั้งสองวิธีสามารถคัดแยกประเภทข้อมูลได้ดีกว่าวิธีที่เกี่ยวข้องทั้งหมด

ตารางที่ 4.14 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Ecoli_om บนมาตรวัดเอฟ

%Minority	Variance	%Sampling	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
6.32	7.65	100	58.8%	72.9%	**78.5%**	74.7%	76.7%
		200	58.8%	72.6%	73.2%	**76.9%**	76.1%
		300	58.8%	75.5%	74.3%	**76.6%**	76.2%
		400	58.8%	75.9%	72.7%	**77.1%**	77.1%
		500	58.8%	77.0%	71.5%	**77.1%**	76.8%
Frequency of 1st Rank			0	0	1	4	0

	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
RSMOTE1 (W/E/L)	5/0/0	1/4/0	4/0/1	0/5/0	0/5/0
RSMOTE2 (W/E/L)	5/0/0	1/4/0	4/0/1	0/5/0	0/5/0

จากตารางที่ 4.14 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Ecoli_om บนมาตรวัดค่าเอฟ วิธีที่นำเสนอ RSMOTE1 ให้ผลรวมผลลัพธ์ที่ดีที่สุดมากกว่าวิธีอื่น คือ ให้ผลลัพธ์ที่ดีที่สุด 4 การทดลอง จากทั้งหมด 5 การทดลอง ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 200-500 และ วิธี BSMOTE ให้ผลลัพธ์ที่ดีที่สุดเพียง 1 การทดลอง ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 100

การวัดผลเปรียบเทียบด้วยวิธีการสถิติทดสอบที (t-test) จากตารางที่ 4.14 ส่วนล่าง วิธีที่นำเสนอทั้ง RSMOTE1 และ RSMOTE2 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 5, 1 และ 4 การทดลอง บนวิธี C4.5, SMOTE และ BSMOTE ตามลำดับ จึงสรุปได้ว่าวิธีที่นำเสนอทั้งสองวิธีสามารถคัดแยกประเภทข้อมูลได้ดีกว่าวิธีที่เกี่ยวข้องทั้งหมด

4.2.2 มาตรวัดพื้นที่ใต้เส้นโค้ง และสถิติทดสอบที

เป็นกราฟที่ใช้ในการแสดงผลระหว่าง อัตราความถูกต้องที่เป็นบวก (true positive rate, TP rate) และ อัตราส่วนความผิดพลาดที่เป็นบวก (false positive rate, FP rate) ในกรณีที่พื้นที่ใต้โค้งที่มีค่ามากกว่าแสดงถึงการจำแนกที่มีประสิทธิภาพสูงกว่า โดยจะทำการทดลองทั้งสิ้น 10 ครั้ง (ใช้ 10-fold cross validation) ในแต่ละเปอร์เซ็นต์การสังเคราะห์ข้อมูลประเภทเล็ก (%Sampling) แล้วนำค่าที่ได้มาเฉลี่ยผลการทดลองแสดงดังตารางที่ 4.15 - 4.24 ซึ่งแสดงผลการเปรียบเทียบคุณภาพของการจำแนกประเภทข้อมูลทั้งวิธีที่เกี่ยวข้องและวิธีที่พัฒนา รวมทั้งหมด 5 วิธี) สามารถอ่านค่าตาราง ได้ดังนี้

1. ค่าที่แสดงตัวหนาในตาราง แทนผลลัพธ์ที่ดีที่สุดในแต่ละ%Sampling
2. %Minority แทนเปอร์เซ็นต์ข้อมูลประเภทเล็กในชุดข้อมูล และ %Sampling แทนเปอร์เซ็นต์ในการสังเคราะห์ข้อมูลประเภทเล็ก เช่น ที่ %Sampling 100 ในกรณีมีข้อมูลประเภทเล็ก 10 ตัวอย่าง จะสังเคราะห์ข้อมูลประเภทเล็กเพิ่มขึ้นอีก 10 ตัวอย่าง เป็นต้น
3. Variance ความแปรปรวนของชุดข้อมูล เพื่อดูว่าชุดข้อมูลมีการกระจายตัวแบบใด สามารถคำนวณได้ตามสมการที่ 3.1 ในบทที่ 3 ถ้าค่าความแปรปรวนมีค่ามากแสดงว่าชุดข้อมูลมีการกระจายตัวที่มาก ไม่เหมาะสมกับวิธีที่นำเสนอ
4. C4.5 แทนวิธีการที่ไม่มีการสร้างตัวอย่างสังเคราะห์เพิ่มแก่ข้อมูลต้นฉบับ, SMOTE แทนวิธีสโมท, BSMOTE แทนวิธีการสร้างข้อมูลสังเคราะห์บริเวณขอบ, RSMOTE1 แทนวิธีที่นำเสนอแต่ใช้ Threshold เป็นค่าเฉลี่ย และ RSMOTE2 แทนวิธีที่นำเสนอแต่ใช้ Threshold เป็นค่ามัธยฐาน
5. Frequency of 1st Rank แทนผลลัพธ์รวมที่เป็นคำตอบดีที่สุด เช่น ตารางที่ 4.15 ที่คอลัมน์วิธี C4.5 มีค่า Frequency of 1st Rank เป็น 5 แสดงว่าที่ชุดข้อมูล Ionosphere วิธี C4.5 ให้ผลลัพธ์ที่ดีที่สุด 5 การทดลองจากทั้งหมด 5 การทดลอง บนมาตรวัดพื้นที่ใต้เส้นโค้ง
6. W/E/L (Win/Equal/Lose Significant [9]) เป็นค่าที่บอกว่าวิธีที่เปรียบเทียบ มากกว่า/เท่ากัน/น้อยกว่า อย่างมีนัยสำคัญทางสถิติที่ความเชื่อมั่น 95% หรือไม่ โดยใช้วิธีการสถิติทดสอบที (t-test)

Win Significant (Win: W) ค่าเปรียบเทียบระหว่างสองวิธีให้ผลลัพธ์ที่มากกว่าและแตกต่างอย่างมีนัยสำคัญ

Equal Significant (Equal: E) ค่าเปรียบเทียบระหว่างสองวิธีให้ผลลัพธ์ที่แตกต่างอย่างไม่มีนัยสำคัญ

Lose Significant (Lose: L) ค่าเปรียบเทียบระหว่างสองวิธีให้ผลลัพธ์ที่น้อยและแตกต่างอย่างมีนัยสำคัญ

เช่น ตารางที่ 4.15 ที่แถว RSMOTE1 กับ คอลัมน์ SMOTE มีค่า W/El/L เท่ากับ 3/2/0 แสดงว่าวิธีที่นำเสนอ RSMOTE1 มีผลลัพธ์ที่ดีกว่าและแตกต่างอย่างมีนัยสำคัญอยู่ 3 การทดลอง, ผลลัพธ์ที่

แตกต่างอย่างไม่มีนัยสำคัญอยู่ 2 การทดลองและ ผลลัพธ์ที่น้อยกว่าและแตกต่างอย่างมีนัยสำคัญอยู่ 0 ทดลองจากทั้งหมด 5 การทดลอง เมื่อเทียบกับวิธี SMOTE

ตารางที่ 4.15 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Ionosphere บนมาตรวัดพื้นที่ใต้เส้นโค้ง

%Minority	Variance	%Sampling	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
35.8	35.63	100	**90.6%**	86.2%	86.5%	87.1%	86.3%
		200	**90.6%**	86.6%	86.9%	86.6%	85.7%
		300	**90.6%**	86.5%	86.7%	86.3%	85.3%
		400	**90.6%**	86.1%	85.8%	86.4%	85.0%
		500	**90.6%**	85.2%	85.7%	86.2%	85.5%
Frequency of 1st Rank			5	0	0	0	0

	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
RSMOTE1 (W/E/L)	0/4/1	3/2/0	0/5/0	0/5/0	1/4/0
RSMOTE2 (W/E/L)	0/1/4	1/2/2	0/4/1	0/4/1	0/5/0

จากตารางที่ 4.15 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Ionosphere บนมาตรวัดพื้นที่ใต้เส้นโค้ง ที่เปอร์เซ็นต์สังเคราะห์ข้อมูลที่ 100 - 500 วิธี C4.5 ให้ผลลัพธ์ในการคัดแยกประเภทข้อมูลที่ดีที่สุด คือให้ผลรวมผลลัพธ์ที่ดีที่สุด 5 จากทั้งหมด 5 การทดลอง ในขณะที่วิธีที่นำเสนอ RSMOTE1 ให้ผลลัพธ์รองลงมา และวิธีที่นำเสนอ RSMOTE2 ให้ผลลัพธ์ในการคัดแยกประเภทข้อมูลด้อยที่สุดจากทั้งหมด 5 วิธี

การวัดผลเปรียบเทียบด้วยวิธีการสถิติทดสอบที (t-test) สาเหตุที่ต้องเปรียบเทียบว่าผลลัพธ์ที่ได้แตกต่างอย่างมีนัยสำคัญหรือไม่นั้น เนื่องมาจากการดูเพียงค่าผลรวมผลลัพธ์ที่ดีที่สุดไม่เพียงพอที่จะบอกว่าได้ว่าวิธีที่นำเสนอนั้นสามารถเพิ่มประสิทธิภาพในการคัดแยกประเภทข้อมูลได้จริง อาจจะให้ผลลัพธ์ที่เท่ากันก็ได้ จึงต้องนำมาเปรียบเทียบว่าผลลัพธ์ที่ได้นั้นต้องแตกต่างอย่างมีนัยสำคัญจริง ซึ่งจากตารางที่ 4.15 ส่วนล่าง วิธีที่นำเสนอ RSMOTE1 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญเพียง 3 การทดลอง บนวิธี SMOTE เท่านั้น และให้ประสิทธิภาพต่ำกว่าอย่างมีนัยสำคัญเพียง 1 การทดลอง บนวิธี C4.5 ส่วนวิธีที่นำเสนอ RSMOTE2 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญเพียง 1 การทดลอง บนวิธี SMOTE และให้ประสิทธิภาพต่ำกว่าอย่างมีนัยสำคัญ 4, 2 และ 1 การทดลอง เมื่อเปรียบเทียบกับวิธีการ C4.5, SMOTE และ BSMOTE ตามลำดับ จึงสรุปได้ว่าวิธีที่

นำเสนอ RSMOTE1 สามารถคัดแยกประเภทข้อมูลได้ดีกว่าวิธี SMOTE เท่านั้น แต่คัดแยกได้ด้อยกว่าในวิธี C4.5 ส่วนวิธีที่นำเสนอ RSMOTE2 คัดแยกประเภทข้อมูลได้ด้อยที่สุด

ตารางที่ 4.16 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Pima Indians Diabetes บนมาตรวัดพื้นที่ใต้เส้นโค้ง

%Minority	Variance	%Sampling	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
34.77	741,156	100	71.7%	71.1%	72.9%	**73.2%**	72.5%
		200	71.7%	71.3%	**72.5%**	72.3%	72.0%
		300	71.7%	70.8%	**72.6%**	72.5%	72.3%
		400	71.7%	70.5%	**72.7%**	72.5%	72.0%
		500	71.7%	71.1%	**73.1%**	72.4%	71.6%
Frequency of 1st Rank			0	0	4	1	0

	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
RSMOTE1 (W/E/L)	5/0/0	5/0/0	0/5/0	0/5/0	1/4/0
RSMOTE2 (W/E/L)	4/1/0	2/3/0	0/2/3	0/4/1	0/5/0

จากตารางที่ 4.16 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Pima Indians Diabetes บนมาตรวัดพื้นที่ใต้เส้นโค้ง วิธี BSMOTE ให้ผลลัพธ์ในการคัดแยกประเภทข้อมูลที่ดีที่สุด ส่วนวิธีที่นำเสนอ RSMOTE1 ให้ผลลัพธ์รองลงมา โดยที่ปัญหา Pima Indians Diabetes มีค่าความแปรปรวน (Variance) ของชุดข้อมูลมากกกว่า 1000 แสดงว่าชุดข้อมูลที่นำมาสังเคราะห์ข้อมูลนั้นมีการกระจายตัวของข้อมูลที่สูงมากไม่เหมาะกับวิธีที่นำเสนอทั้ง RSMOTE1 และ RSMOTE2

การวัดผลเปรียบเทียบด้วยวิธีการสถิติทดสอบที (t-test) จากตารางที่ 4.16 ส่วนล่าง วิธีที่นำเสนอ RSMOTE1 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 5 การทดลอง บนวิธี C4.5 และ SMOTE เท่านั้น แต่ให้ประสิทธิภาพพอๆกัน กับวิธี BSMOTE ส่วนวิธีที่นำเสนอ RSMOTE2 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 4 และ2 การทดลอง บนวิธี C4.5 และ SMOTE แต่ให้ประสิทธิภาพต่ำกว่าอย่างมีนัยสำคัญ 3 การทดลอง บนวิธี BSMOTE จึงสรุปได้ว่าวิธีที่นำเสนอทั้งสองวิธีสามารถคัดแยกประเภทข้อมูลได้ดีกว่าวิธี C4.5 และ SMOTE แต่วิธีที่นำเสนอ RSMOTE2 สามารถคัดแยกประเภทข้อมูลได้ด้อยกว่าวิธี BSMOTE

ตารางที่ 4.17 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Seeds บนมาตรวัดพื้นที่ใต้เส้นโค้ง

%Minority	Variance	%Sampling	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
33.33	0.55	100	83.2%	**92.4%**	89.3%	89.7%	89.3%
		200	83.2%	**91.1%**	89.4%	89.6%	89.2%
		300	83.2%	88.8%	89.1%	**89.9%**	88.6%
		400	83.2%	89.5%	88.9%	**89.9%**	87.9%
		500	83.2%	89.6%	89.0%	**89.9%**	88.5%
Frequency of 1st Rank			0	2	0	3	0

	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
RSMOTE1 (W/E/L)	5/0/0	0/4/1	3/2/0	0/5/0	2/3/0
RSMOTE2 (W/E/L)	5/0/0	0/4/1	0/5/0	0/3/2	0/5/0

จากตารางที่ 4.17 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Seeds บนมาตรวัดพื้นที่ใต้เส้นโค้ง วิธีที่นำเสนอ RSMOTE1 ให้ผลรวมผลลัพธ์ในการคัดแยกประเภทข้อมูลที่ดีที่สุด คือ ให้ผลลัพธ์ที่ดีที่สุด 3 การทดลอง จากทั้งหมด 5 การทดลอง ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 300-500 โดยวิธี SMOTE ให้ผลลัพธ์ที่ดีที่สุดเพียง 2 การทดลอง ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 100-200 ส่วนวิธีที่นำเสนอ RSMOTE2 ให้ผลลัพธ์ในการคัดแยกดีกว่าวิธี C4.5 เท่านั้น

การวัดผลเปรียบเทียบด้วยวิธีการสถิติทดสอบที (t-test) จากตารางที่ 4.17 ส่วนล่าง วิธีที่นำเสนอ RSMOTE1 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 5 และ 3 การทดลอง บนวิธี C4.5 และ BSMOTE แต่ให้ประสิทธิภาพต่ำกว่าอย่างมีนัยสำคัญเพียง 1 การทดลอง บนวิธี SMOTE ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 100 ส่วนวิธีที่นำเสนอ RSMOTE2 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 5 การทดลอง บนวิธี C4.5 เท่านั้น แต่ให้ประสิทธิภาพต่ำกว่าอย่างมีนัยสำคัญเพียง 1 การทดลอง บนวิธี SMOTE จึงสรุปได้ว่าวิธีที่นำเสนอ RSMOTE1 สามารถคัดแยกประเภทข้อมูลได้ดีกว่าวิธี C4.5 และ BSMOTE แต่คัดแยกได้ด้อยกว่าในวิธี SMOTE ส่วนวิธีที่นำเสนอ RSMOTE2 สามารถคัดแยกประเภทข้อมูลได้ดีกว่าวิธี C4.5 เท่านั้น แต่คัดแยกได้ด้อยกว่าในวิธี SMOTE

ตารางที่ 4.18 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Vehicle บนมาตรวัดพื้นที่ใต้เส้นโค้ง

%Minority	Variance	%Sampling	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
30.75	637	100	91.6%	92.2%	91.9%	92.5%	**92.6%**
		200	91.6%	91.5%	92.0%	**92.7%**	92.3%
		300	91.6%	92.2%	91.9%	**92.6%**	92.0%
		400	91.6%	91.8%	92.0%	**92.7%**	**92.7%**
		500	91.6%	92.0%	92.1%	**92.7%**	92.3%
Frequency of 1st Rank			0	0	0	4	2

	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
RSMOTE1 (W/E/L)	5/0/0	2/3/0	1/4/0	0/5/0	1/4/0
RSMOTE2 (W/E/L)	3/2/0	2/3/0	2/3/0	0/4/1	0/5/0

จากตารางที่ 4.18 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Vehicle บนมาตรวัดพื้นที่ใต้เส้นโค้ง วิธีที่นำเสนอ RSMOTE1 ให้ผลรวมผลลัพธ์ที่ดีที่สุดมากกว่าวิธีอื่น คือ ให้ผลลัพธ์ที่ดีที่สุด 4 การทดลอง จากทั้งหมด 5 การทดลอง ที่เปอร์เซ็นต์สังเคราะห์ข้อมูลที่ 200-500 ส่วนวิธีที่นำเสนอ RSMOTE2 ให้ผลลัพธ์ที่ดีที่สุด 2 การทดลอง ที่เปอร์เซ็นต์สังเคราะห์ข้อมูลที่ 100 กับ 400

การวัดผลเปรียบเทียบด้วยวิธีการสถิติทดสอบที (t-test) จากตารางที่ 4.18 ส่วนล่าง วิธีที่นำเสนอ RSMOTE1 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 5, 2 และ 1 การทดลอง บนวิธี C4.5, SMOTE และ BSMOTE ตามลำดับ ส่วนวิธีที่นำเสนอ RSMOTE2 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 3, 2 และ 2 การทดลอง บนวิธี C4.5, SMOTE และ BSMOTE ตามลำดับ จึงสรุปได้ว่าวิธีที่นำเสนอทั้งสองวิธีสามารถคัดแยกประเภทข้อมูลได้ดีกว่าวิธีที่เกี่ยวข้องทั้งหมด

ตารางที่ 4.19 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Contraceptive บนมาตรวัดพื้นที่ใต้เส้นโค้ง

%Minority	Variance	%Sampling	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
29.21	10.18	100	42.9%	63.8%	**64.1%**	**64.1%**	**64.1%**
		200	42.9%	64.2%	63.9%	**64.4%**	64.1%
		300	42.9%	62.9%	63.8%	**64.4%**	63.9%
		400	42.9%	62.8%	63.2%	**64.3%**	63.9%
		500	42.9%	62.7%	63.0%	**64.3%**	63.7%
Frequency of 1st Rank			0	0	1	5	1

	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
RSMOTE1 (W/E/L)	5/0/0	3/2/0	2/3/0	0/5/0	0/5/0
RSMOTE2 (W/E/L)	5/0/0	1/4/0	1/4/0	0/5/0	0/5/0

จากตารางที่ 4.19 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Contraceptive บนมาตรพื้นที่ใต้เส้นโค้ง วิธีที่นำเสนอ RSMOTE1 ให้ผลรวมผลลัพธ์ที่ดีที่สุดมากกว่าวิธีอื่น คือ ให้ผลลัพธ์ที่ดีที่สุด 5 การทดลอง จากทั้งหมด 5 การทดลอง ส่วนวิธีที่นำเสนอ RSMOTE2 ให้ผลลัพธ์ที่ดีที่สุดเพียง 1 การทดลอง ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 100 เช่นเดียวกันกับวิธี BSMOTE

การวัดผลเปรียบเทียบด้วยวิธีการสถิติทดสอบที (t-test) จากตารางที่ 4.19 ส่วนล่าง วิธีที่นำเสนอ RSMOTE1 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 5, 3 และ 2 การทดลอง บนวิธี C4.5, SMOTE และ BSMOTE ตามลำดับ ส่วนวิธีที่นำเสนอ RSMOTE2 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 5, 1 และ 1 การทดลอง บนวิธี C4.5, SMOTE และ BSMOTE ตามลำดับ จึงสรุปได้ว่าวิธีที่นำเสนอทั้งสองวิธีสามารถคัดแยกประเภทข้อมูลได้ดีกว่าวิธีที่เกี่ยวข้องทั้งหมด

ตารางที่ 4.20 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Haberman บนมาตรวัดพื้นที่ใต้เส้นโค้ง

%Minority	Variance	%Sampling	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
26.47	300	100	57.2%	65.7%	**66.0%**	63.8%	65.4%
		200	57.2%	65.0%	64.8%	**65.3%**	64.9%
		300	57.2%	62.5%	64.2%	**64.6%**	64.5%
		400	57.2%	63.4%	63.8%	**64.8%**	64.3%
		500	57.2%	63.1%	63.7%	**64.5%**	64.1%
Frequency of 1st Rank			0	0	1	4	0

	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
RSMOTE1 (W/E/L)	5/0/0	1/3/1	1/3/1	0/5/0	0/5/0
RSMOTE2 (W/E/L)	5/0/0	0/5/0	0/5/0	0/5/0	0/5/0

จากตารางที่ 4.20 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Haberman บนมาตรวัดพื้นที่ใต้เส้นโค้ง วิธีที่นำเสนอ RSMOTE1 ให้ผลรวมผลลัพธ์ที่ดีที่สุดมากกว่าวิธีอื่น คือ ให้ผลลัพธ์ที่ดีที่สุด 4 การทดลอง จากทั้งหมด 5 การทดลอง ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 200-500 และ วิธี BSMOTE ให้ผลลัพธ์ที่ดีที่สุดเพียง 1 การทดลอง ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 100

การวัดผลเปรียบเทียบด้วยวิธีการสถิติทดสอบที (t-test) จากตารางที่ 4.20 ส่วนล่าง วิธีที่นำเสนอ RSMOTE1 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 5, 1 และ 1 การทดลอง บนวิธี C4.5, SMOTE และ BSMOTE ตามลำดับ แต่ให้ประสิทธิภาพต่ำกว่าอย่างมีนัยสำคัญเพียง 1 การทดลอง บนวิธี SMOTE และ BSMOTE ส่วนวิธีที่นำเสนอ RSMOTE2 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 5 การทดลอง บนวิธี C4.5 เท่านั้น จึงสรุปได้ว่าวิธีที่นำเสนอทั้งสองวิธีสามารถคัดแยกประเภทข้อมูลได้ดีกว่าวิธี C4.5 และคัดแยกประเภทข้อมูลได้พอๆกัน กับวิธี SMOTE และ BSMOTE

ตารางที่ 4.21 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Breast Tissue บนมาตรวัดพื้นที่ใต้เส้นโค้ง

%Minority	Variance	%Sampling	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
24.71	1.95	100	91.7%	91.5%	**92.5%**	92.2%	91.2%
		200	91.7%	93.0%	93.2%	**93.5%**	**93.5%**
		300	91.7%	93.2%	93.4%	**94.2%**	**94.2%**
		400	91.7%	91.8%	91.3%	94.2%	**94.7%**
		500	91.7%	92.7%	92.5%	94.2%	**94.7%**
Frequency of 1st Rank			0	0	1	2	4

	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
RSMOTE1 (W/E/L)	5/0/0	3/2/0	3/2/0	0/5/0	0/5/0
RSMOTE2 (W/E/L)	4/1/0	2/3/0	2/3/0	0/5/0	0/5/0

จากตารางที่ 4.21 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Breast Tissue บนมาตรวัดพื้นที่ใต้เส้นโค้ง วิธีที่นำเสนอ RSMOTE1 ให้ผลลัพธ์ที่ดีที่สุด 2 การทดลอง จากทั้งหมด 5 การทดลอง ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 200 กับ 300 ส่วนวิธีที่นำเสนอ RSMOTE2 ให้ผลรวมผลลัพธ์ที่ดีที่สุดมากกว่าวิธีอื่น คือ ให้ผลลัพธ์ที่ดีที่สุด 4 การทดลอง จากทั้งหมด 5 การทดลอง ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 200-500 และ วิธี BSMOTE ให้ผลลัพธ์ที่ดีที่สุดเพียง 1 การทดลอง ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 100

การวัดผลเปรียบเทียบด้วยวิธีการสถิติทดสอบที (t-test) จากตารางที่ 4.21 ส่วนล่าง วิธีที่นำเสนอ RSMOTE1 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 5, 3 และ 3 การทดลอง บนวิธี C4.5, SMOTE และ BSMOTE ตามลำดับ ส่วนวิธีที่นำเสนอ RSMOTE2 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 4, 2 และ 2 การทดลอง บนวิธี C4.5, SMOTE และ BSMOTE ตามลำดับ จึงสรุปได้ว่าวิธีที่นำเสนอทั้งสองวิธีสามารถคัดแยกประเภทข้อมูลได้ดีกว่าวิธีที่เกี่ยวข้องทั้งหมด

ตารางที่ 4.22 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Satimage บนมาตรวัดพื้นที่ใต้เส้นโค้ง

%Minority	Variance	%Sampling	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
10.75	640,827	100	75.4%	76.9%	**77.3%**	77.0%	77.0%
		200	75.4%	77.3%	**77.9%**	77.3%	77.4%
		300	75.4%	78.3%	**78.7%**	78.3%	78.2%
		400	75.4%	78.3%	**78.5%**	78.3%	78.1%
		500	75.4%	78.2%	**78.8%**	78.2%	77.9%
Frequency of 1st Rank			0	0	5	0	0

	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
RSMOTE1 (W/E/L)	5/0/0	0/5/0	0/3/2	0/5/0	0/5/0
RSMOTE2 (W/E/L)	5/0/0	0/5/0	0/4/1	0/5/0	0/5/0

จากตารางที่ 4.22 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Satimage บนมาตรวัดพื้นที่ใต้เส้นโค้ง วิธี BSMOTE ให้ผลลัพธ์ในการคัดแยกประเภทข้อมูลที่ดีที่สุด คือ ให้ผลลัพธ์ที่ดีที่สุด 5 การทดลอง จากทั้งหมด 5 การทดลอง ส่วนวิธีที่นำเสนอ RSMOTE1 ให้ผลลัพธ์รองลงมา โดยที่ปัญหา Satimage มีค่าความแปรปรวน (Variance) ของชุดข้อมูลมากกกว่า 1000 แสดงว่าชุดข้อมูลที่นำมาสังเคราะห์ข้อมูลนั้นมีการกระจายตัวของข้อมูลที่สูงมากไม่เหมาะกับวิธีที่นำเสนอทั้ง RSMOTE1 และ RSMOTE2

การวัดผลเปรียบเทียบด้วยวิธีการสถิติทดสอบที (t-test) จากตารางที่ 4.22 ส่วนล่าง วิธีที่นำเสนอ RSMOTE1 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 5 บนวิธี C4.5 เท่านั้น แต่ให้ประสิทธิภาพต่ำกว่าอย่างมีนัยสำคัญ 2 การทดลอง บนวิธี BSMOTE ส่วนวิธีที่นำเสนอ RSMOTE2 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 5 บนวิธี C4.5 แต่ให้ประสิทธิภาพต่ำกว่าอย่างมีนัยสำคัญเพียง 1 การทดลอง บนวิธี BSMOTE จึงสรุปได้ว่าวิธีที่นำเสนอทั้งสองวิธีสามารถคัดแยกประเภทข้อมูลได้ดีกว่าวิธี C4.5 และคัดแยกประเภทข้อมูลได้พอๆกัน กับวิธี SMOTE แต่คัดแยกประเภทข้อมูลได้ด้อยกว่าวิธี BSMOTE

ตารางที่ 4.23 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Glass_3 บนมาตรวัดพื้นที่ใต้เส้นโค้ง

%Minority	Variance	%Sampling	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
8.63	11.29	100	**76.3%**	73.5%	72.1%	69.3%	72.3%
		200	76.3%	75.5%	72.9%	77.1%	**78.6%**
		300	76.3%	75.8%	71.6%	**78.3%**	**78.3%**
		400	76.3%	76.5%	70.3%	78.0%	**79.0%**
		500	76.3%	75.2%	69.2%	76.6%	**77.8%**
Frequency of 1st Rank			1	0	0	1	4

	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
RSMOTE1 (W/E/L)	4/0/1	2/3/0	4/1/0	0/5/0	0/4/1
RSMOTE2 (W/E/L)	4/0/1	1/4/0	4/1/0	1/4/0	0/5/0

จากตารางที่ 4.23 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Glass_3 บนมาตรวัดค่าพื้นที่ใต้เส้นโค้ง วิธีที่นำเสนอ RSMOTE2 ให้ผลรวมผลลัพธ์ที่ดีที่สุดมากกว่าวิธีอื่น คือ ให้ผลลัพธ์ที่ดีที่สุด 4 การทดลอง จากทั้งหมด 5 การทดลอง ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 200-500 ส่วนวิธีที่นำเสนอ RSMOTE1 ให้ผลลัพธ์ที่ดีที่สุดเพียง 1 การทดลอง ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 300 และ วิธี C4.5 ให้ผลลัพธ์ที่ดีที่สุดเพียง 1 การทดลอง ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 100

การวัดผลเปรียบเทียบด้วยวิธีการสถิติทดสอบที (t-test) จากตารางที่ 4.23 ส่วนล่าง วิธีที่นำเสนอ RSMOTE1 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 4, 2 และ 4 การทดลอง บนวิธี C4.5, SMOTE และ BSMOTE ตามลำดับ ส่วนวิธีที่นำเสนอ RSMOTE2 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 4, 1 และ 4 การทดลอง บนวิธี C4.5, SMOTE และ BSMOTE ตามลำดับ จึงสรุปได้ว่าวิธีที่นำเสนอทั้งสองวิธีสามารถคัดแยกประเภทข้อมูลได้ดีกว่าวิธีที่เกี่ยวข้องทั้งหมด

ตารางที่ 4.24 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Ecoli_om บนมาตรวัดพื้นที่ใต้เส้นโค้ง

%Minority	Variance	%Sampling	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
6.32	7.65	100	74.3%	86.0%	87.4%	88.5%	**89.8%**
		200	74.3%	87.3%	84.9%	90.2%	**90.6%**
		300	74. 3%	89.8%	85.7%	90.2%	**90.4%**
		400	74.3%	89.9%	83.4%	**90.7%**	**90.7%**
		500	74.3%	90.0%	82.8%	**90.7%**	90.5%
Frequency of 1st Rank			0	0	0	2	5

	C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
RSMOTE1 (W/E/L)	5/0/0	2/3/0	4/1/0	0/5/0	0/4/1
RSMOTE2 (W/E/L)	5/0/0	2/3/0	5/0/0	1/4/0	0/5/0

จากตารางที่ 4.24 ผลการเปรียบเทียบการจำแนกประเภทข้อมูลของปัญหา Ecoli_om บนมาตรวัดค่าพื้นที่ใต้เส้นโค้ง วิธีที่นำเสนอ RSMOTE1 ให้ผลลัพธ์ที่ดีที่สุด 2 การทดลอง จากทั้งหมด 5 การทดลอง ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 400 กับ 500 ส่วนวิธีที่นำเสนอ RSMOTE2 ให้ผลรวมผลลัพธ์ที่ดีที่สุดมากกว่าวิธีอื่น คือ ให้ผลลัพธ์ที่ดีที่สุด 4 การทดลอง จากทั้งหมด 5 การทดลอง ที่เปอร์เซ็นต์การสังเคราะห์ข้อมูลที่ 100-400

การวัดผลเปรียบเทียบด้วยวิธีการสถิติทดสอบที (t-test) จากตารางที่ 4.24 ส่วนล่าง วิธีที่นำเสนอ RSMOTE1 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 5, 2 และ 4 การทดลอง บนวิธี C4.5, SMOTE และ BSMOTE ตามลำดับ ส่วนวิธีที่นำเสนอ RSMOTE2 ให้ประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ 5, 2 และ 5 การทดลอง บนวิธี C4.5, SMOTE และ BSMOTE ตามลำดับ จึงสรุปได้ว่าวิธีที่นำเสนอทั้งสองวิธีสามารถคัดแยกประเภทข้อมูลได้ดีกว่าวิธีที่เกี่ยวข้องทั้งหมด

4.3 การวิเคราะห์ผลการทดลอง

จากตารางที่ 4.25 และ 4.26 เป็นตารางที่รวมค่า Frequency 1st rank ของแต่ละวิธี ทั้งวิธีเกี่ยวข้องและวิธีที่นำเสนอ ว่าให้ผลการเปรียบเทียบการจำแนกประเภทข้อมูลเพียงใดบนชุดข้อมูล เช่น ตารางที่ 4.25 สรุปผลการทดลอง Frequency 1st rank บนมาตรวัดเอฟ แถว Ionosphere คอลัมน์ C4.5 มีค่า Frequency 1st rank เท่ากับ 5 แสดงว่า วิธี C4.5 ให้ผลลัพธ์ในการคัดแยกประเภทข้อมูลดีที่สุด 5 การทดลองจากทั้งหมด 5 การทดลอง (จากการทดลองที่100, 200, 300, 400 และ 500 %Sampling) สามารถดูข้อมูลอย่างละเอียดได้ที่ตาราง 4.5 จากข้อมูลสรุปผลการทดลองเห็นได้ว่าค่ามาตรวัดเอฟและพื้นที่ใต้เส้นโค้งของวิธีการนำเสนอ RSMOTE1 กับ RMOTE 2 ให้ผลรวมผลลัพธ์ในการคัดแยกประเภทข้อมูลได้ดีที่สุดเป็นจำนวน 26 การทดลองจากทั้งหมด 50 การทดลอง และ 11 การทดลองจากทั้งหมด 50 การทดลองบนมาตรวัดเอฟ ส่วนมาตรวัดพื้นที่ใต้เส้นโค้งให้ผลรวมผลลัพธ์ในการคัดแยกประเภทข้อมูลได้ดีที่สุดเป็นจำนวน 22 การทดลองจากทั้งหมด 50 การทดลอง และ 16 การทดลองจากทั้งหมด 50 การทดลอง ตามลำดับ ในขณะที่วิธีสโมทให้ผลรวมผลลัพธ์ดีที่สุดเพียง 2 การทดลองบนมาตรวัดค่าเอฟ และ 2 การทดลองบนมาตรวัดพื้นที่ใต้เส้นโค้ง และวิธี BSMOTE ให้ผลรวมผลลัพธ์ดีที่สุด 11 การทดลองบนมาตรวัดค่าเอฟ และ 12 การทดลองบนมาตรวัดพื้นที่ใต้เส้นโค้ง อีกทั้งยังแนบวิธีการที่ไม่มีการสร้างตัวอย่างสังเคราะห์เพิ่มแก่ข้อมูลต้นฉบับ (C4.5) เพื่อใช้เปรียบเทียบกับวิธีการต่างๆว่าสามารถเพิ่มประสิทธิภาพได้ปริมาณเท่าใด

จุดเด่นของวิธีที่นำเสนอ ในกรณีที่ตัวอย่างข้อมูลประเภทเล็กมีจำนวนน้อยๆ อาทิ เช่น มีจำนวนน้อยกว่า 100 ตัวอย่าง วิธีการที่นำเสนอจะให้ประสิทธิภาพสูงกว่าวิธีการอื่นในมาตรวัดของ ค่าเอฟ และ พื้นที่ใต้เส้นโค้ง แต่ในกรณีที่ชุดข้อมูลที่มีตัวอย่างข้อมูลประเภทเล็กเป็นจำนวนมาก วิธี BSMOTE จะให้ผลลัพธ์ที่มีประสิทธิภาพสูงกว่า สาเหตุที่ทำให้วิธีที่นำเสนอมีประสิทธิภาพที่สูงกว่าวิธีการอื่นๆ ในชุดข้อมูลที่มีตัวอย่างข้อมูลประเภทเล็กเป็นจำนวนน้อยๆ เนื่องมาจากวิธีที่นำเสนอนั้นมุ่งเน้นที่จะช่วยเหลือตัวอย่างข้อมูลประเภทเล็กที่คัดแยกประเภทข้อมูลไม่ถูกต้อง ให้สามารถคัดแยกประเภทข้อมูลให้ถูกต้องมากขึ้น ซึ่งแตกต่างจากวิธี BSMOTE ที่มุ่งเน้นสร้างข้อมูลสังเคราะห์ประเภทข้อมูลเล็กบริเวณใกล้ขอบ ในกรณีที่ชุดข้อมูลมีจำนวนข้อมูลประเภทเล็กปริมาณน้อยมากๆ แล้วนั้น อาจทำให้ข้อมูลประเภทเล็กแทบทั้งหมดถูกระบุว่าเป็นข้อมูลบริเวณขอบ ทำให้วิธีการที่นำเสนอนั้นมีประสิทธิภาพสูงกว่าวิธีการอื่นๆ ในชุดข้อมูล Breast Tissue, Glass_3, และ Ecoli_om ที่มีข้อมูลประเภทเล็ก 21, 17, และ 20 ตัวอย่าง อย่างมีนัยสำคัญด้วยการทดสอบที (t-test) จึงทำให้เชื่อได้ว่าวิธีการที่นำเสนอเหมาะสำหรับชุดข้อมูลที่ไม่สมดุลที่มีตัวอย่างข้อมูลประเภทเล็กเป็นจำนวนน้อยๆ คือมีจำนวนข้อมูลประเภทเล็กไม่เกิน 333 ข้อมูล และมีเปอร์เซ็นต์ข้อมูลประเภทเล็กไม่เกิน 33.33 เปอร์เซ็นต์ อีกทั้งต้องมีค่า Varience ของชุดข้อมูลน้อยกว่า 1000

สาเหตุหลักที่วิธีการที่นำเสนอมีประสิทธิภาพสูงกว่าวิธีการ BSMOTE ในกรณีที่มีประเภทข้อมูลเล็กมีปริมาณน้อยๆ เพราะวิธีการที่นำเสนอนั้นมีการปรับให้เป็นปรกติ (normalization) ให้ตัวอย่างข้อมูล ดังแสดงในสมการที่ 3.7 และ 3.8 ในบทที่ 3 เพื่อให้ค่าน้ำหนักรวมของประเภทข้อมูลประเภทเล็กเท่ากับข้อมูลประเภทใหญ่ แตกต่างกับวิธีการ BSMOTE ให้ค่านำหนักของแต่ละตัวอย่างเท่าๆกัน ดังนั้นในกรณีที่ข้อมูลมีปริมาณน้อยๆ จะทำให้วิธีการที่นำเสนอรองรับกับข้อมูลที่มีปริมาณน้อยได้ แต่ในทางกลับกันอาจทำให้วิธีการ BSMOTE ระบุถึงประเภทข้อมูลได้ผิดพลาดอาทิ เช่น ระบุข้อมูลที่เป็นแกนข้อมูลว่าเป็นข้อมูลบริเวณขอบ และข้อมูลบริเวณขอบเป็นข้อมูลรบกวน (noise) เป็นต้น

การวัดผลเปรียบเทียบด้วยวิธีการสถิติทดสอบที (t-test) ในตารางที่ 4.27 บนมาตรวัดเอฟ วิธีการที่นำเสนอ RSMOTE1 นั้นมีประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ เมื่อเปรียบเทียบกับวิธีการ C4.5, SMOTE และ BSMOTE 44, 21, และ 21 การทดลองตามลำดับ และมีประสิทธิภาพต่ำกว่าเพียง 2, 4, และ 5 การทดลองตามลำดับ เท่านั้น

การวัดผลเปรียบเทียบด้วยวิธีการสถิติทดสอบที ในตารางที่ 4.28 บนมาตรวัดพื้นที่ใต้เส้นโค้ง วิธีการที่นำเสนอ RSMOTE1 นั้นมีประสิทธิภาพสูงกว่าอย่างมีนัยสำคัญ เมื่อเปรียบเทียบกับวิธีการ C4.5, SMOTE และ BSMOTE 44, 21, และ 18 การทดลองตามลำดับ และมีประสิทธิภาพต่ำกว่าเพียง 2, 2, และ 3 การทดลองตามลำดับ เท่านั้น

จากผลการทดลองที่กล่าวมาแล้วนั้นส่งผลให้เชื่อได้ว่าวิธีการที่นำเสนอเหมาะสำหรับข้อมูลที่ไม่สมดุลที่มีปริมาณข้อมูลประเภทเล็กจำนวนน้อยๆ และให้ผลเทียบเคียงกับวิธีการอื่นๆในข้อมูลที่มีข้อมูลประเภทเล็กมากเพียงพอ แต่ให้ผลที่ต่ำกว่าในชุดข้อมูลที่มีค่าความแปรปรวนมากกว่า 1000 เช่น ชุดข้อมูล Pima และ Satimage

การกำหนดค่าน้อยที่สุดที่ยอมรับได้ (Threshold) โดยใช้ค่าเฉลี่ยเป็นวิธี RSMOTE1 หรือค่ามัธยฐานเป็นวิธี RSMOTE2 จะให้ค่าเทียบเคียงกันทั้งบนมาตรวัดค่าเอฟและพื้นที่ใต้เส้นโค้ง ดังแสดงในตาราง 4.29 RSMOTE1 ให้ผลลัพธ์รวมเทียบเคียงกับ RSMOTE2 ให้ค่าแตกต่างอย่างไม่มีนัยสำคัญ (Equal Significant) 44 และ 43 การทดลอง ตามลำดับ จากทั้งหมด 50 การทดลอง แต่วิธี RSMOTE1 ใช้เวลาในการทดลองน้อยกว่า วิธีRSMOTE2 เนื่องจากวิธี RSMOTE2 ต้องนำค่าความรู้ของตัวอย่างสังเคราะห์ทั้งหมดมาเรียงเพื่อหาค่ามัธยฐาน ในขณะที่วิธี RSMOTE1 ใช้การหาค่าเฉลี่ยของค่าความรู้ ไม่ต้องนำค่าความรู้ของตัวอย่างสังเคราะห์ทั้งหมดมาเรียง ทำให้ใช้เวลาในการทดลองน้อยกว่า จึงบอกได้ว่าวิธี RSMOTE1 ดีกว่า วิธี RSMOTE2 อีกทั้งยังให้ประสิทธิภาพด้อยกว่าอย่างมีนัยสำคัญเพียง 1 การทดลองจาก 50 การทดลองบนมาตรวัดเอฟ และ 2 การทดลองจาก 50 การทดลองบนมาตรวัดพื้นที่ใต้เส้นโค้ง ซึ่งเป็นผลการทดลองที่ได้จากชุดข้อมูล 10 ชุดที่นำมาทดลอง

ตารางที่ 4.25 สรุปผลการทดลอง Frequency of 1st Rank บนมาตรวัดเอฟ

#	Data set	%Minority	#Minority	Frequency of 1st Rank				
				C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
1	Ionosphere	35.8	126	5	0	0	1	0
2	Pima	34.77	268	0	0	4	1	0
3	Seeds	33.33	70	0	1	0	4	0
4	Vehicle	30.75	199	0	0	0	5	1
5	Contraceptive	29.21	333	0	0	1	4	0
6	Haberman	26.47	81	0	0	1	4	1
7	Breast Tissue	24.71	21	0	0	1	2	4
8	Satimage	10.75	625	0	1	3	0	1
9	Glass_3	8.63	17	1	0	0	1	4
10	Ecoli_om	6.32	20	0	0	1	4	0
Total				6	2	11	26	11

ตารางที่ 4.26 สรุปผลการทดลอง Frequency of 1st Rank บนมาตรวัดพื้นที่ใต้เส้นโค้ง

#	Data set	%Minority	#Minority	Frequency of 1st Rank				
				C4.5	SMOTE	BSMOTE	RSMOTE1	RSMOTE2
1	Ionosphere	35.8	126	5	0	0	0	0
2	Pima	34.77	268	0	0	4	1	0
3	Seeds	33.33	70	0	2	0	3	0
4	Vehicle	30.75	199	0	0	0	4	2
5	Contraceptive	29.21	333	0	0	1	5	1
6	Haberman	26.47	81	0	0	1	4	0
7	Breast Tissue	24.71	21	0	0	1	2	4
8	Satimage	10.75	625	0	0	5	0	0
9	Glass_3	8.63	17	0	1	0	1	4
10	Ecoli_om	6.32	20	0	0	0	2	5
Total				5	3	12	22	16

<u>หมายเหตุ</u> Total มีค่ารวมเกินกว่า 50 เนื่องมาจากที่เปอร์เซ็นต์สังเคราะห์ข้อมูลเดียวกันแล้ว อาจมีวิธีมากกว่า 1 วิธีที่ให้ผลลัพธ์การคัดแยกข้อมูลที่ดีที่สุดหรือให้ผลลัพธ์ที่ดีที่สุดเท่ากัน

ตารางที่ 4.27 สรุปผลการทดลอง Win/Equal/Lose Significant บนมาตรวัดเอฟ

#	Data set	%Minority	#Minority	Variance	RSMOTE1			RSMOTE2		
					C4.5	SMOTE	BSMOTE	C4.5	SMOTE	BSMOTE
1	Ionosphere	35.8	126	35.63	0/4/1	1/4/0	0/5/0	0/1/4	0/3/2	0/2/3
2	Pima	34.77	268	741,156	5/0/0	4/1/0	0/4/1	4/1/0	3/2/0	0/4/1
3	Seeds	33.33	70	0.55	5/0/0	1/3/1	4/1/0	5/0/0	0/3/2	0/5/0
4	Vehicle	30.75	199	637	5/0/0	3/2/0	3/2/0	3/2/0	1/3/1	0/5/0
5	Contraceptive	29.21	333	10.18	5/0/0	3/2/0	2/3/0	5/0/0	3/2/0	1/4/0
6	Haberman	26.47	81	300	5/0/0	2/2/1	1/3/1	5/0/0	1/4/0	0/5/0
7	Breast Tissue	24.71	21	1.95	5/0/0	3/2/0	3/2/0	4/1/0	3/2/0	2/3/0
8	Satimage	10.75	625	640,827	5/0/0	0/4/1	0/3/2	5/0/0	0/5/0	0/5/0
9	Glass_3	8.63	17	11.29	4/0/1	3/1/1	4/1/0	4/0/1	4/1/0	4/1/0
10	Ecoli_om	6.32	20	7.65	5/0/0	1/4/0	4/0/1	5/0/0	1/4/0	4/0/1
Win/Equal/Lose					44/4/2	21/25/4	21/24/5	40/5/5	16/29/5	11/34/5
Frequency of 1st rank					48	46	45	45	45	45
Total					50	50	50	50	50	50

ตารางที่ 4.28 สรุปผลการทดลอง Win/Equal/Lose Significant บนมาตรวัดพื้นที่ใต้เส้นโค้ง

#	Data set	%Minority	#Minority	Variance	RSMOTE1			RSMOTE2		
					C4.5	SMOTE	BSMOTE	C4.5	SMOTE	BSMOTE
1	Ionosphere	35.8	126	35.63	0/4/1	3/2/0	0/5/0	0/1/4	1/2/2	0/4/1
2	Pima	34.77	268	741,156	5/0/0	5/0/0	0/5/0	4/1/0	2/3/0	0/2/3
3	Seeds	33.33	70	0.55	5/0/0	0/4/1	3/2/0	5/0/0	0/4/1	0/5/0
4	Vehicle	30.75	199	637	5/0/0	2/3/0	1/4/0	3/2/0	2/3/0	2/3/0
5	Contraceptive	29.21	333	10.18	5/0/0	3/2/0	2/3/0	5/0/0	1/4/0	1/4/0
6	Haberman	26.47	81	300	5/0/0	1/3/1	1/3/1	5/0/0	0/5/0	0/5/0
7	Breast Tissue	24.71	21	1.95	5/0/0	3/2/0	3/2/0	4/1/0	2/3/0	2/3/0
8	Satimage	10.75	625	640,827	5/0/0	0/5/0	0/3/2	5/0/0	0/5/0	0/4/1
9	Glass_3	8.63	17	11.29	4/0/1	2/3/0	4/1/0	4/0/1	1/4/0	4/1/0
10	Ecoli_om	6.32	20	7.65	5/0/0	2/3/0	4/1/0	5/0/0	2/3/0	5/0/0
Win/Equal/Lose					44/4/2	21/27/2	18/29/3	40/5/5	10/37/3	14/31/5
Frequency of 1st rank					48	48	47	45	47	45
Total					50	50	50	50	50	50

ตารางที่ 4.29 ผลการเปรียบเทียบ Win/Equal/Lose Significant ของวิธี RSMOTE1 กับ RSMOTE2

				RSMOTE1 : RSMOTE2	
#	Data set	%Minority	#Minority	F-Measure	AUC
1	Ionosphere	35.8	126	1/4/0	1/4/0
2	Pima	34.77	268	0/5/0	1/4/0
3	Seeds	33.33	70	2/3/0	2/3/0
4	Vehicle	30.75	199	2/3/0	1/4/0
5	Contraceptive	29.21	333	0/5/0	0/5/0
6	Haberman	26.47	81	0/5/0	0/5/0
7	Breast Tissue	24.71	21	0/5/0	0/5/0
8	Satimage	10.75	625	0/5/0	0/5/0
9	Glass_3	8.63	17	0/4/1	0/4/1
10	Ecoli_om	6.32	20	0/5/0	0/4/1
Win/Equal/Lose				5/44/1	5/43/2
Total				50	50

บทที่ 5
สรุปและข้อเสนอแนะ

5.1 สรุป

ในงานวิจัยชิ้นนี้นำเสนอวิธีการเพิ่มตัวอย่างข้อมูลโดยการสังเคราะห์ข้อมูลประเภทเล็ก โดยพิจารณาจากค่าความรู้ตัวอย่างข้อมูลสังเคราะห์ที่มีค่าความรู้สูงๆ แสดงถึงตัวอย่างสังเคราะห์ที่ช่วยให้ข้อมูลประเภทเล็กที่คัดแยกผิด ให้สามารถคัดแยกประเภทข้อมูลได้อย่างถูกต้อง ในการเตรียมการทดลองนั้นได้ใช้ข้อมูลทดสอบจากฐานข้อมูล UCI 10 ชุดข้อมูล ทดสอบเปรียบเทียบกับวิธีการ SMOTE และ BSMOTE โดยใช้มาตรวัดเอฟ และพื้นที่ใต้เส้นโค้ง ผลการทดลองแสดงให้เห็นว่าวิธีการที่นำเสนอ RSMOTE1 นั้นสามารถช่วยให้คัดแยกข้อมูลที่ไม่สมดุลให้มีประสิทธิภาพสูงขึ้นบนมาตรวัดเอฟ และค่าพื้นที่ใต้โค้ง และมีประสิทธิภาพสูงกว่าวิธีการที่เปรียบเทียบอย่างมีนัยสำคัญ 86 การทดลองจาก 150 การทดลองบนมาตรวัดเอฟ ซึ่งประกอบด้วย 44 การทดลองเมื่อเปรียบเทียบกับวิธี C4.5, 21 การทดลองเมื่อเปรียบเทียบกับวิธี SMOTE และ21 การทดลองเมื่อเปรียบเทียบกับวิธี BSMOTE และ 83 การทดลองจาก 150 การทดลองบนค่าพื้นที่ใต้เส้นโค้ง ซึ่งประกอบด้วย 44 การทดลองเมื่อเปรียบเทียบกับวิธี C4.5, 21 การทดลองเมื่อเปรียบเทียบกับวิธี SMOTE และ18 การทดลองเมื่อเปรียบเทียบกับวิธี BSMOTE ทั้งยังให้ประสิทธิภาพที่ดีในกรณีข้อมูลมีปริมาณจำนวนข้อมูลประเภทเล็กเป็นจำนวนน้อยได้อย่างประสิทธิภาพที่ดีกว่าเมื่อเปรียบเทียบกับวิธีการอื่นๆ จึงสรุปได้ว่าวิธีการนี้ได้เสนอวิธีการสร้างตัวอย่างข้อมูลสังเคราะห์เพื่อแก้ไขปัญหาความไม่สมดุลของข้อมูล ได้อย่างมีประสิทธิภาพ และมีประสิทธิภาพสูงกว่าวิธีการอื่นๆอย่างมีนัยสำคัญทางสถิติ ยกเว้นชุดข้อมูล Pima กับ Satimage ที่มี Variance เกิน 1000 วิธี BSMOTE จะให้ประสิทธิภาพในการคัดแยกประเภทข้อมูลดีกว่า

5.2 ข้อเสนอแนะ

1. วิธีการที่นำเสนอนั้นยังรองรับเพียงข้อมูลที่มีสองประเภทข้อมูล ควรปรับปรุงวิธีการที่นำเสนอให้สามารถใช้ได้กับชุดข้อมูลที่มีประเภทข้อมูลมากกว่า 2 ประเภท (multiclass) โดยไม่ต้องมาแปลงให้ชุดข้อมูลเหลือเพียงสองประเภทข้อมูล (binary classification)
2. การหาค่าน้อยที่สุดที่ยอมรับได้ (Threshold) ในการเลือกตัวอย่างสังเคราะห์ที่เหมาะสม ซึ่งในงานวิจัยนี้ได้ใช้ค่าน้อยที่สุดที่ยอมรับเป็นค่าเฉลี่ย และค่ามัธยฐานของค่าความรู้
3. การหาจำนวนเพื่อนบ้านที่ใกล้ที่สุด (k-NN) เพื่อใช้ในการสังเคราะห์ข้อมูลประเภทเล็ก โดยในงานวิจัยนี้ใช้ค่า k = 5 ตามวิธีสโมท (SMOTE) และการสร้างข้อมูลสังเคราะห์บริเวณขอบ (BSMOTE)

เอกสารอ้างอิง

[1] Bradley, A. (1997). The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. **Pattern Recognition 30**. 1997. vol. 6 pp. 1145–1159

[2] Buckland, M. and Gey, F. The Relationship between Recall and Precision. **Journal of the American Society for Information Science 1994.** vol. 45(1) pp. 12–19.

[3] Chawla, N., Bowyer, K., Hall, L. and Kegelmeyer, W. SMOTE: Synthetic Minority Over-Sampling Technique. **Journal of Artificial Intelligence Research 16**. May 2002. pp. 321–357.

[4] David D. Lewis and Jason Catlett. Heterogeneous Uncertainty Sampling for Supervised Learning. **Proceedings of the 11th International Conference on Machine Learning ICML'94** vol.11 pp. 148-156.

[5] Han, H., Wang, W. and Mao. B. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In: Huang, D.-S., Zhang, X.-P., Huang, G.-B. (eds.) **ICIC 2005.** LNCS vol. 3644 pp. 878–887.

[6] Kubat, m., Holte, R., and Matwin, S. Machine Learning for the Detection of Oil Spills in Satellite Radar Images. **Machine Learning 1998.** vol. 30 pp.195–215

[7] Nitesh V.Chawla, Nathalie Japkowicz and Aleksander Kolcz. Editorial: Special Issue on Learning from Imbalanced Data Sets. **SIGKDD Explorations 6**. 2004 vol.1 pp. 1-6

[8] Nitesh V.Chawla. Data Mining for Imbalanced Datasets: An Overview doi:10.1007/978-0-387-09823-4_45 In: Maimon, Oded; Rokach, Lior (Eds) Data Mining and Knowledge Discovery Handbook. **Springer 2010.** ISBN 978-0-387-09823-4 pp. 875-886

[9] Xiannian Fan, Ke Tang and Thomas Weise Margin-Based Over-Sampling Method for Learning From Imbalanced Datasets. **Proceedings of the 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining 2011 (PAKDD'11)**

[10] Yang, Q. and Wu, X. 10 challenging problems in data mining research. **International Journal of Information Technology and Decision Making 2006.** vol. 5(4) pp. 597–604 http://dblp.unitrier.de/db/journals /ijitdm/ijitdm5.html

[11] ZhaohuiZheng, Xiaoyun Wu and Rohini Srihari. Feature Selection for Text Categorization on Imbalanced Data. **SIGKDD Explorations 2004**. vol. 6(1) pp. 80-89

ภาคผนวก

ผลงานที่ได้รับตีพิมพ์

รายงานสืบเนื่องจากการประชุมวิชาการระดับชาติ

Proceedings of the 5th Science Research Conference

วิทยาศาสตร์วิจัย ครั้งที่ 5

4-5 มีนาคม 2556

The 5th Science RESEARCH Conference

ณ อาคารเทคโนโลยีสารสนเทศและการสื่อสาร มหาวิทยาลัยพะเยา

ISBN : 978-616-91198-8-3

วิธีการเพิ่มจำนวนข้อมูลโดยใช้การย้อนกลับไปดูเพื่อนบ้านเพื่อเพิ่มข้อมูลส่วนน้อย

Reverse K - NN SMOTE: Over - Sampling Method to add The Synthetic Minority Instances

กัลยา พันธุมะผล[1,*] และ นวลสวาท หิรัญสกลวงศ์[1]
Kalaya Phuntumapol[1,*] and Nualsawat Hiransakolwong[1]

[1]ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าเจ้าคุณทหาร
ถนนฉลองกรุง เขตลาดกระบัง กรุงเทพฯ 10520
โทร 0-2329-8400 ถึง 8411 โทรสาร 0-2329-8412 E-mail: kalaya.tm@gmail.com

บทคัดย่อ

ข้อมูลไม่สมดุล (Imbalanceddataset) คือ ข้อมูลที่มีการกระจายตัวที่ไม่เท่ากันในแต่ละประเภทข้อมูล (class) กล่าวคือ อัตราส่วนของประเภทข้อมูลโน้มเอียงไปยังประเภทข้อมูลประเภทใดประเภทหนึ่ง ปัจจุบันเราสามารถพบข้อมูลไม่สมดุลในงานจริงจำนวนมาก ในงานวิจัยชิ้นนี้เรามุ่งเน้นเฉพาะการคัดแยกข้อมูลแบบ 2 ประเภท (binaryclassification) ซึ่งประกอบด้วยประเภทข้อมูลเล็ก (minorityclass) และประเภทข้อมูลใหญ่ (majorityclass) โดยกำหนดให้ประเภทข้อมูลเล็กคือ ประเภทข้อมูลที่มีอัตราส่วนข้อมูลจำนวนน้อยๆ และประเภทข้อมูลใหญ่คือ ประเภทข้อมูลที่มีอัตราส่วนข้อมูลจำนวนมากๆ อาทิ เช่น ในข้อมูลใดๆ อาจประกอบด้วย ประเภทข้อมูลเล็ก 0.5% และประเภทข้อมูลใหญ่ 95.5% เป็นต้น ปัญหาหลักของข้อมูลไม่สมดุลคือโมเดลการคัดแยกข้อมูลจะโน้มเอียงไปยังประเภทข้อมูลใหญ่ ส่งผลให้คัดแยกข้อมูลทั้งหมดเป็นประเภทข้อมูลใหญ่ ทั้งที่ในความเป็นจริงแล้วประกอบด้วยประเภทข้อมูลเล็กด้วย วิธีที่ใช้แก้ปัญหาข้อมูลไม่สมดุลก็คือ ปรับการกระจายตัวของประเภทข้อมูล โดยการสร้างข้อมูลสังเคราะห์ส่วนน้อย (syntheticminoritydata) ให้อัตราส่วนข้อมูลส่วนน้อยใกล้เคียงกับข้อมูลส่วนมาก ในการวิจัยชิ้นนี้นำเสนอวิธีการเพิ่มจำนวนข้อมูลโดยใช้วิธีย้อนกลับไปดูเพื่อนบ้านในการเลือกข้อมูลส่วนน้อยที่จะเพิ่มเข้าไปตามค่าความรู้ (informationvalue) ในการทดลองได้นำวิธีการที่นำเสนอวัดผลกับวิธีการสร้างข้อมูลสังเคราะห์ส่วนน้อยประเภทอื่นๆ อาทิ เช่น สโมท (SMOTE) และ การสร้างข้อมูลสังเคราะห์บริเวณขอบ (Borderline-SMOTE) จากผลการทดลองแสดงให้เห็นว่าวิธีการที่นำเสนอสามารถเพิ่มประสิทธิภาพการคัดแยกข้อมูลได้ดีกับชุดข้อมูลส่วนใหญ่ เมื่อใช้มาตรวัดค่าเอฟ (F-Measure) และพื้นที่ใต้กราฟ (AUC)

คำสำคัญ: ปัญหาจำนวนข้อมูลไม่เท่ากัน/ การเพิ่มจำนวนข้อมูล/ SMOTE

บทนำ

ปัญหาข้อมูลที่ไม่สมดุล [1] นั้นในงานวิจัยนี้จะพิจารณาเพียงสองประเภทเท่านั้น คือ ประเภทข้อมูลเล็ก กับประเภทข้อมูลใหญ่ ประเภทข้อมูลเล็กมีจำนวนข้อมูลเป็นอัตราส่วนจำนวนน้อย ในขณะที่ประเภทข้อมูลใหญ่นั้นกับมีอัตราส่วนข้อมูลมาก อาทิ เช่น ในข้อมูลใดๆ อาจประกอบด้วย ประเภทข้อมูลเล็ก 0.5% และ ประเภทข้อมูลใหญ่ 95.5% เป็นต้น ในการวัดประสิทธิภาพโมเดลการคัดแยกประเภทข้อมูลนั้นจะใช้ค่าความแม่นยำ (accuracy) ซึ่งค่าความแม่นยำนั้นจะโน้มเอียงไปสู่ประเภทข้อมูลใหญ่ สาเหตุจากประเภทข้อมูลใหญ่นั้นจะมีปริมาณข้อมูลจำนวนมาก ยังผลให้โมเดลที่สร้างขึ้นจากข้อมูลที่ไม่สมดุลนั้นจะจำแนกข้อมูลทั้งหมดเป็นประเภทข้อมูลใหญ่ทั้งหมด ซึ่งปัญหาเหล่านี้สามารถพบได้ในข้อมูลจริง อาทิ เช่น การตรวจสอบการรั่วไหลของน้ำมันจากภาพเรดาร์ของดาวเทียม [2], การคัดแยกประเภทข้อความ (text classification) [3], การสืบค้นสารสนเทศ (information retrieval), การคัดกรองงาน (data filtering) [4] และอื่นๆ ยังผลให้ข้อมูลที่ไม่สมดุลนั้นเป็นปัญญาที่สำคัญมากในการทำเหมืองข้อมูล (data mining) [5]

เนื่องมาจากค่าความแม่นยำและอัตราส่วนความผิดพลาด (error rate) มีค่าโน้มเอียงตามข้อมูลส่วนมาก ดังนั้นงานวิจัยฉบับนี้จึงใช้มาตรวัดพื้นที่ใต้เส้นโค้ง (Area Under Curve, AUC) [6] ของกราฟรูปแบบของตัวรับสัญญาณ (Receiver Operator Characteristic, ROC) [6] ซึ่งรูปแบบของตัวรับสัญญาณเป็นกราฟที่ใช้ในการแสดงผลระหว่าง อัตราความถูกต้องที่เป็นบวก (true positive rate, TP rate) และอัตราส่วนความผิดพลาดที่เป็นบวก (false positive rate, FP rate) ในกรณีที่พื้นที่ใต้โค้งที่มีค่ามากกว่าแสดงถึงการจำแนกที่มีประสิทธิภาพสูงกว่าในทางกลับกันพื้นที่ใต้โค้งที่มีค่าน้อยกว่าแสดงถึงการจำแนกที่มีประสิทธิภาพต่ำกว่า และอีกค่าหนึ่งที่นิยมใช้ก็คือ ค่ามาตรวัดเอฟ (F-Measure) [7] เป็นค่าเฉลี่ยระหว่างค่าค้นคืน (recall) และค่าความถูกต้อง (precision)

*Corresponding author. E-mail: kalaya.tm@gmail.com

จากการศึกษานั้นมีวิธีหลักอยู่ 2 วิธีที่ใช้ในการแก้ปัญหาความไม่สมดุลของประเภทข้อมูล วิธีที่หนึ่งคือการเพิ่มจำนวนข้อมูล (over-sampling) และวิธีที่สองคือการลดจำนวนข้อมูล (under-sampling) [8] ข้อเสียของวิธีการลดจำนวนข้อมูล ก็คือข้อมูลที่มีประโยชน์อาจถูกลบออกจากชุดข้อมูลได้ แต่ในทางกลับกันวิธีการเพิ่มจำนวนข้อมูลนั้นอาจจะสร้างข้อมูลที่ซ้ำซ้อนกันกับข้อมูลเดิม จึงมีการนำเสนอการสร้างข้อมูลสังเคราะห์ส่วนน้อย อาทิ เช่น สโมท (Synthetic Minority Oversampling Technique, SMOTE) [9] สโมทเป็นการเพิ่มประเภทข้อมูลส่วนน้อยโดยสร้างข้อมูลใหม่อย่างสุ่มระหว่างเส้นตรงที่ลากผ่านข้อมูลสองตัวที่อยู่ใกล้กัน โดยเลือกข้อมูลสองตัวนั้นจากวิธีการหาเพื่อนบ้าน *k* ตัวอย่างที่ใกล้ที่สุด (*k*-NN) ที่เป็นประเภทข้อมูลส่วนน้อย ข้อดีของสโมทคือเพิ่มขอบเขตของข้อมูลส่วนน้อย ลดความซ้ำซ้อนของข้อมูลอีกทั้งยังช่วยปรับปรุงการทำนายประเภทข้อมูลส่วนน้อยให้ดีขึ้น และไม่ทำให้ค่าความแม่นยำของข้อมูลลดลงอีกด้วย การสร้างข้อมูลสังเคราะห์บริเวณขอบ (Borderline-SMOTE) เป็นการเลือกเฉพาะข้อมูลที่อยู่ใกล้ขอบของข้อมูลประเภทน้อย มาสังเคราะห์ข้อมูลโดยใช้วิธีสโมท ส่วนข้อมูลที่อยู่ไกลจากขอบและอยู่บนแก่นข้อมูลประเภทน้อยจะถูกกรองทิ้ง ข้อเสียของการสร้างข้อมูลสังเคราะห์บริเวณขอบถ้าชุดข้อมูลมีการกระจายตัวที่ห่างมาก จะทำให้ไม่สามารถที่จะหาข้อมูลที่อยู่ใกล้ขอบของข้อมูลประเภทน้อยได้ ดังนั้น ในงานวิจัยนี้จึงเสนอวิธีการเลือกข้อมูลที่พัฒนาจากวิธีสโมท โดยนำข้อมูลถูกสังเคราะห์ที่สร้างจากวิธีสโมท มาคัดแยกตามค่าความรู้ ซึ่งคำนวณมาจากการย้อนกลับไปดูเพื่อนบ้าน เพื่อข้อมูลที่ถูกเลือกมานั้นไม่มีความซ้ำซ้อนกับข้อมูลเดิม

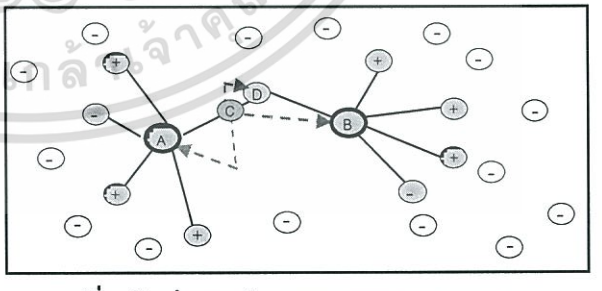
วิธีการ

แนวความคิดเบื้องต้นของวิธีการที่นำเสนอคือ เพิ่มตัวอย่างเฉพาะข้อมูลประเภทข้อมูลน้อย และคงตัวอย่างข้อมูลส่วนมากที่คัดแยกไว้ดีแล้ว สำหรับข้อมูลใดๆ การที่จะดูว่าข้อมูลนั้นได้ถูกคัดแยกไว้ดีแล้วหรือไม่ ให้หาเพื่อนบ้านที่ใกล้ที่สุด k ตัวอย่าง และทำการโหวตประเภทข้อมูลจากประเภทข้อมูลของเพื่อนบ้านทั้งหมด ถ้าประเภทข้อมูลที่ได้จากการโหวตว่าตรงกับประเภทของข้อมูลเริ่มต้นจะสรุปว่าข้อมูลนั้นได้ถูกคัดแยกไว้ดีแล้ว ในทางกลับกันถ้าประเภทข้อมูลที่ได้จากการโหวตไม่ตรงกับประเภทของข้อมูลเริ่มต้นจะสรุปว่าข้อมูลนั้นได้ถูกคัดแยกไว้ไม่ดี

ในงานวิจัยชิ้นนี้ ใช้วิธีการหาเพื่อนบ้าน k ตัวอย่างที่ใกล้ที่สุดแบบกลับด้าน(reverse K-NN, rkNN) สำหรับข้อมูลสังเคราะห์ใดๆ จะค้นกลับหาว่าตัวอย่างสังเคราะห์นั้นเป็นเพื่อนบ้าน k ตัวอย่างที่ใกล้ที่สุด(kNN) แก่ข้อมูลใดบ้างข้อมูลนั้นก็จะเป็น kตัวอย่างที่ใกล้ที่สุดแบบกลับด้านของตัวอย่างสังเคราะห์ที่กล่าวมา ดังแสดงในรูปที่ 1 ข้อมูลประเภทน้อยที่ถูกสังเคราะห์ C ซึ่งสร้างขึ้นด้วยวิธีการสโมท เมื่อคำนวณหาเพื่อนบ้านที่ใกล้ที่สุด k ตัวอย่าง บนข้อมูล A, B และ D พบว่า มีข้อมูล C ประกอบอยู่ด้วย จึงสามารถระบุได้ว่า A, B และ D เป็นเพื่อนบ้าน k ตัวอย่างที่ใกล้ที่สุดแบบกลับด้านของ C เช่นกัน จากรูปที่ 1 วงกลมเครื่องหมายบวกแสดงถึงข้อมูลประเภทน้อย วงกลมเครื่องหมายลบแสดงถึงข้อมูลประเภทมาก, เส้นทึบแสดงถึงความสัมพันธ์แบบ kNN และเส้นปะแสดงถึงความสัมพันธ์แบบ rkNN

การสร้างข้อมูลสังเคราะห์ ในงานวิจัยนี้ได้ใช้วิธีการสโมทในการสร้าง โดยกำหนดให้ทุกรอบจะสร้างข้อมูลสังเคราะห์เพียง 25% ของเปอร์เซ็นการเพิ่มข้อมูลประเภทน้อยซึ่งกำหนดโดยผู้ใช้เมื่อได้จำนวนข้อมูลประเภทน้อยที่ต้องสร้างแล้วให้นำมาสังเคราะห์ข้อมูลด้วยวิธีการสโมทเป็นจำนวน 10 เท่า ของจำนวนข้อมูลประเภทน้อยนั้น เพื่อให้มีข้อมูลสังเคราะห์ที่จำนวนเพียงพอให้สามารถเลือกตัวอย่างข้อมูลสังเคราะห์ที่ดีที่สุดมาเพิ่มแก่ข้อมูลต้นฉบับ โดยส่วนที่เหลือจะถูกกรองทิ้ง

Algorithm: Reverse K-NN SMOTE (k=5)
Input: a set of all training set instances: D, |D| = N
a set of all training set minority instances: Min
a set of all training set majority instances: Maj
where |Maj| + |Min| = |D|
Percentage of over-sampling: Percentage,
i.e., 100, 200, 300, so on
Output: set of synthetic minority instances L



รูปที่ 1 ตัวอย่างการทำ reverse-*k*NN ของ node C

1. Define L = Ø
2. Calculate distance (x_{ij}) for all i in D and for all j in its k nearest neighbor, and then find their average distances ($\bar{x}$)Variance = $\sum_{i=1}^{N}\sum_{j=1}^{k}\frac{(x_{ij}-\bar{x})^2}{N-1}$ (1)
3. Do sampling 25% per iteration until reach the Percentage
   3.1 Calculate number of iteration numIteration = Percentage/25(2)
   3.2 Calculate number of sampling
   numSampling = $\lceil |Min| * 0.25 \rceil$ (3)

4. Generate synthetic minority data 10 times of numSampling with SMOTE
   Candidate = set of synthetic data that was generated by SMOTE
5. D = D Union L
6. For each instance D
   Find set of *k*-nearest neighbors of D_i (*k*-NN_i)
7. Find set of reverse *k*-NN (r*k*NN_j) of each $Candidate_j$
   For $Candidate_j$ in Candidate {For each D_i in D {If $Candidate_j$ is in *k*-NN_i { Add D_i to r*k*NN_j }}}
8. Calculate support (Support) For each D_i in D
   {Count number of minority instance in *k*-NN_i ($|Min_{knn\_i}|$) Count number of majority instance in *k*-NN_i ($|Maj_{knn\_i}|$)
   If ($|Min_{knn\_i}| \leq \lfloor (|Min_{knn\_i}| + |Maj_{knn\_i}|)/2 \rfloor$) {$Support_i$= 1} (4)
   else { $Support_i = v_1 |minor| + v_2$ } (5)

   where $v_1 = \frac{1}{\lfloor \frac{k}{2} \rfloor - k}$, $v_2 = \frac{-k}{\lfloor \frac{k}{2} \rfloor - k}$ (6)
9. Calculate information value ($Info_i$)
   For $Candidate_j$ in Candidate {$Info_i$ = 0For each $D_n \in r kNN_j$ {
   if ($D_n$is minority class) { $Info_i = Info_i + (S_n \times \frac{|Min| + |Maj|}{|Min| \times 2}$)} (7)
   if (D_n is majority class) { $Info_i = Info_i - (S_n \times \frac{|Min| + |Maj|}{|Maj| \times 2}$)} (8)
10. Sort Candidate with Information value by descending order
11. Calculate threshold
    If (average(Info) < 0){ *threshold* = 0 } else { threshold = average(Info) } (9)
12. Prune all candidate that information value lower than the threshold
13. Add candidate to L
14. go to the 4rd step until reach to numIteration
15. return L

สำหรับทุกๆ ชุดข้อมูล ถ้าค่าความแปรปรวน (variance) นั้นมีค่ามากกว่า 1,000 แสดงว่าชุดข้อมูลที่นำมาสังเคราะห์ข้อมูลนั้นมีการกระจายตัวของข้อมูลที่กระจายตัวสูงมาก เราจะทำการเตรียมข้อมูลโดยแปลงข้อมูลทั้งหมดให้อยู่ช่วง [0,1] หลังจากการเตรียมข้อมูลแล้วนั้นจะสร้างข้อมูลสังเคราะห์ส่วนน้อยเป็นจำนวน 10 เท่าของ numSampling มาเลือกตัวอย่างที่ดีที่สุดจำนวนหนึ่ง โดยนำข้อมูลสังเคราะห์ทั้งหมดจะนำมาคำนวณค่าความรู้ ดังแสดงในสมการ 7 และ 8 โดยที่ค่าความรู้จะเพิ่มขึ้นสูงเมื่อตัวอย่างเพื่อนบ้านเป็นข้อมูลประเภทน้อยที่ไม่สามารถคัดแยกได้ถูกต้อง แต่จะโดนลดค่าเมื่อเพื่อนบ้านเป็นข้อมูลประเภทใหญ่

เพื่อคำนวณค่าความรู้นั้นเบื้องต้นต้องคำนวณค่าความช่วยเหลือ (support, S) ของทุกข้อมูลที่มีข้อมูลสังเคราะห์นั้นว่าเป็นหนึ่งในเพื่อนบ้าน k ตัวอย่างที่ใกล้ที่สุด (k-NN) หรือกล่าวคือการหาเพื่อนบ้าน k ตัวอย่างที่ใกล้ที่สุดแบบกลับด้านของข้อมูลสังเคราะห์ โดยค่าความช่วยเหลือแสดงได้ดังสมการที่ 4, 5 และ 6 ตามลำดับ ค่าความช่วยเหลือจะมีค่าเท่ากับ 1 เมื่อจำนวนเพื่อนบ้าน kตัวอย่างที่ใกล้ที่สุด (k-NN) เกินกว่าครึ่งเป็นข้อมูลประเภทใหญ่ดังแสดงในสมการที่ 4 และมีค่าลดลงเป็นกราฟเส้นตรงตามลำดับดังแสดงในสมการที่ 5 และ 6 เมื่อคำนวณค่าความรู้แก่ตัวอย่างสังเคราะห์ทั้งหมดแล้ว จึงทำการเรียงตัวอย่างสังเคราะห์จากมากไปน้อยตามค่าความรู้ และเลือกเฉพาะตัวอย่างสังเคราะห์ที่มีค่าสูงกว่า ค่าน้อยที่สุดที่ยอมรับได้ (threshold) ดังแสดงในสมการที่ 9 ซึ่งวิธีการที่กล่าวมาจะสร้างข้อมูลสังเคราะห์และคัดเลือกข้อมูลไปเรื่อยๆ จนกว่าครบรอบที่กำหนด

ผลและอภิปราย

งานวิจัยนี้ใช้มาตรวัดเอฟ (F-Measure) และพื้นที่ใต้กราฟ (AUC) ในการวัดประสิทธิภาพของการคัดแยกประเภทข้อมูลบน 10 ชุดข้อมูลจากคลังข้อมูล UCI [11] ดังแสดงในตารางที่ 1 โดยชุดข้อมูลถูกเรียงลำดับด้วยอัตราส่วนของข้อมูลประเภทน้อยจากค่ามากไปน้อยตามลำดับงานวิจัยใกล้เคียงที่ใช้เปรียบเทียบกับงานวิจัยที่นำเสนอ คือ สโมท (SMOTE) และการสร้างข้อมูลสังเคราะห์บริเวณขอบ (Borderline-SMOTE) กำหนดให้ค่า k=5 เหมือนในงานวิจัย [9] และ [10] การแบ่งข้อมูลทดลองและข้อมูลเรียนรู้นั้นได้ใช้วิธีการสลับข้อมูล 10 รอบ (10-fold

cross-validation) โดยแบ่งข้อมูลเป็น 10 ส่วน ใช้ 1 ส่วนเป็นข้อมูลทดสอบ ข้อมูลที่เหลือเป็นข้อมูลการเรียนรู้ สลับข้อมูลทดสอบไปเรื่อยจนครบ 10 รอบ โดยที่ข้อมูลเรียนรู้นั้นจะถูกเรียนโดยใช้ต้นไม้การตัดสินใจ C4.5 (Decision tree, C4.5) และวัดประสิทธิภาพบนข้อมูลทดสอบด้วยมาตรวัดเอฟ และพื้นที่ใต้กราฟบน 10 ชุดข้อมูล ค่าเฉลี่ยของการคำนวณทั้ง 10 ครั้ง ของค่ามาตรวัดเอฟและพื้นที่ใต้กราฟที่แสดงในตารางที่ 2

จากตารางที่ 2 จะเห็นได้ว่าค่ามาตรวัดเอฟและพื้นที่ใต้กราฟของวิธีการนำเสนอ (RSMOTE) ให้ผลลัพธ์ในการคัดแยกประเภทข้อมูลได้ดีที่สุดเป็นจำนวน 17 จากทั้งหมด 30 กรณี และ 18 จากทั้งหมด 30 กรณี ตามลำดับ ส่วนวิธีสโมทจะเป็นวิธีที่ให้ผลลัพธ์ดีที่สุด 2 กรณีบนมาตรวัดค่าเอฟและ 1 กรณีบนมาตรวัดพื้นที่ใต้กราฟ จึงทำให้เชื่อได้ว่าวิธีที่นำเสนอนั้นได้เพิ่มประสิทธิภาพการคัดแยกประเภทข้อมูลจากวิธีการพื้นฐานหรือสโมท ได้อย่างชัดเจน ในกรณีของสร้างข้อมูลสังเคราะห์บริเวณขอบ (BSmote) นั้นเห็นได้ชัดว่าในกรณีที่เปอร์เซ็นต์การสร้างข้อมูลสังเคราะห์ที่ 300% และ 500% วิธีการที่นำเสนอให้ประสิทธิภาพการคัดแยกข้อมูลที่คงแต่วิธีการสร้างข้อมูลสังเคราะห์บริเวณขอบมีค่าลดลง แต่ในชุดข้อมูลที่มีการกระจายตัวของข้อมูลสูงอย่าง Pima และ Satimage ถ้าใช้วิธีที่นำเสนอในการเพิ่มตัวอย่างข้อมูลแล้วจะให้ผลลัพธ์ในการคัดแยกประเภทข้อมูลที่น้อยกว่าวิธีสร้างข้อมูลสังเคราะห์บริเวณขอบ จะเห็นได้ว่าวิธีการที่นำเสนอจะให้ผลลัพธ์ในการคัดแยกประเภทข้อมูลที่ดีกว่าเมื่ออัตราสัดส่วนของประเภทข้อมูลส่วนน้อยมีจำนวนน้อย อาทิ เช่น ข้อมูล Glass_3 และ Ecoli_om โดยที่ตารางที่ 2 ได้แนบวิธีการที่ไม่มีการสร้างตัวอย่างสังเคราะห์เพิ่มแก่ข้อมูลต้นฉบับ (C4.5) เพื่อใช้เปรียบเทียบกับวิธีการต่าง ๆ ว่าสามารถเพิ่มประสิทธิภาพได้ปริมาณเท่าใด

ตารางที่ 1 แสดงรายละเอียดของชุดข้อมูล

Dataset	number of Examples	number of Attributes	Class label (minority : majority)	Percentage of minority class
Ionosphere	351	35	bad:good	35.8%
Pima	768	8	1:0	34.77%
Seeds	210	8	1:2	33.33%
Vehicle	846	19	van:other	30.75%
Contraceptive	1473	10	2:other	29.21%
Haberman	306	3	2:1	26.47%
BreastTissue	106	10	car:other	24.71%
Satimage	6435	37	4:other	10.75%
Glass_3	214	10	3:other	8.63%
Ecoli_om	336	8	om:other	6.32%

ตารางที่ 2 แสดงค่า F-measure และ AUC ที่ได้จากการทดลอง โดยค่าที่เป็นตัวสีแดงเป็นค่าที่ให้ผลดีที่สุดจาก 4 วิธี

			F-Measure				AUC			
Dataset	%Minority	%Sampling	C4.5	SMOTE	BSmote	RSMOTE	C4.5	SMOTE	BSmote	RSMOTE
Ionosphere	35.8%	100	83.3%	82.2%	82.6%	83.3%	90.6%	86.2%	86.5%	87.1%
		300	83.3%	82.3%	82.5%	82.1%	90.6%	86.5%	86.7%	86.3%
		500	83.3%	80.4%	81.2%	82.1%	90.6%	85.1%	85.7%	86.2%
Pima	34.77%	100	62.8%	63.3%	65.3%	65.4%	71.7%	71.1%	72.9%	73.2%
		300	62.8%	63.4%	65.2%	64.7%	71.7%	70.8%	72.6%	72.3%
		500	62.8%	63.8%	65.8%	64.7%	71.7%	71.1%	73.1%	72.4%
Seeds	33.33%	100	77.5%	89.3%	84.9%	85.5%	83.2%	92.4%	84.9%	89.7%
		300	77.5%	84.7%	84.2%	85.9%	83.2%	88.8%	84.2%	89.9%
		500	77.5%	85.3%	84.3%	86.0%	83.2%	89.6%	84.3%	89.9%
Vehicle	30.75%	100	86.6%	87.4%	87.3%	88.1%	91.6%	92.2%	91.9%	92.6%
		300	86.6%	87.1%	86.8%	88.0%	91.6%	92.2%	91.9%	92.4%
		500	86.6%	86.7%	86.9%	88.4%	91.6%	92.0%	92.1%	92.7%
Contraceptive	29.21%	100	42.9%	44.0%	44.5%	44.4%	63.4%	63.8%	64.1%	64.1%
		300	42.9%	42.6%	44.0%	44.9%	63.4%	62.9%	63.8%	64.4%
		500	42.9%	42.5%	42.9%	44.7%	63.4%	62.7%	63.0%	64.2%
Haberman	26.47%	100	36.3%	50.0%	50.2%	47.3%	57.2%	65.7%	66.0%	63.8%
		300	36.3%	47.1%	48.5%	48.9%	57.2%	62.5%	64.2%	64.6%
		500	36.3%	47.8%	48.0%	49.4%	57.2%	63.1%	63.7%	64.9%
BreastTissue	24.71%	100	82.6%	83.6%	84.6%	84.2%	91.7%	91.5%	92.5%	92.2%
		300	82.6%	85.4%	85.7%	86.4%	91.7%	93.2%	93.4%	94.2%
		500	82.6%	84.9%	84.6%	86.4%	91.7%	92.7%	92.5%	94.2%
Satimage	10.75%	100	55.7%	57.6%	57.4%	56.5%	75.4%	76.9%	77.3%	76.6%
		300	55.7%	58.2%	58.7%	57.4%	75.4%	78.3%	78.7%	77.8%
		500	55.7%	57.4%	58.2%	57.5%	75.4%	78.2%	78.8%	77.7%
Glass_3	8.63%	100	52.1%	50.3%	48.4%	44.1%	76.3%	73.5%	72.1%	69.3%
		300	52.1%	52.0%	46.8%	57.3%	76.3%	75.8%	71.6%	78.2%
		500	52.1%	50.2%	42.1%	55.4%	76.3%	75.2%	69.2%	76.6%
Ecoli_om	5.85%	100	58.8%	72.9%	78.5%	74.6%	74.3%	86.0%	87.4%	88.4%
		300	58.8%	75.5%	74.2%	76.5%	74.3%	89.8%	85.7%	90.2%
		500	58.8%	77.0%	71.4%	77.1%	74.3%	90.0%	82.8%	90.7%
			4	**2**	**8**	**17**	**4**	**1**	**8**	**18**

บทสรุป

การเพิ่มตัวอย่างข้อมูลโดยการสังเคราะห์ข้อมูลนั้นเป็นแนวทางหนึ่งที่ใช้แก้ปัญหาข้อมูลที่ไม่สมดุลกัน ในงานวิจัยชิ้นนี้นำเสนอวิธีการหาเพื่อนบ้าน kตัวอย่างที่ใกล้ที่สุดแบบกลับด้าน (reverse K-NN, rk-NN) ร่วมกับการสร้างข้อมูลสังเคราะห์ในการปรับข้อมูลที่ไม่สมดุล ซึ่งวิธีที่นำเสนอนั้นจะคำนวณค่าความรู้ เพื่อที่จะเลือกตัวอย่างข้อมูลสังเคราะห์ที่ช่วยคัดแยกข้อมูลประเภทข้อมูลส่วนน้อยที่คัดแยกผิด ให้สามารถคัดแยกประเภทข้อมูลได้อย่างถูกต้องได้ จากการทดลองข้อมูลทั้ง 10 ชุด โดยใช้ค่ามาตรวัดเอฟ และค่ามาตรวัดพื้นที่ใต้เส้นโค้ง ผลการทดลองแสดงให้เห็นว่า วิธีที่นำเสนอนี้มีประสิทธิภาพในการเพิ่มตัวอย่างข้อมูลได้ดีกว่าทั้งวิธีสโมทที่เป็นวิธีพื้นฐาน, วิธีสร้างข้อมูลสังเคราะห์บริเวณขอบและข้อมูลต้นฉบับที่ไม่มีการเพิ่มข้อมูลสังเคราะห์กับชุดข้อมูลส่วนใหญ่ เมื่อใช้มาตรวัดค่าเอฟ

(F-Measure) และพื้นที่ใต้กราฟ (AUC)สิ่งที่สามารถพัฒนาคือ การหาค่าน้อยที่สุดที่ยอมรับได้ในการเลือกตัวอย่างสังเคราะห์ที่เหมาะสม ซึ่งในงานวิจัยนี้ได้ใช้ค่า threshold เป็นค่าเฉลี่ยของค่าความรู้

เอกสารอ้างอิง

Bradley, A. (1997). The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. **Pattern Recognition,** 30(6), 1145–1159.

Buckland, M. and Gey, F. (1994). The Relationship between Recall and Precision. **Journal of the American Society for Information Science,** 45(1), 12–19.

Chawla, N., Bowyer, K., Hall, L. and Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over- Sampling Technique. **Journal of Artificial Intelligence Research,** 16, 321–357.

David D. Lewis and Jason Catlett (1994). **Heterogeneous Uncertainty Sampling for Supervised Learning.** Proceedings of the 11th International Conference on Machine Learning ICML'94, 11, 148-156.

Han, H., Wang, W. and Mao, B. (2005). Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In: Huang, D.-S., Zhang, X.-P., Huang, G.-B. (2005). **ICIC 2005. LNCS,** 3644, 878–887.

Kubat, m., Holte, R., and Matwin, S. Machine (1998). Learning for the Detection of Oil Spills in Satellite Radar Images. **Machine Learning,** 30, 195–215.

NiteshV.Cha (2010). Data Mining for Imbalanced Datasets: An Overview doi:10.1007/978-0-387-09823-4_45 In: Maimon, Oded; Rokach, Lior (Eds) Data Mining and Knowledge Discovery Handbook. **Springer,** ISBN 978-0-387-09823-4, 875-886

NiteshV.Chawla, Nathalie Japkowicz and AleksanderKolcz (2004). Editorial: Special Issue on Learning from Imbalanced Data Sets. **SIGKDD Explorations,** 6(1), 1-6.

UCI machine learning repository, Retrieved June 2, 2012, from http://archive.ics.uci.edu/ml

Yang, Q. and Wu, X. (2006). 10 challenging problems in data mining research. **International Journal of Information Technology and Decision Making,** 5(4), 597–604 from http://dblp.unitrier.de/db/journals/ijitdm/ijitdm5.html

ZhaohuiZheng, Xiaoyun Wu and RohiniSrihari (2004). Feature Selection for Text Categorization on Imbalanced Data. **SIGKDD Explorations,** 6(1), 80-89.

ประวัติผู้เขียน

ชื่อ – สกุล	นางสาวกัลยา พันธุมะผล
วัน เดือน ปี เกิด	26 ธันวาคม 2527
ที่อยู่	105/12 ซอย 5 หมู่บ้านนักกีฬา เขตสะพานสูง จังหวัดกรุงเทพเทพมหานคร 10250

ประวัติการศึกษา

2550	จบการศึกษาปริญญาวิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าลาดกระบัง