เทคนิคการสุ่มเพิ่มตัวอย่างข้างน้อยสังเคราะห์และเทคนิคการสุ่มลดตัวอย่างข้างมากสำหรับปัญหาความไม่ดุลระหว่างกลุ่ม

นาย ปณต ทรงวัฒนศิริ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาการคณนา ภาควิชาคณิตศาสตร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2553

SYNTHETIC MINORITY OVER-SAMPLING AND MAJORITY UNDER-SAMPLING TECHNIQUES FOR CLASS IMBALANCED PROBLEMS

Mr. Panote Songwattanasiri

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Computational Science
Department of Mathematics,
Faculty of Science,
Chulalongkorn University
Academic Year 2010

หัวข้อวิทยานิพนธ์	เทคนิคการสุ่มเพิ่มตัวอย่างข้างน้อยสังเคราะห์และเทคนิคการสุ่มลดตัวอย่างข้างมากสำหรับปัญหาความไม่ดุลระหว่างกลุ่ม
โดย	นาย ปณต ทรงวัฒนศิริ
สาขาวิชา	วิทยาการคณนา
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร. กรุง สินอภิรมย์สราญ

คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะวิทยาศาสตร์

(ศาสตราจารย์ ดร. สุพจน์ หารหนองบัว)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(ผู้ช่วยศาสตราจารย์ ดร. จารุโลจน์ จงสถิตย์วัฒนา)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(ผู้ช่วยศาสตราจารย์ ดร. กรุง สินอภิรมย์สราญ)

..... กรรมการ

(อาจารย์ ดร. บุญฤทธิ์ อินทิยศ)

..... กรรมการภายนอกมหาวิทยาลัย

(อาจารย์ ดร. กมล เกียรติเรืองกมลา)

ปณต ทรงวัฒนศิริ : เทคนิคการสุ่มเพิ่มตัวอย่างข้างน้อยสังเคราะห์และเทคนิคการสุ่มลดตัวอย่างข้างมากสำหรับปัญหาความไม่ดุลระหว่างกลุ่ม. อ. ที่ปรึกษาวิทยานิพนธ์หลัก : ผศ.ดร. กรุง สินอภิรมย์สราญ , 73 หน้า.

เทคนิคการสุ่มเพิ่มตัวอย่างข้างน้อยสังเคราะห์และเทคนิคการสุ่มลดตัวอย่างข้างมากสำหรับปัญหาความไม่ดุลระหว่างกลุ่ม (SMOUTE) เป็นกระบวนจัดการข้อมูลก่อนการสร้างตัวแบบสำหรับการแก้ปัญหาความไม่ดุลระหว่างกลุ่ม SMOUTE เป็นการผสานระหว่าง SMOTE ซึ่งเป็นเทคนิคการเพิ่มจำนวนแบบสุ่ม (Over-sampling technique) โดยเพิ่มจำนวนข้อมูลของไมนอริตี้คลาสกับการใช้เทคนิคการลดแบบสุ่ม (Under-sampling technique) ลดจำนวนข้อมูลของมาจอริตี้คลาส ในส่วนของการลดแบบสุ่ม เราใช้ขั้นตอนวิธีค่าเฉลี่ย k (k-means algorithm) เพื่อแบ่งข้อมูลของมาจอริตี้คลาสออกเป็น k กลุ่ม และลดจำนวนข้อมูลของมาจอริตี้คลาสบริเวณใกล้เคียงกับเซนทรอยด์ (Centroid) แต่ละตัว เราใช้ตัวแบบ C4.5 ตัวแบบการแบ่งประเภทเบย์อย่างง่าย (Naïve Bayes) และตัวแบบเพอร์เซ็ปตรอนหลายชั้น (Multilayer perceptron) เป็นตัวแยกประเภท (Classifiers) ผลการทดสอบพบว่า SMOUTE มีความแม่นยำในการทำนายข้อมูลไมนอริตี้กว่า SMOTE และความเร็วของขั้นตอนวิธีของ SMOUTE เร็วกว่าขั้นตอนวิธีของ SMOTE สำหรับข้อมูลขนาดใหญ่

ภาควิชา.....คณิตศาสตร์..... ลายมือชื่อนิสิต.....

สาขาวิชา.....วิทยาการคณนา..... ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก.....

ปีการศึกษา.....2553.....

5172347123 : MAJOR COMPUTATIONAL SCIENCE

KEYWORDS : CLASS IMBALANCED PROBLEM / OVER-SAMPLING / UNDER-SAMPLING / SMOTE / K-MEANS ALGORITHM

PANOTE SONGWATTANASIRI : SYNTHETIC MINORITY OVER-SAMPLING AND MAJORITY UNDER-SAMPLING TECHNIQUES FOR CLASS IMBALANCED PROBLEMS. THESIS ADVISOR : ASST. PROF. KRUNG SINAPIROMSARAN, PH.D., 73 pp.

Synthetic minority over-sampling and majority under-sampling techniques for class imbalanced problems (SMOUTE) is the data preprocessing for handling the class imbalanced problem. SMOUTE uses synthetics minority over-sampling technique (SMOTE) to insert the minority class instances and uses under-sampling technique to purge the majority class instances. For under-sampling, we use *k*-means algorithm to partition the majority class instances into *k* clusters then we drop some majority class instances around centroids. We perform experiments based on three classifiers, C4.5, Naïve Bayes and multilayer perceptron. Our results show that classifiers using SMOUTE are correctly grouped the minority class better than SMOTE. Moreover, the speed of SMOUTE is much faster than that of SMOTE for large datasets.

Department : Mathematics ..... Student's Signature .....

Field of Study : Computational Science .... Advisor's Signature .....

Academic Year : 2010 .....

กิตติกรรมประกาศ

ผู้วิจัยขอกราบขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร. กรุง สินอภิรมย์สราญ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ได้กรุณาให้ความรู้ คำแนะนำ และคำปรึกษาต่างๆที่ทำให้วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยดี

ขอกราบขอบพระคุณผู้ช่วยศาสตราจารย์ ดร. จารุโรจน์ จงสถิตวัฒนา ประธานกรรมการ อาจารย์ ดร. บุญฤทธิ์ อินทิยศ กรรมการ และอาจารย์ ดร. กมล เกียรติเรืองกมลา กรรมการภายนอกมหาวิทยาลัย ที่ได้ให้คำปรึกษา คำแนะนำและแก้ไขข้อบกพร่องต่างๆ ในงานวิจัยนี้ ซึ่งทำให้วิทยานิพนธ์ฉบับนี้มีความสมบูรณ์มากยิ่งขึ้น

ขอกราบขอบพระคุณบิดาและมารดา ตลอดจนพี่น้องในครอบครัวและเพื่อนๆ ทุกคนที่คอยเป็นกำลังใจและช่วยเหลือผู้วิจัยมาโดยตลอด

สารบัญ

สารบัญตาราง

สารบัญภาพ

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การแยกประเภท (Classification) เป็นสาขาหนึ่งของการทำเหมืองข้อมูล (Data mining) การแยกประเภทถูกใช้เพื่อการทำนายข้อมูล และระบุคลาสเป้าหมาย (Target class) ของข้อมูลที่รับเข้ามาใหม่ ตัวแบบที่ใช้ในการทำนายหรือตัวแยกประเภท (Classifier) ถูกสร้างขึ้นโดยอิงจากชุดข้อมูลฝึกหัด (Training set) และทดสอบประสิทธิภาพของตัวแบบด้วยชุดข้อมูลทดสอบ (Test set) [1] อย่างไรก็ดี นักวิจัยยังไม่มีข้อสรุปสำหรับตัวแบบแยกประเภทที่สามารถใช้ได้กับทุกปัญหา ดังนั้นจึงมีการคิดค้นวิธีการเฉพาะเพื่อสร้างตัวแบบแยกประเภทที่เหมาะสมสำหรับแต่ละโจทย์ที่สนใจ

ปัญหาการแยกประเภทยุคปัจจุบันที่พบส่วนใหญ่เป็นปัญหาที่กลุ่มข้อมูลแต่ละกลุ่มมักมีขนาดไม่ดุลกัน เราเรียกปัญหาดังกล่าวว่าปัญหาความไม่ดุลระหว่างกลุ่ม (Class imbalanced problem) [2] ชุดข้อมูลจะถูกกล่าวว่ามีความไม่ดุลระหว่างกลุ่มก็ต่อเมื่อ ชุดข้อมูลนั้นมีคลาสเป้าหมายคลาสหนึ่งมีจำนวนข้อมูลของคลาสนั้นน้อย หรือไมนอริตี้คลาส เมื่อเทียบกับคลาสเป้าหมายอื่นซึ่งมีจำนวนข้อมูลเยอะกว่ามาก (มาจอริตี้คลาส) โดยเป้าหมายสำคัญในการแก้ปัญหาความไม่ดุลระหว่างกลุ่มคือ การสร้างตัวแบบที่มีความแม่นยำสูงในการทำนายไมนอริตี้คลาส

ตั้งแต่ปี 2000 เป็นต้นมามีงานวิจัยมากมายถูกนำเสนอเพื่อแก้ปัญหาความไม่ดุลระหว่างกลุ่ม โดย N. V. Chawla N. Japkowicz และ A. Kolcz (2004) [3] ได้จำแนกเทคนิคต่างๆเกี่ยวกับงานวิจัยทางด้านนี้ออกเป็น 3 กลุ่มคือ การชักข้อมูล (Sampling), การเรียนรู้แบบคลาสเดียว (One-class learning) และการคัดเลือกลักษณะเฉพาะ (Feature selection)

การชักข้อมูล (Sampling) เกิดจากแนวคิดที่ว่าการกระจายตัวของข้อมูลในแต่ละคลาส (Class distribution) ที่ได้จากการเก็บข้อมูลจริงอาจไม่เหมาะที่จะนำมาสร้างตัวแบบ ดังนั้นการเปลี่ยนรูปแบบการกระจายตัวของข้อมูลให้เหมาะสมจะช่วยเพิ่มประสิทธิภาพของตัวแบบในการทำนายข้อมูล งานวิจัยที่ใช้วิธีการชักข้อมูลส่วนใหญ่แบ่งออกเป็น 3 กลุ่มคือ เทคนิคการชักตัวอย่างแบบเพิ่ม (Over-sampling technique) เทคนิคการชักตัวอย่างแบบลด (Under-sampling technique) และการผสานกันของเทคนิคการชักตัวอย่างแบบเพิ่มกับเทคนิคการชักตัวอย่างแบบลด

กฎการลดจำนวนข้อมูลของบริเวณใกล้เคียงที่สุด (The Condensed Nearest Neighbor Rule หรือ CNN) ถูกนำเสนอโดย P. E. Hart (1968) [4] วิธีการ CNN เป็นหนึ่งในเทคนิคการชักตัวอย่างแบบลด โดยการใช้ขั้นตอนวิธี 1 เนียเรสเนเบอร์ (1-nearest neighbor technique) ตรวจสอบข้อมูลของมาจอริตี้คลาสว่าอยู่ใกล้กับข้อมูลของไมนอริตี้คลาสหรือมาจอริตี้คลาส และทำการสุ่มลบเฉพาะข้อมูลของมาจอริตี้คลาสที่ใกล้ข้อมูลของมาจอริตี้คลาส การประมวลผลเสร็จสิ้นเมื่อจำนวนข้อมูลของมาจอริตี้คลาสเท่ากับจำนวนที่ผู้ใช้กำหนด

สองการดัดแปลงของวิธีการ CNN (Two Modifications of CNN) หรือโทเม็คลิงคส์ (Tomek links) ถูกนำเสนอโดย I. Tomek (1976) [6] โทเม็คลิงคส์เป็นการนำวิธีการ CNN มาดัดแปรโดยการใช้เทคนิค 1 เนียเรสเนเบอร์ ตรวจสอบข้อมูลของมาจอริตี้คลาสว่าอยู่ใกล้กับข้อมูลของไมนอริตี้คลาสหรือมาจอริตี้คลาสเช่นเดียวกับวิธีการ CNN แต่โทเม็คลิงคส์สุ่มลบเฉพาะข้อมูลของมาจอริตี้คลาสที่มีข้อมูลของไมนอริตี้คลาสเป็นข้อมูลที่อยู่ใกล้ที่สุด ซึ่งเป็นวิธีการตรงกันข้ามกับวิธีการ CNN การประมวลผลเสร็จสิ้นเมื่อจำนวนข้อมูลของมาจอริตี้คลาสเท่ากับจำนวนที่ผู้ใช้กำหนด

เทคนิคการชักตัวอย่างสังเคราะห์ไมนอริตี้แบบเพิ่ม (The Synthetic Minority Over-sampling TEchnique หรือ SMOTE) ถูกพัฒนาโดย N. V. Chawla K. W. Bowyer L. O. Hall และ W. P. Kegelmayer (2002) [7] วิธีการ SMOTE ใช้ขั้นตอนวิธี k เนียเรสเนเบอร์ (K-nearest neighbor algorithm) ในการเพิ่มข้อมูลของไมนอริตี้คลาสบนตำแหน่งระหว่างข้อมูลของไมนอริตี้คลาสด้วยกัน โดยการประมวลผลของวิธีการ SMOTE สิ้นสุดเมื่อจำนวนข้อมูลของไมนอริตี้คลาสที่สร้างขึ้นใหม่มีจำนวนเท่ากับจำนวนที่ผู้ใช้กำหนด ผู้พัฒนาวิธีการ SMOTE ประยุกต์วิธีการนี้กับตัวแยกประเภท (Classifier) ได้แก่ C4.5 และ ริปเปอร์ (Ripper) และแสดงประสิทธิภาพของ SMOTE ในการทำนายข้อมูลของไมนอริตี้คลาสโดยเปรียบเทียบกับการใช้เทคนิคการสุ่มชักตัวอย่างแบบลด (Random under-sampling technique) และการใช้เพียงตัวแยกประเภทในการทำนายโดยไม่ใช้เทคนิคใดช่วย ผลการทดสอบแสดงให้เห็นว่าวิธีการ SMOTE สามารถทำนายข้อมูลของไมนอริตี้คลาสถูกต้องมากกว่าอีกสองวิธี

ในปี 2005 H. Han W. Y. Wan และ B. H. Mao นำวิธีการ SMOTE มาดัดแปลงเป็นวิธีการใหม่ชื่อว่า บอร์เดอร์ไลน์ SMOTE (Borderline SMOTE หรือ B-SMOTE) [8] พวกเขาให้นิยามกับข้อมูลของไมนอริตี้คลาส โดยการแบ่งข้อมูลของไมนอริตี้คลาสเป็น 3 ส่วนคือ

1. ส่วนสัญญาณรบกวน (Noise) หรือข้อมูลของไมนอริตี้คลาสที่มีข้อมูลที่อยู่ใกล้เคียงเป็นข้อมูลของมาจอริตี้คลาสเป็นจำนวนมาก
2. ส่วนปลอดภัย (Safe) หรือข้อมูลของไมนอริตี้คลาสที่มีข้อมูลที่อยู่ใกล้เคียงเป็นข้อมูลของไมนอริตี้คลาสเป็นจำนวนมาก

3. ส่วนแบ่งอาณาเขต (Borderline) หรือข้อมูลของไมนอริตี้คลาสที่มีข้อมูลที่อยู่ใกล้เคียงเป็นข้อมูลของไมนอริตี้คลาสและข้อมูลของมาจอริตี้คลาสในสัดส่วนใกล้เคียงกัน

วิธีการ B-SMOTE เพิ่มจำนวนข้อมูลของไมนอริตี้ด้วยวิธีการ SMOTE แต่ B-SMOTE เพิ่มจำนวนข้อมูลของไมนอริตี้คลาสที่อยู่บนส่วนแบ่งอาณาเขตเท่านั้น เพื่อเพิ่มความหนาแน่นของกลุ่มข้อมูลไมนอริตี้คลาสเฉพาะบริเวณที่ใกล้กับข้อมูลมาจอริตี้คลาส ผลการทดลองของพวกเขาแสดงให้เห็นว่าค่าพรีสิชัน (Precision) และค่า F (F-value) ของการทำนายข้อมูลไมนอริตี้คลาสด้วยวิธีการ B-SMOTE มีความแม่นยำกว่าวิธีการ SMOTE และวิธีการสุ่มชักตัวอย่างแบบเพิ่ม (Random over-sampling)

เซฟเลเวล SMOTE (Safe-level SMOTE หรือ SL-SMOTE) ถูกนำเสนอโดย C. Bunkhumpornpat K. Sinapiromsaran และ C. Lursinsap (2009) [9] SL-SMOTE ถูกดัดแปลงมาจากวิธีการ SMOTE พร้อมกับค่าระดับความปลอดภัย SL-SMOTE แบ่งระดับความปลอดภัยของข้อมูลไมนอริตี้คลาสออกเป็น 4 กรณีคือ กรณีที่ข้อมูลไมนอริตี้คลาสมีข้อมูลบริเวณใกล้เคียงเป็นข้อมูลไมนอริตี้คลาสทั้งหมด กรณีที่ข้อมูลไมนอริตี้คลาสมีข้อมูลบริเวณใกล้เคียงเป็นข้อมูลมาจอริตี้คลาสทั้งหมด กรณีที่ข้อมูลไมนอริตี้คลาสมีข้อมูลบริเวณใกล้เคียงเป็นข้อมูลไมนอริตี้คลาสจำนวนมากกว่าข้อมูลมาจอริตี้คลาส กรณีที่ข้อมูลไมนอริตี้คลาสมีข้อมูลบริเวณใกล้เคียงเป็นข้อมูลไมนอริตี้คลาสจำนวนน้อยกว่าข้อมูลมาจอริตี้คลาส เช่นเดียวกับวิธีการ B-SMOTE สร้างตัวสังเคราะห์ไมนอริตี้ระหว่างข้อมูลไมนอริตี้คลาสสองตัว แต่วิธีการ SL-SMOTE สร้างตัวสังเคราะห์ไมนอริตี้ให้อยู่บริเวณใกล้เคียงกับข้อมูลไมนอริตี้คลาสที่มีระดับปลอดภัยสูงกว่าข้อมูลไมนอริตี้คลาสอีกตัว จากผลการทดลอง ผู้วิจัยสรุปว่าค่าพรีสิชันและค่า F ของการทำนายข้อมูลของไมนอริตี้คลาสของวิธีการ SL-SMOTE มีความแม่นยำกว่าวิธีการ SMOTE และวิธีการ B-SMOTE

การศึกษาพฤติกรรมของหลายวิธีการเพื่อการปรับดุลชุดข้อมูลฝึกหัดสำหรับการเรียนรู้ของเครื่อง (A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data) ถูกนำเสนอโดย G. E. A. Batista R. C. Prati และ M. C. Monard (2004) [10] พวกเขานำวิธีการ SMOTE ผสานกับวิธีการโทเม็คลิงคส์ 2 วิธีคือ วิธีการ SMOTE+Tomek links และวิธีการ SMOTE+ENN สำหรับ SMOTE+Tomek links เริ่มโดยการใช้การสังเคราะห์แบบสุ่มเช่นเดียวกับวิธีการ SMOTE เพิ่มจำนวนข้อมูลของไมนอริตี้คลาส หลังจากนั้นใช้วิธีการโทเม็คลิงคส์ลบข้อมูลของมาจอริตี้คลาสที่อยู่ใกล้กับข้อมูลของไมนอริตี้คลาสและลบข้อมูลของไมนอริตี้คลาสที่อยู่ใกล้กับข้อมูลของมาจอริตี้คลาส วิธีการ SMOTE+ENN เริ่มโดยการใช้การสังเคราะห์แบบสุ่ม เช่นเดียวกับวิธีการ SMOTE เพิ่มจำนวนข้อมูลของไมนอริตี้คลาส หลังจากนั้นใช้วิธีการโทเม็คลิงคส์เช่นเดียวกับวิธีการแรกแต่เปลี่ยนจากการใช้เทคนิค 1 เนียเรสเนเบอร์เป็น 3 เนียเรสเนเบอร์ (3-nearest neighbor technique) แล้วเลือกลบข้อมูลของมาจอริตี้คลาสที่มีข้อมูลบริเวณ

ใกล้เคียงที่มีสัดส่วนของจำนวนข้อมูลของไมนอริตี้คลาสเท่ากับสองในสามของจำนวนข้อมูลทั้งหมดและลบข้อมูลของไมนอริตี้คลาสที่มีข้อมูลบริเวณใกล้เคียงเป็นสัดส่วนของข้อมูลของมาจอริตี้คลาสเท่ากับสองในสามของจำนวนข้อมูลทั้งหมด พวกเขาใช้ตัวแบบ C4.5 เป็นตัวแยกประเภทในการทดลองและสรุปผลการทดลองว่าเทคนิคการสุ่มชักตัวอย่างแบบเพิ่ม (Random over-sampling technique) ซึ่งเป็นวิธีพื้นฐานของการชักตัวอย่างแบบเพิ่มและไม่มีความซับซ้อนในการคำนวณมีประสิทธิภาพมากกว่าเทคนิคการชักตัวอย่างแบบเพิ่มที่มีความซับซ้อนในการคำนวณ

การเรียนรู้แบบคลาสเดียว (One-class learning) เป็นการสร้างตัวแบบโดยอิงจากข้อมูลของคลาสเป้าหมายเพียงคลาสเดียวซึ่งแตกต่างเทคนิคการชักข้อมูล เนื่องจากการชักข้อมูลเป็นกระบวนการจัดการข้อมูลก่อนการสร้างตัวแบบ ในขณะที่การเรียนรู้แบบคลาสเดียวเป็นกระบวนการระหว่างการสร้างตัวแบบ งานวิจัยที่ใช้เทคนิคการเรียนรู้แบบคลาสเดียวส่วนใหญ่ เป็นการนำเครื่องเวกเตอร์เกื้อหนุน (Support vector machines หรือ SVM) มาดัดแปลงให้เหมาะกับการแก้ปัญหาความไม่ดุลระหว่างกลุ่ม

วิธีการ SVM เอนเซมเบิลหลายวิธีผสานกับวิธีการชักตัวอย่างแบบลดสำหรับการแก้ปัญหาความไม่ดุลระหว่างกลุ่ม (Several SVM Ensemble Methods Integrated with Under-Sampling for Imbalanced Data Learning) ถูกนำเสนอโดย Z. Y. Lin Z. F. Hao X. W. Yang และ X. L. Liu (2009) [11] ผู้วิจัยใช้เทคนิคแบ็กกิง (Bagging) และเทคนิคบูสทิงซึ่งเป็นวิธีการเอนเซมเบิลในการประยุกต์กับ SVM 4 วิธีคือ วิธีการแบ็กกิงแบบอสมมาตรธรรมดา (Normal Asymmetric Bagging Ensemble หรือ NABagE) วิธีการแบ็กกิงโดยการแบ่งกลุ่ม (Cluster Based Asymmetric Bagging Ensemble หรือ CBABagE) วิธีการบูสทิงแบบอสมมาตรธรรมดา (Normal Asymmetric Boosting Ensemble หรือ NABstE) และวิธีการบูสทิงแบบอสมมาตรดัดแปร (Modified Asymmetric Boosting Ensemble หรือ MABstE) โดย 2 วิธีแรกเป็นการใช้เทคนิคแบ็กกิงสุ่มเลือกกลุ่มข้อมูลของมาจอริตี้จากจำนวนข้อมูลของมาจอริตี้คลาสทั้งหมด โดยมีจำนวนข้อมูลของมาจอริตี้คลาสที่ถูกสุ่มมาตามจำนวนที่ผู้ใช้กำหนด และสร้างตัวแบบจากกลุ่มข้อมูลของมาจอริตี้คลาสที่ถูกสุ่มมากับข้อมูลของไมนอริตี้คลาส ทำวิธีการเดิมไปเรื่อยๆตามจำนวนการทำซ้ำที่ผู้ใช้กำหนด แล้วนำตัวแบบทั้งหมดมาคำนวณใหม่เป็นตัวแบบที่เหมาะสมที่สุด ส่วนวิธีการ NABstE และ MABstE ใช้วิธีการแบ่งกลุ่มข้อมูลของมาจอริตี้แบบเดียวกับวิธีการแรก แต่ใช้เทคนิคบูสทิงโดยการให้ค่าถ่วงน้ำหนัก (Weight) กับข้อมูลของมาจอริตี้คลาส เพื่อทำการปรับค่าถ่วงน้ำหนักของข้อมูลของมาจอริตี้คลาสไปเรื่อยๆจนกระทั่งได้กลุ่มข้อมูลของมาจอริตี้คลาสที่มีความสำคัญในการสร้างตัวแบบ ทำวิธีการเดิมไปเรื่อยๆตามจำนวนการทำซ้ำที่ผู้ใช้กำหนด แล้วนำตัวแบบทั้งหมดมาคำนวณใหม่เป็นตัวแบบที่เหมาะสมที่สุด จากการทดลองของ

ผู้วิจัย พวกเขาสรุปว่าวิธีการ MABstE เป็นวิธีการที่ดีที่สุด เนื่องจากวิธีการนี้สามารถหลีกเลี่ยงปัญหาความจำเพาะเกิน (Over fitting) ของตัวแบบ

การคัดเลือกลักษณะเฉพาะ (Feature selection) เป็นการเลือกลักษณะประจำ (Attribute) ที่มีความสำคัญสำหรับการแยกประเภทข้อมูล เพื่อลดจำนวนของลักษณะประจำที่ไม่จำเป็นในการสร้างตัวแยกประเภท เนื่องจากการสร้างตัวแบบจากชุดข้อมูลที่มีจำนวนของลักษณะประจำมากส่งผลให้สูญเสียเวลาในการสร้างตัวแบบมาก และมีความแม่นยำในการทำนายต่ำ มีงานวิจัยหลายแขนงใช้การคัดเลือกลักษณะเฉพาะเพื่อแก้ปัญหาความไม่ดุลระหว่างกลุ่มเช่น การประมวลผลภาพ (Image processing) การแยกประเภทข้อความ (Text classification) ชีวสารสนเทศศาสตร์ (Bioinformatics) เป็นต้น

ในปี 2004 Z. Zheng X. wu และ R. Srihari นำเสนอการคัดเลือกลักษณะเฉพาะสำหรับการจำแนกข้อความบนข้อมูลที่มีความไม่ดุลระหว่างกลุ่ม (Feature Selection for Text Categorization on Imbalanced Data) [12] โดยใช้ตัวคัดเลือกลักษณะเฉพาะ 6 ตัว ได้แก่ ค่าการเพิ่มคุณค่าของข้อมูล (Information gain) ไคกำลังสอง (Chi-square) สัมประสิทธิ์สหสัมพันธ์ (Correlation coefficient) อัตราส่วนออดส์ (Odds ratio) อัตราส่วนออดส์กำลังสอง (OR-square หรือ ORS) และ ค่าการเพิ่มคุณค่าของข้อมูลด้วยเครื่องหมาย (Signed IG หรือ SIG) โดย 2 วิธีการหลังเป็นวิธีการที่ผู้วิจัยนำเสนอ วิธีการของพวกเขาเริ่มโดยการกำหนดจำนวนของลักษณะประจำที่จะใช้ทั้งหมดในการสร้างตัวแบบ และกำหนดจำนวนของลักษณะประจำของไมนอริตี้คลาส คัดเลือกลักษณะประจำที่มีความสำคัญของไมนอริตี้คลาสและมาจอริตี้คลาสด้วยตัวคัดเลือกลักษณะเฉพาะทั้ง 6 ตัว โดยการเลือกจำนวนลักษณะประจำที่มีความสำคัญของไมนอริตี้คลาสตามจำนวนที่ผู้ใช้กำหนด และเลือกจำนวนลักษณะประจำที่มีความสำคัญของมาจอริตี้คลาสเท่ากับจำนวนของลักษณะประจำทั้งหมดที่ผู้ใช้กำหนดลบด้วยจำนวนของลักษณะประจำของไมนอริตี้คลาส หลังจากนั้นจึงนำลักษณะประจำที่ถูกเลือกมาใช้สร้างตัวแบบแยกประเภทด้วยการแบ่งประเภทเบย์อย่างง่าย (Naïve Bayes) และ การวิเคราะห์ถดถอยลอจิสติก (Logistic regression) จากการทดลองของผู้วิจัย พวกเขาสรุปว่าตัวคัดเลือกลักษณะเฉพาะ ORS และ SIG สามารถคัดเลือกลักษณะประจำที่ใช้สร้างตัวแบบแยกประเภทที่มีประสิทธิภาพมากที่สุด

ในงานวิจัยนี้เราใช้วิธีการชักข้อมูลสำหรับการแก้ปัญหาความไม่ดุลระหว่างกลุ่ม โดยใช้วิธีการผสานกันของเทคนิคการชักตัวอย่างแบบเพิ่มกับเทคนิคการชักตัวอย่างแบบลด สำหรับเทคนิคการชักตัวอย่างแบบเพิ่ม เราใช้วิธีการเดียวกับ SMOTE เหมือนกับวิธีการ SMOTE+Tomek links และวิธีการ SMOTE+ENN แต่เราลบเฉพาะข้อมูลมาจอริตี้คลาสโดยพยายามหลีกเลี่ยงการลบข้อมูลของมาจอริตี้คลาสบริเวณใกล้กับข้อมูลไมนอริตี้คลาส และนำชุดข้อมูลที่ได้ไปสร้างตัวแบบแยกประเภท ด้วยวิธีการดังกล่าวจะสามารถรักษาความแม่นยำในการทำนายข้อมูลของไมนอ

ริตี้คลาสและข้อมูลของมาจอริตี้คลาส และสามารถลดขนาดของชุดข้อมูลซึ่งเป็นการลดระยะเวลาในการประมวลผลสำหรับการสร้างตัวแบบแยกประเภทอีกด้วย

1.2 วัตถุประสงค์ของการวิจัย

ผู้วิจัยต้องการพัฒนาเทคนิคการสุ่มเพิ่มตัวอย่างข้างน้อยสังเคราะห์และเทคนิคการสุ่มลดตัวอย่างข้างมากสำหรับปัญหาความไม่ดุลระหว่างกลุ่มโดยการสร้างตัวจำลองให้มีขนาดเล็กกว่า SMOTE และมีความแม่นยำในการทำนายข้อมูลของไมนอริตี้คลาสใกล้เคียงกันหรือมากกว่า SMOTE

1.3 ประโยชน์ที่คาดว่าจะได้รับ

สามารถแก้ปัญหาความไม่ดุลระหว่างกลุ่มด้วยการสร้างตัวจำลองที่มีขนาดเล็กกว่าวิธีการ SMOTE และมีความแม่นยำใกล้เคียงกัน

1.4 ลำดับขั้นตอนในการเสนอผลการวิจัย

ในบทที่ 2 เรากล่าวถึงข้อตกลงที่ใช้ในงานวิจัยนี้ นิยามของคลาสเป้าหมาย และวิธีการที่ถูกนำไปประยุกต์ใช้ในการพัฒนางานวิจัยนี้ รวมถึงวิธีการที่ใช้ในการสร้างตัวแบบ และวิธีการทดสอบการประสิทธิภาพของตัวแบบ

ในบทที่ 3 เรากล่าวถึงรายละเอียด ขั้นตอนวิธี และตัวอย่างการประมวลผลของงานวิจัยนี้

ในบทที่ 4 เราแสดงผลการทดสอบประสิทธิภาพ และความรวดเร็วในการประมวลผลของขั้นตอนวิธี SMOUTE โดยการเปรียบเทียบกับวิธีการ SMOTE

ในบทสุดท้าย เรากล่าวถึงข้อสรุป ปัญหาและงานวิจัยที่น่าสนใจในอนาคต

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

ในงานวิจัยนี้เราพิจารณาเฉพาะชุดข้อมูลที่มีคลาสเป้าหมาย (Target class) สองคลาส และมีลักษณะประจำเชิงตัวเลข (Numerical attributes) เท่านั้น คลาสที่มีจำนวนของข้อมูลขนาดเล็กกว่าถูกเรียกว่าไมนอริตี้คลาส (Minority class) หรือคลาสบวก (Positive class) สำหรับคลาสที่มีขนาดใหญ่กว่าถูกเรียกว่ามาจอริตี้คลาส (Majority class) หรือคลาสลบ (Negative class)

2.1 เทคนิคการชักตัวอย่างแบบเพิ่มและเทคนิคการชักตัวอย่างแบบลด

2.1.1 เทคนิคการชักตัวอย่างแบบเพิ่ม (Over-sampling technique)

เทคนิคการชักตัวอย่างแบบเพิ่ม คือ การเพิ่มจำนวนของข้อมูลที่เป็นคลาสบวกจนกระทั่งจำนวนข้อมูลของคลาสบวกเท่ากับจำนวนที่ผู้ใช้กำหนดไว้ หลักการพื้นฐานของการชักตัวอย่างแบบเพิ่มคือการสุ่มชักตัวอย่างแบบเพิ่ม (Random over-sampling) ซึ่งเป็นการสร้างข้อมูลของคลาสบวกโดยสุ่มเลือกข้อมูลของคลาสบวก และทำการเพิ่มจำนวนข้อมูลของคลาสบวกที่ตำแหน่งใกล้เคียงกับข้อมูลของคลาสบวกที่ถูกสุ่มเลือกมา ทำเช่นนี้ไปเรื่อยๆ จนกระทั่งจำนวนของคลาสบวกเท่ากับจำนวนที่ผู้ใช้กำหนด [11]

2.1.2 เทคนิคการชักตัวอย่างแบบลด (Under-sampling technique)

เทคนิคการชักตัวอย่างแบบลด คือ การลดจำนวนของข้อมูลที่เป็นคลาสลบจนกระทั่งจำนวนข้อมูลของคลาสลบเท่ากับจำนวนที่ผู้ใช้กำหนด หลักการพื้นฐานของเทคนิคการชักตัวอย่างแบบลดคือการสุ่มชักตัวอย่างแบบลด (Random under-sampling) ซึ่งเป็นการลบข้อมูลของคลาสลบโดยสุ่มเลือกข้อมูลของคลาสลบ ทำเช่นนี้ไปเรื่อยๆ จนกระทั่งจำนวนข้อมูลของคลาสลบมีเท่ากับจำนวนที่ผู้ใช้กำหนด [11]

ทั้งเทคนิคการชักตัวอย่างแบบเพิ่มและเทคนิคการชักตัวอย่างแบบลดเป็นกระบวนจัดการข้อมูลก่อนการสร้างตัวแบบ โดยงานวิจัยนี้ เราใช้ทั้งสองเทคนิคเป็นแนวคิดพื้นฐานสำหรับการแก้ปัญหาความไม่ดุลระหว่างกลุ่ม

2.2 **เทคนิคการชักตัวอย่างสังเคราะห์ไมนอริตี้แบบเพิ่ม** (Synthetic Minority Over-sampling TEchnique **หรือ** SMOTE)

เทคนิคการชักตัวอย่างสังเคราะห์ไมนอริตี้แบบเพิ่ม (SMOTE) [1] คือกระบวนจัดการข้อมูลก่อนการสร้างตัวแบบสำหรับการแก้ปัญหาความไม่ดุลระหว่างกลุ่ม วิธีการ SMOTE ใช้ขั้นตอนวิธี k เนียเรสเนเบอร์ (K-nearest neighbor algorithm) [13] เพื่อการสังเคราะห์ข้อมูลของคลาสบวกที่เพิ่มขึ้นมาจนได้ชุดข้อมูลใหม่ และนำชุดข้อมูลที่ได้ไปใช้เป็นชุดข้อมูลฝึกหัดสำหรับการสร้างตัวแบบแยกประเภท

ขั้นตอนวิธีของ SMOTE เริ่มโดยผู้ใช้กำหนดจำนวนข้อมูลของคลาสบวกที่ต้องการสร้างเป็นจำนวน T ตัว หาข้อมูลของคลาสบวกจำนวน k ตัวที่ใกล้ที่สุดกับข้อมูลของคลาสบวกแต่ละตัว เลือกข้อมูลของคลาสบวกหนึ่งตัวมาเป็นข้อมูลของคลาสบวกที่ถูกพิจารณา ทำการสุ่มเลือกข้อมูลของคลาสบวกที่อยู่ใกล้กับข้อมูลของคลาสบวกที่ถูกพิจารณามาหนึ่งตัว และสร้างตัวสังเคราะห์โดยสุ่มตำแหน่งบนเส้นระยะทางระหว่างข้อมูลทั้งสองตัว ทำการสร้างตัวสังเคราะห์ซ้ำไปเรื่อยๆจนครบ T ครั้ง แล้วทำการเปลี่ยนข้อมูลของคลาสบวกที่ถูกพิจารณา และทำการสร้างตัวสังเคราะห์เป็นจำนวน Tครั้ง ทำซ้ำไปเรื่อยๆจนกระทั่งข้อมูลของคลาสบวกทั้งหมดถูกใช้เป็นข้อมูลของคลาสบวกที่ถูกพิจารณาจึงสิ้นสุดการประมวลผล

ขั้นตอนวิธีของ SMOTE มีดังนี้

Let *N* = (*int*)(*N* /100)(* *The amount of SMOTE must be integer from multiples of 100.* *)

k = Number of nearest neighbors

numattrs = Number of attributes

Sample[ ][ ]: array for original minority class instances

newindex: keeps a count of number of synthetic instances generated, initialized to 0

Synthetic[ ][ ]: array for synthetic instances (* *Compute k nearest neighbors for each minority class sample only.* *)

dif = Distant between the pair of the minority class instances

gap = Random number between 0 and 1

Algorithm SMOTE(*T, N, k*)

Input: Number of minority class instances *T*; Amount of SMOTE *N* %; Number of the nearest neighbors *k*

Output: (*N* /100) * *T* synthetic minority class instances

1. (* *If N is less than 100%, randomize the minority class instances as only a random percent of them will be SMOTEd.* *)
2. if (*N* <100) {
3. then Randomize the *T* minority class instances
4. *T* = (*N* /100) * *T*
5. *N* = 100
6. }
7. For *i* ⟵ 1 to *T* {
8. Compute *k* nearest neighbors for *i*, and save the indices in the *nnarray*
9. Populate(*N*, *i*, *nnarray*)
10. }

Populate(*N, i, nnarray*) (* *Function to generate the synthetic instances.* *)

1. while(*N* == 0) {
2. Choose a random number between 1 and *k*, call it *nn*. This step chooses one of the *k* nearest neighbors of *i*.
3. For *attr* ⟵ 1 to *numattrs* {
4. Compute: *dif* = *Sample*[*nnarray*[*nn*]][*attr*] − *Sample*[*i*][*attr*]
5. Compute: *gap*
6. *Synthetic*[*newindex*][*attr*] = *Sample*[*i*][*attr*] + *gap* * *dif*
7. }
8. *newindex*++
9. *N* = *N* − 1
10. }
11. return (* *End of Populate.* *)

2.3 ขั้นตอนวิธีค่าเฉลี่ย k (K-means clustering algorithm)

การเกาะกลุ่มข้อมูล (Clustering) เป็นการเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning) โดยรวมข้อมูลที่มีลักษณะประจำใกล้เคียงกันออกเป็นกลุ่มและนำกลุ่มข้อมูลที่ได้ไปวิเคราะห์ มีวิธีการหลายวิธีในการวัดความใกล้เคียงกันของชุดข้อมูล โดยงานวิจัยนี้ใช้วิธีการเกาะ

กลุ่มข้อมูลแบบไม่แบ่งลำดับชั้น (Non-hierarchical clustering) ซึ่งเป็นวิธีการแบ่งข้อมูลออกเป็นกลุ่มตามจำนวนที่ผู้ใช้กำหนด และแต่ละกลุ่มไม่มีข้อมูลทับซ้อนกัน [1]

ขั้นตอนวิธีค่าเฉลี่ย k เป็นขั้นตอนวิธีที่นิยมใช้สำหรับการแก้ปัญหาการเกาะกลุ่มแบบไม่แบ่งลำดับชั้น เนื่องจากขั้นตอนวิธีค่าเฉลี่ย k ใช้ระยะเวลาในการคำนวณน้อยกว่าขั้นตอนวิธีสำหรับการเกาะกลุ่มข้อมูลแบบอื่น [13] ขั้นตอนวิธีค่าเฉลี่ย k เริ่มต้นโดยผู้ใช้กำหนดจำนวนกลุ่มข้อมูลที่ต้องการแบ่ง และสุ่มเลือกข้อมูลมาเป็นเซนทรอยด์ (Centroid) จากชุดข้อมูล จำนวนเซนทรอยด์ที่ถูกสุ่มมีเท่ากับจำนวนของกลุ่มที่ผู้ใช้ต้องการแบ่ง ข้อมูลแต่ละตัวในชุดข้อมูลถูกแบ่งไปอยู่กลุ่มเดียวกับเซนทรอยด์ที่อยู่ใกล้ที่สุด โดยใช้ฟังก์ชันระยะทาง (Distance function) เช่น การวัดระยะแบบยูคลิด (Euclidean distance) การวัดระยะแบบแมนฮัตตัน (Manhattan distance) การวัดระยะแบบเชบิเชฟ (Chebychev distance) [14] หลังจากการแบ่งกลุ่ม เซนทรอยด์แต่ละตัวถูกคำนวณใหม่โดยการหาค่าเฉลี่ยจากข้อมูลทุกตัวในกลุ่ม การเกาะกลุ่มข้อมูลใหม่และการคำนวณเซนทรอยด์ใหม่ถูกทำซ้ำไปเรื่อยๆจนกระทั่งไม่มีข้อมูลตัวใดเปลี่ยนกลุ่ม หรือจำนวนรอบการทำซ้ำเท่ากับจำนวนที่ผู้ใช้กำหนดไว้

Algorithm *K*-means(*K*, *T*)

Input: The number of the instances groups *K*, the number of the iterations *T*

Output: The negative instances are separated into *K*-groups.

1. *N* = the number of negative instances
2. Random choose the *K* initial centroids
3. *flag* = 1;
4. While(*i* != *T* and *flag* != 0){
5. *flag* = 0;
6. For each insatance-*i* in N{
7. assign the instance-*i* to the group with the closest centroid.
8. *i* = *i*+1;
9. if instance-*i* is changed the group then *flag* = 0;
10. }
11. }

ในงานวิจัยนี้เราใช้ขั้นตอนวิธีค่าเฉลี่ย k เพื่อแบ่งกลุ่มข้อมูลของคลาสลบ โดยเลือกค่า k ด้วยการวิเคราะห์องค์ประกอบหลัก (Principle component analysis) และลบข้อมูลของคลาสลบ

ที่อยู่บริเวณใกล้เคียงกับแต่ละเซนทรอยด์ โดยเหลือข้อมูลของคลาสลบบริเวณขอบของแต่ละกลุ่มสร้างตัวแบบเท่านั้น

2.4 **ตัวแยกประเภท** (The classifiers)

ตัวแยกประเภทคือตัวแบบสำหรับการแยกประเภทข้อมูลตามคลาสเป้าหมายที่ผู้ใช้กำหนด โดยการนำชุดข้อมูลฝึกหัดมาสร้างเป็นตัวแบบ และทดสอบประสิทธิภาพของตัวแบบด้วยชุดข้อมูลทดสอบ มีนักวิจัยได้คิดค้นแนวคิดในการสร้างตัวแบบหลากหลายวิธีเช่น ต้นไม้การตัดสิน (Decision tree) การแบ่งประเภทเบย์อย่างง่าย (Naïve Bayes) โครงข่ายประสาทเทียม (Neural network) เครื่องเวกเตอร์เกื้อหนุน (Support vector machines หรือ SVM) การวิเคราะห์จำแนกประเภท (Discriminant analysis) เป็นต้น ในงานวิจัยนี้ เราใช้ตัวแยกประเภท 3 แบบในการประยุกต์กับเทคนิคการสุ่มเพิ่มตัวอย่างข้างน้อยสังเคราะห์และเทคนิคการสุ่มลดตัวอย่างข้างมากสำหรับปัญหาความไม่ดุลระหว่างกลุ่มคือ C4.5 การแบ่งประเภทเบย์อย่างง่าย (Naïve Bayes) และเพอร์เซ็ปตรอนหลายชั้น (Multilayer perceptron)

2.4.1 C4.5

C4.5 เป็นหนึ่งในวิธีการสร้างตัวแบบโดยใช้ต้นไม้การตัดสินใจ (Decision tree) ตามลักษณะประจำของข้อมูลในการจำแนกประเภท C4.5 เริ่มต้นโดยการเลือกลักษณะประจำเพื่อแยกประเภทข้อมูลออกเป็นบัพ (node) โดยการคำนวณค่าเอนโทรปี (Entropy) หรือ $\mathrm{H}(D)$ ดังสมการ 2.1

$$H(D) = -\sum_{i=1}^{c} p(x_i) \times \log_2 p(x_i) \tag{2.1}$$

$p(x_i)$ คือ อัตราส่วนของจำนวนข้อมูลของคลาสที่ i ต่อจำนวนทั้งหมด
c คือ จำนวนคลาสทั้งหมด
D คือ ชุดข้อมูลฝึกหัดที่นำมาสร้างตัวแบบ C4.5

ค่าเอนโทรปีของลักษณะประจำแต่ละค่าถูกคำนวณจากสมการ 2.2

$$H_A(D) = \sum_{j=1}^{n} \frac{|D_j|}{|D|} H(D_j) \tag{2.2}$$

A คือ ชื่อของลักษณะประจำ

n คือ จำนวนค่าความแตกต่างของข้อมูลในลักษณะประจำ A

D_j คือ จำนวนข้อมูลชนิดที่มีค่าความแตกต่างเป็นตัวที่ j ของลักษณะประจำ A

และคำนวณหาค่าการเพิ่มค่าของข้อมูล (Information gain) ดังสมการ 2.3 โดยลักษณะประจำที่มีค่าการเพิ่มค่าของข้อมูลสูงที่สุดถูกใช้เป็นตัวแยกประเภทข้อมูล

$$Gain(A) = H(D) - H_A(D) \tag{2.3}$$

การคำนวณค่าเอนโทรปีทำได้กับลักษณะประจำที่มีข้อมูลแบบเป็นกลุ่ม (Categorical) เท่านั้น สำหรับลักษณะประจำที่มีข้อมูลแบบต่อเนื่อง (Continuous) การคำนวณค่าเอนโทรปีเริ่มต้นโดยการเรียงลำดับของข้อมูล และใช้ค่ากึ่งกลางระหว่างข้อมูลแต่ละตัวมาแบ่งด้วยเครื่องหมายมากกว่า น้อยกว่าและเท่ากับ เพื่อแยกข้อมูลออกเป็นกลุ่ม คำนวณค่าเอนโทรปีของข้อมูลแต่ละกลุ่มเพื่อหาค่าที่น้อยที่สุด และนำไปเปรียบเทียบกับเอนโทรปีของลักษณะประจำอื่น

หลังจากการคัดแยกประเภทข้อมูลออกเป็นหลายบัพแล้ว แต่ละบัพหยุดการคำนวณค่าเอนโทรปีก็ต่อเมื่อข้อมูลในบัพนั้นมีคลาสเป้าหมาย (Target class) เป็นชนิดเดียวกันทั้งหมด หรือมีจำนวนข้อมูลของคลาสเป้าหมายกลุ่มใดกลุ่มหนึ่งน้อยเกินไป หรือความสูงของต้นไม้เกินค่าที่ผู้ใช้กำหนด ซึ่งเราเรียกว่า ใบ (Leaf node)

ตัวอย่าง 2.1 ชุดข้อมูลตัวอย่างในการสร้างแบบจำลองต้นไม้การตัดสินใจ มีจำนวนข้อมูลทั้งหมด 10 ตัว จำนวนลักษณะประจำ 3 ลักษณะประจำคือผู้ครองสิทธิที่อยู่อาศัย (Home Owner) สถานภาพ (Marital Status) และรายได้ต่อปี (Annual income) มีคลาสเป้าหมาย 2 คลาสคือผู้กู้ยืมที่ผิดสัญญา (Defaulted borrower=Yes) จำนวน 3 คน และผู้กู้ยืมที่ไม่ผิดสัญญา (Defaulted borrower=No) จำนวน 7 คน โดยตัวอย่าง 2.1 เราใช้การแบ่งข้อมูลแบบต้นไม้ทวิภาค (binary tree) คือการแบ่งบัพของต้นไม้ออกเป็น 2 บัพต่อการแบ่งข้อมูลหนึ่งครั้ง

ตาราง 2.1 ชุดข้อมูลฝึกหัดสำหรับตัวอย่าง 2.1

TID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	High	No
2	No	Married	Medium	No
3	No	Single	Low	No
4	Yes	Married	High	No
5	No	Divorced	Medium	Yes
6	No	Married	Low	No
7	Yes	Divorced	High	No
8	No	Single	Medium	Yes
9	No	Married	Low	No
10	No	Single	Medium	Yes

คำนวณค่าเอนโทรปีของชุดข้อมูลด้วยสมการ 2.1

$$\mathrm{H}(D) = -(7/10)(\log_2(7/10))-(3/10)/(\log_2(3/10)) = 0.8813$$

คำนวณค่าการเพิ่มค่าของข้อมูลของลักษณะประจำทุกตัว

$$\mathrm{H}_{\mathrm{Owner}}(D) = (3/10)[0-(3/3)(\log_2(3/3))]+(7/10)[-(3/7)(\log_2(3/7))-(4/7)(\log_2(4/7))] = 0.6897$$

$$\mathrm{Gain}_{\mathrm{Owner}} = 0.8813-0.6897 = 0.1916$$

ในกรณีของลักษณะประจำสถานภาพ เราแบ่งข้อมูลเป็น 2 กลุ่มคือ กลุ่มโสดและหม้าย กับกลุ่มสมรส

$$\mathrm{H}_{\mathrm{Status}}(D) = (4/10)[0-(4/4)(\log_2(4/4))]+(6/10)[-(3/6)(\log_2(3/6))-(3/6)(\log_2(3/6))] = 0.5$$

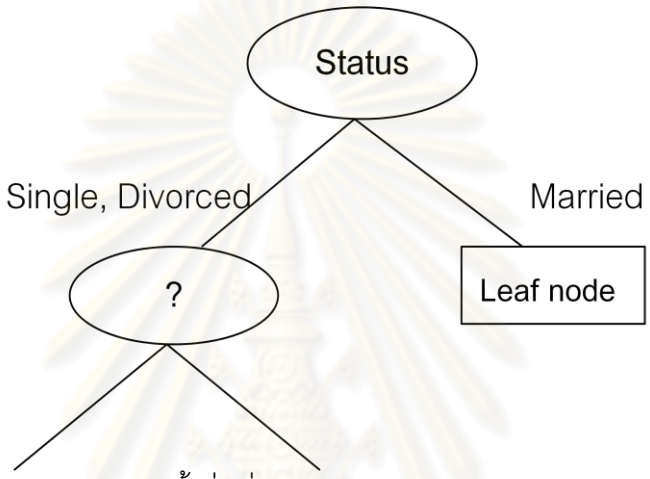
$$\mathrm{Gain}_{\mathrm{Status}} = 0.8813-0.5 = 0.3813$$

ในกรณีของลักษณะประจำรายได้ต่อปี เราแบ่งข้อมูลเป็น 2 กลุ่มคือ กลุ่มรายได้ต่ำและรายได้ปานกลาง กับกลุ่มรายได้สูง

$H_{Income}(D) = (3/10)[0-(3/3)(\log_2(3/3))]+(7/10)[-(3/7)(\log_2(3/7))-(4/7)(\log_2(4/7))] = 0.6897$

$Gain_{Income} = 0.8813-0.6897 = 0.1916$

ค่าการเพิ่มค่าของข้อมูลของลักษณะประจำสถานภาพมีค่าสูงที่สุด ดังนั้นเราเลือกลักษณะประจำสถานภาพเป็นตัวแบ่งข้อมูล



รูป 2.1 การแบ่งบัพครั้งที่หนึ่งของต้นไม้การตัดสินใจของตัวอย่าง 2.1

ตาราง 2.2 การแบ่งบัพครั้งที่หนึ่งของต้นไม้การตัดสินใจของตัวอย่าง 2.1

TID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	High	No
2	No	Married	Medium	No
3	No	Single	Low	No
4	Yes	Married	High	No
5	No	Divorced	Medium	Yes
6	No	Married	Low	No
7	Yes	Divorced	High	No
8	No	Single	Medium	Yes
9	No	Married	Low	No
10	No	Single	Medium	Yes

TID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	High	No
3	No	Single	Low	No
5	No	Divorced	Medium	Yes
7	Yes	Divorced	High	No
8	No	Single	Medium	Yes
10	No	Single	Medium	Yes

TID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
2	No	Married	Medium	No
4	Yes	Married	High	No
6	No	Married	Low	No
9	No	Married	Low	No

จากรูป 2.1 บัพทางขวามือมีคลาสเป้าหมายเป็นคลาสเดียวกันทั้งหมด ดังนั้นเราจะแบ่งบัพทางซ้ายมือเท่านั้น หาค่าเอนโทรของชุดข้อมูลในบัพทางซ้ายมือ และคำนวณค่าการเพิ่มค่าของข้อมูลของลักษณะประจำทุกตัว

$$H(D) = -(3/6)(\log_2(3/6))-(3/6)(\log_2(3/6)) = 1$$

$$H_{Owner}(D) = (2/6)(0)+(4/6)(0.8113) = 0.5409$$

$$Gain_{Owner} = 1-0.5409 = 0.4591$$

$$H_{Status}(D) = (3/6)(0.9184)+(3/6)(0.9184)$$

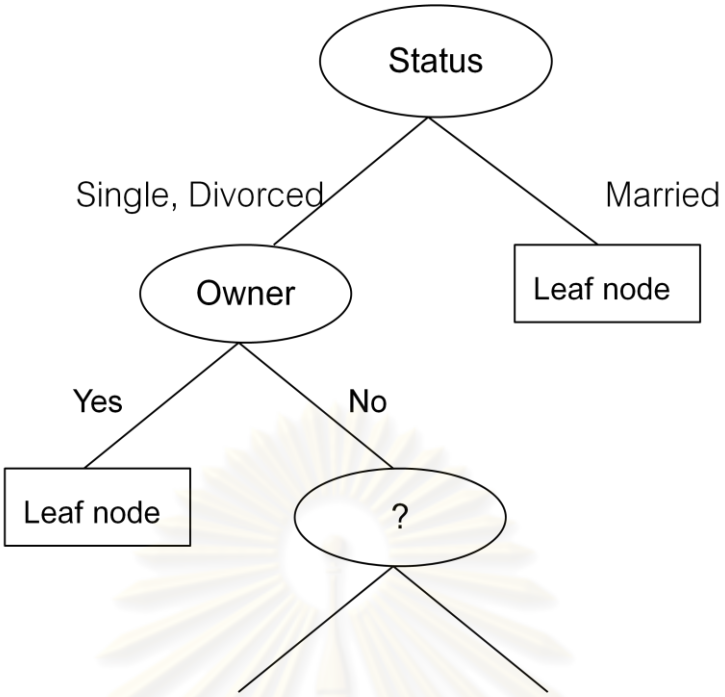
$$Gain_{Status} = 1-0.9184 = 0.0816$$

ในกรณีของลักษณะประจำรายได้ต่อปี เราแบ่งข้อมูลเป็น 2 กลุ่มคือ กลุ่มรายได้ต่ำและรายได้ปานกลาง กับกลุ่มรายได้สูง

$$H_{Income}(D) = (4/6)(0.8113)+(2/6)(1) = 0.8733$$

$$Gain_{Income} = 1-0.8733 = 0.1267$$

เลือกลักษณะประจำผู้ครองสิทธิที่อยู่อาศัยเป็นตัวแบ่งข้อมูล



รูป 2.2 การแบ่งบัพครั้งที่สองของต้นไม้การตัดสินใจของตัวอย่าง 2.1

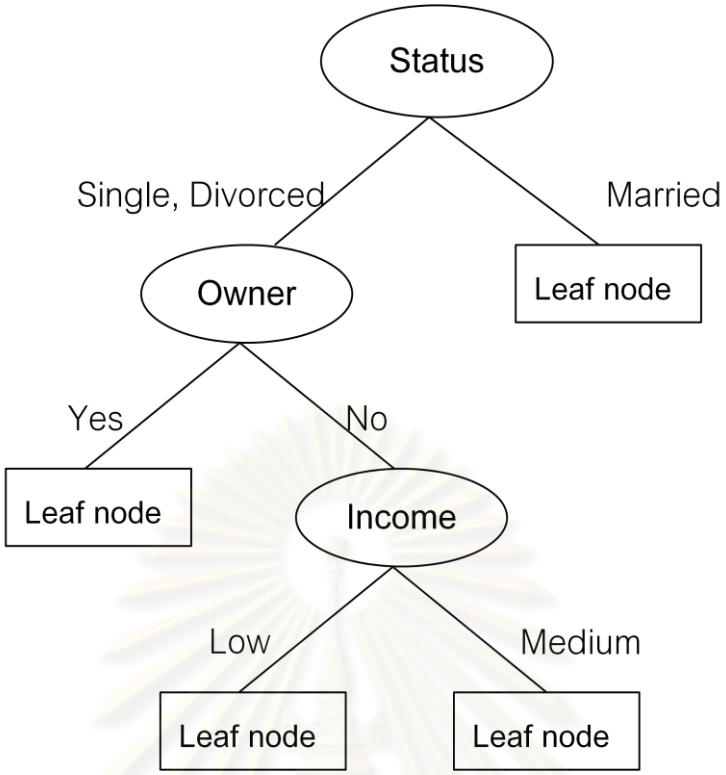
จากรูป 2.2 บัพทางซ้ายมือมีคลาสเป้าหมายเป็นคลาสเดียวกันทั้งหมด ดังนั้นเราจะแบ่งบัพทางขวามือเท่านั้น และลักษณะประจำผู้ครองสิทธิที่อยู่อาศัยในบัพทางขวามือมีเพียงชนิดเดียว จึงไม่สามารถใช้ในการแบ่งข้อมูลได้อีก ดังนั้นเราจะคำนวณค่าการเพิ่มค่าของข้อมูลของลักษณะประจำสถานภาพและลักษณะประจำรายได้ต่อปีเท่านั้น

$\mathrm{H}(D) = 0.8113$

$\mathrm{Gain}_{\mathrm{Status}} = 0.8113\text{-}0.6888 = 0.1225$

$\mathrm{Gain}_{\mathrm{Income}} = 0.8113\text{-}0 = 0.8113$

เลือกลักษณะประจำรายได้ต่อปีเป็นตัวแบ่งข้อมูล



รูป 2.3 การแบ่งบัพครั้งที่สามของต้นไม้การตัดสินใจของตัวอย่าง 2.1

จากรูป 2.3 การสร้างตัวแบบ C4.5 เสร็จสิ้นเนื่องจากทั้งสองบัพมีคลาสเป้าหมายภายในบัพเป็นคลาสเดียวกันทั้งหมด

2.4.2 การแบ่งประเภทเบย์อย่างง่าย (Naïve Bayes)

การแบ่งประเภทเบย์อย่างง่ายใช้หลักความน่าจะเป็นของทฤษฎีของเบย์ (Bayes' Theorem) ดังสมการ 2.4 และสมมติให้เหตุการณ์ (ลักษณะประจำ) ทุกแบบเป็นอิสระต่อกัน [13]

$$P(H|E) = \frac{[P(E|H) \times P(H)]}{P(E)} \quad (2.4)$$

P(H) คือ ความน่าจะเป็นที่จะเกิดเหตุการณ์ H

P(H|E) คือ คือความน่าจะเป็นที่จะเกิดเหตุการณ์ H เมื่อเหตุการณ์ E เกิดขึ้นแล้ว

P(E|H) คือ คือความน่าจะเป็นที่จะเกิดเหตุการณ์ E เมื่อเหตุการณ์ H เกิดขึ้นแล้ว

เมื่อพิจารณาทฤษฎีของเบย์ พร้อมสมบัติเป็นอิสระต่อกันที่กล่าวมาในตอนต้น เราสามารถแสดงการคัดแยกประเภทข้อมูลที่มีเหตุการณ์มากกว่า 1 เหตุการณ์ ได้ดังสมการต่อไปนี้

$$P(H|E_1, E_2, \dots, E_n) = \frac{P(E_1|H) \times P(E_2|H) \times \dots \times P(E_n|H) \times P(H)}{P(E_1, E_2, \dots, E_n)} \quad (2.5)$$

เมื่อ $E_1, E_2, \dots, E_n$ คือเหตุการณ์ที่เป็นอิสระต่อกัน

ตัวอย่าง 2.2 คือชุดข้อมูลตัวอย่างการสร้างตัวแบบการแบ่งประเภทเบย์อย่างง่าย โดยทำนายการตีเทนนิส จากเหตุการณ์ 4 เหตุการณ์คือ สภาพอากาศ (outlook) อุณหภูมิ (temperature), ความชื้น (humidity) และแรงลม (windy) ดังตาราง 2.7

ตาราง 2.3 ชุดข้อมูลฝึกหัดสำหรับตัวอย่าง 2.2

Outlook	**Temperature**	**Humidity**	**Windy**	**Play Tennis**
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cold	normal	false	yes
rainy	cold	normal	true	no
overcast	cold	normal	true	yes
sunny	mild	high	false	no
sunny	cold	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

ในกรณีมีข้อมูลใหม่ที่เราต้องการทำนายการเล่นเทนนิสเมื่อมี สภาพอากาศ = overcast อุณหภูมิ = cold ความชื้น = normal และ แรงลม = true จะสามารถทำนายการเล่นเทนนิส ดังสมการ

$$P(H|E_1, E_2, E_3, E_4) = P(E_1|H) \times P(E_2|H) \times P(E_3|H) \times P(E_4|H) \times P(H)$$

$$P(H|E_1, E_2, E_3, E_4) = (0.444 \times 0.222 \times 0.666 \times 0.333 \times 0.692) = 0.015$$

H คือ เหตุการณ์ที่การเล่นเทนนิส = yes
E_1 คือ เหตุการณ์ที่สภาพอากาศ = overcast
E_2 คือ เหตุการณ์ที่อุณหภูมิ = cold
E_3 คือ เหตุการณ์ที่ความชื้น = normal
E_4 คือ เหตุการณ์ที่แรงลม = true

สามารถทำนายการไม่เล่นเทนนิส ดังสมการดังต่อไปนี้

$$P(H|E_1, E_2, E_3, E_4) = P(E_1|H) \times P(E_2|H) \times P(E_3|H) \times P(E_4|H) \times P(H)$$

$$P(H|E_1, E_2, E_3, E_4) = (0 \times 0.250 \times 0.250 \times 0.750 \times 0.307) = 0$$

H คือ เหตุการณ์ที่การเล่นเทนนิส = no
E_1 คือ เหตุการณ์ที่สภาพอากาศ = overcast
E_2 คือ เหตุการณ์ที่อุณหภูมิ = cold
E_3 คือ เหตุการณ์ที่ความชื้น = normal
E_4 คือ เหตุการณ์ที่แรงลม = true

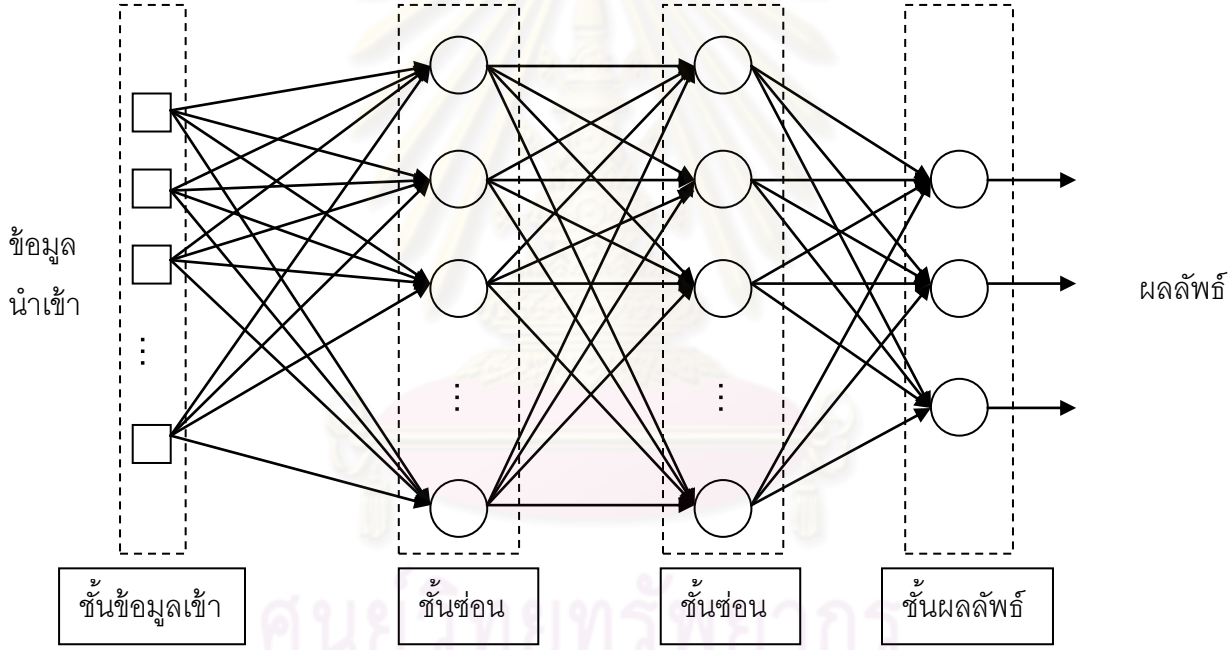
ดังนั้นในกรณี สภาพอากาศ = overcast อุณหภูมิ = cold ความชื้น = normal และ แรงลม = true จะสามารถทำนายการเล่นเทนนิส = yes

2.4.3 **เพอร์เซ็ปตรอนหลายชั้น** (Multilayer perceptron)

เพอร์เซ็ปตรอนหลายชั้นเป็นวิธีการหนึ่งของโครงข่ายประสาทเทียม (Neural network) ซึ่งมีแนวคิดมาจากการทำงานของสมองของสิ่งมีชีวิต [1] เพอร์เซ็ปตรอนหลายชั้นถูกพัฒนาเพื่อใช้กับชุดข้อมูลที่ไม่สามารถแยกประเภทได้ด้วยฟังก์ชันเชิงเส้น (Linear activation function) เพอร์เซ็ปตรอนหลายชั้นมีองค์ประกอบหลัก 3 ส่วนคือ ชั้นข้อมูลเข้า ชั้นซ่อน และชั้นผลลัพธ์ การทำงานของเพอร์เซ็ปตรอนหลายชั้นแบ่งเป็นสองส่วนคือการส่งผ่านไปข้างหน้า (Forward Pass) และการส่งผ่านย้อนกลับ (Backward Pass) [15]

เพอร์เซ็ปตรอนหลายชั้นเริ่มต้นโดยการส่งข้อมูลเข้าสู่โครงข่ายประสาทเทียมที่ชั้นข้อมูลเข้า และส่งผ่านข้อมูลเข้าสู่ชั้นซ่อน (จำนวนของชั้นซ่อนถูกกำหนดโดยผู้ใช้) จนกระทั่งถึงชั้นผลลัพธ์ซึ่งอยู่ในส่วนการส่งผ่านไปข้างหน้า ส่วนการส่งผ่านย้อนกลับค่าถ่วงน้ำหนักการเชื่อมต่อถูกปรับเปลี่ยนให้สอดคล้องกับกฎการแก้ข้อผิดพลาด (Error-Correction) ที่นิยามจากผลต่างของผลตอบที่แท้จริง (Actual Response) กับผลตอบเป้าหมาย (Target Response) เกิดเป็นสัญญาณผิดพลาด (Error Signal) ซึ่งสัญญาณผิดพลาดนี้จะถูกส่งย้อนกลับเข้าสู่โครงข่ายประสาทเทียมในทิศทางตรงกันข้ามกับการเชื่อมต่อ และค่าถ่วงน้ำหนักของการเชื่อมต่อจะถูกปรับจนกระทั่งผลต่างของผลตอบที่แท้จริงกับผลตอบเป้าหมายมีค่าน้อยกว่าที่ผู้ใช้กำหนด

ในงานวิจัยนี้ เราใช้ขั้นตอนการส่งค่าย้อนกลับ (Backpropagation) ซึ่งเป็นขั้นตอนวิธีที่นิยมที่สุดของเพอร์เซ็ปตรอนหลายชั้น



รูป 2.4 โครงสร้างของตัวแยกประเภทโครงข่ายประสารทเทียม

2.5 **ตัววัดประสิทธิภาพ** (The performance measures)

คอนฟิวชันเมทริกซ์ (Confusion matrix) ถูกใช้เพื่อประเมินประสิทธิภาพของตัวแยกประเภท ค่า TP คือจำนวนข้อมูลของคลาสบวกที่ถูกทำนายถูกต้อง ค่า FN คือจำนวนข้อมูลของคลาสบวกที่ทำนายว่าเป็นคลาสลบ ค่า FP คือจำนวนข้อมูลของคลาสลบที่ทำนายว่าเป็นคลาสบวก ค่า TN คือจำนวนข้อมูลของคลาสลบที่ทำนายถูกต้อง [13]

ตาราง 2.4 ตารางคอนฟิวชันเมทริกซ์

	Predicted Positive	Predicted Negative
Actual Positive	**TP**: True Positive	**FN**: False Negative
Actual Negative	**FP**: False Positive	**TN**: True Negative

การวัดค่าความแม่นยำของคลาสบวก

รีคอล+ (Recall+) TP/(TP+FN) หรือ ความไว (Sensitivity) คือความสามารถของตัวแบบในการทำนายข้อมูลของคลาสบวกได้ถูกต้องจากจำนวนข้อมูลของคลาสบวกแท้จริงทั้งหมด

พรีสิชัน+ (Precision+) = TP/(TP+FP) คือความสามารถของตัวแบบในการทำนายข้อมูลของคลาสบวกได้ถูกต้องจากจำนวนข้อมูลที่ถูกทำนายว่าเป็นคลาสบวกทั้งหมด

ค่า F+ (F-value+) = $((1+\beta)^2 \cdot \text{Recall} \cdot \text{Precision})/(\beta^2 \cdot \text{Recall}+\text{Precision})$, ค่าพารามิเตอร์$\beta$ เป็นตัวกำหนดความสำคัญของค่ารีคอล+ และค่าพรีสิชัน+ โดยค่าพารามิเตอร์β จะอยู่ในช่วง 0 ถึง 1 ถ้าค่า β เป็น 0 ค่า F+ จะแปรผันตรงกับค่ารีคอล+ แต่ถ้าค่า β เป็น 1 ค่า F+ จะแปรผันตรงกับค่าพรีสิชัน+ ในงานวิจัยนี้เรากำหนดให้ค่าβ เท่ากับ 1

AUC+ (พื้นที่ใต้เส้นโค้ง หรือ Area Under the Curve) คือหนึ่งในตัววัดที่ได้รับความนิยมในการวัดประสิทธิภาพของตัวแบบที่ถูกใช้กับปัญหาความไม่ดุลระหว่างกลุ่ม AUC+ เป็นพื้นที่ใต้เส้นโค้งของ ROC (Receiver Operating Characteristic) ซึ่งแสดงความสัมพันธ์ทางประสิทธิภาพของค่า TP และค่า FP แนวแกน X แสดงอัตราของการทำนายข้อมูลที่เป็นคลาสบวกผิด (FP/(TN+FP)) แนวแกน Y แสดงอัตราของการทำนายข้อมูลที่เป็นคลาสบวกถูก (TP/(TP+FN))

การวัดค่าความแม่นยำของคลาสลบ

รีคอล- (Recall-) TN/(TN+FP) หรือ ความไว คือความสามารถของตัวแบบในการทำนายข้อมูลของคลาสลบได้ถูกต้องจากจำนวนข้อมูลของคลาสลบแท้จริงทั้งหมด

พรีสิชัน- (Precision-) = TN/(TN+FN) คือความสามารถของตัวแบบในการทำนายข้อมูลของคลาสลบได้ถูกต้องจากจำนวนข้อมูลที่ถูกทำนายว่าเป็นคลาสลบทั้งหมด

ค่า F- (F-value-) = $((1+\beta)^2 \cdot \text{Recall} \cdot \text{Precision})/(\beta^2 \cdot \text{Recall}+\text{Precision})$, ค่าพารามิเตอร์$\beta$ เป็นตัวกำหนดความสำคัญของค่ารีคอล- และค่าพรีสิชัน- โดยค่าพารามิเตอร์β จะอยู่ในช่วง 0 ถึง 1

ถ้าค่า β เป็น 0 ค่า F- จะแปรผันตรงกับค่ารีคอล- แต่ถ้าค่า β เป็น 1 ค่า F- จะแปรผันตรงกับค่าพรีสิชัน- ในงานวิจัยนี้เรากำหนดให้ค่าβ เท่ากับ 1

AUC- (พื้นที่ใต้เส้นโค้ง หรือ Area Under the Curve) เป็นพื้นที่ใต้เส้นโค้งของ ROC (Receiver Operating Characteristic) ซึ่งแสดงความสัมพันธ์ทางประสิทธิภาพของค่า TN และค่า FN แนวแกน X แสดงอัตราของการทำนายข้อมูลที่เป็นคลาสลบผิด (FN/(TP+FN)) แนวแกน Y แสดงอัตราของการทำนายข้อมูลที่เป็นคลาสลบถูก (TN/(TN+FP)) ค่า AUC- มีค่าเท่ากับค่า AUC+ เนื่องจาก (FN/(TP+FN)) = 1-(TP/(TP+FN)) และ (TN/(TN+FP)) = 1-(FP/(TN+FP))

เราใช้วิธีการ SMOTE และขั้นตอนวิธีค่าเฉลี่ย k มาประยุกต์ใช้ในวิธีการ SMOUTE และนำชุดข้อมูลที่ได้จากวิธีการ SMOUTE มาใช้เป็นชุดข้อมูลฝึกหัดในการสร้างตัวแบบแยกประเภท C4.5 การแบ่งประเภทเบย์อย่างง่าย และเพอร์เซ็ปตรอนหลายชั้น จากนั้นตัวแบบแยกประเภทจะถูกวัดประสิทธิภาพด้วยตัววัดประสิทธิภาพพรีสิชัน+ พรีสิชัน- รีคอล+ รีคอล- ค่า F+ ค่า F- AUC+ และ AUC- เพื่อเปรียบเทียบผลการทำนายข้อมูลของวิธีการ SMOUTE กับวิธีการ SMOTE

บทที่ 3
เทคนิคการสุ่มเพิ่มตัวอย่างข้างน้อยสังเคราะห์และเทคนิคการสุ่มลดตัวอย่างข้างมากสำหรับปัญหาความไม่ดุลระหว่างกลุ่ม

ในบทนี้เรานำเสนอรายละเอียดและขั้นวิธีของเทคนิคการสุ่มเพิ่มตัวอย่างข้างน้อยสังเคราะห์และเทคนิคการสุ่มลดตัวอย่างข้างมากสำหรับปัญหาความไม่ดุลระหว่างกลุ่ม (SMOUTE) รวมทั้งการลงจุด 2 มิติของชุดข้อมูลก่อนและหลังการใช้ SMOUTE ในงานวิจัยนี้ วิธีการ SMOTE หมายถึงการเพิ่มจำนวนข้อมูลของคลาสบวกจนกระทั่งมีขนาดใกล้เคียงกับจำนวนข้อมูลของคลาสลบด้วยวิธีการ SMOTE และวิธีการ SMOUTE หมายถึงการเพิ่มจำนวนข้อมูลของคลาสบวกและลดจำนวนข้อมูลของคลาสลบจนกระทั่งทั้งสองคลาสมีจำนวนข้อมูลใกล้เคียงกัน การคำนวณหาจำนวนข้อมูลในงานวิจัยนี้ เราปัดทศนิยมของค่าที่คำนวณให้เป็นจำนวนเต็มทั้งหมด

3.1 รายละเอียดของเทคนิคการสุ่มเพิ่มตัวอย่างข้างน้อยสังเคราะห์และเทคนิคการสุ่มลดตัวอย่างข้างมากสำหรับปัญหาความไม่ดุลระหว่างกลุ่ม (SMOUTE)

เทคนิคการสุ่มเพิ่มตัวอย่างข้างน้อยสังเคราะห์และเทคนิคการสุ่มลดตัวอย่างข้างมากสำหรับปัญหาความไม่ดุลระหว่างกลุ่ม (SMOUTE) คือกระบวนจัดการข้อมูลก่อนการสร้างตัวแบบสำหรับการแก้ปัญหาความไม่ดุลระหว่างกลุ่ม SMOUTE ถูกพัฒนามาจากเทคนิคการชักตัวอย่างสังเคราะห์ไมนอริตี้แบบเพิ่ม (SMOTE) [1] เนื่องจากการเพิ่มจำนวนข้อมูลของคลาสบวกด้วย SMOTE เพียงอย่างเดียวอาจทำให้ชุดข้อมูลมีจำนวนข้อมูลเพิ่มขึ้นเป็นสองเท่าจากชุดข้อมูลเดิม และบริเวณที่ถูกสร้างตัวสังเคราะห์ของคลาสบวกอาจมีความหนาแน่นของข้อมูลของคลาสบวกมากเกินไป ส่งผลให้ข้อมูลของคลาสลบที่อยู่ใกล้เคียงกับบริเวณที่มีการสร้างตัวสังเคราะห์ของคลาสบวกถูกทำนายพลาดได้ การผสานของการเพิ่มจำนวนข้อมูลของคลาสบวกและการลดจำนวนข้อมูลของคลาสลบสามารถช่วยให้ข้อมูลของคลาสลบที่อยู่ใกล้เคียงกับบริเวณที่มีการสร้างตัวสังเคราะห์ของคลาสบวกไม่ถูกลดความสำคัญลงไป โดยการลบข้อมูลของคลาสลบบริเวณที่มีความหนาแน่นของข้อมูลของคลาสลบมากแทนการเพิ่มความหนาแน่นของข้อมูลของคลาสบวกจะเป็นการลดความสำคัญของข้อมูลของคลาสลบที่อยู่ห่างจากข้อมูลคลาสบวก และเพิ่มความสำคัญให้กับข้อมูลของคลาสลบที่อยู่ใกล้กับข้อมูลของคลาสบวกแทน

ในการนำเสนอผลงานเทคนิคการสุ่มเพิ่มตัวอย่างข้างน้อยสังเคราะห์และเทคนิคการสุ่มลดตัวอย่างข้างมากสำหรับปัญหาความไม่ดุลระหว่างกลุ่ม (SMOUTE: Synthetic Minority Over-sampling and majority Under-sampling TEchniques for class imbalanced problem) [16] เรากำหนดจำนวนของการแบ่งกลุ่มข้อมูลของคลาสลบเท่ากับจำนวนของลักษณะประจำ ซึ่งพบว่าการแบ่งกลุ่มของจำนวนข้อมูลของคลาสลบเท่ากับจำนวนของลักษณะประจำ ให้ผลไม่แตกต่างกับการแบ่งกลุ่มข้อมูลออกเป็น 10 กลุ่มขึ้นไป ในขณะที่การใช้วิธีการวิเคราะห์องค์ประกอบหลัก (Principle component analysis) อาจให้จำนวนกลุ่มข้อมูลของคลาสลบที่เหมาะสมกว่า

การวิเคราะห์องค์ประกอบหลักช่วยในการตัดสินใจเลือกจำนวนของกลุ่มที่เราต้องการแบ่งโดยการเลือกลักษณะประจำที่มีค่าไอเกนสูงสองค่ามาวางในแนวแกน X และแนวแกน Y ดังนั้นเราสามารถสังเกตเห็นบริเวณที่มีความหนาแน่นของข้อมูลของคลาสลบได้ และเราสามารถประมาณจำนวนกลุ่มที่ควรเลือกใช้สำหรับการแบ่งข้อมูลของคลาสลบ การประมาณค่า k ด้วยการวิเคราะห์องค์ประกอบหลัก ช่วยให้ความแม่นยำในการทำนายข้อมูลเพิ่มขึ้น เนื่องจากเราสามารถหลีกเลี่ยงการลบข้อมูลของคลาสลบที่สำคัญได้ ในกรณีที่ไม่สามารถเลือกจำนวนกลุ่มข้อมูลด้วยวิธีการวิเคราะห์องค์ประกอบหลัก ในกรณีที่การใช้วิธีการวิเคราะห์องค์ประกอบหลักไม่สามารถใช้ในเลือกจำนวนของกลุ่มข้อมูล เราจะกำหนดให้ค่า k เป็น 10 เนื่องจากการแบ่งกลุ่มข้อมูลออกเป็น 10 กลุ่มขึ้นไปจะให้ความแม่นยำในการทำนายข้อมูลคงที่

วิธีการของเราเริ่มโดยการใช้ SMOTE สร้างข้อมูลสังเคราะห์ของคลาสบวกจนกระทั่งจำนวนข้อมูลสังเคราะห์ของคลาสบวกเท่ากับค่า T เมื่อ T คือเปอร์เซ็นต์ของจำนวนข้อมูลสังเคราะห์ของคลาสบวกที่ผู้ใช้กำหนด

ในขั้นตอนของการลดจำนวนข้อมูลของคลาสลบ เราลดจำนวนข้อมูลของคลาสลบจนกระทั่งจำนวนข้อมูลของคลาสลบเท่ากับจำนวนข้อมูลของคลาสบวก โดยการคำนวณเปอร์เซ็นต์ในการลดจำนวนข้อมูลของคลาสลบ เราใช้สมการ 3.1 ในการคำนวณ

$$p_U = \left(1 - \frac{n_S}{n_M}\right) \times 100\% \qquad (3.1)$$

p_U คือ เปอร์เซ็นต์ของจำนวนข้อมูลของคลาสลบที่ต้องการลด

n_S คือ จำนวนข้อมูลของคลาสบวกหลังจากใช้ SMOTE

n_M คือ จำนวนข้อมูลของคลาสลบทั้งหมด หรือจำนวนข้อมูลของไมนอริตี้เมื่อใช้ SMOTE เพียงอย่างเดียว (เพิ่มจำนวนของคลาสบวกจนกระทั่งมีจำนวนใกล้เคียงกับคลาสลบ)

เราใช้ขั้นตอนวิธีค่าเฉลี่ย k แบ่งข้อมูลของคลาสลบออกเป็น k กลุ่ม เราคำนวณจำนวนข้อมูลของคลาสลบที่ต้องการลดในแต่ละกลุ่มที่ถูกแบ่งจากสมการ 3.2 สำหรับทุก $i \in \{1, ..., N\}$

$$c_i = \frac{g_i}{m} \times u \tag{3.2}$$

c_i คือ จำนวนข้อมูลของคลาสลบที่ต้องการลบภายในกลุ่มที่ i

g_i คือ จำนวนข้อมูลของคลาสลบภายในกลุ่มที่ i

m คือ จำนวนข้อมูลของคลาสลบทั้งหมด

u คือ จำนวนข้อมูลของคลาสลบที่ต้องการลบที่ผู้ใช้กำหนด

SMOUTE สิ้นสุดการประมวลผลเมื่อจำนวนข้อมูลของคลาสลบถูกลดจนได้เท่ากับจำนวนที่ผู้ใช้กำหนด

3.2 ขั้นตอนวิธีของเทคนิคการสุ่มเพิ่มตัวอย่างข้างน้อยสังเคราะห์และเทคนิคการสุ่มลดตัวอย่างข้างมากสำหรับปัญหาความไม่ดุลระหว่างกลุ่ม (SMOUTE)

ขั้นตอนวิธีของ SMOUTE มีดังนี้

Let *M* be a majority class dataset of D

K be a number of the *k*-means clusters

numG[*i*] = the number of instances in group-*i* ; *i*∈{1,2,…,*K*}

numM = the number of the major class instances

numUG[*i*] = the number of under-sampling instances in group-*i* ; *i*∈{1,2,…,*K*}

numU = A user's parameter for a number of the under-sampling instances

K-means(*M*, *K*) runs *k*-means clustering algorithm for partition *M* into *K* clusters

Algorithm: SMOUTE

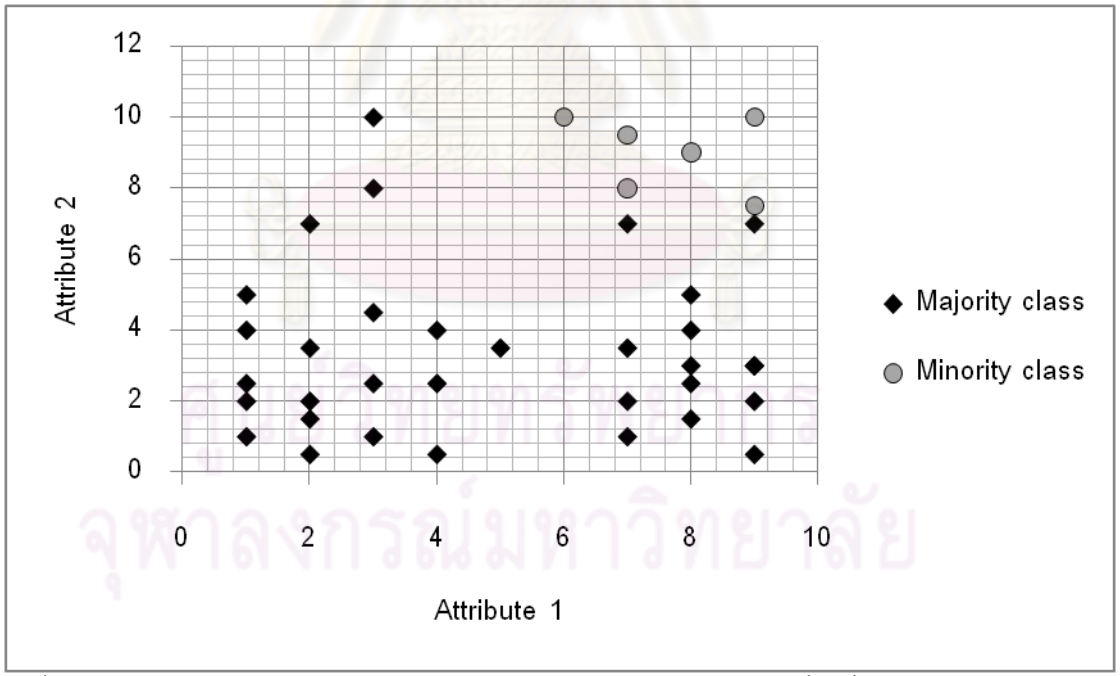
Input: a set *D* of all instances after over-sampling by SMOTE

Output: a modified set *D*

1. Extract *M* from *D*
2. K-means(*M*, *K*)
3. For each group-*i*; *i* ∈ {1,2,…,*K*} {

4. numUG[*i*] = numG[*i*]/numM*numU
5. For each centroid *c*[*i*] {
6. *k* = numUG[*i*]
7. For each *k*-nearest neighbors of c[*i*] in group-*i* {
8. *u* = the selected *k*-nearest neighbor of *c*[*i*]
9. *D* = *D* – {*u*}
10. }
11. }
12. }
13. return *D*

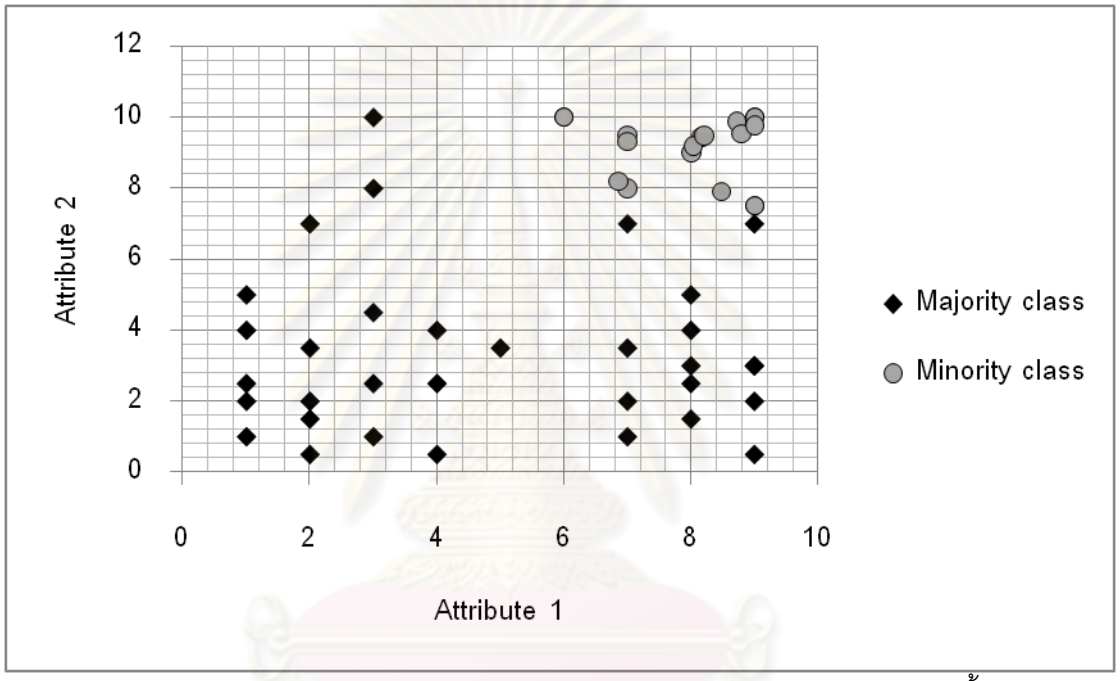
ตัวอย่าง 3.1 ชุดข้อมูลตัวอย่างมี 2 ลักษณะประจำ และมีจำนวนข้อมูลทั้งหมด 39 ตัวแบ่งเป็นข้อมูลคลาสลบ 33 ตัวและข้อมูลคลาสบวก 6 ตัว



รูปที่ 3.1 ชุดข้อมูลตัวอย่าง 3.1 วงกลมแทนข้อมูลของคลาสบวกและสี่เหลี่ยมขนมเปียกปูนแทนข้อมูลของคลาสลบ

เริ่มต้นเรากำหนดจำนวนตัวสังเคราะห์ไมนอริตี้ที่เราต้องการสร้างมีจำนวนเป็น 200% ของข้อมูลของคลาสบวก และเราใช้ SMOTE เพิ่มจำนวนข้อมูลของคลาสบวกโดยใช้ขั้นตอนวิธี k

เนเรียสเนเบอร์หาข้อมูลของคลาสบวกที่ใกล้เคียงที่สุดจำนวน k ตัว (กำหนดให้ k = 5) ของข้อมูลของคลาสบวกแต่ละตัว ตัวอย่างเช่น ข้อมูลของคลาสบวกที่ตำแหน่ง (15, 9) มีข้อมูลที่ใกล้ที่สุดคือ (6, 20) (7, 16) (7, 19) (8, 18) และ (9, 20) เราสุ่มเลือกข้อมูลมาหนึ่งตัวคือ (7, 19) และสุ่มเลือกตำแหน่งมาหนึ่งมาหนึ่งตำแหน่งระหว่างตำแหน่ง (15, 9) และ (7, 19) เพื่อสร้างข้อมูลสังเคราะห์ไมนอริตี้ ดังนั้นเราได้ข้อมูลสังเคราะห์ไมนอริตี้ที่ตำแหน่ง (8.479, 15.808) เราสร้างข้อมูลสังเคราะห์ไมนอริตี้จนกระทั่งข้อมูลสังเคราะห์ไมนอริตี้มีจำนวนเท่ากับที่เรากำหนด (12 ตัว)



รูป 3.2 ชุดข้อมูลตัวอย่าง 3.1 หลังจากการใช้ SMOTE สร้างตัวสังเคราะห์ไมนอริตี้ 200%

จากรูป 3.2 เราใช้ SMOTE เพิ่มจำนวนข้อมูลของคลาสบวกอีก 200% ดังนั้นจำนวนข้อมูลของคลาสบวกจะมีจำนวนเป็น 18 ตัว (3 เท่าของจำนวนข้อมูลเดิม) เราคำนวณเปอร์เซนต์ของจำนวนข้อมูลของคลาสลบที่ต้องการลบออกด้วยสมการ 3.1

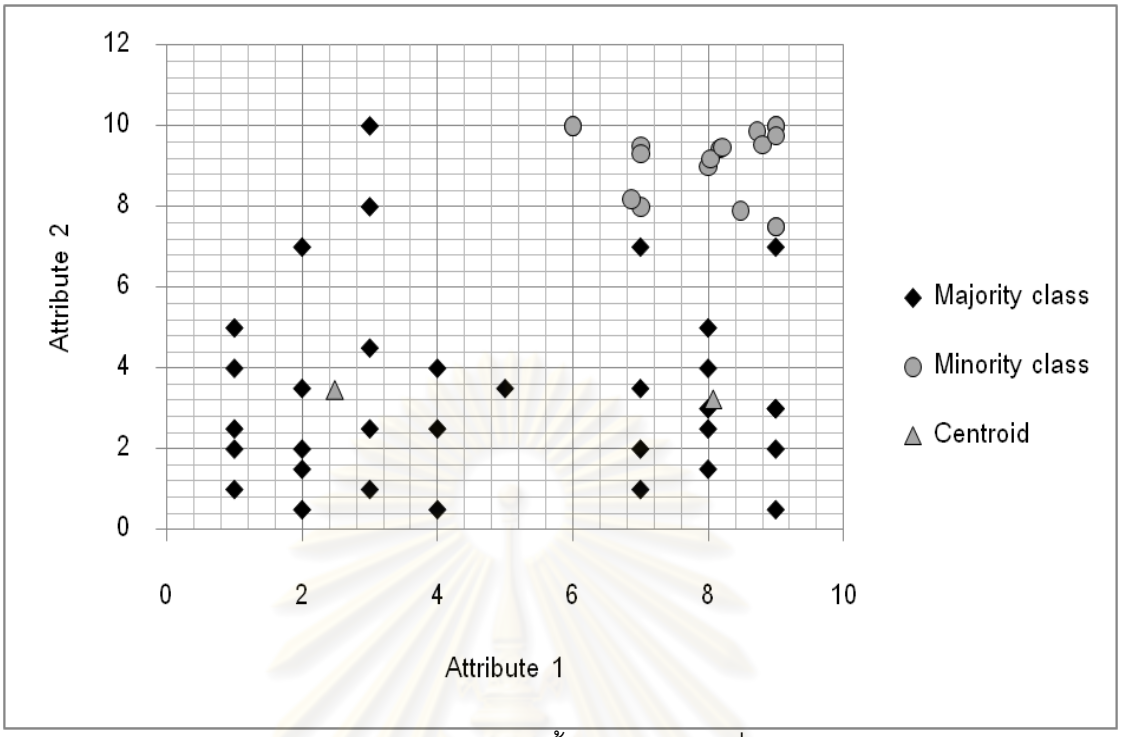
ในกรณีที่ n_M คือ จำนวนข้อมูลของคลาสลบทั้งหมด จะได้

(1-18/33)x100% = 45.45%

ในกรณีที่ n_M คือ จำนวนข้อมูลของไมนอริตี้เมื่อใช้ SMOTE เพียงอย่างเดียว จะได้

(1-18/33)x100% = 43.75%

คิดเป็นจำนวนข้อมูลของคลาสลบโดยประมาณ 14 ตัว จากนั้นเราใช้ขั้นตอนวิธีค่าเฉลี่ย k แบ่งข้อมูลของคลาสลบออกเป็น k= 2



รูป 3.3 ชุดข้อมูลตัวอย่าง 3.1 หลังจากการใช้ ขั้นตอนวิธีค่าเฉลี่ย k แบ่งข้อมูลของคลาสลบออกเป็น 2 กลุ่ม และสามเหลี่ยมแทนเซนทรอยด์ของแต่ละกลุ่ม

จากรูป 3.3 เราใช้ขั้นตอนวิธีค่าเฉลี่ย k แบ่งข้อมูลของคลาสลบออกเป็น 2 กลุ่ม เซนทรอยด์ 2 จุดเป็นตัวแทนกลุ่มของจำนวนข้อมูลคลาสลบทั้ง 2 กลุ่ม โดยกลุ่มแรก และกลุ่มที่สองมีจำนวนข้อมูลของคลาสลบเป็น 20 ตัวและ 13 ตัวตามลำดับ

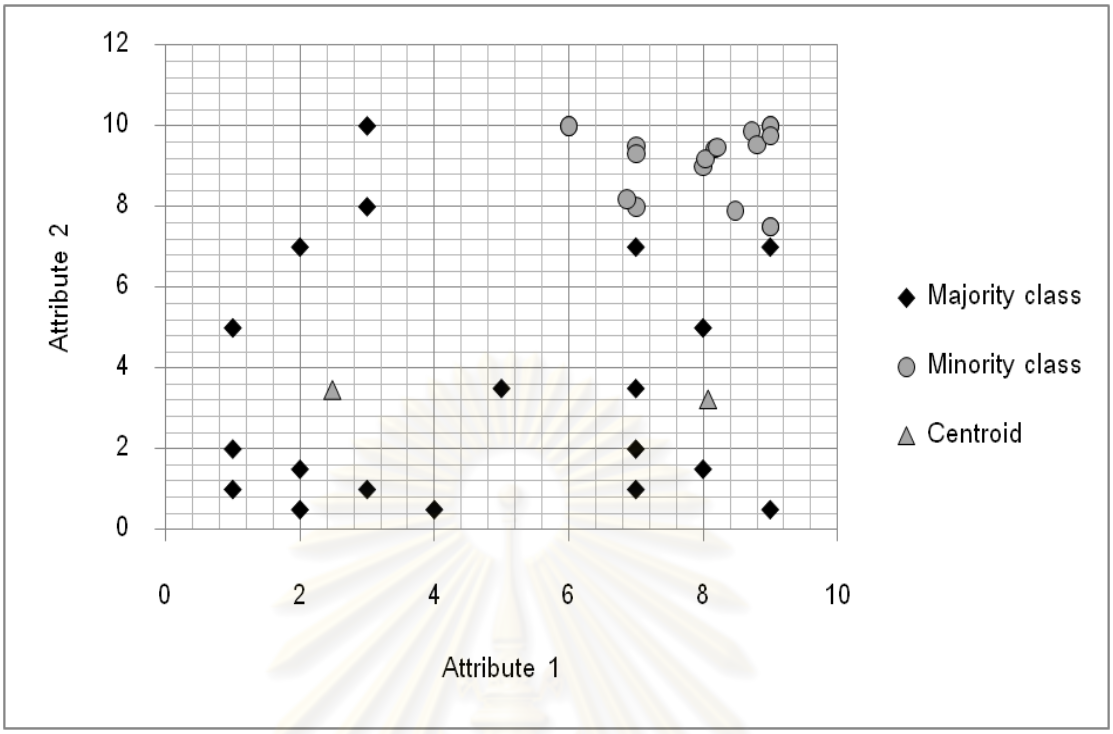
จากสมการ 3.2 จำนวนข้อมูลของคลาสลบที่เราต้องลบออกในกลุ่มแรกคือ

(20/33)x14 = 8.48

จากสมการ 3.2 จำนวนข้อมูลของคลาสลบที่เราต้องลบออกในกลุ่มที่สองคือ

(13/33)x14 = 5.52

ดังนั้นในกลุ่มแรกข้อมูลของคลาสลบจะถูกลบ 8 ตัว และกลุ่มที่สองข้อมูลของคลาสลบจะถูกลบ 5 ตัว



รูป 3.4 ชุดข้อมูลตัวอย่าง 3.1 หลังจากการใช้ SMOUTE ลบจำนวนข้อมูลของคลาสลบ 13 ตัว

จากรูป 3.4 ข้อมูลของคลาสลบโดยรอบเซนทรอยด์ถูกลบออกทั้งหมด เหลือข้อมูลของคลาสลบเฉพาะบริเวณขอบของกลุ่มเท่านั้น ชุดข้อมูลที่ได้จากการใช้วิธีการ SMOUTE จะถูกนำไปใช้เป็นชุดข้อมูลฝึกหัดเพื่อสร้างตัวแบบแยกประเภท

จากตัวอย่าง 3.1 วิธีการ SMOUTE ลบข้อมูลเฉพาะบริเวณที่อยู่ใกล้กับเซนทรอยด์ เนื่องจากบริเวณโดยรอบเซนทรอยด์เป็นบริเวณที่มีความหนาแน่นของข้อมูลของคลาสลบสูง แทนที่การเพิ่มความหนาแน่นให้กับข้อมูลของคลาสบวกด้วยวิธีการ SMOTE เนื่องจากบริเวณที่มีความหนาแน่นของข้อมูลสูงอาจส่งผลให้ตัวแยกประเภทที่ถูกสร้างจากชุดข้อมูลดังกล่าวเกิดปัญหาความจำเพาะเกิน (Overfitting) ซึ่งทำให้ตัวแบบแยกประเภทมีค่าความแม่นยำเมื่อใช้กับชุดข้อมูลทดสอบน้อย

บทที่ 4

ผลการวิเคราะห์ข้อมูล

ในบทนี้ เราแสดงผลการทดลองของเทคนิคการสุ่มเพิ่มตัวอย่างข้างน้อยสังเคราะห์และเทคนิคการสุ่มลดตัวอย่างข้างมากสำหรับปัญหาความไม่ดุลระหว่างกลุ่ม (SMOUTE) เปรียบเทียบกับวิธีการ SMOTE โดยการนำวิธีการทั้งสองวิธีมาประยุกต์กับตัวแยกประเภท 3 แบบ คือ C4.5 การแบ่งประเภทเบย์อย่างง่าย และเพอร์เซ็ปตรอนหลายชั้น มีตัววัดประสิทธิภาพของตัวแบบคือ พรีสิชัน+ พรีสิชัน- รีคอล+ รีคอล- ค่า F+ ค่า F- AUC+ และ AUC- จำนวนข้อมูล 70% ของชุดข้อมูลถูกใช้เป็นข้อมูลฝึกหัด และจำนวนข้อมูลอีก 30% ของชุดข้อมูลถูกใช้เป็นชุดข้อมูลทดสอบ เราแบ่งผลการทดลองเป็น 2 ส่วนคือ ประสิทธิภาพในการทำนายข้อมูลของคลาสเป้าหมาย ความรวดเร็วในการประมวลผลของวิธีการ

4.1 รายละเอียดของชุดข้อมูล

ในงานวิจัยนี้ เราใช้ชุดข้อมูล 4 ชุดในการทดลอง มีรายละเอียดดังนี้

4.1.1 ชุดข้อมูลการอยู่รอดของฮาเบอร์แมน (Haberman's survival dataset หรือ ชุดข้อมูล Haberman)

ชุดข้อมูล Haberman เป็นข้อมูลของจำนวนการรอดชีวิตของผู้ป่วยโรคมะเร็งหลังเข้ารับการผ่าตัดในโรงพยาบาลแห่งหนึ่ง ชุดข้อมูล Haberman มีลักษณะประจำ 3 ตัว จำนวนข้อมูลทั้งหมด 306 ตัว แบ่งเป็นข้อมูลของผู้ป่วยที่เสียชีวิตภายในระยะเวลา 5 ปีหลังเข้ารับการผ่าตัดมี 81 ตัว และของผู้ป่วยที่เสียชีวิตภายในระยะเวลา 5 ปีหลังเข้ารับการผ่าตัดมี 225 ตัว [17] เรากำหนดให้ข้อมูลของผู้ป่วยที่เสียชีวิตภายในระยะเวลา 5 ปีหลังเข้ารับการผ่าตัดเป็นคลาสบวก และข้อมูลของผู้ป่วยที่เสียชีวิตภายในระยะเวลา 5 ปีหลังเข้ารับการผ่าตัดเป็นคลาสลบ

ตาราง 4.1 จำนวนข้อมูลของคลาสบวกและคลาสลบของชุดข้อมูล Haberman

Haberman	instances	%instances
Minority class	81	26.47
Majority class	225	73.53
Total	306	100

4.1.2 **ชุดข้อมูลภาพดาวเทียม** (Satellite image dataset **หรือ ชุดข้อมูล** satimage)

ชุดข้อมูล satimage เป็นชุดข้อมูลภาพถ่ายจากดาวเทียมมีจำนวนลักษณะประจำ 35 ตัว มีจำนวนข้อมูล 6435 ตัว และจำนวนคลาสเป้าหมาย 6 คลาส [18] ในงานวิจัยนี้เรากำหนดให้คลาส 4 ซึ่งมีจำนวนข้อมูล 626 ตัวเป็นข้อมูลของคลาสบวก และคลาสอื่นอีก 5 คลาสซึ่งมีจำนวนข้อมูลทั้งหมด 5809 ตัวถูกรวมเป็นคลาสลบ

ตาราง 4.2 จำนวนข้อมูลของคลาสบวกและคลาสลบของชุดข้อมูล satimage

satimage	instances	%instances
Minority class	626	9.73
Majority class	5809	90.27
Total	6435	100

4.1.3 **ชุดข้อมูล** ecoli (Ecoli dataset)

ชุดข้อมูล ecoli เป็นชุดข้อมูลสำหรับการจำแนกองค์ประกอบของ ecoli มีจำนวนลักษณะประจำ 7 ตัว จำนวนข้อมูล 336 ตัว และจำนวนคลาสเป้าหมาย 8 คลาส [18][19] ในงานวิจัยนี้เรากำหนดให้เยื่อหุ้มเซลล์ชั้นนอก (Outer membrane) ซึ่งมีจำนวนข้อมูล 20 ตัวเป็นข้อมูลของคลาสบวก และคลาสอื่นอีก 7 คลาสซึ่งมีจำนวนข้อมูลทั้งหมด 316 ตัวถูกรวมเป็นคลาสลบ

ตาราง 4.3 จำนวนข้อมูลของคลาสบวกและคลาสลบของชุดข้อมูล ecoli

Ecoli	instances	%instances
Minority class	20	5.95
Majority class	316	94.05
Total	336	100

4.1.4 **ชุดข้อมูลกระสวยอวกาศ** (shuttle dataset)

ชุดข้อมูลกระสวยอวกาศ มีจำนวนลักษณะประจำ 8 ตัว จำนวนข้อมูล 58000 ตัว และจำนวนคลาสเป้าหมาย 7 คลาส [18][20] ในงานวิจัยนี้เรากำหนดให้ Fpv open ซึ่งมีจำนวนข้อมูล

171 ตัวเป็นข้อมูลของคลาสบวก และคลาสอื่นอีก 6 คลาสซึ่งมีจำนวนข้อมูลทั้งหมด 57829 ตัวถูกรวมเป็นคลาสลบ

ตาราง 4.4 จำนวนข้อมูลของคลาสบวกและคลาสลบของชุดข้อมูลกระสวยอวกาศ

shuttle	instances	%instances
Minority class	171	0.29
Majority class	57829	99.71
Total	58000	100

4.2 ผลการวิจัย

4.2.1 ผลการทดสอบประสิทธิภาพในการทำนายข้อมูลของวิธีการ SMOUTE และวิธีการ SMOTE

ในส่วนนี้ เราแสดงผลการทดสอบประสิทธิภาพในการทำนายข้อมูลของวิธีการ SMOUTE เปรียบเทียบกับวิธีการ SMOTE โดยมีการกำหนดสัญลักษณ์ดังนี้

Negative คือ ค่าการทำนายข้อมูลของคลาสลบ

Positive คือ ค่าการทำนายข้อมูลของคลาสบวก

O=r% คือ r เปอร์เซ็นต์ของจำนวนข้อมูลของคลาสบวกที่ถูกเพิ่มด้วยวิธีการ SMOTE จนมีจำนวนข้อมูลของคลาสบวกใกล้เคียงกับจำนวนข้อมูลของคลาสลบ

OU=r% คือ r เปอร์เซ็นต์ของจำนวนข้อมูลของคลาสบวกที่ถูกเพิ่มด้วยวิธีการ SMOTE และลบจำนวนข้อมูลของคลาสลบจนกระทั่งข้อมูลทั้งสองคลาสมีจำนวนใกล้เคียงกัน (ใช้วิธีการ SMOUTE)

กราฟค่าพรีสิชัน (Precision) แสดงค่าพรีสิชัน+ (Precision+) และค่าพรีสิชัน- (Precision-)

กราฟค่ารีคอล (Recall) แสดงค่ารีคอล+ (Recall+) และค่ารีคอล- (Recall-)

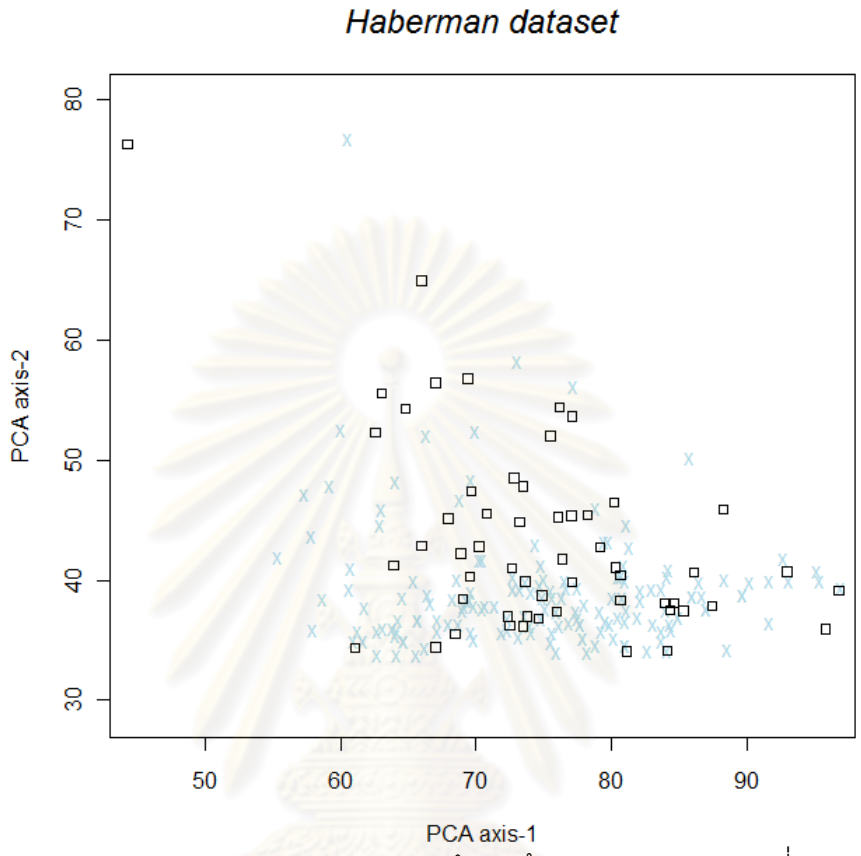
กราฟค่า F (F value) แสดงค่า F+ (F-value+) และค่า F- (F-value-)

กราฟค่า AUC แสดงค่า AUC+ และ AUC-

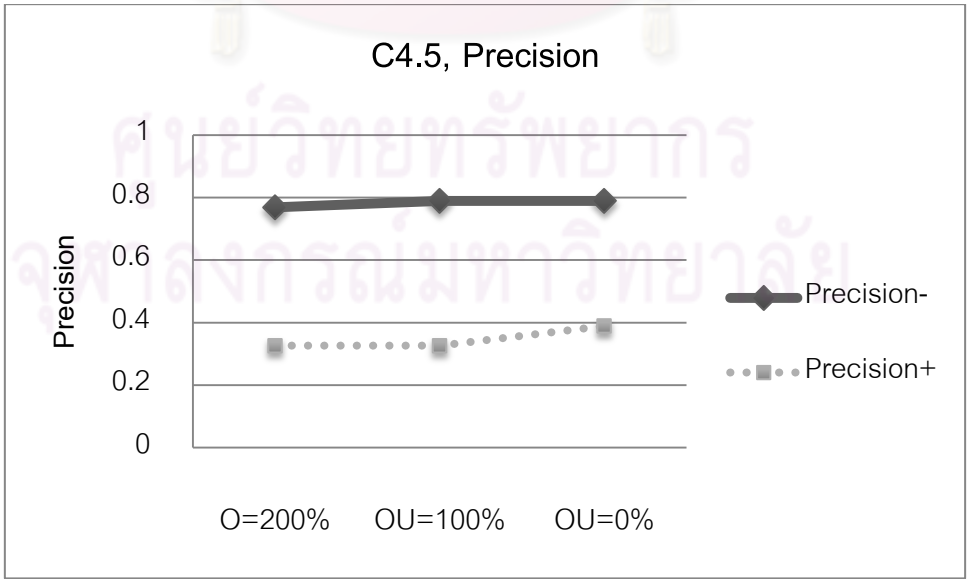
4.2.1.1 ชุดข้อมูล Haberman

สำหรับชุดข้อมูล Haberman เราใช้วิธีการ SMOTE เพิ่มจำนวนข้อมูลของคลาสบวก 200% (O=200%) เปรียบเทียบกับวิธีการ SMOUTE ที่เพิ่มจำนวนข้อมูลของคลาสบวก 100% ลด

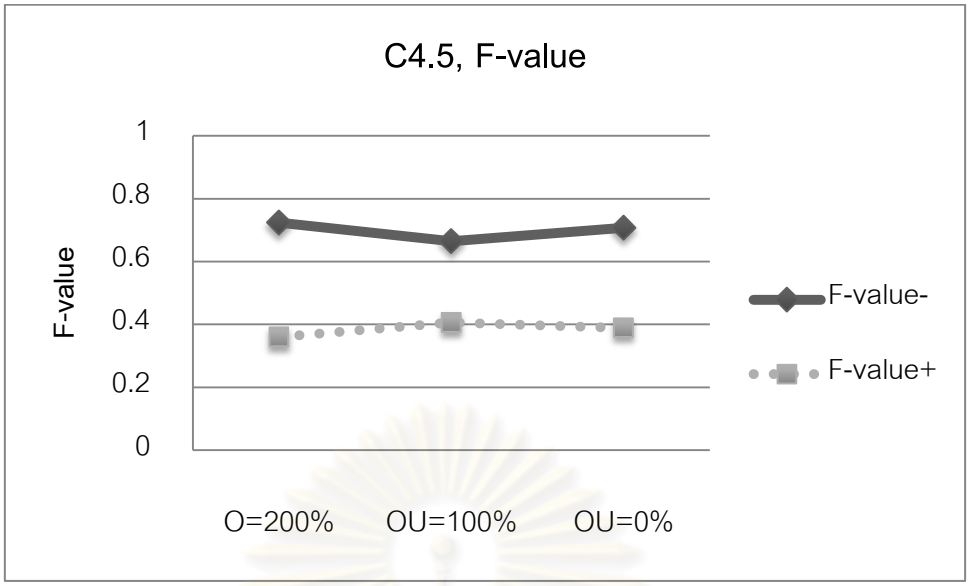
จำนวนของคลาสลบ 33.33% (OU=100%) และวิธีการ SMOUTE ที่เพิ่มจำนวนข้อมูลของคลาสบวก 0% ลดจำนวนของคลาสลบ 66.67% (OU=0%)



รูป 4.1 กราฟการกระจายตัวของชุดข้อมูลฝึกหัด Haberman ชุดที่ 1

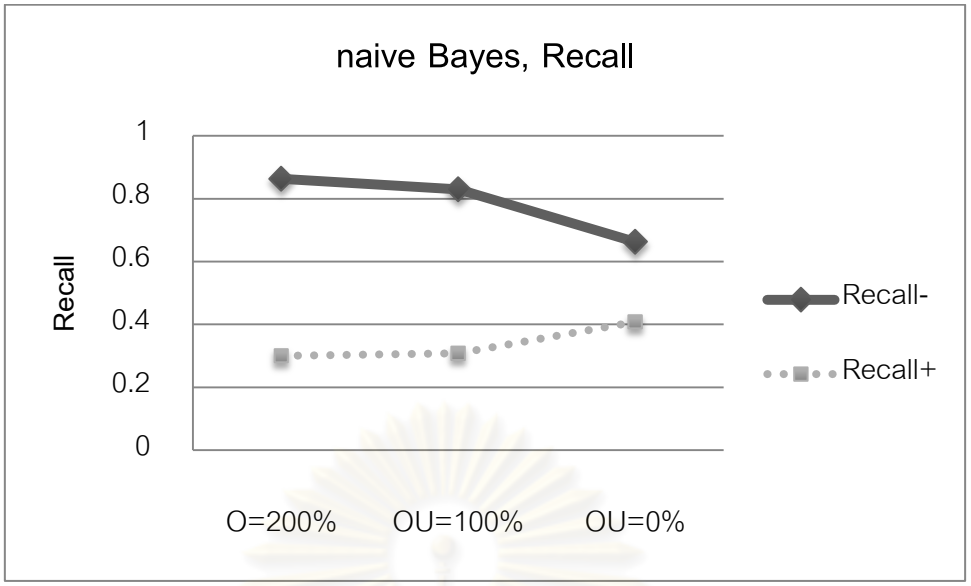


รูป 4.2 ค่าพรีสิชัน+ และค่าพรีสิชัน- ของผลการทำนายข้อมูลด้วย C4.5 บนชุดข้อมูล Haberman

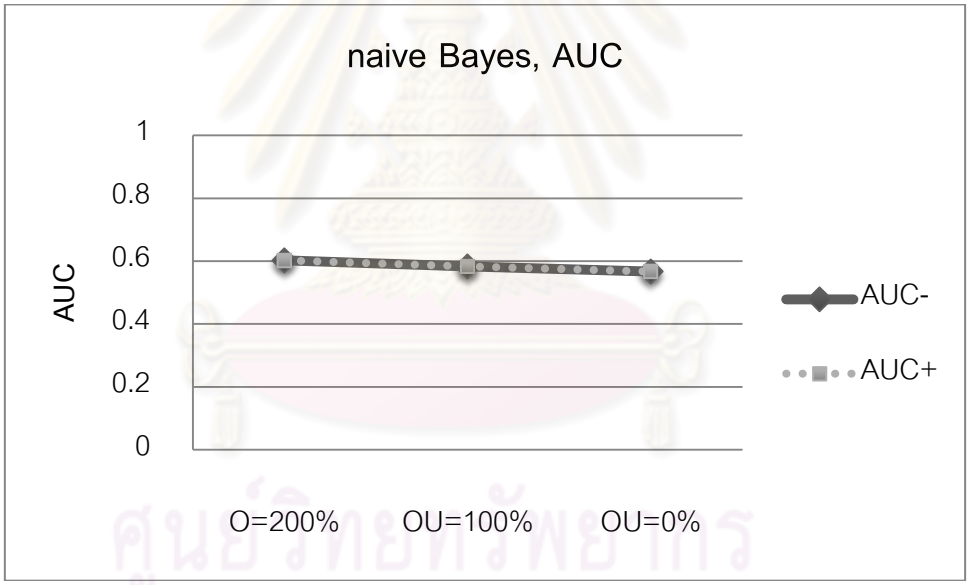


รูป 4.3 ค่า F+ และค่า F- ของผลการทำนายข้อมูลด้วย C4.5 บนชุดข้อมูล Haberman

จากรูป 4.2 และ 4.3 ประสิทธิภาพในการทำนายข้อมูลของคลาสลบของวิธีการ SMOTE และวิธีการ SMOUTE ที่ OU=0% มีความแม่นยำกว่าวิธีการ SMOUTE ที่ OU=100% เล็กน้อย ในขณะที่ผลการทำนายข้อมูลของคลาสบวกของวิธีการ SMOUTE ที่ OU=0% มีความแม่นยำมากที่สุด ชุดข้อมูลฝึกหัด Haberman มีจำนวนข้อมูลน้อยและข้อมูลของคลาสบวกของชุดข้อมูลฝึกหัด Haberman มีการกระจายตัวของข้อมูลสูง (รูป 4.1) ทำให้ตัวสังเคราะห์ของคลาสบวกที่ถูกสร้างด้วยวิธีการ SMOTE มีการกระจายตัวสูงเช่นกัน เนื่องจาก C4.5 ใช้หลักการแบ่งข้อมูลแบบเชิงเส้น (Linearly separable) [1] ดังนั้นการใช้วิธีการ SMOTE ทำให้การแบ่งข้อมูลของคลาสบวกคลาดเคลื่อนมากกว่าการใช้วิธีการ SMOUTE ลดจำนวนข้อมูลของคลาสลบ



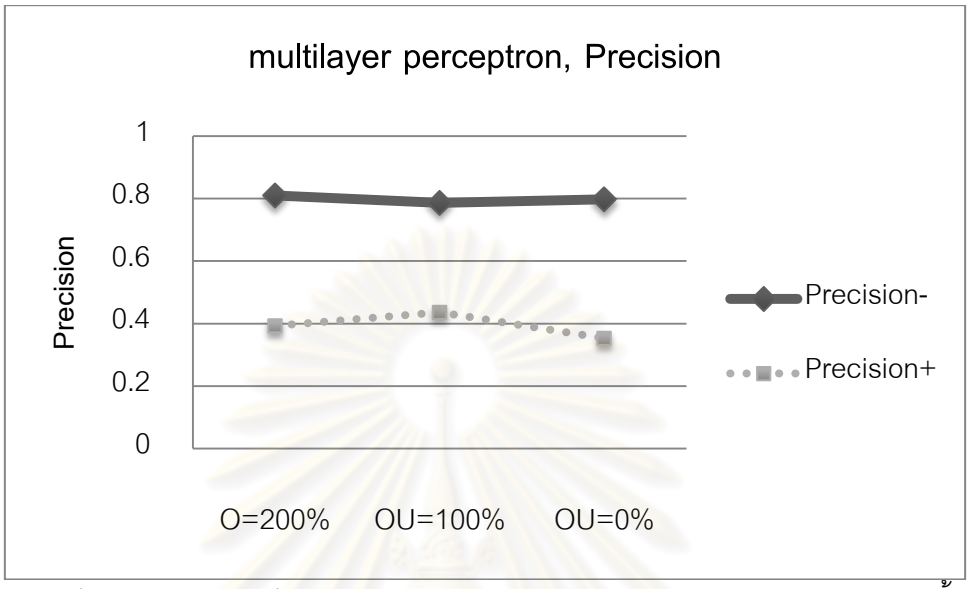
รูป 4.4 ค่ารีคอล+ และค่ารีคอล- ของผลการทำนายข้อมูลด้วยการแบ่งประเภทเบย์อย่างง่ายบนชุดข้อมูล Haberman



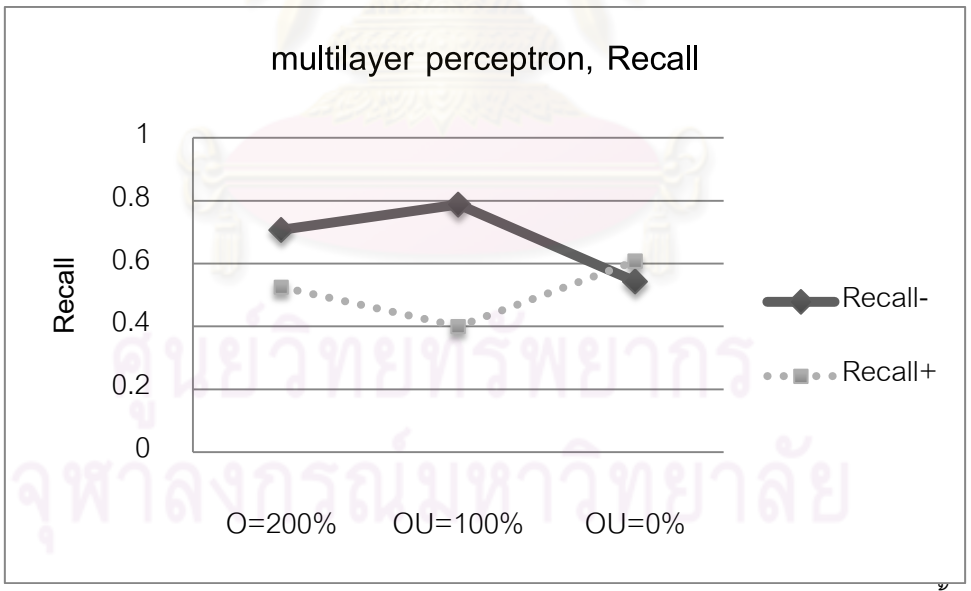
รูป 4.5 ค่า AUC+ และค่า AUC- ของผลการทำนายข้อมูลด้วยการแบ่งประเภทเบย์อย่างง่ายบนชุดข้อมูล Haberman

จากรูป 4.3 และ 4.4 ประสิทธิภาพในการทำนายข้อมูลของคลาสลบของวิธีการ SMOTE มีความแม่นยำกว่าวิธีการ SMOUTE ที่ OU=100% เล็กน้อย และวิธีการ SMOUTE ที่ OU=0% มีประสิทธิภาพในการทำนายข้อมูลของคลาสลบน้อยที่สุด เนื่องจากชุดข้อมูลฝึกหัด Haberman มีจำนวนข้อมูลน้อยและข้อมูลของคลาสบวกของชุดข้อมูลฝึกหัด Haberman มีการกระจายตัวของ

ข้อมูลสูง การใช้วิธีการ SMOUTE โดยการลบข้อมูลของคลาสลบเพียงอย่างเดียวอาจลบข้อมูลที่มีความสำคัญเมื่อใช้การแบ่งประเภทเบย์อย่างง่ายเป็นตัวแบบแยกประเภท



รูป 4.6 ค่าพรีสิชัน+ และค่าพรีสิชัน- ของผลการทำนายข้อมูลด้วยเพอร์เซ็ปตรอนหลายชั้นบนชุดข้อมูล Haberman



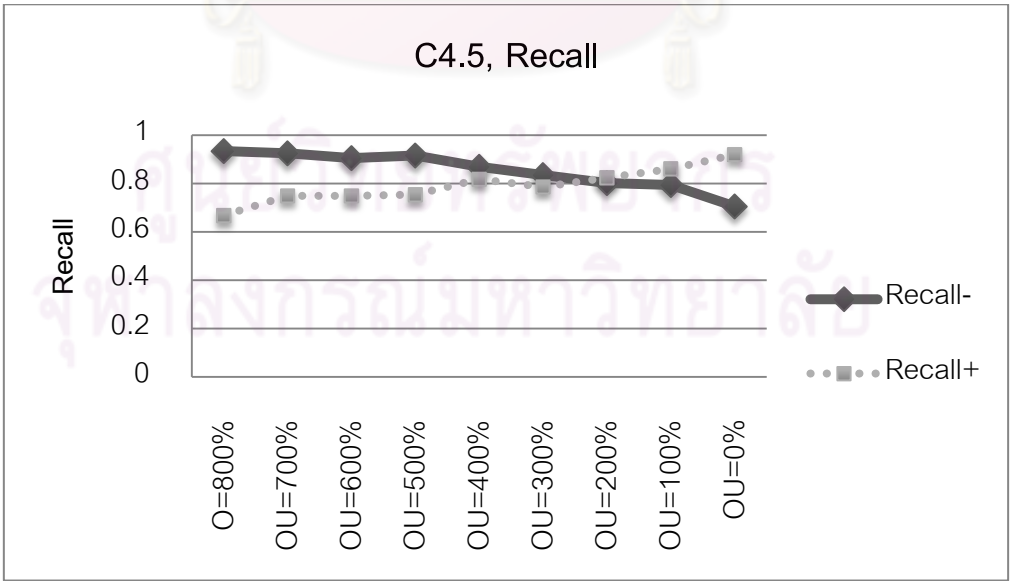
รูป 4.7 ค่ารีคอล+ และค่ารีคอล- ของผลการทำนายข้อมูลด้วยเพอร์เซ็ปตรอนหลายชั้นบนชุดข้อมูล Haberman

จากรูป 4.6 และ 4.7 ประสิทธิภาพในการทำนายข้อมูลของคลาสลบของวิธีการ SMOTE มีความแม่นยำกว่าวิธีการ SMOUTE ที่ OU=100% เล็กน้อย และวิธีการ SMOUTE ที่ OU=0% มี

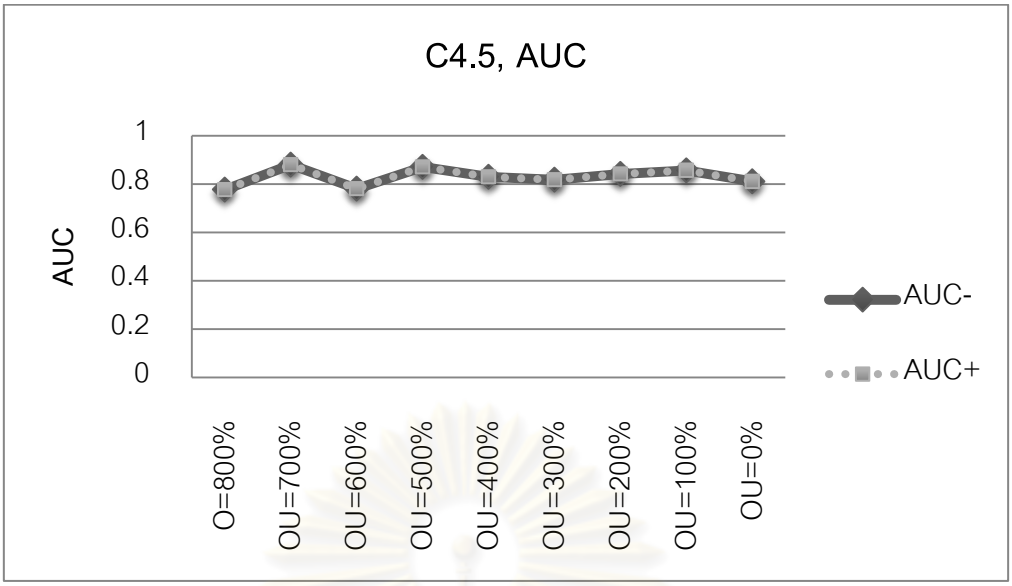
ความแม่นยำในการทำนายข้อมูลของคลาสลบน้อยที่สุด เนื่องจากชุดข้อมูลฝึกหัด Haberman มีจำนวนข้อมูลน้อยและข้อมูลของคลาสบวกของชุดข้อมูลฝึกหัด Haberman มีการกระจายตัวของข้อมูลสูง การใช้วิธีการ SMOUTE โดยการลบข้อมูลของคลาสลบเพียงอย่างเดียวอาจลบข้อมูลที่มีความสำคัญเมื่อใช้เพอร์เซ็ปตรอนหลายชั้นเป็นตัวแบบแยกประเภท

4.2.2 **ชุดข้อมูล** satimage

สำหรับชุดข้อมูล satimage เราใช้วิธีการ SMOTE เพิ่มจำนวนข้อมูลของคลาสบวก 800% (O=800%) เปรียบเทียบกับวิธีการ SMOUTE ที่เพิ่มจำนวนข้อมูลของคลาสบวก 700% ลดจำนวนของคลาสลบ 11.11% (OU=700%) วิธีการ SMOUTE ที่เพิ่มจำนวนข้อมูลของคลาสบวก 600% ลดจำนวนของคลาสลบ 22.22% (OU=600%) วิธีการ SMOUTE ที่เพิ่มจำนวนข้อมูลของคลาสบวก 500% ลดจำนวนของคลาสลบ 33.33% (OU=500%) วิธีการ SMOUTE ที่เพิ่มจำนวนข้อมูลของคลาสบวก 400% ลดจำนวนของคลาสลบ 44.44% (OU=400%) วิธีการ SMOUTE ที่เพิ่มจำนวนข้อมูลของคลาสบวก 300% ลดจำนวนของคลาสลบ 55.56% (OU=300%) วิธีการ SMOUTE ที่เพิ่มจำนวนข้อมูลของคลาสบวก 200% ลดจำนวนของคลาสลบ 66.67% (OU=200%) วิธีการ SMOUTE ที่เพิ่มจำนวนข้อมูลของคลาสบวก 100% ลดจำนวนของคลาสลบ 77.78% (OU=100%) และวิธีการ SMOUTE ที่เพิ่มจำนวนข้อมูลของคลาสบวก 0% ลดจำนวนของคลาสลบ 88.89% (OU=0%)

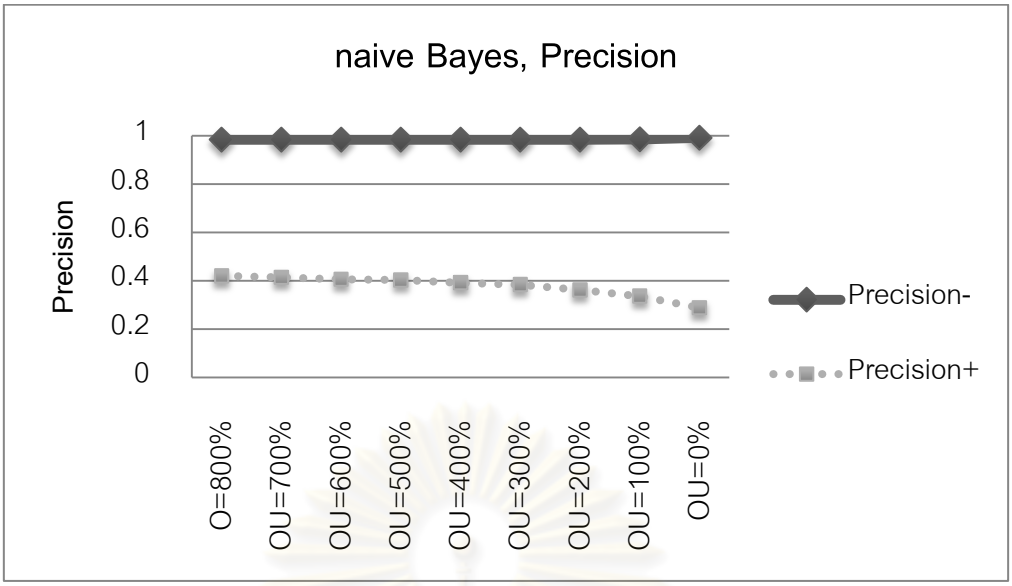


รูป 4.8 ค่ารีคอล+ และค่ารีคอล- ของผลการทำนายข้อมูลด้วย C4.5 บนชุดข้อมูล satimage

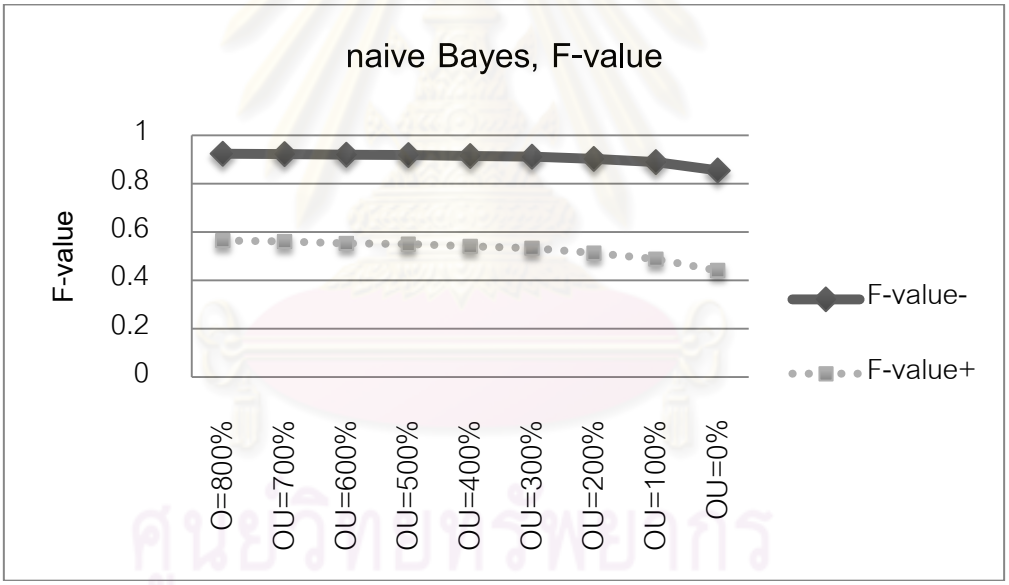


รูป 4.9 ค่า AUC+ และค่า AUC- ของผลการทำนายข้อมูลด้วย C4.5 บนชุดข้อมูล satimage

จากรูป 4.8 และ 4.9 ประสิทธิภาพในการทำนายข้อมูลของคลาสลบและคลาสบวกด้วยวิธีการ SMOUTE ที่ OU=700% OU=600% และ OU=500% แตกต่างกันเล็กน้อยเมื่อเปรียบเทียบกับวิธีการ SMOTE และประสิทธิภาพในการทำนายข้อมูลของคลาสลบด้วยวิธีการ SMOUTE ที่ OU=400% OU=300% OU=200% OU=100% และ OU=0% ลดลง ในขณะที่ค่าการทำนายข้อมูลของคลาสบวกเพิ่มขึ้นเมื่อเปรียบเทียบกับวิธีการ SMOTE เนื่องจากชุดข้อมูล satimage มีบริเวณที่มีความหนาแน่นของจำนวนข้อมูลของคลาสลบบางกลุ่มอยู่ใกล้กับกลุ่มข้อมูลของคลาสบวก ดังนั้นการใช้วิธีการ SMOUTE ลดจำนวนข้อมูลของคลาสลบในบริเวณที่มีความหนาแน่นมากอาจส่งผลให้ข้อมูลของคลาสลบที่อยู่ใกล้กับกลุ่มข้อมูลของคลาสบวกถูกทำนายว่าเป็นคลาสบวก และทำให้การทำนายข้อมูลของคลาสบวกแม่นยำมากขึ้น



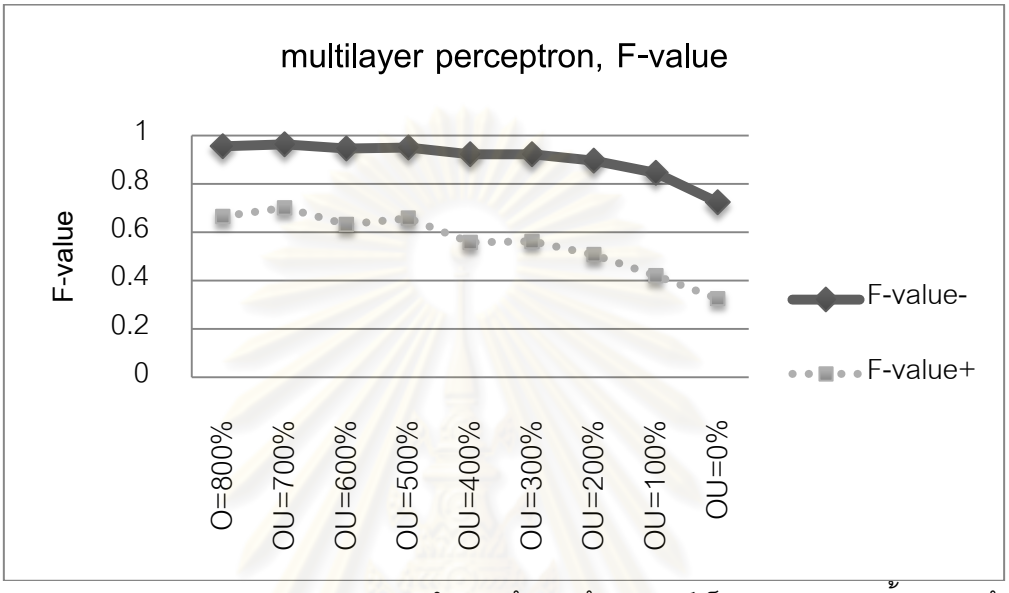
รูป 4.10 ค่าพรีสิชัน+ และค่าพรีสิชัน- ของผลการทำนายข้อมูลด้วยการแบ่งประเภทเบย์อย่างง่ายบนชุดข้อมูล satimage



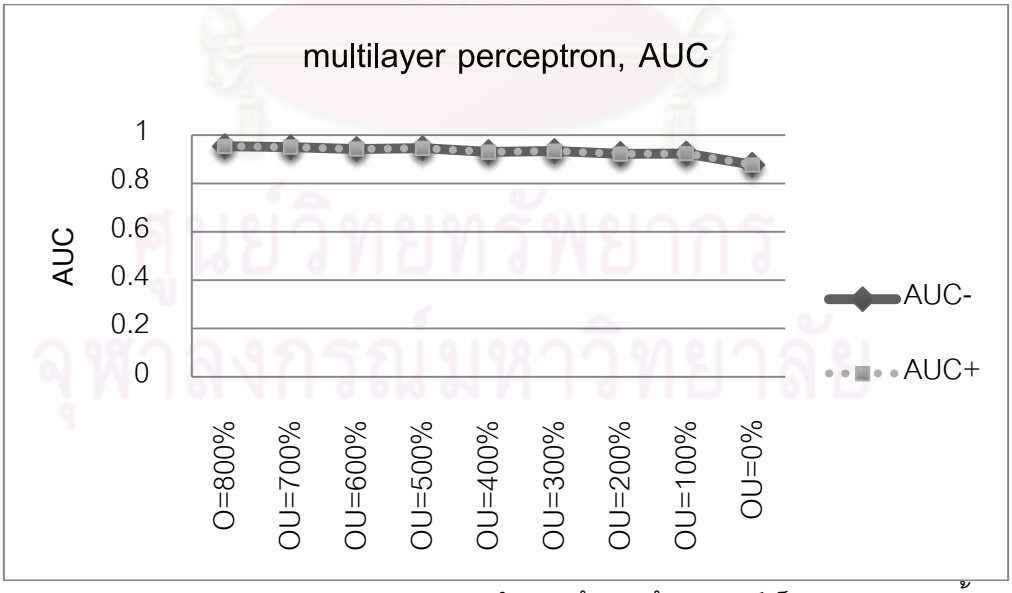
รูป 4.11 ค่า F+ และค่า F- ของผลการทำนายข้อมูลด้วยการแบ่งประเภทเบย์อย่างง่ายบนชุดข้อมูล satimage

จากรูป 4.10 และ 4.11 ประสิทธิภาพในการทำนายข้อมูลของคลาสลบและคลาสบวกด้วยวิธีการ SMOUTE ที่ OU=700% OU=600% OU=500% และ OU=400% แตกต่างกันเล็กน้อยเมื่อเปรียบเทียบกับวิธีการ SMOTE และค่าการทำนายข้อมูลของคลาสลบและคลาสบวกด้วยวิธีการ SMOUTE ที่ OU=300% OU=200% OU=100% และ OU=0% ลดลงเล็กน้อย แสดงให้เห็นว่าชุดข้อมูล satimage มีความเป็นอิสระต่อกันของลักษณะประจำแต่ละตัวมาก ทำให้ตัว

แบบแยกประเภทของการแบ่งประเภทเบย์อย่างง่ายสามารถทำนายสามารถทำนายข้อมูลของคลาสลบได้แม่นยำมาก และการทำนายข้อมูลของคลาสลบและคลาสบวกไม่เปลี่ยนแปลงมากนักเมื่อเปรียบเทียบกับวิธีการ SMOTE แม้ว่าจำนวนข้อมูลของคลาสลบที่อยู่ใกล้บริเวณบวกจะถูกลดลงมาก



รูป 4.12 ค่า F+ และค่า F- ของผลการทำนายข้อมูลด้วยเพอร์เซ็ปตรอนหลายชั้นบนชุดข้อมูล satimage

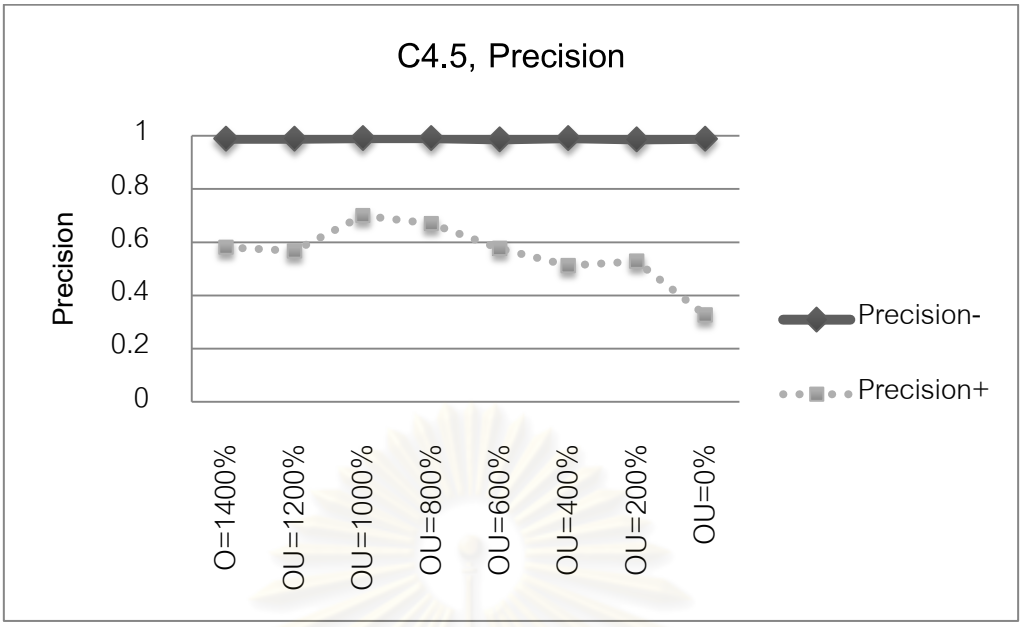


รูป 4.13 ค่า AUC+ และค่า AUC- ของผลการทำนายข้อมูลด้วยเพอร์เซ็ปตรอนหลายชั้นบนชุดข้อมูล satimage

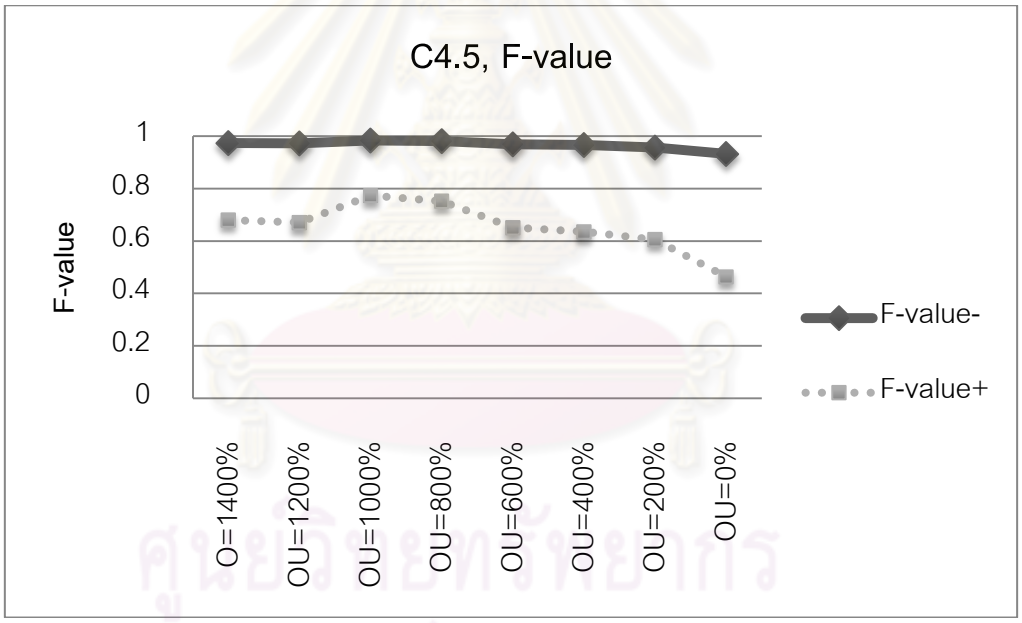
จากรูป 4.12 และ 4.13 ประสิทธิภาพในการทำนายข้อมูลของคลาสลบและคลาสบวกด้วยวิธีการ SMOUTE ที่ OU=700% OU=600% และ OU=500% ลดลงเล็กน้อยเมื่อเปรียบเทียบกับวิธีการ SMOTE และค่าการทำนายข้อมูลของคลาสลบและคลาสบวกด้วยวิธีการ SMOUTE ที่ OU=400% OU=300% OU=200% OU=100% และ OU=0% ลดลงมากเมื่อเปรียบเทียบกับวิธีการ SMOTE เนื่องจากชุดข้อมูล satimage มีบริเวณที่มีความหนาแน่นของจำนวนข้อมูลของคลาสลบบางกลุ่มอยู่ใกล้กับกลุ่มข้อมูลของคลาสบวก และตัวแยกประเภทเพอร์เซ็ปตรอนหลายชั้นใช้หลักการแบ่งแบบไม่เชิงเส้น (Non-linearly separable) [1] ดังนั้นการใช้วิธีการ SMOUTE ลดจำนวนข้อมูลของคลาสลบในบริเวณที่มีความหนาแน่นมากอาจส่งผลให้ข้อมูลของคลาสลบที่อยู่ใกล้กับกลุ่มข้อมูลของคลาสบวกถูกทำนายว่าเป็นคลาสบวก

4.2.3 **ชุดข้อมูล** ecoli

สำหรับชุดข้อมูล ecoli เราใช้วิธีการ SMOTE เพิ่มจำนวนข้อมูลของคลาสบวก 1400% (O=1400%) เปรียบเทียบกับวิธีการ SMOUTE ที่เพิ่มจำนวนข้อมูลของคลาสบวก 1200% ลดจำนวนของคลาสลบ 13.33% (OU=1200%) วิธีการ SMOUTE ที่เพิ่มจำนวนข้อมูลของคลาสบวก 1000% ลดจำนวนของคลาสลบ 26.67% (OU=1000%) วิธีการ SMOUTE ที่เพิ่มจำนวนข้อมูลของคลาสบวก 800% ลดจำนวนของคลาสลบ 40.00% (OU=800%) วิธีการ SMOUTE ที่เพิ่มจำนวนข้อมูลของคลาสบวก 600% ลดจำนวนของคลาสลบ 53.33% (OU=600%) วิธีการ SMOUTE ที่เพิ่มจำนวนข้อมูลของคลาสบวก 400% ลดจำนวนของคลาสลบ 66.67% (OU=400%) วิธีการ SMOUTE ที่เพิ่มจำนวนข้อมูลของคลาสบวก 200% ลดจำนวนของคลาสลบ 80.00% (OU=200%) และวิธีการ SMOUTE ที่เพิ่มจำนวนข้อมูลของคลาสบวก 0% ลดจำนวนของคลาสลบ 93.33% (OU=0%)



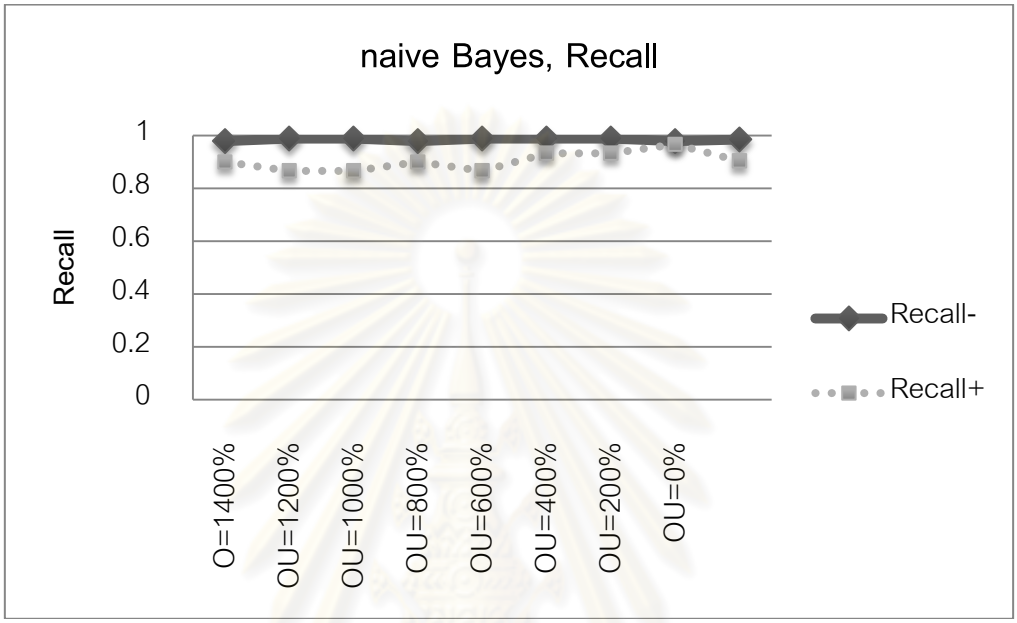
รูป 4.14 ค่าพรีสิชัน+ และค่าพรีสิชัน- ของผลการทำนายข้อมูลด้วย C4.5 บนชุดข้อมูล ecoli



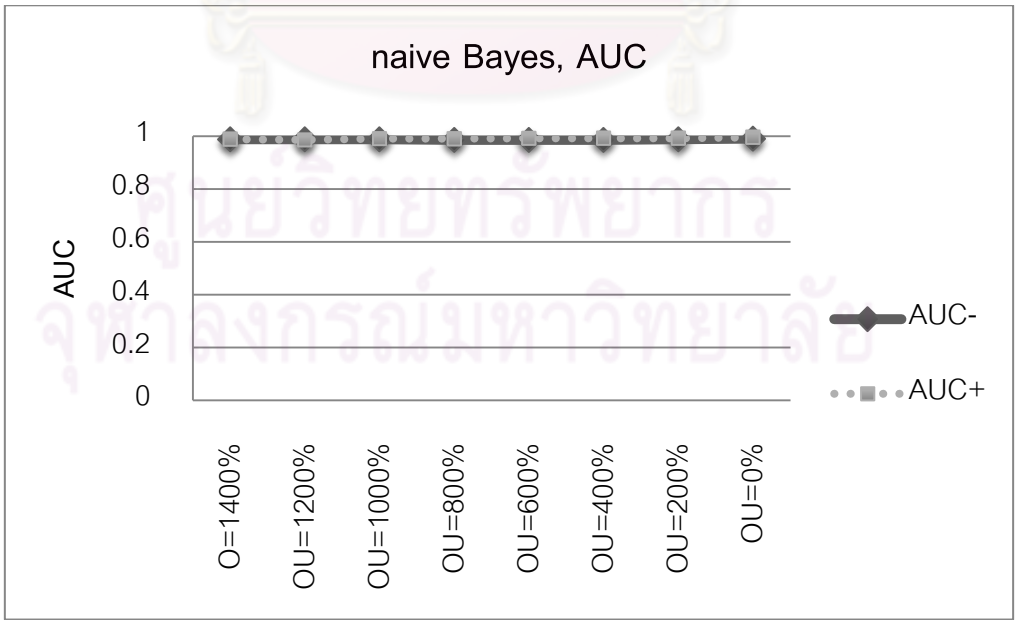
รูป 4.15 ค่า F+ และค่า F- ของผลการทำนายข้อมูลด้วย C4.5 บนชุดข้อมูล ecoli

จากรูป 4.14 และ 4.15 ประสิทธิภาพในการทำนายข้อมูลของคลาสลบด้วยวิธีการ SMOUTE และวิธีการ SMOTE ไม่มีความแตกต่างกัน ในขณะที่วิธีการ SMOUTE ที่ระดับ OU=1200% OU=1000% และ OU=800% ให้ค่าการทำนายข้อมูลของคลาสบวกแม่นยำกว่าวิธีการ SMOTE สำหรับวิธีการ SMOUTE ที่ระดับ OU=600% OU=400% OU=200% และ OU=0% ให้ค่าการทำนายข้อมูลของคลาสบวกแม่นยำน้อยกว่าวิธีการ SMOTE เนื่องจากชุดข้อมูล ecoli มีจำนวนข้อมูลน้อย และตัวแยกประเภท C4.5 ใช้หลักการแบ่งแบบเชิงเส้น ดังนั้นการ

ลดจำนวนข้อมูลของคลาสลบในแต่ละกลุ่มเป็นจำนวนมากอาจส่งผลให้ข้อมูลของคลาสลบที่อยู่ใกล้กับบริเวณขอบของแต่ละกลุ่มถูกลบไปด้วย ทำให้การทำนายข้อมูลของคลาสบวกคลาดเคลื่อนเป็นข้อมูลของคลาสลบมากขึ้น ในขณะที่ข้อมูลของคลาสลบส่วนมากมีการกระจายตัวห่างจากข้อมูลคลาสบวกมาก ทำให้การทำนายข้อมูลของคลาสลบส่วนมากไม่ผิดพลาด

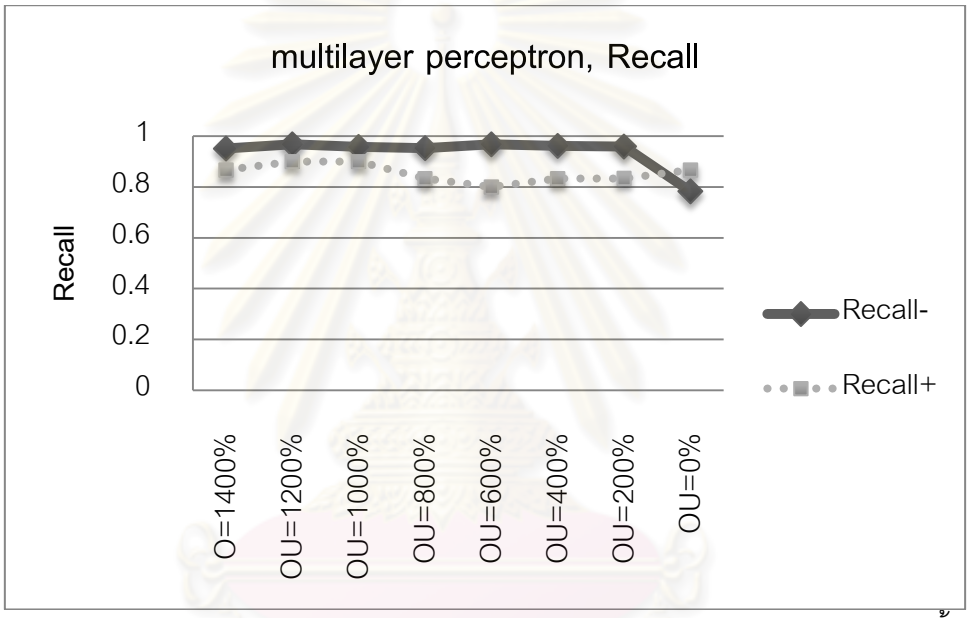


รูป 4.16 ค่ารีคอล+ และค่ารีคอล- ของผลการทำนายข้อมูลด้วยการแบ่งประเภทเบย์อย่างง่ายบนชุดข้อมูล ecoli

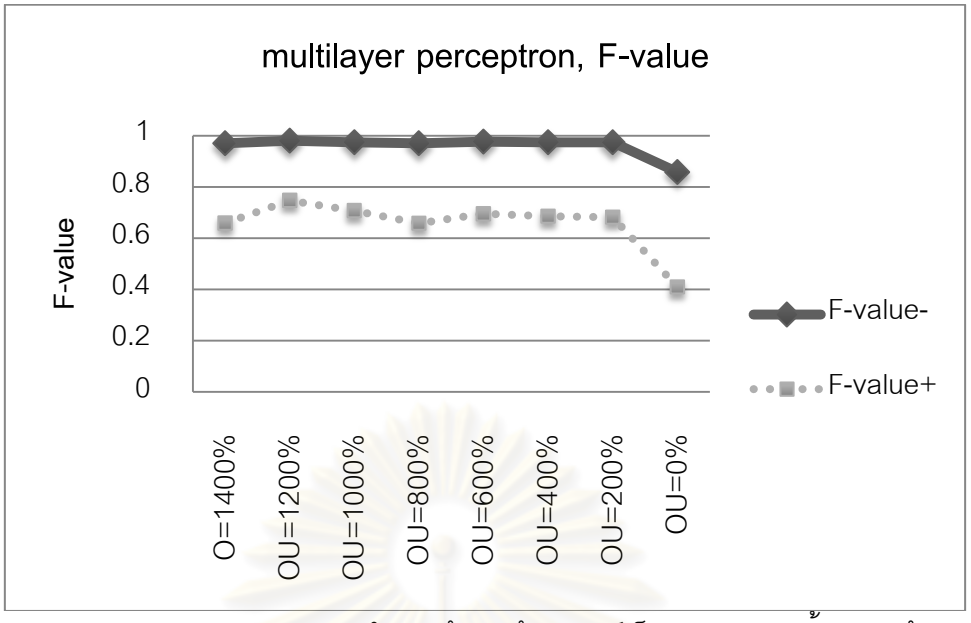


รูป 4.17 ค่า AUC+ และค่า AUC- ของผลการทำนายข้อมูลด้วยการแบ่งประเภทเบย์อย่างง่ายบนชุดข้อมูล ecoli

จากภาพ 4.16 และ 4.17 ประสิทธิภาพในการทำนายข้อมูลของคลาสลบด้วยวิธีการ SMOUTE และวิธีการ SMOTE ไม่มีความแตกต่างกัน ในขณะที่ค่าการทำนายข้อมูลของคลาสบวกด้วยวิธีการ SMOUTE และวิธีการ SMOTE มีความแตกต่างกันเล็กน้อย การลดจำนวนข้อมูลของคลาสลบในแต่ละกลุ่มเป็นจำนวนมากอาจส่งผลให้ข้อมูลของคลาสลบที่อยู่ใกล้กับบริเวณขอบของแต่ละกลุ่มถูกลบ แต่สำหรับตัวแยกประเภทการแบ่งประเภทเบย์อย่างง่ายใช้หลักความน่าจะเป็นในการทำนายข้อมูล ดังนั้นข้อมูลของคลาสลบบริเวณขอบของแต่ละกลุ่มอาจไม่ใช่ข้อมูลที่มีความสำคัญสำหรับวิธีการแบ่งประเภทเบย์อย่างง่าย ทำให้การทำนายข้อมูลของคลาสลบและคลาสบวกมีความแม่นยำสูง



รูป 4.18 ค่ารีคอล+ และค่ารีคอล- ของผลการทำนายข้อมูลด้วยเพอร์เซ็ปตรอนหลายชั้นบนชุดข้อมูล ecoli



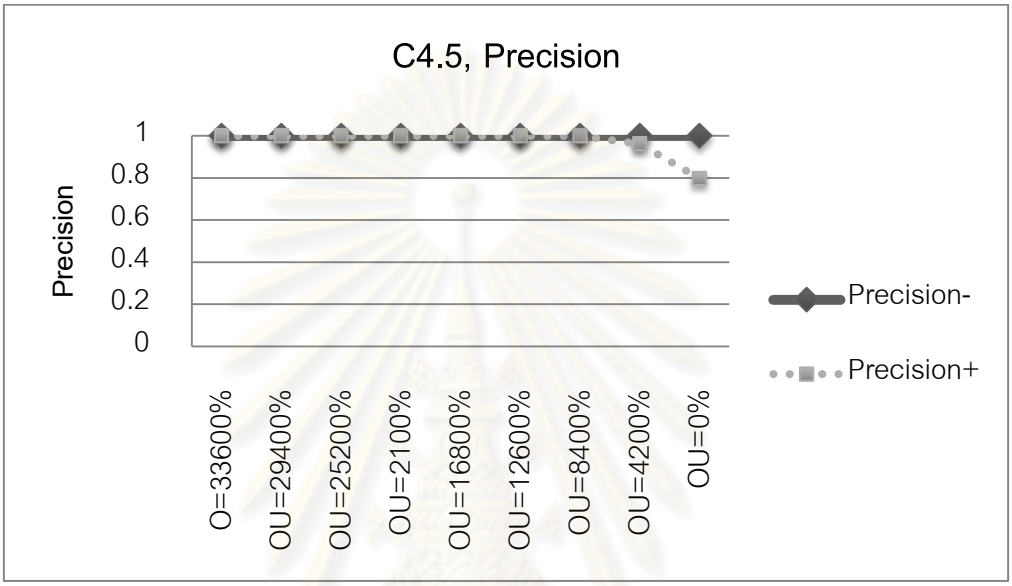
รูป 4.19 ค่า F+ และค่า F- ของผลการทำนายข้อมูลด้วยเพอร์เซ็ปตรอนหลายชั้นบนชุดข้อมูล ecoli

จากรูป 4.18 และ 4.19 ประสิทธิภาพในการทำนายข้อมูลของคลาสลบด้วยวิธีการ SMOUTE และวิธีการ SMOTE ไม่มีความแตกต่างกัน ในขณะที่ค่าการทำนายข้อมูลของคลาสบวกด้วยวิธีการ SMOUTE และวิธีการ SMOTE มีความแตกต่างกันเล็กน้อย ยกเว้นกรณีที่ใช้วิธีการ SMOUTE ที่ระดับ OU=0% ซึ่งเป็นการลบข้อมูลของคลาสลบโดยไม่มีการเพิ่มจำนวนข้อมูลของคลาสบวก มีค่าการทำนายข้อมูลของคลาสบวกน้อยกว่าวิธีการ SMOTE อย่างเห็นได้ชัด เนื่องจากการลดจำนวนข้อมูลของคลาสลบในแต่ละกลุ่มเป็นจำนวนมากอาจส่งผลให้การกระจายตัวของชุดข้อมูลเปลี่ยนแปลงมาก ทำให้การทำนายข้อมูลของคลาสบวกด้วยเพอร์เซ็ปตรอนหลายชั้นคลาดเคลื่อนมากขึ้น ในขณะที่ข้อมูลของคลาสลบส่วนมากมีการกระจายตัวห่างจากข้อมูลคลาสบวกมาก ทำให้การทำนายข้อมูลของคลาสลบส่วนมากไม่ผิดพลาด

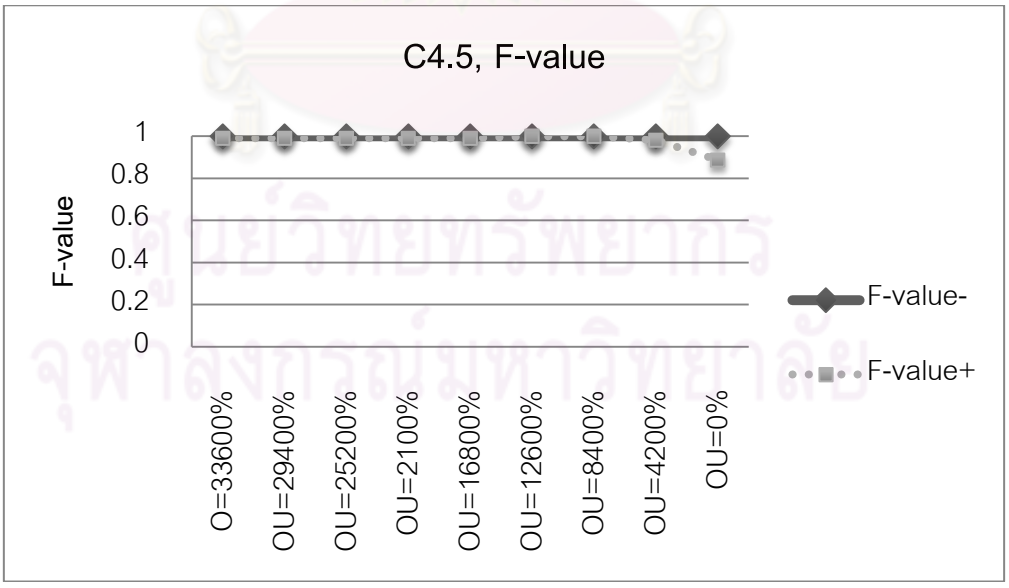
4.2.4 **ชุดข้อมูล** shuttle

สำหรับชุดข้อมูล shuttle เราใช้วิธีการ SMOTE เพิ่มจำนวนข้อมูลของคลาสบวก 33600% (O=33600%) เปรียบเทียบกับวิธีการ SMOUTE ที่เพิ่มจำนวนข้อมูลของคลาสบวก 29400% ลดจำนวนของคลาสลบ 12.46% (OU=29400%) วิธีการ SMOUTE ที่เพิ่มจำนวนข้อมูลของคลาสบวก 25200% ลดจำนวนของคลาสลบ 24.36% (OU=25200%) วิธีการ SMOUTE ที่เพิ่มจำนวนข้อมูลของคลาสบวก 21000% ลดจำนวนของคลาสลบ 37.39% (OU=21000%) วิธีการ SMOUTE ที่เพิ่มจำนวนข้อมูลของคลาสบวก 16800% ลดจำนวนของคลาสลบ 49.85%

(OU=16800%) วิธีการ SMOUTE ที่เพิ่มจำนวนข้อมูลของคลาสบวก 12600% ลดจำนวนของคลาสลบ 62.31% (OU=12600%) วิธีการ SMOUTE ที่เพิ่มจำนวนข้อมูลของคลาสบวก 8400% ลดจำนวนของคลาสลบ 74.78% (OU=8400%) วิธีการ SMOUTE ที่เพิ่มจำนวนข้อมูลของคลาสบวก 4200% ลดจำนวนของคลาสลบ 87.24% (OU=4200%) และวิธีการ SMOUTE ที่เพิ่มจำนวนข้อมูลของคลาสบวก 0% ลดจำนวนของคลาสลบ 99.70% (OU=0%)

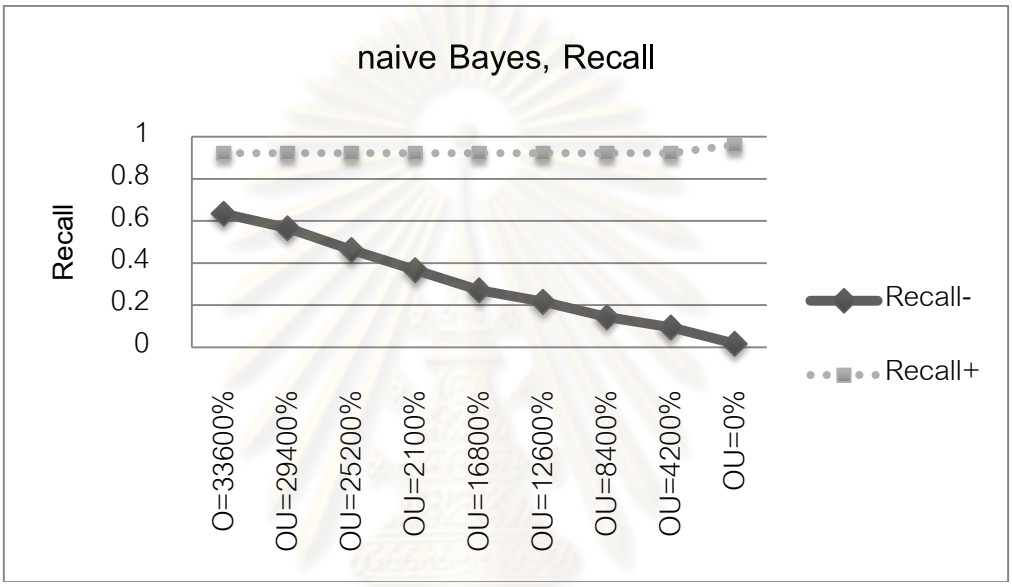


รูป 4.20 ค่าพรีสิชัน+ และค่าพรีสิชัน- ของผลการทำนายข้อมูลด้วย C4.5 บนชุดข้อมูล shuttle

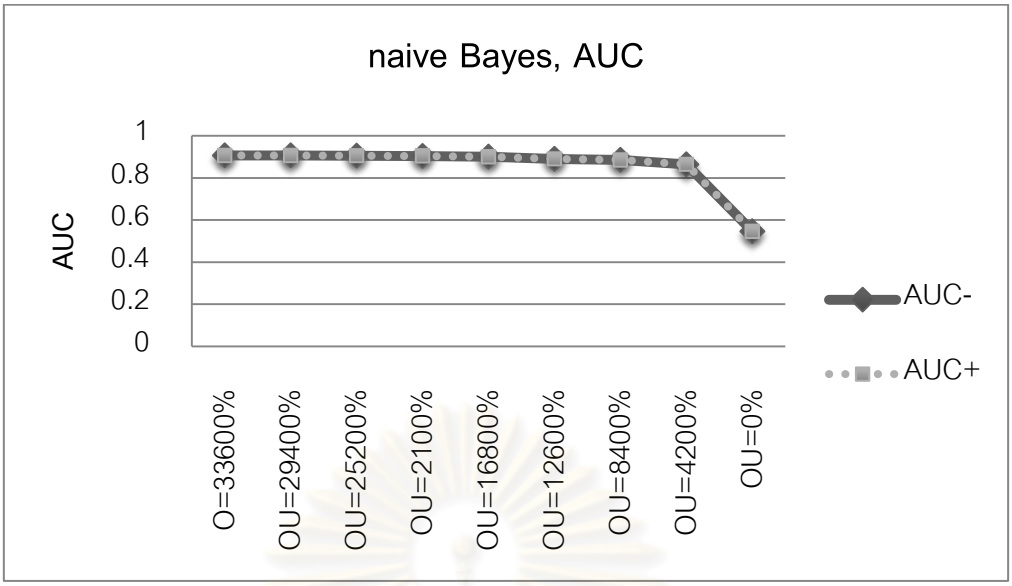


รูป 4.21 ค่า F+ และค่า F- ของผลการทำนายข้อมูลด้วย C4.5 บนชุดข้อมูล shuttle

จากรูป 4.20 และ 4.21 ประสิทธิภาพในการทำนายข้อมูลของคลาสลบและคลาสบวกด้วยวิธีการ SMOUTE และวิธีการ SMOTE ไม่มีความแตกต่างกัน ยกเว้นวิธีการ SMOUTE ที่ OU=0% ให้ค่าการทำนายข้อมูลของคลาสบวกลดลง เนื่องจากชุดข้อมูล shuttle มีจำนวนข้อมูลมาก และมีการกระจายตัวของกลุ่มข้อมูลของคลาสบวกแยกกับกลุ่มข้อมูลของคลาสลบค่อนข้างชัดเจน ส่งผลให้การทำนายข้อมูลด้วยตัวแบบแยกประเภท C4.5 มีความแม่นยำสูงมาก แม้ว่าจะลดข้อมูลของคลาสลบเป็นจำนวนมากก็ตาม

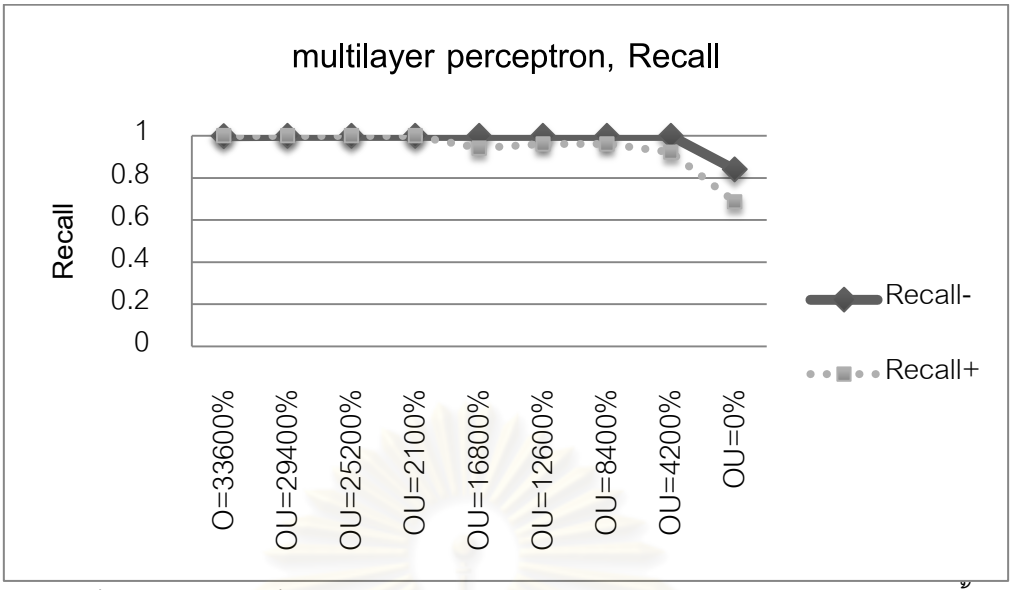


รูป 4.22 ค่ารีคอล+ และค่ารีคอล- ของผลการทำนายข้อมูลด้วยการแบ่งประเภทเบย์อย่างง่ายบนชุดข้อมูล shuttle

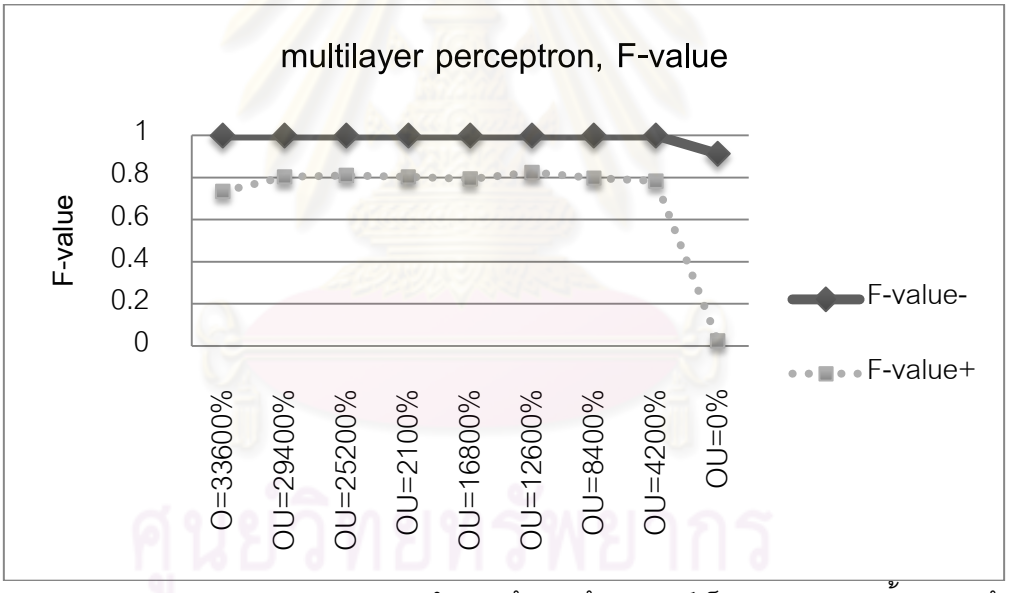


รูป 4.23 ค่า AUC+ และค่า AUC- ของผลการทำนายข้อมูลด้วยการแบ่งประเภทเบย์อย่างง่ายบนชุดข้อมูล shuttle

จากรูป 4.22 และ 4.23 ประสิทธิภาพในการทำนายข้อมูลของคลาสบวกด้วยวิธีการ SMOUTE และวิธีการ SMOTE ไม่มีความแตกต่างกัน ในขณะที่ประสิทธิภาพในการทำนายข้อมูลของคลาสลบด้วยวิธีการ SMOUTE และวิธีการ SMOTE ลดลงมาก เนื่องจากชุดข้อมูล shuttle อาจมีความสัมพันธ์กันของลักษณะประจำแต่ละตัวมาก เมื่อลดจำนวนข้อมูลของคลาสลบจำนวนมาก ส่งผลให้ความแม่นยำในการทำนายข้อมูลของคลาสลบลดลงมากเมื่อใช้การแบ่งประเภทเบย์อย่างง่ายเป็นตัวแบบแยกประเภท



รูป 4.24 ค่ารีคอล+ และค่ารีคอล- ของผลการทำนายข้อมูลด้วยเพอร์เซ็ปตรอนหลายชั้นบนชุดข้อมูล shuttle



รูป 4.25 ค่า F+ และค่า F- ของผลการทำนายข้อมูลด้วยเพอร์เซ็ปตรอนหลายชั้นบนชุดข้อมูล shuttle

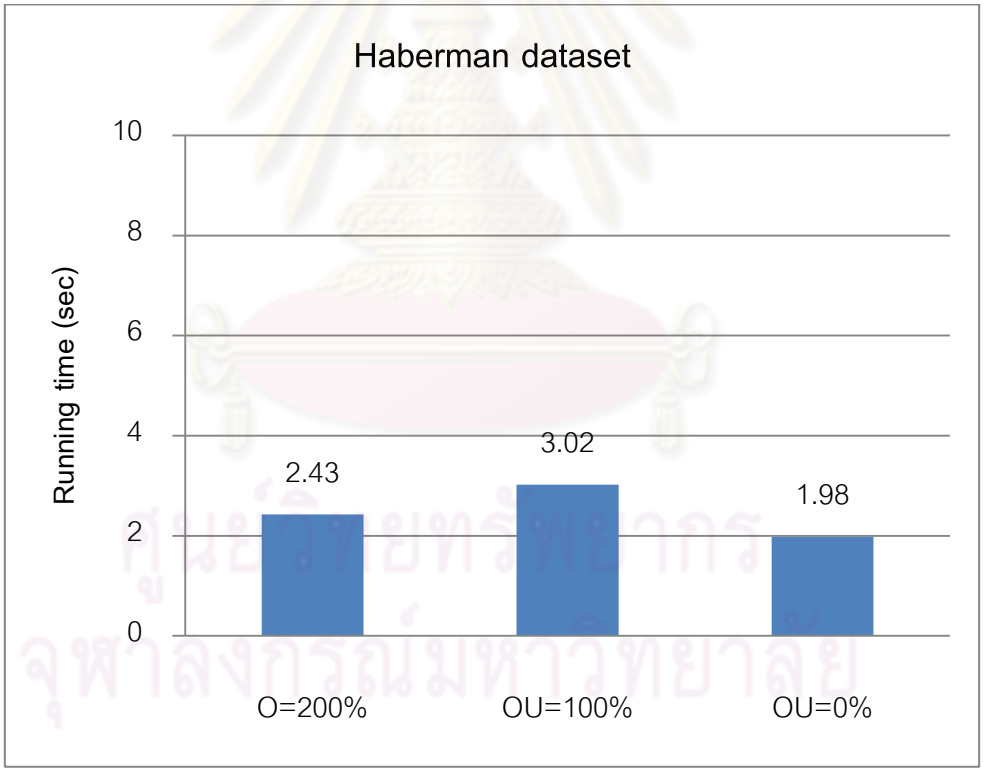
จากรูป 4.24 และ 4.25 ประสิทธิภาพในการทำนายข้อมูลของคลาสลบและคลาสบวกด้วยวิธีการ SMOUTE และวิธีการ SMOTE ไม่มีความแตกต่างกัน ยกเว้นวิธีการ SMOUTE ที่ OU=0% ให้ค่าการทำนายข้อมูลของคลาสบวกลดลงมาก เนื่องจากชุดข้อมูล shuttle มีจำนวนข้อมูลมาก และมีการกระจายตัวของกลุ่มข้อมูลของคลาสบวกแยกกับกลุ่มข้อมูลของคลาสลบค่อนข้างชัดเจน ส่งผลให้การทำนายข้อมูลด้วยตัวแบบแยกประเภทเพอร์เซ็ปตรอนหลายชั้นมี

ความแม่นยำสูง ในขณะที่การปรับเปลี่ยนรูปแบบการกระจายตัวของข้อมูลของคลาสบวกด้วยวิธีการ SMOTE สามารถเพิ่มความแม่นยำในการทำนายข้อมูลของคลาสบวกเมื่อใช้เพอร์เซ็ปตรอนหลายชั้นเป็นตัวแบบแยกประเภท

4.2.2 ผลการทดสอบความเร็วในการประมวลผลของวิธีการ SMOUTE และวิธีการ SMOTE

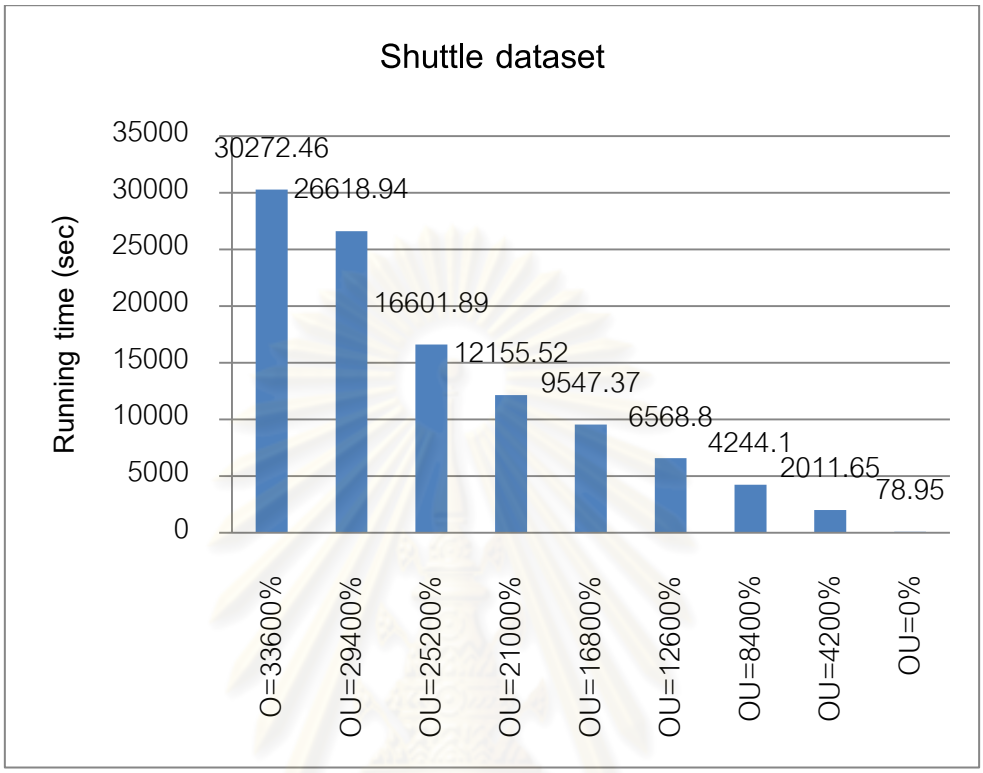
O=r% คือ r เปอร์เซ็นต์ของจำนวนข้อมูลของคลาสบวกที่ถูกเพิ่มด้วยวิธีการ SMOTE จนมีจำนวนข้อมูลของคลาสบวกใกล้เคียงกับจำนวนข้อมูลของคลาสลบ

OU=r% คือ r เปอร์เซ็นต์ของจำนวนข้อมูลของคลาสบวกที่ถูกเพิ่มด้วยวิธีการ SMOTE และลบจำนวนข้อมูลของคลาสลบจนกระทั่งข้อมูลทั้งสองคลาสมีจำนวนใกล้เคียงกัน (ใช้วิธีการ SMOUTE)



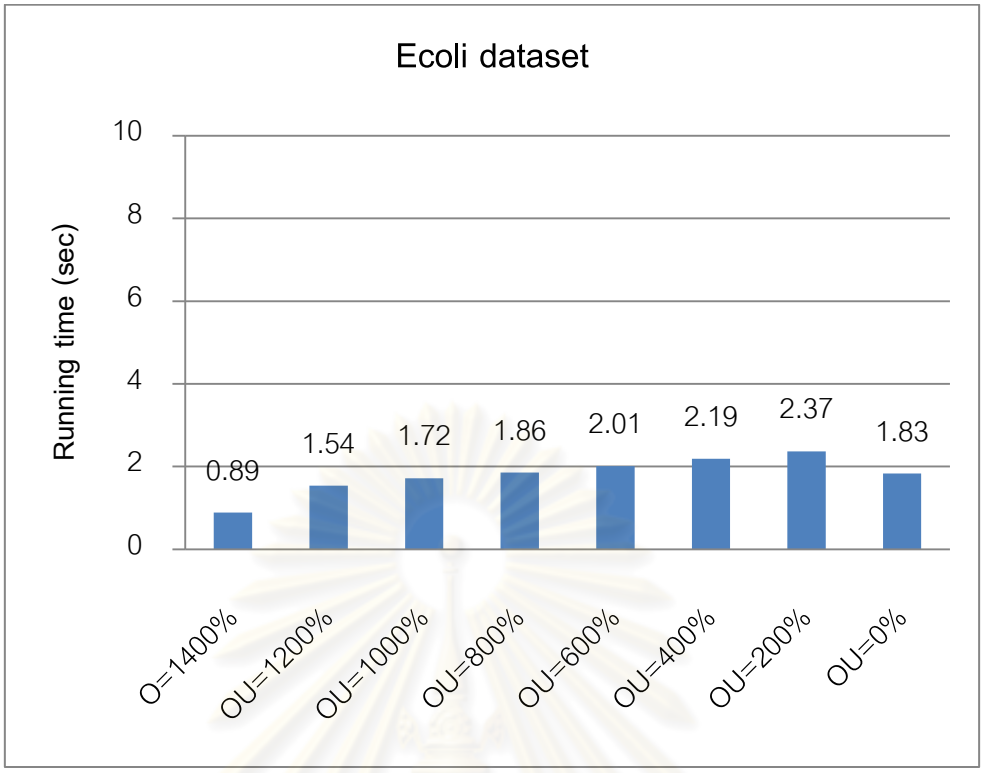
รูป 4.26 ระยะเวลาในการประมวลผลของวิธีการ SMOUTE และวิธีการ SMOUTE บนชุดข้อมูล Haberman

จากรูป 4.26 ระยะเวลาในการประมวลผลของวิธีการ SMOTE (O=200%) วิธีการ SMOUTE ที่ OU=100% และวิธีการ SMOUTE ที่ OU=0% คือ 2.43 3.02 และ 1.98 วินาที ตามลำดับ



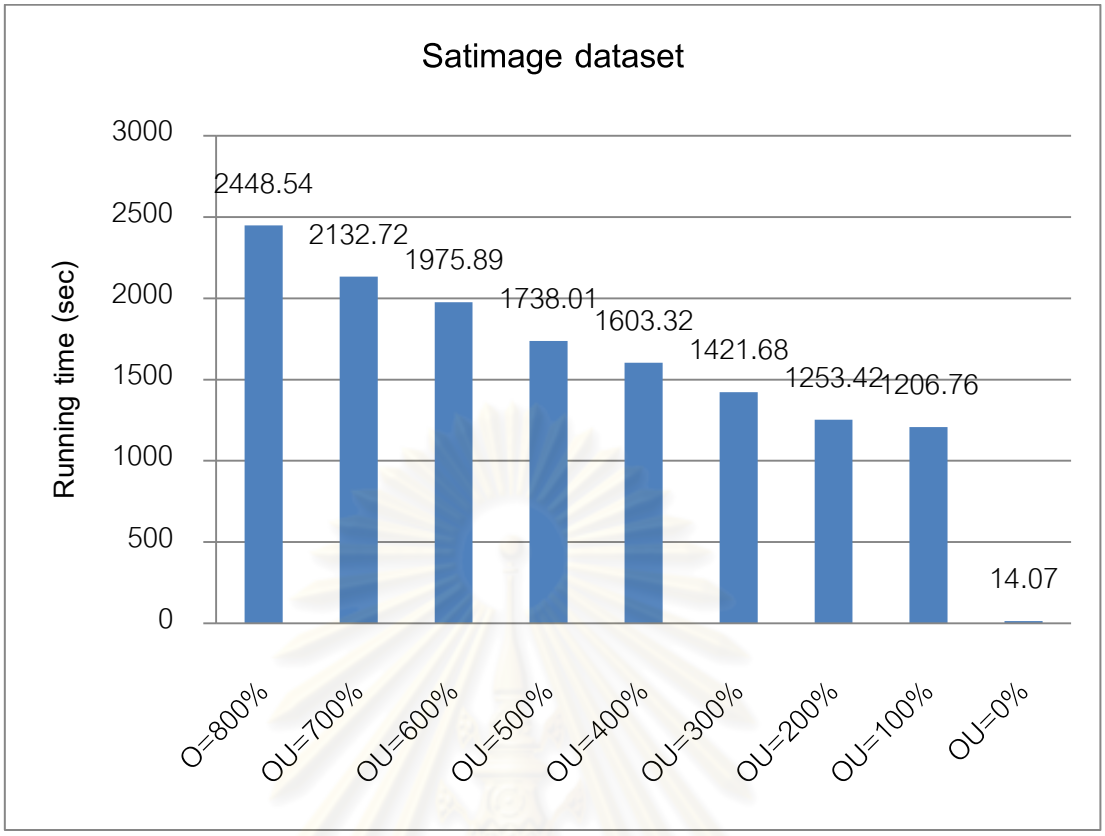
รูป 4.27 ระยะเวลาในการประมวลผลของวิธีการ SMOUTE และวิธีการ SMOUTE บนชุดข้อมูล satimage

จากรูป 4.27 ระยะเวลาในการประมวลผลของวิธีการ SMOTE (O=800%) วิธีการ SMOUTE ที่ OU=700% วิธีการ SMOUTE ที่ OU=600% วิธีการ SMOUTE ที่ OU=500% วิธีการ SMOUTE ที่ OU=400% วิธีการ SMOUTE ที่ OU=300% วิธีการ SMOUTE ที่ OU=200% วิธีการ SMOUTE ที่ OU=100% และวิธีการ SMOUTE ที่ OU=0% คือ 2448.54 2132.72 1975.89 1738.01 1603.32 1421.68 1253.42 1206.76 และ 14.07 วินาทีตามลำดับ



รูป 4.28 ระยะเวลาในการประมวลผลของวิธีการ SMOUTE และวิธีการ SMOUTE บนชุดข้อมูล ecoli

จากรูป 4.28 ระยะเวลาในการประมวลผลของวิธีการ SMOTE (O=1400%) วิธีการ SMOUTE ที่ OU=1200% วิธีการ SMOUTE ที่ OU=1000% วิธีการ SMOUTE ที่ OU=800% วิธีการ SMOUTE ที่ OU=600% วิธีการ SMOUTE ที่ OU=400% วิธีการ SMOUTE ที่ OU=200% และวิธีการ SMOUTE ที่ OU=0% คือ 0.89 1.54 1.72 1.86 2.01 2.19 2.37 และ 1.83 วินาที ตามลำดับ

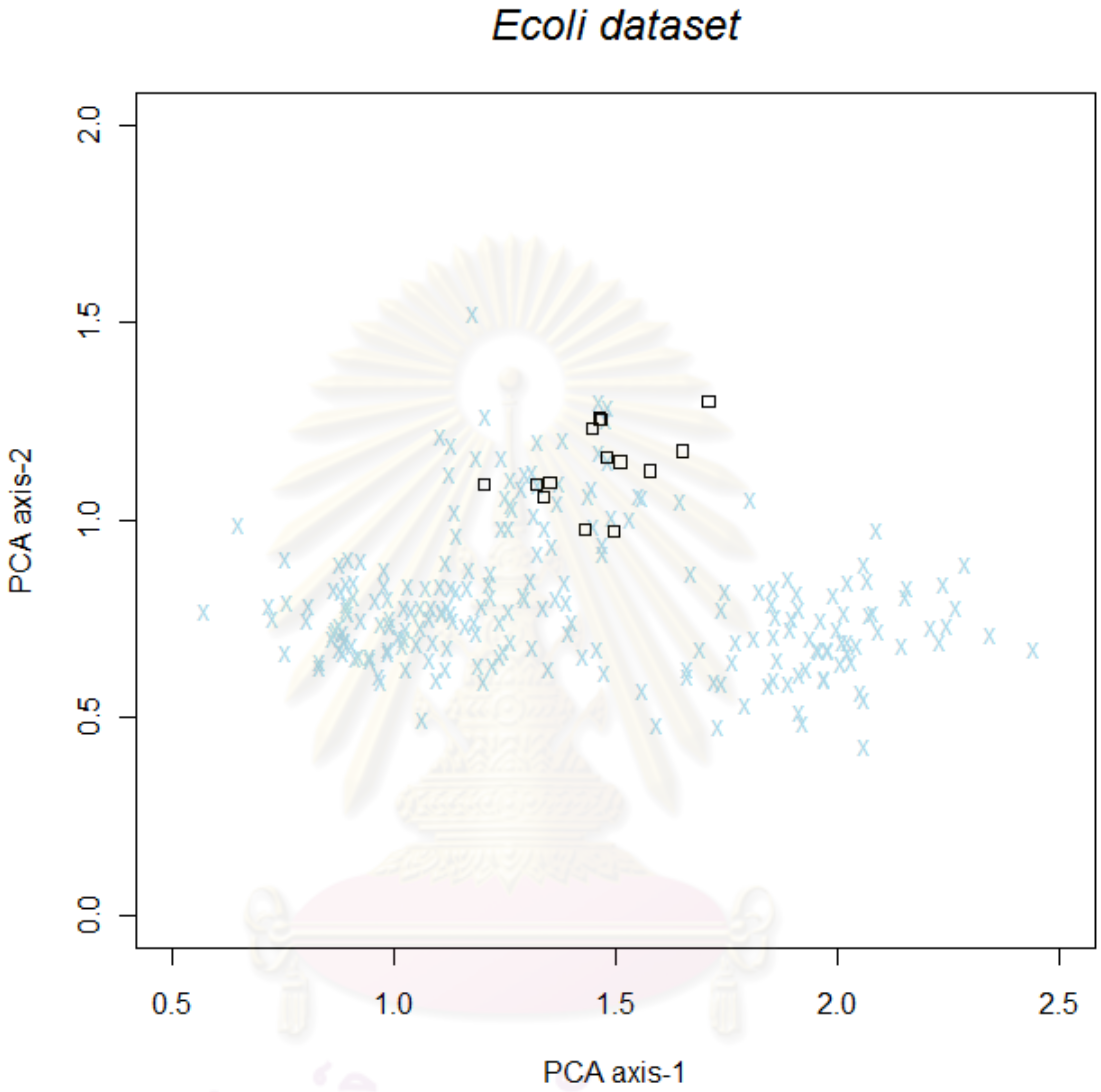


รูป 4.29 ระยะเวลาในการประมวลผลของวิธีการ SMOUTE และวิธีการ SMOUTE บนชุดข้อมูล shuttle

จากรูป 4.28 ระยะเวลาในการประมวลผลของวิธีการ SMOTE (O=33600%) วิธีการ SMOUTE ที่ OU=29400% วิธีการ SMOUTE ที่ OU=25200% วิธีการ SMOUTE ที่ OU=21000% วิธีการ SMOUTE ที่ OU=16800% วิธีการ SMOUTE ที่ OU=12600% วิธีการ SMOUTE ที่ OU=8400% SMOUTE ที่ OU=4200% และวิธีการ SMOUTE ที่ OU=0% และ 1.83 คือ 30272.46 26618.94 16601.89 12155.52 9547.37 6568.8 4244.1 2011.65 และ 78.95 วินาทีตามลำดับ

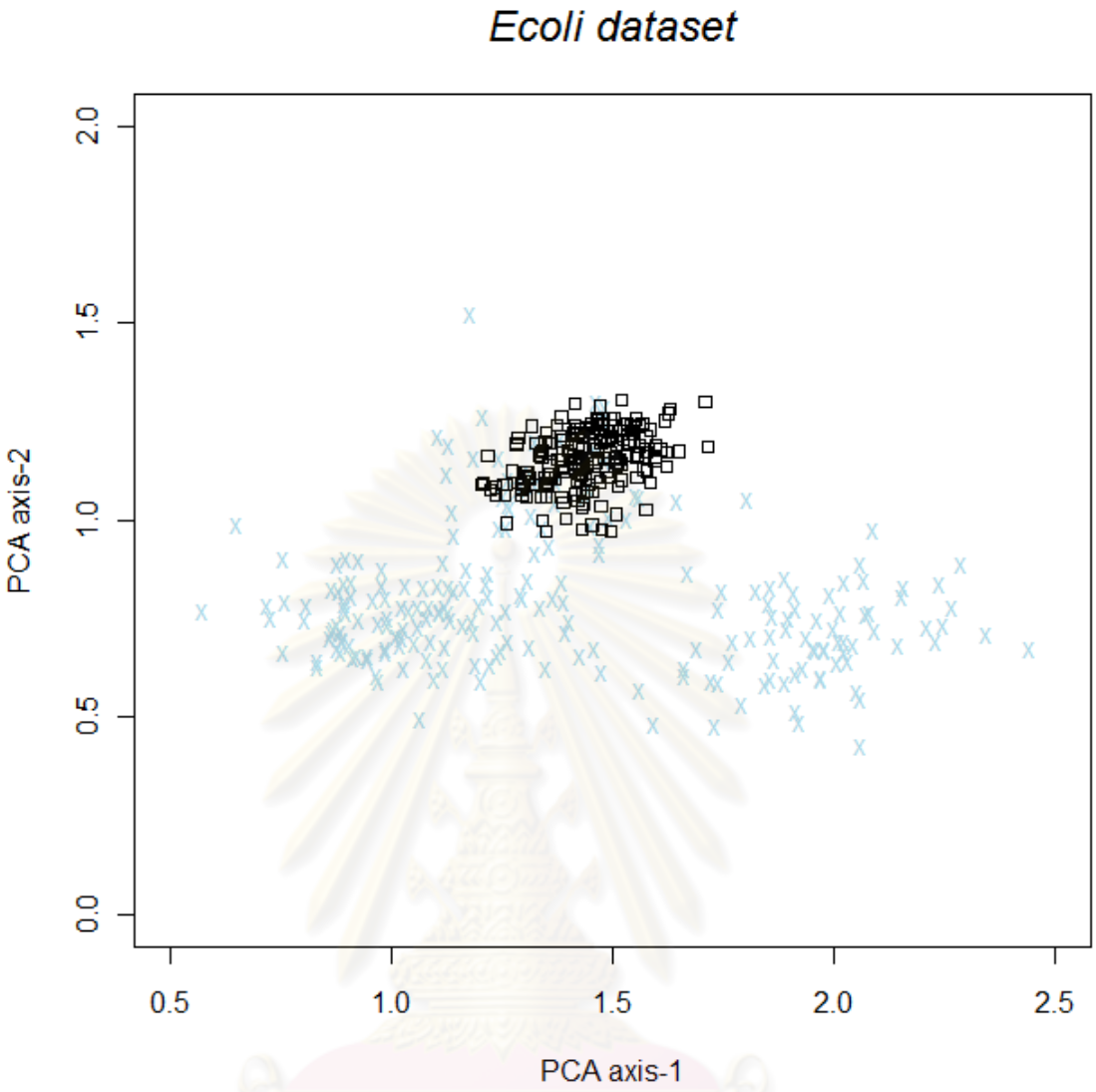
สำหรับชุดข้อมูล satimage และชุดข้อมูล shuttle วิธีการ SMOUTE ลดจำนวนการเพิ่มจำนวนของตัวสังเคราะห์ของคลาสบวกที่ถูกลดลงและแทนที่ด้วยการลดจำนวนข้อมูลของคลาสลบ ทำให้ระยะเวลาในการประมวลผลด้วยวิธีการ SMOUTE เร็วกว่าวิธีการ SMOTE และเมื่อลดเปอร์เซ็นต์ของการเพิ่มจำนวนของตัวสังเคราะห์ของคลาสบวกลงก็จะทำให้ความเร็วในการประมวลผลของวิธีการ SMOUTE เร็วขึ้นเช่นกัน

4.3 ภาพประกอบการประมวลผลของวิธีการ SMOUTE และวิธีการ SMOTE บนชุดข้อมูล ecoli

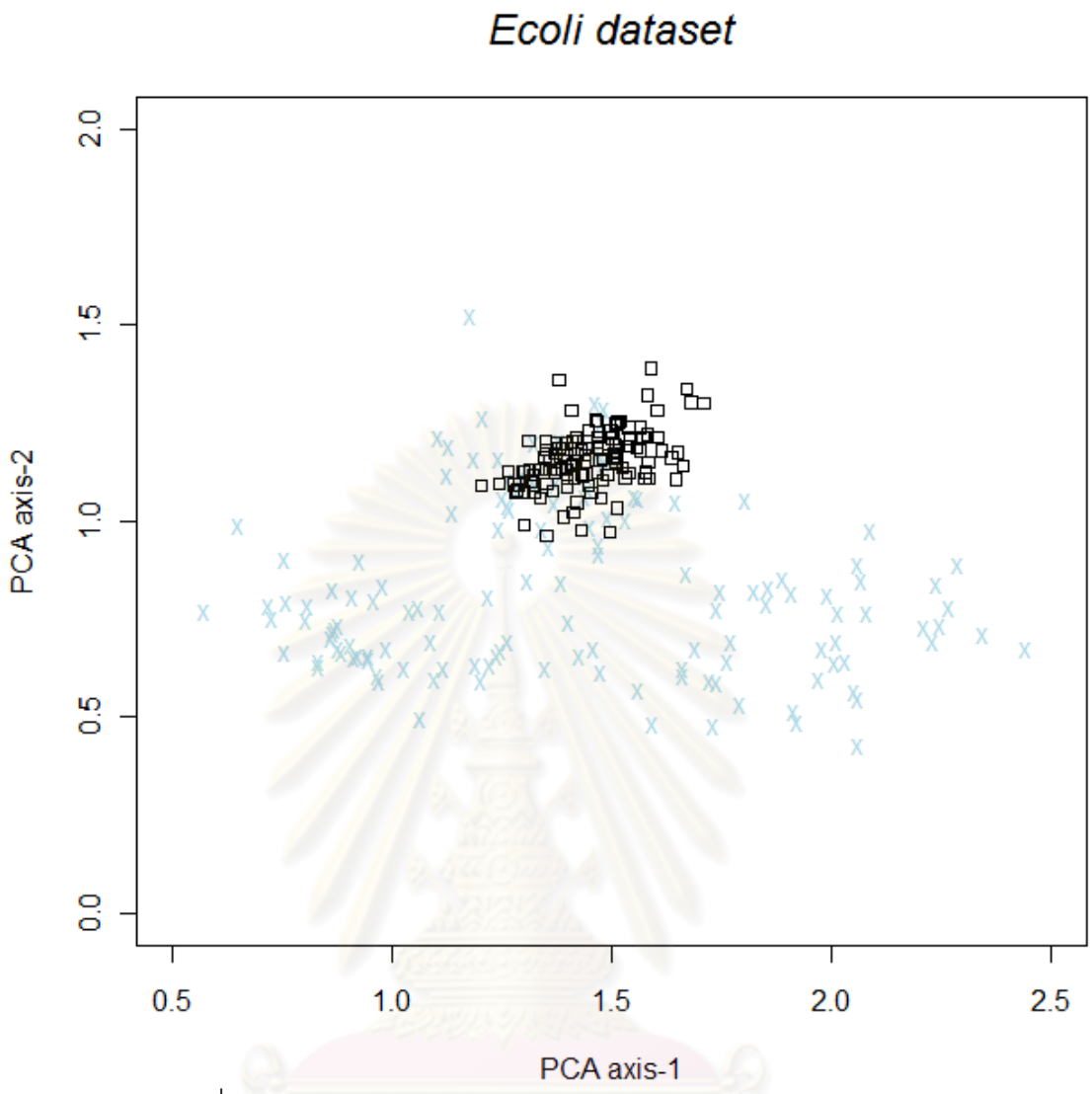


รูปที่ 4.30 ชุดข้อมูลฝึกหัด ecoli ชุดที่หนึ่ง

จากรูปที่ 4.30 สี่เหลี่ยมแทนข้อมูลไมนอริตี้คลาสและกากบาทแทนข้อมูลมาจอริตี้คลาส แนวแกน X คือแนวแกนที่ได้จากการวิเคราะห์องค์ประกอบหลักแนวแกนที่หนึ่ง และแนวแกน Y คือแนวแกนที่ได้จากการวิเคราะห์องค์ประกอบหลักแนวแกนที่สอง



รูปที่ 4.31 ชุดข้อมูล ecoli หลังจากการใช้ SMOTE (O=1400%)



รูปที่ 4.32 ชุดข้อมูล ecoli หลังจากการใช้ SMOUTE (OU=800%)

จากรูป 4.31 ชุดข้อมูลฝึกหัด ecoli หลังจากใช้วิธีการ SMOTE (O=1400%) มีความหนาแน่นของจำนวนข้อมูลของคลาสบวกมากเกินไป ในขณะที่ข้อมูลของคลาสลบมีการกระจายตัวในลักษณะไม่แตกต่างจากชุดข้อมูลฝึกหัดแรกเริ่ม (รูป 4.30)

จากรูป 4.32 ชุดข้อมูลฝึกหัด ecoli หลังจากใช้วิธีการ SMOUTE (OU=800%) มีความหนาแน่นของจำนวนข้อมูลของคลาสบวกน้อยกว่าชุดข้อมูลฝึกหัด ecoli หลังจากใช้วิธีการ SMOTE และลดจำนวนข้อมูลของคลาสลบบริเวณที่อยู่ไกลจากข้อมูลของคลาสบวกซึ่งเป็นข้อมูลที่มีความสำคัญน้อย การเปลี่ยนรูปแบบการกระจายตัวของข้อมูลของคลาสลบและลดความหนาแน่นของจำนวนข้อมูลของคลาสบวกส่งผลให้การทำนายข้อมูลของคลาสบวกถูกต้องมากขึ้น เนื่องจากสามารถลดปัญหาความจำเพาะเกินในการสร้างตัวแบบแยกประเภท

บทที่ 5

สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

เทคนิคการสุ่มเพิ่มตัวอย่างข้างน้อยสังเคราะห์และเทคนิคการสุ่มลดตัวอย่างข้างมากสำหรับปัญหาความไม่ดุลระหว่างกลุ่ม (SMOUTE) เป็นกระบวนจัดการข้อมูลก่อนการสร้างตัวแบบแยกประเภท โดยการใช้เทคนิคการชักตัวอย่างแบบเพิ่มและเทคนิคการชักตัวอย่างแบบลดเพื่อจัดการกับปัญหาความไม่ดุลระหว่างกลุ่ม ในวิทยานิพนธ์นี้ เราเปรียบเทียบประสิทธิภาพในการทำนายข้อมูลของคลาสเป้าหมาย และเปรียบเทียบความเร็วในการประมวลผลของวิธีการ SMOUTE กับวิธีการ SMOTE โดยนำชุดข้อมูลที่ได้จากการประมวลผลด้วยวิธีการทั้งสองวิธีมาใช้เป็นชุดข้อมูลฝึกหัดสำหรับการสร้างตัวแบบแยกประเภทด้วย C4.5 การแบ่งประเภทเบย์อย่างง่าย และเพอร์เซ็ปตรอนหลายชั้น

จากผลการทดลอง เมื่อใช้วิธีการ SMOUTE โดยเปรียบเทียบกับวิธีการ SMOTE ในการสร้างชุดข้อมูลพบว่า บนชุดข้อมูลที่มีบริเวณข้อมูลที่มีความหนาแน่นของคลาสลบอยู่ไกลจากข้อมูลของคลาสบวก ส่งผลให้วิธีการ SMOUTE มีประสิทธิภาพในการทำนายข้อมูลของคลาสลบและคลาสบวกสูง ในทางตรงกันข้าม หากชุดข้อมูลมีบริเวณข้อมูลที่มีความหนาแน่นของคลาสลบอยู่ใกล้กับข้อมูลของคลาสบวก จะส่งผลให้ประสิทธิภาพในการทำนายข้อมูลของคลาสลบลดลง แต่เมื่อวัดด้วยค่า AUC+ และ AUC- ประสิทธิภาพในการทำนายข้อมูลของวิธีการ SMOUTE และวิธีการ SMOTE ไม่แตกต่างกัน และระยะเวลาในการประมวลผลของวิธีการ SMOUTE เร็วกว่าวิธีการ SMOTE มากเมื่อใช้กับชุดข้อมูลที่มีจำนวนข้อมูลมาก

วิธีการ SMOUTE สามารถลดจำนวนของชุดข้อมูลฝึกหัดได้ ซึ่งเป็นประโยชน์สำหรับชุดข้อมูลที่มีจำนวนข้อมูลมาก เนื่องจากชุดข้อมูลฝึกหัดที่มีจำนวนข้อมูลมากอาจทำให้ซอฟต์แวร์ที่ใช้สำหรับการสร้างตัวแบบแยกประเภทมีพื้นที่ของหน่วยความจำไม่เพียงพอต่อการเก็บข้อมูลทั้งหมด ดังนั้นการลดจำนวนข้อมูลของชุดข้อมูลฝึกหัดจะทำให้พื้นที่ของหน่วยความจำเพียงพอต่อการรองรับชุดข้อมูลฝึกหัด

วิธีการ SMOUTE ที่มีการลบจำนวนข้อมูลของคลาสลบน้อยกว่าครึ่งหนึ่งของจำนวนข้อมูลของคลาสลบทั้งหมดให้ค่าการทำนายข้อมูลของคลาสลบและคลาสบวกใกล้เคียงกับวิธีการ SMOTE เมื่อวัดด้วยตัววัดพรีสิชัน- รีคอล- และค่า F- วิธีการ SMOUTE ที่มีการลบจำนวนข้อมูลของคลาสลบมากกว่าครึ่งหนึ่งของจำนวนข้อมูลของคลาสลบทั้งหมดมีประสิทธิภาพในการทำนายข้อมูลของคลาสลบน้อยลง แต่ประสิทธิภาพในการทำนายข้อมูลของคลาสบวกมากขึ้นเมื่อวัดด้วยตัววัดพรีสิชัน+ รีคอล+ และค่า F+ แต่เมื่อวัดด้วยตัววัด AUC+ และ AUC- วิธีการ SMOUTE และวิธีการ SMOTE ไม่มีความแตกต่างกัน การลบข้อมูลของคลาสลบเพียงอย่างเดียวโดยไม่เพิ่ม

จำนวนข้อมูลของคลาสบวก (วิธีการ SMOUTE ที่ OU=0%) ส่งผลให้ประสิทธิภาพในการทำนายข้อมูลของคลาสลบและคลาสบวกลดลงมากเมื่อวัดด้วยตัววัดประสิทธิภาพพรีสิชัน+ พรีสิชัน- รีคอล+ รีคอล- ค่า F+ ค่า F- AUC+ และ AUC- โดยเฉพาะเมื่อใช้เพอร์เซ็ปตรอนหลายชั้นเป็นตัวแบบแยกประเภท เนื่องจากการลบข้อมูลของคลาสลบบริเวณใกล้เคียงกับเซนทรอยด์เพียงเล็กน้อย จะทำให้ตำแหน่งของเซนทรอยด์เปลี่ยนแปลงเพียงเล็กน้อย ในขณะที่การลบข้อมูลของคลาสลบมากเกินไป จะทำให้ตำแหน่งของเซนทรอยด์เปลี่ยนแปลงไปจากตำแหน่งเดิมมาก ซึ่งหมายถึงโครงสร้างของกลุ่มข้อมูลมีการเปลี่ยนแปลงมากเช่นกัน

เมื่อเปรียบเทียบผลการทดลองวิธีการ SMOUTE กับตัวแบบแยกประเภท C4.5 การแบ่งประเภทเบย์อย่างง่าย และเพอร์เซ็ปตรอนหลายชั้น การลดจำนวนข้อมูลของคลาสลบด้วยวิธีการ SMOUTE เป็นจำนวนมากบนชุดข้อมูลที่มีความมีความสัมพันธ์กันของลักษณะประจำแต่ละตัวมาก จะส่งผลให้ประสิทธิภาพในการทำนายข้อมูลของคลาสลบลดลงมากเมื่อใช้วิธีการ SMOUTE ร่วมกับการแบ่งประเภทเบย์อย่างง่าย ในขณะที่วิธีการ SMOUTE ที่ประยุกต์กับตัวแบบแยกประเภท C4.5 และตัวแบบแยกประเภทเพอร์เซ็ปตรอนหลายชั้น ประสิทธิภาพในการทำนายข้อมูลของคลาสลบและคลาสบวกไม่แตกต่างกันทุกชุดข้อมูล เนื่องจากการลบข้อมูลของคลาสลบบริเวณที่มีความหนาแน่นสูงและเลือกใช้ข้อมูลของคลาสลบบริเวณขอบของแต่ละกลุ่มเป็นชุดข้อมูลฝึกหัด SMOUTE จึงเหมาะสำหรับตัวแบบแยกประเภทที่ใช้หลักการแบ่งข้อมูลเชิงเส้น (Linearly separable) และตัวแบบแยกประเภทที่ใช้หลักการแบ่งข้อมูลไม่เชิงเส้น (Non-linearly separable)

อย่างไรก็ดี การเลือกจำนวนกลุ่มที่ใช้ในการแบ่งกลุ่มข้อมูลของคลาสลบด้วยวิธีการวิเคราะห์องค์ประกอบหลักไม่สามารถเลือกกลุ่มที่เหมาะสมได้เมื่อบริเวณที่มีความหนาแน่นของข้อมูลของคลาสลบอยู่ใกล้กับข้อมูลของคลาสบวก และการลบข้อมูลของคลาสลบ ณ บริเวณดังกล่าวส่งผลให้ประสิทธิภาพในการทำนายข้อมูลของคลาสลบลดลง

ในอนาคต วิธีการ SMOUTE สามารถพัฒนาให้มีประสิทธิภาพในการทำนายข้อมูลได้ดียิ่งขึ้น โดยการเลือกลบข้อมูลของคลาสลบเฉพาะกลุ่มข้อมูลของคลาสลบที่มีเซนทรอยด์อยู่ไกลจากเซนทรอยด์ของกลุ่มข้อมูลของคลาสบวก และการเลือกลบข้อมูลของคลาสลบเพียงบางตัวบริเวณรอบเซนทรอยด์ของแต่ละกลุ่ม เพื่อรักษารูปแบบการกระจายตัวของข้อมูลของคลาสลบไม่ให้เปลี่ยนแปลงมากเกินไป และในกรณีที่มีคลาสเป้าหมายมากกว่าสองคลาส เราสามารถพัฒนาวิธีการ SMOUTE ได้โดยใช้วิธีการเกาะกลุ่ม (Clustering) รวมกลุ่มข้อมูลเป็นกลุ่มๆสำหรับแต่ละคลาสเป้าหมาย แล้วค่อยทำการลบข้อมูลบริเวณที่อยู่ใกล้กับเซนทรอยด์ของแต่ละกลุ่ม ด้วยวิธีการดังกล่าวจะทำให้โครงสร้างของกลุ่มข้อมูลของคลาสเป้าหมายแต่ละคลาสไม่เปลี่ยนแปลงมากเกินไป

รายการอ้างอิงหนังสือ

[1] Shmueli, G., Patel, R.N., and Bruce, C.P., Data Mining for Business Intelligence. New Jersey: John Wiley & Sons, 2007.

[2] Altincay, H., and Ergun, C., "Clustering Based Under-Sampling for Improveing Speaker Verification Decisionss Using AdaBoost", SSPR&SPR 2004 (2004), 698-706.

[3] Chawla, N. V., Japkowicz, N., and Kolcz, A., "Editorial: Special Issue on Learning from Imbalanced Data Sets", ACM SIGKDD Explorations Newsletter (2004), 1-6.

[4] Hart, P. E., "The Condensed Nearest Neighbor Rule, "IEEE Transactions on Information Theory IT-14 (1968), 515–516.

[5] Tomek, I., "Two Modifications of CNN," IEEE Transactions on Systems Man and Communications SMC-6 (1976), 769–772.

[6] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmayer, W. P., "SMOTE: Synthetic Minority Over-Sampling Technique," Journal of Artificial Intelligent Research (2002,) 321–357.

[7] Han, H., Wang, W.Y., and Mao, B.H., "Borderline-SMOTE: A new Over-Sampling Method in Imbalanced Data Sets Learning," International Conference on Intelligent Computing (2005), 878-887.

[8] Bunkhumpornpat, C., C., Sinapiromsaran, K., and Lursinsap, C., "Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem," Lecture Notes in Computer Science (2009), 475-48.

[9] Batista, G. E. A., Prati R. C., and Monard, M. C., "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," ACM SIGKDD Explorations Newsletter (2004), 20-29.

[10] Lin, Z. Y., Hao, Z. F., Yang, X. W., and Liu, X. L., "Several SVM Ensemble Methods Integrated with Under-Sampling for Imbalanced Data Learning," Lecture Notes in Computer Science (2009), 536–544.

[11] Estabrooks, A., Jo, T., and Japkowicz, N., "A Multiple Resampling Method for Learning from Imbalanced Data Sets," Computational Intelligence (2004), 18-36.

[12] Zheng, Z., Wu, X., and Srihari, R., "Feature selection fortext categorization on imbalanced data," SIGKDD Ex-plorations (2004), 80-89.

[13] Han, J. W., and Kamber, M., Data Mining: Concepts and Techniques, Elsevier Inc, 2006.

[14] Rohani, B., and Nugroho, B., "Manhattan-Chebychev Distance Metric for MIMO Systems," IEICE Technical Committee Submission System Conference (2008), 49-52.

[15] Haykin, S., NEURAL NETWORKS: A Comprehensive Foundation, Second edition, Prentice Hall International Inc, 1999.

[16] Songwattanasiri, P., and Sinapiromsaran, K., "SMOUTE:Synthetics Minority Over-sampling and Under-sampling Teniques for class imbalanced problem," Annual International Conference on Computer Science Education: Innovation & Technology (CSEIT) 2010 (2010), 78-83.

[17] Haberman, "S. J. Generalized Residuals for Log-Linear Models," Proceedings of the 9th International Biometrics Conference, (1976): 104-122.

[18] Frank, A., and Asuncion, A., UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science (2010).

[19] King, R. D., Karwath, A., Clare, A., and Dehaspe, L., G78/6609, BIF08765, GR/L62849 by PharmaDM, Ambachtenlaan, 54/D, B-3001 Leuven, Belgium (2000).

[20] Catlett, J., Basser Department of Computer Science, University of Sydney, N.S.W., Australia.

ภาคผนวก

ภาคผนวก

ในส่วนนี้ เราแสดงผลการทดสอบประสิทธิภาพในการทำนายข้อมูลของวิธีการ SMOUTE เปรียบเทียบกับวิธีการ SMOTE โดยมีการกำหนดสัญลักษณ์ดังนี้

O=r% คือ r เปอร์เซ็นต์ของจำนวนข้อมูลของคลาสบวกที่ถูกเพิ่มด้วยวิธีการ SMOTE จนมีจำนวนข้อมูลของคลาสบวกใกล้เคียงกับจำนวนข้อมูลของคลาสลบ

OU=r% คือ r เปอร์เซ็นต์ของจำนวนข้อมูลของคลาสบวกที่ถูกเพิ่มด้วยวิธีการ SMOTE และลบจำนวนข้อมูลของคลาสลบจนกระทั่งข้อมูลทั้งสองคลาสมีจำนวนใกล้เคียงกัน (ใช้วิธีการ SMOUTE)

ตาราง ก-1 ผลการทำนายชุดข้อมูล Haberman ด้วย C4.5

Haberman	จำนวนข้อมูล	C4.5							
		พรีสิชัน		รีคอล		ค่า F		AUC	
		-	+	-	+	-	+	-	+
O=200%	329	.769	.326	.690	.417	.724	.360	.558	.558
OU=100%	220	.790	.326	.582	.558	.665	.406	.554	.554
OU=0%	110	.790	.391	.664	.494	.705	.391	.570	.570

ตาราง ก-2 ผลการทำนายชุดข้อมูล Haberman ด้วยการแบ่งประเบย์อย่างง่าย

Haberman	จำนวนข้อมูล	การแบ่งประเบย์อย่างง่าย							
		พรีสิชัน		รีคอล		ค่า F		AUC	
		-	+	-	+	-	+	-	+
O=200%	329	0.763	0.407	0.863	0.300	0.816	0.349	0.602	0.602
OU=100%	220	0.758	0.381	0.830	0.308	0.797	0.338	0.583	0.583
OU=0%	110	0.761	0.325	0.658	0.412	0.700	0.350	0.562	0.562

ตาราง ก-3 ผลการทำนายชุดข้อมูล Haberman ด้วยเพอร์เซ็ปตรอนหลายชั้น

Haberman	จำนวนข้อมูล	เพอร์เซ็ปตรอนหลายชั้น							
		พรีสิชัน		รีคอล		ค่า F		AUC	
		-	+	-	+	-	+	-	+
O=200%	329	0.810	0.393	0.707	0.525	0.751	0.441	0.658	0.658
OU=100%	220	0.786	0.436	0.788	0.400	0.783	0.400	0.645	0.645
OU=0%	110	0.799	0.356	0.539	0.611	0.611	0.420	0.593	0.593

ตาราง ก-4 ผลการทำนายชุดข้อมูล satimage ด้วย C4.5

Satimage	จำนวนข้อมูล	C4.5							
		พรีสิชัน		รีคอล		ค่า F		AUC	
		-	+	-	+	-	+	-	+
O=800%	8081	0.963	0.58	0.934	0.668	0.948	0.585	0.777	0.777
OU=700%	7126	0.972	0.671	0.925	0.749	0.948	0.611	0.88	0.88
OU=600%	6235	0.971	0.68	0.905	0.749	0.937	0.568	0.781	0.781
OU=500%	5345	0.972	0.671	0.916	0.754	0.943	0.594	0.871	0.871
OU=400%	4454	0.978	0.686	0.869	0.818	0.92	0.539	0.83	0.83
OU=300%	3563	0.973	0.721	0.835	0.786	0.899	0.473	0.818	0.818
OU=200%	2672	0.977	0.681	0.802	0.824	0.881	0.449	0.842	0.842
OU=100%	1781	0.982	0.681	0.794	0.861	0.878	0.455	0.857	0.857
OU=0%	890	0.988	0.012	0.705	0.92	0.823	0.394	0.811	0.811

ตาราง ก-5 ผลการทำนายชุดข้อมูล satimage ด้วยการแบ่งประเภทเบย์อย่างง่าย

Satimage	จำนวนข้อมูล	การแบ่งประเบย์อย่างง่าย							
		พรีสิชัน		รีคอล		ค่า F		AUC	
		-	+	-	+	-	+	-	+
O=800%	8081	0.984	0.42	0.871	0.866	0.924	0.565	0.917	0.929
OU=700%	7126	0.984	0.414	0.869	0.866	0.923	0.561	0.917	0.929
OU=600%	6235	0.984	0.406	0.864	0.866	0.92	0.553	0.917	0.929
OU=500%	5345	0.984	0.403	0.862	0.866	0.919	0.55	0.919	0.93
OU=400%	4454	0.984	0.392	0.855	0.872	0.915	0.541	0.921	0.931
OU=300%	3563	0.984	0.384	0.85	0.872	0.912	0.533	0.922	0.932
OU=200%	2672	0.984	0.362	0.834	0.877	0.903	0.513	0.931	0.932
OU=100%	1781	0.985	0.336	0.812	0.888	0.89	0.488	0.932	0.933
OU=0%	890	0.991	0.288	0.752	0.936	0.855	0.441	0.933	0.933

ตาราง ก-6 ผลการทำนายชุดข้อมูล satimage ด้วยเพอร์เซ็ปตรอนหลายชั้น

Satimage	จำนวนข้อมูล	เพอร์เซ็ปตรอนหลายชั้น							
		พรีสิชัน		รีคอล		ค่า F		AUC	
		-	+	-	+	-	+	-	+
O=800%	8081	0.977	0.571	0.936	0.797	0.956	0.665	0.954	0.954
OU=700%	7126	0.975	0.643	0.954	0.77	0.964	0.701	0.95	0.951
OU=600%	6235	0.979	0.515	0.917	0.818	0.947	0.632	0.943	0.943
OU=500%	5345	0.983	0.537	0.921	0.85	0.951	0.658	0.946	0.946
OU=400%	4454	0.983	0.415	0.87	0.856	0.923	0.558	0.93	0.93
OU=300%	3563	0.984	0.415	0.869	0.866	0.923	0.562	0.934	0.934
OU=200%	2672	0.988	0.352	0.82	0.909	0.896	0.507	0.922	0.922
OU=100%	1781	0.987	0.273	0.741	0.909	0.846	0.42	0.924	0.924
OU=0%	890	0.995	0.195	0.569	0.973	0.724	0.325	0.877	0.877

ตาราง ก-7 ผลการทำนายชุดข้อมูล ecoli ด้วย C4.5

Ecoli	จำนวนข้อมูล	C4.5							
		พรีสิชัน		รีคอล		ค่า F		AUC	
		-	+	-	+	-	+	-	+
O=1400%	432	0.989	0.581	0.9596	0.8332	0.9738	0.6798	0.8786	0.8786
OU=1200%	374	0.989	0.5672	0.9574	0.8332	0.9728	0.6704	0.8966	0.8966
OU=1000%	316	0.9912	0.6998	0.9768	0.8666	0.9838	0.7734	0.9098	0.9098
OU=800%	259	0.9912	0.6698	0.9724	0.8666	0.9816	0.752	0.9064	0.9064
OU=600%	201	0.9868	0.5764	0.953	0.8	0.9692	0.651	0.8574	0.8574
OU=400%	144	0.991	0.512	0.9426	0.8666	0.966	0.635	0.918	0.918
OU=200%	86	0.9862	0.5282	0.9296	0.8	0.9564	0.6056	0.873	0.873
OU=0%	28	0.9884	0.3268	0.883	0.8332	0.932	0.4634	0.8582	0.8582

ตาราง ก-8 ผลการทำนายชุดข้อมูล ecoli ด้วยการแบ่งประเภทเบย์อย่างง่าย

Ecoli	จำนวนข้อมูล	การแบ่งประเบย์อย่างง่าย							
		พรีสิชัน		รีคอล		ค่า F		AUC	
		-	+	-	+	-	+	-	+
O=1400%	432	0.979	0.5624	0.979	0.9	0.979	0.6764	0.988	0.9888
OU=1200%	374	0.979	0.591	0.989	0.8666	0.984	0.6846	0.988	0.9862
OU=1000%	316	0.979	0.6154	0.989	0.8666	0.984	0.7034	0.988	0.9912
OU=800%	259	0.979	0.6396	0.979	0.9	0.979	0.7328	0.986	0.9908
OU=600%	201	0.969	0.6592	0.989	0.8666	0.979	0.7174	0.986	0.9922
OU=400%	144	0.989	0.6166	0.989	0.9332	0.989	0.7232	0.986	0.9926
OU=200%	86	0.989	0.638	0.989	0.9332	0.989	0.7364	0.988	0.9926
OU=0%	28	0.989	0.5776	0.979	0.9666	0.984	0.715	0.991	0.9962

ตาราง ก-9 ผลการทำนายชุดข้อมูล ecoli ด้วยเพอร์เซ็ปตรอนหลายชั้น

Ecoli	จำนวนข้อมูล	เพอร์เซ็ปตรอนหลายชั้น							
		พรีสิชัน		รีคอล		ค่า F		AUC	
		-	+	-	+	-	+	-	+
O=1400%	432	0.9912	0.5428	0.951	0.8666	0.9704	0.659	0.9798	0.9798
OU=1200%	374	0.9934	0.6426	0.968	0.9	0.9804	0.7472	0.9896	0.9896
OU=1000%	316	0.9934	0.595	0.9574	0.9	0.9748	0.7072	0.9746	0.9746
OU=800%	259	0.989	0.559	0.9532	0.8332	0.9704	0.6582	0.982	0.982
OU=600%	201	0.987	0.6332	0.968	0.8	0.9772	0.6952	0.9808	0.9808
OU=400%	144	0.989	0.591	0.9618	0.8332	0.975	0.6858	0.9794	0.9794
OU=200%	86	0.9892	0.6046	0.9594	0.8334	0.974	0.6818	0.9826	0.9826
OU=0%	28	0.9908	0.2914	0.783	0.8666	0.858	0.4094	0.9562	0.9562

ตาราง ก-10 ผลการทำนายชุดข้อมูล shuttle ด้วย C4.5

Shuttle	จำนวนข้อมูล	C4.5							
		พรีสิชัน		รีคอล		ค่า F		AUC	
		-	+	-	+	-	+	-	+
O=33600%	80921	1	1	1	0.98	1	0.99	0.99	0.99
OU=29400%	70835	1	1	1	0.98	1	0.99	0.99	0.99
OU=25200%	60750	1	1	1	0.98	1	0.99	0.99	0.99
OU=21000%	50665	1	1	1	0.98	1	0.99	0.99	0.99
OU=16800%	40580	1	1	1	0.98	1	0.99	0.99	0.99
OU=12600%	30495	1	1	1	1	1	1	1	1
OU=8400%	20410	1	1	1	1	1	1	1	1
OU=4200%	10325	1	0.962	1	1	1	0.981	1	1
OU=0%	240	1	0.797	0.999	1	1	0.887	1	1

ตาราง ก-11 ผลการทำนายชุดข้อมูล shuttle ด้วยการแบ่งประเบย์อย่างง่าย

Shuttle	จำนวนข้อมูล	การแบ่งประเบย์อย่างง่าย							
		พรีสิชัน		รีคอล		ค่า F		AUC	
		-	+	-	+	-	+	-	+
O=33600%	80921	1	1	1	0.98	1	0.99	0.99	0.99
OU=29400%	70835	1	1	1	0.98	1	0.99	0.99	0.99
OU=25200%	60750	1	1	1	0.98	1	0.99	0.99	0.99
OU=21000%	50665	1	1	1	0.98	1	0.99	0.99	0.99
OU=16800%	40580	1	1	1	0.98	1	0.99	0.99	0.99
OU=12600%	30495	1	1	1	1	1	1	1	1
OU=8400%	20410	1	1	1	1	1	1	1	1
OU=4200%	10325	1	0.962	1	1	1	0.981	1	1
OU=0%	240	1	0.797	0.999	1	1	0.887	1	1

ตาราง ก-12 ผลการทำนายชุดข้อมูล shuttle ด้วยเพอร์เซ็ปตรอนหลายชั้น

Shuttle	จำนวนข้อมูล	เพอร์เซ็ปตรอนหลายชั้น							
		พรีสิชัน		รีคอล		ค่า F		AUC	
		-	+	-	+	-	+	-	+
O=33600%		1	1	1	0.98	1	0.99	0.99	0.99
OU=29400%		1	1	1	0.98	1	0.99	0.99	0.99
OU=25200%		1	1	1	0.98	1	0.99	0.99	0.99
OU=21000%		1	1	1	0.98	1	0.99	0.99	0.99
OU=16800%		1	1	1	0.98	1	0.99	0.99	0.99
OU=12600%		1	1	1	1	1	1	1	1
OU=8400%		1	1	1	1	1	1	1	1
OU=4200%		1	0.962	1	1	1	0.981	1	1
OU=0%		1	0.797	0.999	1	1	0.887	1	1

ประวัติผู้เขียนวิทยานิพนธ์

ชื่อ	ปณต ทรงวัฒนศิริ
วัน เดือน ปีที่เกิด	6 กรกฎาคม พ.ศ. 2527
สถานที่เกิด	กรุงเทพฯ ประเทศไทย
ประวัติการศึกษา	ปริญญาตรี วิทยาศาสตร์บัณฑิต มหาวิทยาลัยศรีนครินทรวิโรฒ พ.ศ. 2549

ผลงานตีพิมพ์

Songwattanasiri, P., and Sinapiromsaran, K., “SMOUTE:Synthetics Minority Over-sampling and Under-sampling Teniques for class imbalanced problem,” Annual International Conference on Computer Science Education: Innovation & Technology (CSEIT) 2010 (2010), 78-83.