การจำแนกต้นไม้ตัดสินใจสำหรับชุดข้อมูลไม่สมดุลโดยใช้น้ำหนักต่างกันบนข้อมูลสังเคราะห์

นายสุรพงษ์ เชี่ยวสกุลวัฒนา

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2556

DECISION TREE CLASSIFICATION OF IMBALANCED DATA SETS USING DIFFERENT WEIGHTS ON SYNTHESIZED DATA

Mr. Suraphong Cheawsakunwattana

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Computer Science
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2013

หัวข้อวิทยานิพนธ์	การจำแนกต้นไม้ตัดสินใจสำหรับชุดข้อมูลไม่สมดุลโดยใช้น้ำหนักต่างกันบนข้อมูลสังเคราะห์
โดย	นายสุรพงษ์ เชี่ยวสกุลวัฒนา
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร. สุกรี สินธุภิญโญ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะวิศวกรรมศาสตร์

(ศาสตราจารย์ ดร. บัณฑิต เอื้ออาภรณ์)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ

(ศาสตราจารย์ ดร. บุญเสริม กิจศิริกุล)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(ผู้ช่วยศาสตราจารย์ ดร. สุกรี สินธุภิญโญ)

.....กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร. นัทที นิภานันท์)

.....กรรมการภายนอกมหาวิทยาลัย

(ผู้ช่วยศาสตราจารย์ ดร. รัชฎา คงคะจันทร์)

สุรพงษ์ เชี่ยวสกุลวัฒนา : การจำแนกต้นไม้ตัดสินใจสำหรับชุดข้อมูลไม่สมดุลโดยใช้น้ำหนักต่างกันบนข้อมูลสังเคราะห์. (DECISION TREE CLASSIFICATION OF IMBALANCED DATA SETS USING DIFFERENT WEIGHTS ON SYNTHESIZED DATA) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ. ดร. สุกรี สินธุภิญโญ, 82 หน้า.

การจำแนกต้นไม้ติดสินใจสำหรับชุดข้อมูลไม่สมดุล โดยใช้น้ำหนักต่างกันบนข้อมูลสังเคราะห์ เป็นวิธีการที่ใช้เทคนิค SMOTE ในการเพิ่มจำนวนตัวอย่างกลุ่มน้อยด้วยการสังเคราะห์ข้อมูลกลุ่มน้อยขึ้น เพื่อให้จำนวนตัวอย่างกลุ่มน้อยที่เพิ่มขึ้นมีจำนวนใกล้เคียงกับตัวอย่างกลุ่มมาก และปรับการหาเอนโทรปีใหม่ ซึ่งใช้วิธี C4.5เป็นพื้นฐาน เพื่อจำแนกข้อมูลกลุ่มน้อยได้ดีขึ้น สำหรับการจำแนกข้อมูลแบบสองกลุ่ม ทำการทดสอบแบบไขว้ข้ามสิบกลุ่ม โดยเลือกชุดข้อมูลไม่สมดุลจำนวน 16 ชุดข้อมูลมาทำการทดลอง และเปรียบเทียบผลการทดลองกับอัลกอริทึม C4.5 ที่ใช้เทคนิค SMOTE การทดสอบพบว่าวิธีการที่นำเสนอสามารถจำแนกข้อมูลกลุ่มน้อยได้ดีกว่าวิธีอื่นๆ เมื่อใช้น้ำหนักที่ต่างกันบนข้อมูลสังเคราะห์

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต .....

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก .....

ปีการศึกษา 2556

5370371021 : MAJOR COMPUTER SCIENCE
KEYWORDS: IMBALANCED DATA SETS / MAJORITY CLASS / MINORITY CLASS / SMOTE / SYNTHESIZED DATA

SURAPHONG CHEAWSAKUNWATTANA: DECISION TREE CLASSIFICATION OF IMBALANCED DATA SETS USING DIFFERENT WEIGHTS ON SYNTHESIZED DATA. ADVISOR: ASST. PROF. DR. SUKREE SINTHUPINYO, 82 pp.

Our classification method for an imbalanced data set is based on the decision tree techniques with SMOTE technique. In general, the SMOTE technique will increase the number of minorities by synthesizing new set of the minority class data and augmenting this new data set to the original data set. With the SMOTE technique, the data set becomes almost balanced. We adjust the original entropy function of C4.5 to better handle the newly synthesized data in the augmented data set. In our experiment, we tested our method using standard 10-fold cross validation on 16 imbalanced data sets. All of data sets are two-class data set. The results showed that the presented method performed better than other methods tested in our experiments.

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

กิตติกรรมประกาศ

ขอกราบขอบพระคุณผู้ช่วยศาสตราจารย์ ดร. สุกรี สินธุภิญโญ ซึ่งเป็นอาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ได้กรุณาสละเวลา ดูแลและให้คำแนะนำเกี่ยวกับการทำวิจัย รวมทั้งให้คำปรึกษาชี้แนะการแก้ไขปัญหาต่างๆ ที่เกิดขึ้นระหว่างการทำวิจัย จนวิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี

ขอกราบขอบพระคุณคณะกรรมสอบวิทยานิพนธ์ทุกท่านเป็นอย่างสูง ซึ่งได้แก่ ศาสตราจารย์ ดร. บุญเสริม กิจศิริกุล ผู้ช่วยศาสตราจารย์ ดร. นัททที นิภานันท์ และผู้ช่วยศาสตราจารย์ ดร. รัชฎา คงคะจันทร์ ในการตรวจแก้ไขรูปเล่ม และให้ข้อคิดรวมถึงคำแนะนำอันเป็นประโยชน์ยิ่งต่องานวิจัย และทำให้งานวิจัยมีความสมบูรณ์มากยิ่งขึ้น

ขอขอบพระคุณอาจารย์ประจำภาควิชาวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัยทุกท่านเป็นอย่างสูงที่ให้ความรู้ในการศึกษา ทั้งด้านวิชาการ และด้านอื่นๆ ที่เป็นประโยชน์กับข้าพเจ้า

และในลำดับสุดท้ายขอขอบพระคุณสมาชิกทุกคนในครอบครัว และเพื่อนๆ พี่น้องทุกคนที่คอยช่วยเหลือและให้กำลังใจในทุกๆ ด้านอย่างดีตลอดมา จนทำให้วิทยานิพนธ์ฉบับนี้มีความสมบูรณ์

สารบัญ

หน้า

หน้า

สารบัญตาราง

ตารางที่ หน้า

ตารางที่ หน้า

ตารางที่ หน้า

สารบัญภาพ

ภาพที่ หน้า

ภาพที่ หน้า

ภาพที่ หน้า

ภาพที่ หน้า

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ในปัจจุบันคอมพิวเตอร์เข้ามามีบทบาทกับมนุษย์ในชีวิตประจำวันกับงานในทุกๆ ด้าน การจัดเก็บข้อมูลต่างๆ นั้นสามารถทำได้ง่าย จึงมีข้อมูลเกิดขึ้นจำนวนมากในทุกๆ วัน เช่น ข้อมูลของ Google ข้อมูลการใช้งานโทรศัพท์ ข้อมูลของธนาคารจากการทำธุรกรรมต่างๆ ของลูกค้า ข้อมูลข่าวสารและรูปภาพ เป็นต้น ข้อมูลต่างๆ ที่เกิดขึ้นนั้นล้วนมีความสำคัญ การจัดการกับข้อมูลจำนวนมากด้วยเทคโนโลยีที่มี หรือเคยใช้อาจไม่สามารถทำได้ จึงจำเป็นต้องหาวิธีการใหม่ๆ ในการจัดการกับข้อมูลจำนวนมากที่เกิดขึ้นนี้ เพื่อนำข้อมูลไปใช้ประโยชน์ในด้านต่างๆ โดยมาทำการวิเคราะห์ เพื่อให้ได้สิ่งที่เราต้องการหรือผลลัพธ์ที่เราสนใจ [1]

ข้อมูลจำนวนมากที่เกิดขึ้น โดยทั่วไปจะแบ่งออกเป็น 2 ประเภท ประเภทแรก คือ ข้อมูลกลุ่มน้อย (Minority class) ซึ่งจะเป็นข้อมูลที่มีอยู่จำนวนน้อยในชุดข้อมูลนั้นๆ และอีกประเภท คือ ข้อมูลกลุ่มมาก (Majority class) จะเป็นกลุ่มข้อมูลที่มีอยู่จำนวนมากในชุดข้อมูลนั้น ชุดข้อมูลนี้จะเรียกว่าชุดข้อมูลไม่สมดุล (Imbalanced data set)

การเรียนรู้ของเครื่อง (Machine Learning) จึงถูกนำมาใช้ในการจัดการกับข้อมูลจำนวนมากที่เกิดขึ้นนี้ เพื่อที่จะลดจำนวนคนที่ใช้ในการวิเคราะห์ ลดเวลาที่ใช้ในการวิเคราะห์ข้อมูลจำนวนมากๆ และลดค่าใช้จ่ายต่างๆ ที่เกิดขึ้นในการวิเคราะห์ข้อมูล เพื่อให้คอมพิวเตอร์มีการเรียนรู้และสามารถทำงานที่ต้องการได้อย่างมีประสิทธิภาพ และยังสามารถนำข้อมูลจำนวนมากที่มีอยู่ มาทำการวิเคราะห์เพื่อหาความสัมพันธ์ที่อาจซ่อนอยู่ในข้อมูล ซึ่งวิธีการเรียนรู้สามารถนำมาประยุกต์ใช้งานในด้านต่างๆ ในปัจจุบัน ได้แก่

ด้านการแพทย์	ใช้เพื่อหาสาเหตุของความผิดปกติที่ทำให้เกิดโรค การวินิจฉัยโรค การป้องกันและการรักษาโรค
ด้านการเงิน	ใช้ในการวิเคราะห์บัตรเครดิต การแบ่งกลุ่มลูกค้าเพื่อหาเป้าหมายทางการตลาด การคาดหมายการขึ้นและลงของหุ้น
ด้านการเกษตร	ใช้ในการจำแนกประเภทของโรคพืช
ด้านอาชญวิทยา	ใช้ในการวิเคราะห์หาเจ้าของลายนิ้วมือ
ด้านภูมิศาสตร์	ใช้ในการพยากรณ์สภาพภูมิอากาศ การพยากรณ์การเกิดไฟป่า

วิธีในการเรียนรู้ของเครื่องนั้น มีหลายวิธีที่นำมาใช้ในการจำแนกข้อมูล แต่วิธีหนึ่งที่นิยมนำมาใช้ เนื่องจากสามารถทำความเข้าใจได้ง่าย และมีลักษณะที่ใกล้เคียงกับสิ่งแวดล้อมรอบตัวของมนุษย์ ได้แก่ ต้นไม้ตัดสินใจ (Decision Tree) ซึ่งเป็นแบบจำลองที่มีลักษณะเหมือนโครงสร้างของต้นไม้ที่มีลักษณะของต้นไม้แบบกลับหัว แสดงเส้นทางในการตัดสินใจถึงเหตุการณ์ที่อาจจะเกิดขึ้น โดยแต่ละโหนดจะแทนคุณลักษณะ (attribute) ส่วนกิ่งจะแสดงผลการทดสอบ และโหนดใบ (leaf node) แสดงคลาส ซึ่งเป็นผลลัพธ์ที่ได้จากการจำแนกข้อมูล [2]

จากปัญหาข้างต้นของกรณีที่เกิดข้อมูลที่ไม่สมดุลขึ้นนั้น จะพบว่าเมื่อข้อมูลในชุดข้อมูลมีความแตกต่างกันระหว่างข้อมูล หรือเรียกว่า ข้อมูลไม่สมดุล (imbalanced data) นั้น จะทำให้เมื่อนำข้อมูลไปทำการจำแนกประเภท (classified) ตัวจำแนกประเภท จะทำการจำแนกประเภทข้อมูลไปในกลุ่มข้อมูลกลุ่มมาก จึงมีงานวิจัยต่างๆ มากมาย นำเสนอวิธีการแก้ปัญหาในการจำแนกชุดข้อมูลที่ไม่สมดุลนี้ โดยมีเป้าหมาย คือ ต้องการที่จะจำแนกข้อมูลหรือทำนายตัวอย่างได้แม่นยำมากขึ้น ซึ่งวิธีการในการดำเนินการกับข้อมูล แบ่งเป็น 2 วิธีการ เพื่อทำการปรับข้อมูลที่ไม่สมดุลให้เกิดความสมดุลระหว่างข้อมูล 2 กลุ่ม คือ 1) การสุ่มเพิ่มตัวอย่าง (Random over-sampling) และ 2) การสุ่มลดตัวอย่าง (Random under-sampling)

วิทยานิพนธ์ฉบับนี้ใช้วิธีการปรับหาค่าเอนโทรปีจากอัลกอริทึมพื้นฐาน คือ C4.5 เพื่อเพิ่มประสิทธิภาพในการจำแนกต้นไม้ตัดสินใจสำหรับชุดข้อมูลไม่สมดุลที่มีลักษณะสองกลุ่ม คือข้อมูลกลุ่มมากและข้อมูลกลุ่มน้อย และกำหนดค่าน้ำหนักที่ต่างกันบนข้อมูลกลุ่มน้อยที่มีอยู่เดิมและข้อมูลกลุ่มน้อยที่ถูกสังเคราะห์ขึ้น เพื่อให้สามารถจำแนกประเภทข้อมูลกลุ่มน้อยได้ดีขึ้น

1.2 วัตถุประสงค์ของการวิจัย

งานวิจัยฉบับนี้ต้องการเพิ่มประสิทธิภาพในการจำแนกประเภทของข้อมูลกลุ่มน้อยได้ดีขึ้นในชุดข้อมูลไม่สมดุล โดยทำการปรับการหาค่าเอนโทรปีจากอัลกอริทึมพื้นฐานที่ใช้ในการจำแนกข้อมูล คือ C4.5 เพื่อให้สามารถทำการจำแนกข้อมูลกลุ่มน้อยได้ใกล้เคียงหรือดีกว่าวิธีการ C4.5 ที่นำข้อมูลมาทำการ SMOTE (Synthetic Minority Over-sampling Technique)

1.3 ขอบเขตของการวิจัย

1) ข้อมูลที่ใช้ในการวิจัยนำข้อมูลมาจาก UCI Machine Learning Repository [3] จำนวน 16 ชุดข้อมูล กรณีที่ข้อมูลในชุดข้อมูลมีมากกว่า 2 กลุ่ม เลือกข้อมูลที่มีจำนวนน้อยเป็นกลุ่มที่ 1 ส่วนข้อมูลที่เหลือนำมารวมกันให้เป็นกลุ่มเดียวกันเป็นกลุ่มที่ 2

2) งานวิจัยฉบับนี้ต้องการนำเสนอการจำแนกข้อมูลที่ไม่สมดุลที่มีการแบ่งออกเป็น 2 กลุ่มเท่านั้น

3) ผลการทดลองที่นำเสนอแสดงเป็นค่าเฉลี่ยจากการวิเคราะห์ความถูกต้องในการจำแนกด้วยวิธีไขว้ข้ามสิบกลุ่ม (10 fold cross-validation)

4) การวัดประสิทธิภาพของอัลกอริทึมที่ได้ออกแบบไว้ จะใช้ค่าความระลึก (Recall) ค่าความแม่นยำ (Precision) และค่าเอฟเมเชอร์ (F-measure) เป็นตัววัดในการจำแนก

1.4 ขั้นตอนและวิธีการดำเนินการวิจัย

1) ศึกษางานวิจัยที่เกี่ยวข้องกับชุดข้อมูลไม่สมดุล (Imbalanced data set)

2) ศึกษาทฤษฎีพื้นฐานของการเรียนรู้ด้วยต้นไม้ตัดสินใจ และทฤษฎีพื้นฐานของอัลกอริทึม C4.5

3) ตั้งสมมติฐานที่เกี่ยวข้องในงานวิจัยและระบุปัญหาที่เกี่ยวกับงานวิจัย

4) พัฒนาอัลกอริทึมพื้นฐาน (C4.5) ที่ใช้ในงานวิจัย และออกแบบวิธีการทดลอง

5) ศึกษาเครื่องมือต่างๆ และซอฟต์แวร์ที่เกี่ยวข้องที่จะใช้ในงานวิจัย

6) เลือกชุดข้อมูลจาก UCI Machine Learning Repository จำนวน 16 ชุดข้อมูลเพื่อมาทำการทดสอบ

7) ทดสอบวิธีการที่ได้นำเสนอ ตามสมมติฐานและปัญหาที่ได้ระบุไว้

8) วิเคราะห์ผลการทดลองและสรุปผลที่ได้จากการทดลอง

9) เรียบเรียงและจัดทำวิทยานิพนธ์

1.5 ประโยชน์ที่คาดว่าจะได้รับ

สามารถแก้ปัญหาการจำแนกประเภทข้อมูลที่ไม่สมดุล โดยเฉพาะข้อมูลกลุ่มน้อยได้ดีขึ้นหรือจำแนกได้ใกล้เคียงกับวิธีการ C4.5 ที่ใช้ข้อมูลที่ได้จากการ SMOTE

1.6 ผลงานที่ตีพิมพ์จากวิทยานิพนธ์

ชื่อเรื่อง “Decision Tree Classification of Imbalanced Data Sets Using Different Weight On Synthesized Data” การจำแนกต้นไม้ตัดสินใจสำหรับชุดข้อมูลไม่สมดุล โดยใช้น้ำหนักต่างกันบนข้อมูลสังเคราะห์ โดยสุรพงษ์ เชี่ยวสกุลวัฒนา และ สุกรี สินธุภิญโญ ในการประชุมวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 10 (NCCIT 2014) ณ Angsana laguna จ.ภูเก็ต ประเทศไทย วันที่ 8-9 พฤษภาคม 2557 [4]

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

งานวิจัยนี้พิจารณาวิธีการจำแนกประเภทข้อมูลที่ไม่สมดุล (Imbalanced data) ที่มีลักษณะของข้อมูลแบ่งออกเป็นสองกลุ่มหรือสองคลาสในชุดข้อมูล ซึ่งได้แก่ ข้อมูลกลุ่มมาก (majority class) เป็นข้อมูลที่มีจำนวนตัวอย่างอยู่จำนวนมากในชุดข้อมูล และข้อมูลกลุ่มน้อย (minority class) เป็นข้อมูลที่มีจำนวนตัวอย่างจำนวนน้อยในชุดข้อมูลนั้น

ตารางที่ 2-1 ตัวอย่างชุดข้อมูลการตัดสินใจเพื่อออกไปตีกอล์ฟ

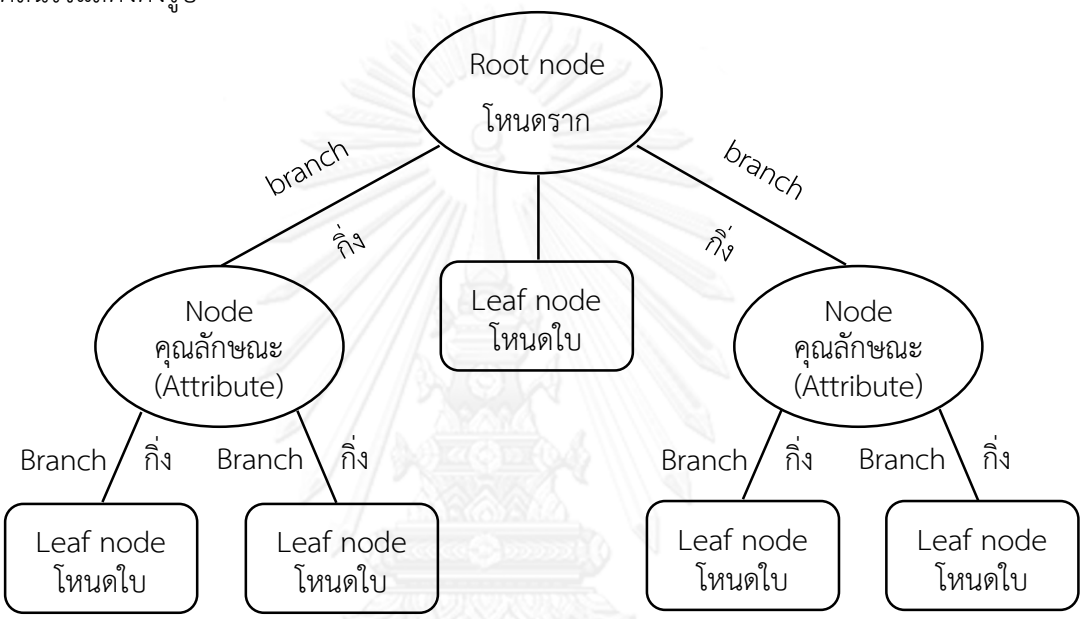
No	Outlook	Temperature	Humidity	Windy	Class
1	sunny	hot	high	False	no
2	sunny	hot	high	True	no
3	overcast	hot	high	False	yes
4	rainy	mild	high	False	yes
5	rainy	cool	normal	False	yes
6	rainy	cool	normal	True	no
7	overcast	cool	normal	True	yes
8	sunny	mild	high	False	no
9	sunny	cool	normal	False	yes
10	rainy	mild	normal	False	yes
11	sunny	mild	normal	True	yes
12	overcast	mild	high	True	yes
13	overcast	hot	normal	False	yes
14	rainy	mild	high	True	no

จากตารางที่ 2-1 เป็นตัวอย่างชุดข้อมูลในการตัดสินใจเพื่อออกไปตีกอล์ฟที่แสดงถึงลักษณะของข้อมูลที่ไม่สมดุล ที่มีลักษณะ 2 กลุ่ม โดยมีข้อมูลกลุ่มมาก คือ yes มีจำนวนตัวอย่าง 9 record และข้อมูลกลุ่มน้อย คือ no มีจำนวนตัวอย่าง 5 record

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 ต้นไม้ตัดสินใจ (Decision tree)

ต้นไม้ตัดสินใจ คือ แบบจำลองที่มีลักษณะเหมือนโครงสร้างของต้นไม้ที่มีลักษณะเป็นต้นไม้กลับหัว แต่ละโหนดจะแสดงคุณลักษณะ (Attribute) ส่วนกิ่งจะแสดงผลในการทดสอบที่ได้ และโหนดใบแสดงคลาสซึ่งจะเป็นผลลัพธ์ที่ได้จากการจำแนกข้อมูล [2] โดยตัวอย่างของต้นไม้ตัดสินใจแสดงดังรูป



ภาพที่ 2-1 โครงสร้างของต้นไม้ตัดสินใจ

โดยการสร้างต้นไม้ตัดสินใจ จะเริ่มจากการพิจารณาที่โหนดรากก่อนเป็นลำดับแรก แล้วจึงพิจารณามาที่กิ่ง และใบต่อไปตามลำดับ ในการพิจารณาเลือกโหนดรากนั้น ต้องคำนวณหาข้อมูลหรือคุณลักษณะที่เหมาะสมที่จะถูกเลือกมาเป็นโหนดรากของต้นไม้ตัดสินใจ โดยเลือกจากค่าการเพิ่มขึ้นของสารสนเทศ จากอัลกอริทึม ID3 และค่าอัตราส่วนเกน (Gain Ratio) จากอัลกอริทึม C4.5 ที่มีค่ามากที่สุดมาเป็นโหนดราก หรือโหนดเริ่มต้น และพิจารณาในส่วนของคุณลักษณะ (Atrribute) ที่เหลือ เพื่อทำการเลือกว่าคุณลักษณะเหมาะสมที่จะถูกเลือกเป็นลำดับต่อไป จะดำเนินการเช่นนี้จนข้อมูลทุกๆ ตัวในโหนดนั้นไม่มีการปะปนกัน เป็นข้อมูลประเภทเดียวกัน หรือจนมีการใช้ทุกคุณลักษณะในการเรียนรู้ครบทุกตัวแล้ว

2.1.2 อัลกอริทึม ID3

เป็นอัลกอริทึมที่ใช้สร้างต้นไม้ตัดสินใจ ซึ่งนิยมนำมาใช้ในการจำแนกข้อมูล โดยมีการใช้ค่าการเพิ่มขึ้นของสารสนเทศ เป็นตัวคัดเลือก ว่าคุณลักษณะใดมีความเหมาะสมที่จะเป็นโหนด

รากหรือโหนดเริ่มต้น และโหนดในลำดับต่อๆ ไป โดยทำการหาค่าเกนมาตรฐานทุกๆ คุณลักษณะในชุดข้อมูล แล้วเลือกค่าเกนมาตรฐานที่มีค่ามากที่สุดมาเป็นโหนดเริ่มต้น โดยค่าเกนมาตรฐานจะคำนวณจากการหาค่าเอนโทรปีก่อนแบ่ง สมการที่ (1) ลบด้วยค่าเอนโทรปีหลังแบ่งด้วยคุณลักษณะที่พิจารณา สมการที่ (2) ซึ่งมีสมการดังต่อไปนี้

$$\text{Entropy}(S) = -\sum_{i=1}^{n} P(y_i) \log_2 P(y_i) \quad (1)$$

$$\text{Entropy}_A(S) = -\sum_{i=1}^{c} \frac{|S_A|}{|S|} \times \text{Entropy}(D_j) \quad (2)$$

กำหนดให้

S คือ ชุดข้อมูลที่นำมาทดสอบทั้งหมด โดย S = $P(y_1)$, $P(y_2)$, $P(y_3)$, ... , $P(y_n)$

$P(y_i)$ คือ อัตราส่วนของจำนวนข้อมูลกลุ่มที่ i ต่อจำนวนข้อมูลทั้งหมด

n คือ จำนวนกลุ่มข้อมูลทั้งหมดในชุดข้อมูล (คลาสที่พิจารณา)

c คือ จำนวนของค่าที่เป็นไปได้ของข้อมูลในคุณลักษณะ A

จากสมการข้างต้นเมื่อทำการคำนวณหาค่าเกน แล้วค่าเกนของคุณลักษณะใดมีค่ามากที่สุด จะถูกเลือกมาเป็นโหนดรากและโหนดในลำดับถัดไปตามลำดับ ทำเช่นนี้จนครบในชุดข้อมูลหรือจนได้ต้นไม้ที่สมบูรณ์

2.1.3 อัลกอริทึม C4.5

เป็นอัลกอริทึมที่พัฒนาต่อมาจากอัลกอริทึม ID3 โดย Quinlan [5] ซึ่งได้แก้ปัญหากรณีที่มีข้อมูลในชุดข้อมูลที่นำมาทดสอบยังมีการกระจายตัวของข้อมูลอยู่มาก หรือยังไม่เป็นกลุ่มเดียวกัน ทำให้เมื่อทำการจำแนกจะเกิดความลำเอียงไปในกลุ่มข้อมูลที่มีอยู่จำนวนมาก จึงเพิ่มค่าสารสนเทศการแบ่งแยก (Split Information) เพื่อใช้ในการคำนวณ และลดปัญหาการเกิดการเอนเอียงในการจำแนกไปทางข้อมูลกลุ่มมากในชุดข้อมูลนั้น ดังสมการที่ (3)

$$\text{SplitInfo}_A(D) = -\sum_{i=1}^{n} \frac{|t_i|}{|T|} \log_2 \frac{|t_i|}{|T|} \quad (3)$$

จากการที่ค่าสารสนเทศการแบ่งแยก (Split Information) บอกถึงลักษณะการกระจายของข้อมูล วิธีแก้ปัญหาความเอนเอียง จึงนำค่าสารสนเทศในการแบ่งแยกหารด้วยค่า

มาตรฐานเกน จะทำให้ได้ค่ามาตรฐานอัตราส่วนเกน (Gain Ratio) ขึ้นเพื่อเป็นตัวที่ใช้เลือกคุณลักษณะที่จะนำมาเป็นโหนดราก หรือโหนดในลำดับต่อไป ซึ่งคำนวณได้จากสมการที่ (4)

$$\text{Gain Ratio (A)} = \frac{\text{Gain (A)}}{\text{SplitInfo}_A\text{ (D)}} \quad (4)$$

เมื่อนำข้อมูลทั้งหมดมาคำนวณหาค่ามาตรฐานอัตราส่วนเกนแล้วเลือกค่าที่มีค่าสูงสุดมาเป็นโหนดเริ่มต้น และทำการสร้างโหนดในระดับต่อไป โดยใช้คุณลักษณะที่เหลือ จนข้อมูลมีการกระจายตัวน้อยที่สุดหรือข้อมูลเป็นกลุ่มเดียวกันแล้ว

2.1.4 การสุ่มเพิ่มตัวอย่างกลุ่มน้อย (Synthetic Minority Over-sampling Technique – SMOTE)

เป็นเทคนิคหรือวิธีการหนึ่งใช้เพื่อเตรียมข้อมูลก่อนที่จะนำไปทำการจำแนกประเภทข้อมูลด้วยวิธีการ C4.5 เป็นวิธีการแก้ปัญหา กรณีที่ต้องการจำแนกข้อมูลไม่สมดุล ซึ่งเป็นข้อมูลที่มีจำนวนตัวอย่างแตกต่างกันมากในแต่ละคลาส เมื่อเรานำข้อมูลที่มีจำนวนแตกต่างกันระหว่างคลาสมาทำการจำแนกประเภท ผลลัพธ์ที่ได้จะทำให้มีการเรียนรู้แต่ข้อมูลกลุ่มมาก และเมื่อทำการจำแนกประเภทก็จะจำแนกไปเป็นข้อมูลกลุ่มมาก ด้วยเหตุนี้การแก้ปัญหาการจำแนกชุดข้อมูลที่ไม่สมดุล จึงเลือกใช้วิธีการเพิ่มจำนวนข้อมูลกลุ่มน้อย โดยการสังเคราะห์เพิ่ม เฉพาะในข้อมูลกลุ่มน้อยเท่านั้น เพื่อให้จำนวนข้อมูลกลุ่มน้อยมีจำนวนใกล้เคียงกับข้อมูลกลุ่มมาก และยังทำให้การจำแนกประเภทในส่วนของข้อมูลกลุ่มน้อยทำได้ดีขึ้น [3, 6]

ซึ่งจากตารางที่ 2-1 ที่แสดงตัวอย่างชุดข้อมูลการตัดสินใจเพื่อออกไปตีกอล์ฟ ของ Quinlan [7] เมื่อนำมาทำการสังเคราะห์ข้อมูลกลุ่มน้อยเพิ่มขึ้น ด้วยเทคนิควิธีการ SMOTE โดยเพิ่มจำนวนกลุ่มน้อยขึ้น 100% ด้วยโปรแกรม Weka [8] จะได้ข้อมูลที่ถูกสังเคราะห์ขึ้นใหม่ดังแสดงในตารางที่ 2-2

ตารางที่ 2-2 ตัวอย่างชุดข้อมูลการตัดสินใจออกไปตีกอล์ฟ ที่ได้ทำการสังเคราะห์ข้อมูลกลุ่มน้อยเพิ่มขึ้นจำนวน 100% ด้วยเทคนิค SMOTE ด้วยโปรแกรม Weka

No	Outlook	Temperature	Humidity	Windy	Class
1	sunny	hot	high	False	no
2	sunny	hot	high	True	no
3	overcast	hot	high	False	yes

4	rainy	mild	high	False	yes
5	rainy	cool	normal	False	yes
6	rainy	cool	normal	True	no
7	overcast	cool	normal	True	yes
8	sunny	mild	high	False	no
9	sunny	cool	normal	False	yes
10	rainy	mild	normal	False	yes
11	sunny	mild	normal	True	yes
12	overcast	mild	high	True	yes
13	overcast	hot	normal	False	yes
14	rainy	mild	high	True	no
15	sunny	hot	high	True	no
16	sunny	hot	high	True	no
17	sunny	hot	high	True	no
18	sunny	hot	high	True	no
19	sunny	hot	high	True	no

เมื่อนำสมการที่ (3) และ (4) มาคำนวณเพื่อหาค่า Gain Ration ของแต่ละคุณลักษณะ จากข้อมูลที่ได้จากการสังเคราะห์ข้อมูลกลุ่มน้อยเพิ่มขึ้น SMOTE ดังนี้

<u>เอนโทรปีก่อนแบ่งแยก</u>

$$\text{Entropy ก่อนแบ่ง} = -\left[\left(\frac{9}{19}\right)\log_2\left(\frac{9}{19}\right)\right] - \left[\left(\frac{10}{19}\right)\log_2\left(\frac{10}{19}\right)\right]$$

$$= 0.5106 + 0.4874$$

$$= 0.9980 \text{ บิต}$$

<u>เอนโทรปีหลังแบ่งด้วยคุณลักษณะต่างๆ</u>

$$\text{Entropy (outlook)} = \left(\frac{10}{19}\right) \times \left[-\left\{\left(\frac{2}{10}\right)\log_2\left(\frac{2}{10}\right)\right\} - \left\{\left(\frac{8}{10}\right)\log_2\left(\frac{8}{10}\right)\right\}\right]$$

$$+\left(\frac{4}{19}\right) \times \left[-\left\{\left(\frac{4}{4}\right)\log_2\left(\frac{4}{4}\right)\right\} - \left\{\left(\frac{0}{4}\right)\log_2\left(\frac{0}{4}\right)\right\}\right]$$

$$+\left(\frac{5}{19}\right) \times \left[-\left\{\left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right)\right\} - \left\{\left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right)\right\}\right]$$

= 0.3800 + 0 + 0.2555

= 0.6355 บิต

Information Gain = Entropy ก่อน – Entropy หลังแบ่งด้วยคุณลักษณะ

Gain (outlook) = 0.9980 – 0.6355

= 0.3625 บิต

SplitInfo (outlook) = 1.4675 บิต

Gain Ratio (outlook) = 0.2470 บิต

สำหรับคุณลักษณะที่เหลือ มีค่าดังนี้

Gain Ratio (temperature)= 0.1161 บิต

Gain Ratio (humidity) = 0.2819 บิต

Gain Ratio (windy) = 0.1701 บิต

จากค่า Gain Ratio ทั้งหมดที่คำนวณได้ จะเลือกค่ามากที่สุด คือ Gain Ratio (humidity) ซึ่งเลือกให้เป็นโหนดเริ่มต้นสำหรับการสร้างต้นไม้ตัดสินใจ และทำการคำนวณซ้ำเช่นนี้เพื่อหาโหนดถัดไปตามลำดับ

2.2 งานวิจัยที่เกี่ยวข้อง

งานวิจัยในหลายปีที่ผ่านมาได้คิดค้นหาวิธีในการจำแนกประเภทของข้อมูลในชุดข้อมูลที่ไม่สมดุล โดยใช้วิธีและเทคนิคที่แตกต่างกัน เช่น ใช้เทคนิคการสุ่มเพิ่มตัวอย่างข้างน้อย (Synthetic Minority Over-sampling Technique – SMOTE) [9] การใช้ต้นไม้ตัดสินใจ การเปลี่ยนค่าเอนโทรปี ซึ่งวิธีการต่างๆ เหล่านี้มีวัตถุประสงค์เพื่อต้องการให้การจำแนกมีประสิทธิภาพที่ดีขึ้น หรือมีความแม่นยำในการจำแนกสูงขึ้น งานวิจัยที่เกี่ยวข้องกับวิธีการดังกล่าวข้างต้น เช่น Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmayer, W. P., [9] ได้เสนอเทคนิคที่เรียกว่า SMOTE ไว้ในปี 2002 โดยจะใช้เทคนิค k-Nearest Neighborhood ในการหาข้อมูลกลุ่มน้อยที่ใกล้ที่สุด k ตัว เพื่อที่จะสร้างตัวอย่างเพิ่มเข้าไปในข้อมูลกลุ่มน้อย โดยการสร้างตัวอย่างเพิ่มเข้าไปอาจจะสร้างเพิ่ม 100%, 200% ของข้อมูลที่เป็นต้นฉบับ แล้วแต่ผู้ใช้กำหนด เพื่อให้ได้จำนวนของตัวอย่างกลุ่มน้อยเพิ่มขึ้น จนมีขนาดใกล้เคียงหรือเท่ากันกับจำนวนตัวอย่างกลุ่มมาก ซึ่งจัดเป็นขั้นตอนในการเตรียมข้อมูลก่อนที่จะทำการจำแนกประเภทด้วยวิธีการต่างๆ

2.2.1 งานวิจัยที่เกี่ยวข้องกับเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยในปัญหาข้อมูลไม่สมดุล

งานวิจัยที่เกี่ยวข้องกับการใช้เทคนิค SMOTE (Synthetic Minority Over-sampling Technique) นั้น เป็นเทคนิคหรือวิธีการหนึ่งที่จะช่วยแก้ปัญหา กรณีที่ต้องการจำแนกข้อมูลไม่สมดุล ซึ่งเป็นข้อมูลที่มีจำนวนตัวอย่างแตกต่างกันมากในแต่ละคลาสหรือกลุ่มตัวอย่างมาทำการจำแนกประเภท เมื่อใช้วิธีการ SMOTE จะทำให้ข้อมูลกลุ่มน้อยนั้น มีจำนวนเพิ่มขึ้น จากการสังเคราะห์ข้อมูลเพิ่มขึ้นมา จนมีจำนวนใกล้เคียงกับข้อมูลกลุ่มมาก หรือตามที่ผู้ใช้กำหนดไว้ เพื่อทำให้การจำแนกไม่เกิดความโน้มเอียงไปในข้อมูลกลุ่มมาก ผลที่ได้จากการจำแนกจะได้ความถูกต้องที่ดีขึ้น จากการมีข้อมูลที่ใช้ในการฝึกที่มีการกระจายข้อมูลเพียงพอ โดยตัวอย่างงานวิจัยในกลุ่มนี้ ได้แก่

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmayer, W. P., [9] ได้เสนอเทคนิคที่เรียกว่า SMOTE ไว้ในปี 2002 โดยจะใช้เทคนิค k-nearest neighborhood ในการหาข้อมูลกลุ่มน้อยที่ใกล้ที่สุด k ตัว เพื่อที่จะสร้างตัวอย่างเพิ่มเข้าไปในข้อมูลกลุ่มน้อย โดยการสร้างตัวอย่างเพิ่มเข้าไปอาจจะสร้างเพิ่ม 100%, 200% ของข้อมูลที่เป็นต้นฉบับ แล้วแต่ผู้ใช้กำหนด เพื่อให้ได้จำนวนของตัวอย่างกลุ่มน้อยเพิ่มขึ้น จนมีขนาดใกล้เคียงหรือเท่ากันกับจำนวนตัวอย่างกลุ่มมาก ซึ่งจัดเป็นขั้นตอนในการเตรียมข้อมูลก่อนที่จะทำการจำแนกประเภทด้วยวิธีการต่างๆ

C. Bunkhumpornpat [10] ใช้เทคนิค SMOTE (Synthetic Minority Over-sampling Technique) ซึ่งเรียกว่า Safe-Level-SMOTE โดยให้ความสนใจกับข้อมูลกลุ่มน้อยที่อยู่ใกล้ๆ กับข้อมูลกลุ่มมากด้วย และจะทำการสังเคราะห์ตัวอย่างกลุ่มน้อยขึ้นในบริเวณที่ได้กำหนดค่าว่าเป็น “ระดับที่ปลอดภัย” โดยระดับที่ปลอดภัยจะกำหนดไว้ 4 กรณี กรณีแรกคือ กรณีที่ข้อมูลกลุ่มน้อยไปอยู่ในบริเวณที่เป็นข้อมูลกลุ่มน้อยด้วยกันทั้งหมด กรณีที่ 2 คือกรณีที่ข้อมูลกลุ่มน้อยไปอยู่ในบริเวณที่ข้อมูลกลุ่มมากอยู่ในบริเวณนั้นทั้งหมด กรณีที่ 3 คือ กรณีที่ข้อมูลกลุ่มน้อยไปอยู่ในบริเวณซึ่งบริเวณนั้นมีการปะปนของข้อมูลกลุ่มน้อยและข้อมูลกลุ่มมาก แต่มีจำนวนข้อมูลกลุ่มน้อยมากกว่าข้อมูลกลุ่มมาก และกรณีที่ 4 คือ กรณีที่ข้อมูลกลุ่มน้อยไปอยู่ในบริเวณที่บริเวณนั้นมีทั้งข้อมูลกลุ่มน้อยและข้อมูลกลุ่มมากอยู่ปะปนกัน แต่จะมีข้อมูลกลุ่มมากอยู่จำนวนมากกว่าจำนวนของข้อมูลกลุ่มน้อย ซึ่งค่าความถูกต้องในการทำนายหรือจำแนกประเภทของข้อมูลของวิธีเซฟเลเวล SMOTE มีประสิทธิภาพดีกว่าวิธีการที่ใช้เทคนิค SMOTE เพียงอย่างเดียว และ ดีกว่า Borderline-SMOTE

P. Songwattanasiri [11] [12] พัฒนาเทคนิคการสุ่มเพิ่มตัวอย่างข้างน้อย และการสุ่มลดตัวอย่างข้างมาก เรียกว่าเทคนิค SMOUTE (Synthetic Minority Over-sampling and majority Under-sampling TEchnique) ซึ่งเป็นกระบวนการในการเตรียมข้อมูลก่อนที่จะนำข้อมูลไปทำการจำแนก เป็นการผสมผสานกันระหว่างเทคนิคการสุ่มเพิ่มตัวอย่างข้อมูลกลุ่มน้อย โดยใช้วิธี k-mean algorithm และการสุ่มลดตัวอย่างกลุ่มมากมากในบริเวณที่ใกล้เคียงกับเซนทรอยด์ โดยมี

วัตถุประสงค์คือ ต้องการเพิ่มจำนวนตัวอย่างข้อมูลกลุ่มน้อยในชุดข้อมูลให้มีจำนวนเพิ่มขึ้นจนใกล้เคียงกับข้อมูลกลุ่มมาก ในกรณีที่จำนวนข้อมูลยังน้อยกว่าจำนวนกลุ่มมากอยู่มาก และการลดจำนวนของตัวอย่างกลุ่มมาก จะกระทำเมื่อจำนวนตัวอย่างกลุ่มมากมีจำนวนมากกว่าจำนวนตัวอย่างกลุ่มน้อย โดยต้องการให้จำนวนตัวอย่างของทั้งสองกลุ่มมีจำนวนใกล้เคียงกันหรือเท่าๆ ใช้ 3 เทคนิคในการทดลอง ได้แก่ C4.5 และ การแบ่งประเภทเบย์ (Naïve Bayes) และตัวแบ่งเพอร์เซ็ปตรอนหลายชั้น (Multilayer perceptron) ผลจากการทดลองพบว่า วิธีการที่นำเสนอ ที่เรียกว่า SMOUTE นั้น มีความแม่นยำในการทำนายข้อมูลกลุ่มน้อยได้ดีกว่าวิธีการที่ใช้เทคนิค SMOTE และความเร็วในการทำงานด้วยเทคนิค SMOUTE เร็วกว่า SMOTE ในข้อมูลที่มีขนาดใหญ่

2.2.2 งานวิจัยที่เกี่ยวข้องกับการปรับเปลี่ยนค่าเอนโทรปีและอื่นๆ

Thach, N,H., Rojanavasu, P., and Pinngern, O., [13] นำเสนอว่าในปัญหาข้อมูลไม่สมดุลเป็นปัญหาที่มีความสำคัญในการเรียนรู้ของเครื่องและในดาต้าไมนิ่ง ระบบการเรียนรู้มีแนวโน้มเอนเอียงไปทางตัวอย่างกลุ่มมาก งานวิจัยนี้นำเสนอการจำแนกข้อมูลไม่สมดุลในเรื่องของความถูกต้องเมื่อใช้ตัวจำแนกในการเรียนรู้ โดยใช้วิธีการที่แตกต่างออกไป ซึ่งมีพื้นฐานมาจากวิธี XCS และ con-sensitive โดยในวิธีการต้องจำแนกตัวอย่างบวกหรือตัวอย่างที่มีอยู่น้อยได้ถูกต้อง ซึ่งให้ความสำคัญมากกว่าตัวอย่างอื่นๆ

Udomthanapong, S., Tamee, K, and Pinngern, O., [14] นำเทคนิค XCS มาพัฒนาใหม่เพื่อใช้แก้ปัญหาในข้อมูลไม่สมดุล นำเสนอเทคนิคที่ใช้การปรับอัตราการยอมรับแต่ละกฎที่จะทำให้เกิดการเรียนรู้ที่สมดุลระหว่างข้อมูลกลุ่มมากและข้อมูลกลุ่มน้อย ผลการทดลองพบว่าเทคนิค XCS สามารถจำแนกข้อมูลที่ไม่สมดุลได้ในทุกระดับ

Lenca, P., Lallich, S., Do, T.N., and Pham, N.K., [15] ในการเรียนรู้ด้วยวิธีการต้นไม้ตัดสินใจ วิธีในการวัดหลายๆ วิธีอาศัยพื้นฐานหรือแนวคิดของแชนนอนเอนโทรปี หลักสำคัญของเอนโทรปีคือ ถ้าค่ามาก แสดงว่ามีการกระจายตัวที่ดี เพื่อการจัดการกับปัญหาข้อมูลไม่สมดุล ได้นำเสนอวิธีการ off-center เอนโทรปี ซึ่งจะนำค่าสูงสุดสำหรับการกระจายที่แก้ไขโดยผู้ใช้ ซึ่งงานวิจัยอื่นๆ ได้นำเสนอวิธีการเอนโทรปีแบบอสมมาตร งานวิจัยนี้นำเสนอแนวคิดของ 3 เอนโทรปี ที่แตกต่างกัน ผลการทดลองโดยการใช้อัลกอริทึม C4.5 ซึ่งเปลี่ยนเพียงฟังก์ชันเดียวเท่านั้น คือ ค่าเอนโทรปี แสดงให้เห็นว่าวิธีการ off-center เอนโทรปี สามารถใช้ในการแก้ปัญหาข้อมูลไม่สมดุลได้ดีกว่าอีก 2 วิธี

อุไรรัตน์ [16] ใช้วิธีการเปลี่ยนหาค่าเอนโทรปี เพื่อทำการจำแนกกลุ่มตัวอย่างกลุ่มน้อยให้ดีขึ้น ด้วยการใช้วิธี C4.5 เป็นพื้นฐานในการสร้างต้นไม้ตัดสินใจ และได้ออกแบบเอนโทรปีใหม่

สำหรับต้นไม้ตัดสินใจ เรียกว่า “เอนโทรปีตัวปรับแบบยึดเกาะ” (Adhesively Modified Entropy: AMIE) โดยจะใช้ตัวปรับที่มีความสามารถในการปรับตัวเอง ซึ่งได้แก้ไขปัญหาในเรื่องของความลำเอียงที่เกิดจากการจำแนกประเภทไม่ถูกต้อง ทำการเปรียบเทียบกับวิธี C4.5 เอนโทรปีแบบอสมมาตร (AE) และเอนโทรปีแบบออกจากศูนย์กลาง (OCE) ซึ่งผลการทดลองแสดงให้เห็นว่าวิธีการที่นำเสนอสามารถจำแนกกลุ่มตัวอย่างกลุ่มน้อยได้ดีกว่าวิธีอื่น สร้างกฎตัวอย่างในกลุ่มที่มีน้อยได้ดี และให้ค่าเอฟเมเชอร์ที่สูงกว่าการทำด้วยวิธี C4.5, AE และ OCE

Boonchuay, K., Sinapiromsaran, K, and Lursinsap, C., [17] ได้กล่าวไว้ว่า ต้นไม้ตัดสินใจ มีแนวโน้มที่จะทำนายหรือจำแนกได้ผิดพลาดในกรณีของจำนวนกลุ่มตัวอย่างกลุ่มน้อยมีอยู่จำนวนน้อยมาก ได้นำเสนออัลกอริทึมที่ใช้ค่าที่แตกต่างกันของข้อมูลกลุ่มน้อยที่จะทำให้ค่ามาตรฐานอัตราส่วนเกน (Gain Ratio) ให้ความสนใจในการกระจายตัวของข้อมูลกลุ่มน้อย อัลกอริทึมที่ใช้จะเลือกจากค่าอัตราส่วนเกนที่มีค่าสูงที่สุดในกลุ่มข้อมูลกลุ่มน้อย หลังจากนั้นจะเลือกจากเอนโทรปีที่มีค่าน้อยที่สุด จากการทดลองกับชุดข้อมูลจาก UCI และ Statlog Respository วิธีการนี้มีประสิทธิภาพที่ดีขึ้นเมื่อเปรียบเทียบกับ C4.5 และให้แนวคิดว่า กรณีของการไปทำการลดตัวอย่างกลุ่มมากที่มีอยู่อาจจะไปลดความสำคัญของตัวอย่างนั้นด้วย

งานวิจัยที่กล่าวข้างต้นซึ่งได้แก่ งานวิจัยที่เกี่ยวข้องกับเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยในปัญหาข้อมูลไม่สมดุล และงานวิจัยที่เกี่ยวข้องกับการปรับเปลี่ยนค่าเอนโทรปี แต่ละเทคนิคจะมีวิธีการที่แตกต่างกัน แต่มีเป้าหมายที่เหมือนกันคือ ต้องการที่จะจำแนกตัวอย่างกลุ่มน้อยได้ดีขึ้นในข้อมูลที่ไม่สมดุล ซึ่งการใช้เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย (SMOTE) กับชุดข้อมูลไม่สมดุลนั้นจะเป็นการเพิ่มเฉพาะตัวอย่างกลุ่มน้อยจากชุดข้อมูลต้นฉบับ โดยจะเพิ่มขึ้นตามจำนวนที่ผู้ใช้ต้องการ เพื่อให้ได้จำนวนตัวอย่างกลุ่มน้อยที่มีจำนวนใกล้เคียงกับตัวอย่างกลุ่มมาก เมื่อไปทำการจำแนกประเภท สามารถจำแนกตัวอย่างกลุ่มน้อยได้ดีขึ้นกว่าวิธีการพื้นฐานที่มีอยู่ เช่น อัลกอริทึม C4.5 ส่วนวิธีการที่ทำการปรับเปลี่ยนค่าเอนโทรปีนั้น เป็นวิธีการหนึ่งที่ใช้วิธีพื้นฐานของอัลกอริทึมต้นไม้ตัดสินใจที่มีอยู่แล้ว เช่น C4.5 นำมาปรับเปลี่ยนวิธีการหาค่าเอนโทรปี หรือออกแบบเอนโทรปีใหม่ ซึ่งจะใช้ชุดข้อมูลต้นฉบับมาทำการจำแนกประเภท โดยไม่มีการเพิ่มหรือลดจำนวนตัวอย่างทั้งตัวอย่างกลุ่มมากและตัวอย่างกลุ่มน้อยแต่อย่างใด มีวัตถุประสงค์เพื่อจำแนกตัวอย่างในกลุ่มที่มีน้อยให้ดีขึ้น ซึ่งงานวิจัยที่นำเสนอนี้เป็นการผสานกันระหว่างวิธีการใช้เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยและการปรับสมการเอนโทรปีของอัลกอริทึม C4.5 โดยการใช้เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยจะทำให้จำนวนตัวอย่างกลุ่มน้อย เมื่อเพิ่มขึ้นจะทำให้มีจำนวนใกล้เคียงกับตัวอย่างกลุ่มมาก และวิธีการปรับสมการเอนโทรปีนั้น จะให้ความสำคัญกับตัวอย่างที่ถูกสังเคราะห์ขึ้นมาใหม่จากการใช้เทคนิคการ

สุ่มเพิ่มตัวอย่างกลุ่มน้อย และมีข้อสันนิษฐานว่า เมื่อกำหนดค่าน้ำหนักที่ต่างกันสำหรับตัวอย่างกลุ่มน้อยที่ถูกสังเคราะห์ขึ้นใหม่ กับตัวอย่างกลุ่มน้อยที่มีอยู่เดิม ในสมการเอนโทรปีที่ได้ออกแบบไว้ จะส่งผลให้สามารถจำแนกตัวอย่างกลุ่มน้อยได้ดีขึ้น เมื่อเทียบกับวิธีพื้นฐานที่ใช้ คือ อัลกอริทึม C4.5

บทที่ 3

วิธีดำเนินการวิจัย

ในการศึกษางานวิจัยนี้ ได้นำเสนอวิธีในการจำแนกต้นไม้ตัดสินใจด้วยวิธีการ C4.5 ซึ่งใช้น้ำหนักต่างกันบนข้อมูลสังเคราะห์ในชุดข้อมูลที่ไม่สมดุล ได้เลือกชุดข้อมูลที่ไม่สมดุลจาก UCI Machine Learning Repository [3] จำนวน 16 ชุดข้อมูล

ในงานวิจัยนี้ ได้ให้คำนิยามของวิธีการ SMOTE หมายถึง การเพิ่มจำนวนตัวอย่างกลุ่มน้อยขึ้น โดยการสังเคราะห์เฉพาะข้อมูลกลุ่มน้อย จำนวน 2 เท่า หรือ 100% ของจำนวนกลุ่มน้อยเดิมที่มีอยู่ เพื่อให้มีจำนวนใกล้เคียงกับจำนวนตัวอย่างกลุ่มมาก ส่วนวิธีการใช้ค่าน้ำหนักที่ต่างกันบนข้อมูลที่ถูกสังเคราะห์ DWSMOTE (Different Weight on Synthetic Minority Over-sampling Technique) คือ การเพิ่มจำนวนตัวอย่างกลุ่มน้อยขึ้น โดยวิธีการสังเคราะห์เช่นเดียวกับวิธีการ SMOTE แต่จะแยกพิจารณาในส่วนของตัวอย่างกลุ่มน้อย คือ ตัวอย่างกลุ่มน้อยที่มีอยู่เดิมกับตัวอย่างกลุ่มน้อยใหม่ที่ถูกสังเคราะห์ขึ้นจากวิธีการ SMOTE แล้วกำหนดค่าน้ำหนักที่ต่างกันให้กับตัวอย่างกลุ่มน้อย ซึ่งการให้ค่าถ่วงน้ำหนัก (weight) สำหรับข้อมูลกลุ่มน้อยและข้อมูลกลุ่มน้อยที่ถูกสังเคราะห์ขึ้นใหม่นั้น เพื่อจะให้ความสำคัญกับข้อมูลกลุ่มน้อยที่ถูกสังเคราะห์ขึ้นใหม่ เนื่องจากถ้าให้น้ำหนักที่ต่างกันอย่างเหมาะสมอาจจะส่งผลให้การจำแนกตัวอย่างกลุ่มน้อยทั้งหมดได้ดีขึ้นกว่าวิธีการ SMOTE

วิธีการให้ค่าน้ำหนักที่ต่างกันบนข้อมูลที่ถูกสังเคราะห์ในข้อมูลกลุ่มน้อยนั้น จะใช้อัลกอริทึม C4.5 ในการจำแนกประเภท เลือกใช้เทคนิคการสังเคราะห์เพิ่มตัวอย่างกลุ่มน้อยขึ้น เพื่อทำให้จำนวนข้อมูลกลุ่มน้อยเพิ่มขึ้นจนมีจำนวนใกล้เคียงกับจำนวนข้อมูลกลุ่มมาก โดยการสังเคราะห์ข้อมูลกลุ่มน้อยให้เพิ่มขึ้น 100% หรือคิดเป็น 2 เท่าของข้อมูลกลุ่มน้อยที่มีอยู่เดิม ซึ่งจัดเป็นขั้นตอนในการเตรียมข้อมูลก่อนที่จะนำชุดข้อมูลไปทำการจำแนกประเภท แล้วจึงนำข้อมูลที่ได้ทำการ SMOTE นั้นมาทำการจำแนกประเภทด้วยวิธีการ DWSMOTE

3.1 วิธีการจำแนกต้นไม้ตัดสินใจสำหรับชุดข้อมูลไม่สมดุลโดยใช้น้ำหนักต่างกันบนข้อมูลสังเคราะห์ (Different Weights on Synthesized data by Synthetic Minority Over-sampling Technique: DWSMOTE)

วิธีการ DWSMOTE เป็นวิธีการที่มีการผสานระหว่างการเตรียมข้อมูลก่อนทำการจำแนกประเภท ด้วยการนำชุดข้อมูลไปทำการ SMOTE ซึ่งได้กล่าวไว้แล้วข้างต้น กับการปรับสมการเอนโทร

ปีของอัลกอริทึม C4.5 โดยเพิ่มอีกพจน์ คือ จำนวนตัวอย่างกลุ่มน้อยที่ถูกสังเคราะห์ขึ้นมาจากการ SMOTE เพื่อให้ความสำคัญกับข้อมูลกลุ่มน้อยในระดับที่ต่างกันและสามารถทำการจำแนกประเภทได้อย่างแม่นยำมากขึ้นในตัวอย่างกลุ่มน้อย ซึ่งสมการเอนโทรปีก่อนแบ่งและเอนโทรปีหลังแบ่งด้วยวิธี DWSMOTE แสดงไว้ในสมการที่ (5) และ (9) ตามลำดับ

ค่าตัวแปรต่างๆ มีดังนี้

S_1 คือ จำนวนตัวอย่างกลุ่มมาก

S_2 คือ จำนวนตัวอย่างกลุ่มน้อยที่มีอยู่เดิม

S_3 คือ จำนวนตัวอย่างกลุ่มน้อยที่ถูกสังเคราะห์ขึ้นใหม่

S คือ จำนวนตัวอย่างทั้งหมด (S_1+ S_2+ S_3)

$$\delta = \left(\frac{S_1}{S_1+S_2+S_3}\right)$$

$$\varepsilon = \left(\frac{S_2}{S_1+S_2+S_3}\right)$$

$$\gamma = \left(\frac{S_3}{S_1+S_2+S_3}\right)$$

$$\varphi = \left(\frac{S_2+S_3}{S_1+S_2+S_3}\right)$$

β คือ ค่าน้ำหนักที่ให้กับตัวอย่างกลุ่มน้อย

α คือ จำนวนเท่าในการให้ค่าน้ำหนักกับตัวอย่างกลุ่มน้อย

T คือ กลุ่มตัวอย่าง i กลุ่ม

X คือ คุณลักษณะ (Attribute) ที่สนใจนำมาพิจารณา

<u>เอนโทรปีก่อนแบ่งแยก</u>

$$\text{DWSMOTE (T)} = -(\delta \log_2 \delta) - [(\alpha\beta)\times(\varepsilon \log_2 \varepsilon)] - [(\beta)\times(\gamma \log_2 \gamma)] \quad (5)$$

สมการที่ (5) มาจาก

<u>เอนโทรปีก่อนแบ่งแยก ของอัลกอริทึม C4.5 (SMOTE) และ DWSMOTE</u> คำนวณได้ดังนี้

Entropy ก่อนแบ่ง (SMOTE)

$$= -\left[\left(\frac{S_1}{S}\right)\log_2\left(\frac{S_1}{S}\right)\right] - \left[\left(\frac{S_2+S_3}{S}\right)\log_2\left(\frac{S_2+S_3}{S}\right)\right] \quad (6)$$

Entropy ก่อนแบ่ง (DWSMOTE)

$$= -\left[\left(\frac{S_1}{S}\right)\log_2\left(\frac{S_1}{S}\right)\right] - \alpha\left[\left(\frac{S_2}{S}\right)\log_2\left(\frac{S_2}{S}\right)\right] - \beta\left[\left(\frac{S_3}{S}\right)\log_2\left(\frac{S_3}{S}\right)\right] \quad (7)$$

จากสมการที่ (7) ค่า α คือ จำนวนเท่าที่กำหนดให้กับตัวอย่างกลุ่มน้อยที่มีอยู่เดิม

β คือ จำนวนเท่าที่กำหนดให้กับตัวอย่างกลุ่มน้อยที่ได้จากการสังเคราะห์เพิ่มขึ้นจากการ SMOTE

จากสมการที่ (6) และ (7) สมการ คือ

$$E\ (SMOTE) = -\left[\left(\frac{S_1}{S}\right)\log_2\left(\frac{S_1}{S}\right)\right] - \left[\left(\frac{S_2+S_3}{S}\right)\log_2\left(\frac{S_2+S_3}{S}\right)\right]$$

$$E\ (DWSMOTE) = -\left[\left(\frac{S_1}{S}\right)\log_2\left(\frac{S_1}{S}\right)\right] - \alpha\left[\left(\frac{S_2}{S}\right)\log_2\left(\frac{S_2}{S}\right)\right] - \beta\left[\left(\frac{S_3}{S}\right)\log_2\left(\frac{S_3}{S}\right)\right]$$

เนื่องจากในพจน์ $-\left[\left(\frac{S_1}{S}\right)\log_2\left(\frac{S_1}{S}\right)\right]$ ของทั้ง 2 สมการมีค่าเท่ากัน จะพิจารณาเฉพาะพจน์หลังเท่านั้น เพื่อทำให้ค่าเอนโทรปีของทั้งสองสมการมีค่าเท่ากัน ต้องการปรับสมการเพื่อหาค่า β ที่จะทำให้เอนโทรปีของ DWSMOTE เท่ากับเอนโทรปี SMOTE สามารถคำนวณได้ดังนี้

$$\alpha\left[\left(\frac{S_2}{S}\right)\log_2\left(\frac{S_2}{S}\right)\right] + \beta\left[\left(\frac{S_3}{S}\right)\log_2\left(\frac{S_3}{S}\right)\right] = \left[\left(\frac{S_2+S_3}{S}\right)\log_2\left(\frac{S_2+S_3}{S}\right)\right]$$

จะได้ $$\left[\log_2\left(\frac{S_2}{S}\right)^{\alpha\left(\frac{S_2}{S}\right)}\right] + \left[\log_2\left(\frac{S_3}{S}\right)^{\beta\left(\frac{S_3}{S}\right)}\right] = \left[\log_2\left(\frac{S_2+S_3}{S}\right)^{\left(\frac{S_2+S_3}{S}\right)}\right]$$

กำหนดให้ $\alpha = 2\beta$ เพื่อหาค่า β จากสมการด้านบน จะทำให้ได้สมการใหม่คือ

$$log_2\left[(\varepsilon^{\alpha\varepsilon}) \times (\gamma^{\beta\gamma})\right] = log_2(\varphi^{\varphi})$$

$$log_2\left[(\varepsilon^{2\beta\varepsilon}) \times (\gamma^{\beta\gamma})\right] = log_2(\varphi^{\varphi})$$

$$log_2\ (\varepsilon^{2\varepsilon} \times \gamma^{\gamma})^{\beta} = log_2(\varphi^{\varphi})$$

$$\beta\ [log_2\ (\varepsilon^{2\varepsilon} \times \gamma^{\gamma})] = log_2(\varphi^{\varphi})$$

$$\beta = \left[\frac{\log_2(\varphi^{\varphi})}{\log_2(\varepsilon^{2\varepsilon} \times \gamma^{\gamma})}\right]$$

เมื่อกำหนดให้ $\alpha = 2$ จะทำให้เมื่อทำการคำนวณหาค่า β แล้วนำค่า β ไปแทนในสมการที่ (5) จะทำให้เอนโทรปีก่อนแบ่งของ DWSMOTE เท่ากับเอนโทรปีก่อนแบ่งของ SMOTE

โดยที่ $(\varepsilon^{2\varepsilon})(\gamma^{\gamma})$ คือ เลขฐานของ log

จะทำให้ได้สมการเพื่อหาค่า β ดังสมการที่ (8) คือ

$$\beta = \left[\log_{(\varepsilon^{2\varepsilon})(\gamma^{\gamma})}(\varphi^{\varphi})\right] \qquad (8)$$

<u>เอนโทรปีหลังแบ่งแยกด้วยคุณลักษณะที่พิจารณา</u>

$$DWSMOTE_A\ (T) = -\sum_{i=1}^{c} \frac{|t_i|}{|T|} \times DWSMOTE\ (T_i)$$

คุณลักษณะที่มีจำนวนตัวอย่างกลุ่มน้อยและจำนวนตัวอย่างกลุ่มมากรวมกันอยู่จำนวนมาก เมื่อนำมาพิจารณาจะส่งผลทำให้ค่าเอนโทรปีที่คำนวณได้มีค่าที่สูง และทำให้เกิดความลำเอียงในการจำแนกข้อมูลได้ จึงใช้ค่าสารสนเทศการแบ่งแยก (Split Information) มาปรับลดค่าเอนโทรปีลง และทำการคำนวณเช่นเดียวกับวิธีการ C4.5 ดังสมการที่ (7)

$$\text{Gain Ratio (T)} = \frac{DWSMOTE\ (T) - DWSMOTE\ (T_i|A)}{\text{Split Information (A)}} \qquad (10)$$

หลังจากนั้นจะเลือกอัตราส่วนเกนมาตรฐาน (Gain Ratio) ที่มีค่ามากที่สุดมาเป็นโหนดราก หรือโหนดเริ่มต้น และโหนดต่อไปตามลำดับ

จากตารางที่ 2-2 แสดงตัวอย่างชุดข้อมูลในการตัดสินใจเพื่อออกไปตีกอล์ฟ ที่ได้ทำการสังเคราะห์ข้อมูลกลุ่มน้อยขึ้น ด้วยวิธีการ SMOTE ทำให้มีข้อมูลกลุ่มน้อยเพิ่มขึ้นอีก 100% ของข้อมูลเดิมนั้น สามารถคำนวณหาค่าต่างๆ ได้ดังนี้

$$\text{ค่า } \beta = \log_{\left[\left(\frac{S_2}{S}\right)^{(\alpha)\left(\frac{S_2}{S}\right)} \times \left(\frac{S_3}{S}\right)^{\left(\frac{S_3}{S}\right)}\right]} \left[\left(\frac{S_2+S_3}{S}\right)^{\left(\frac{S_2+S_3}{S}\right)}\right]$$

ถ้ากำหนดให้ค่า α = 2 จะทำให้ค่าเอนโทรปีที่ได้มีค่าเท่ากับเอนโทรปีของวิธีการ C4.5 ที่ใช้ข้อมูลที่ได้จากการ SMOTE โดยไม่ได้ปรับเอนโทรปีเดิม

$$\text{จะทำให้ค่า } \beta = \log_{\left[\left(\frac{5}{19}\right)^{(2)\left(\frac{5}{19}\right)} \times \left(\frac{5}{19}\right)^{\left(\frac{5}{19}\right)}\right]} \left[\left(\frac{10}{19}\right)^{\left(\frac{10}{19}\right)}\right]$$

$$= 0.3205$$

ค่า β ที่คำนวณได้แต่ละกิ่งจะไม่เท่ากัน เนื่องจากจำนวนตัวอย่างที่เป็นสมาชิกในแต่ละกิ่งจะมีค่าเปลี่ยนไป (ค่า S_1, S_2 และ S_3)

นำค่า β แทนค่าลงในสมการที่ (5) เพื่อหาค่าเอนโทรปีก่อนแบ่งแยก

$$\text{DWSMOTE (T)} = -(\delta \log_2 \delta) - [(\alpha\beta)\times(\varepsilon \log_2 \varepsilon)] - [(\beta)\times(\gamma \log_2 \gamma)]$$

$$= -\left[\left(\frac{S_1}{S}\right)\log_2\left(\frac{S_1}{S}\right)\right] - \left[(\alpha\beta)\times\left\{\left(\frac{S_2}{S}\right)\log_2\left(\frac{S_2}{S}\right)\right\}\right] - \left[(\beta)\times\left\{\left(\frac{S_3}{S}\right)\log_2\left(\frac{S_3}{S}\right)\right\}\right]$$

$$= -\left[\left(\frac{9}{19}\right)\log_2\left(\frac{9}{19}\right)\right] - \left[(0.6411)\left(\frac{5}{19}\right)\log_2\left(\frac{5}{19}\right)\right] - \left[(0.3205)\left(\frac{5}{19}\right)\log_2\left(\frac{5}{19}\right)\right]$$

$$= 0.5106 + 0.3249 + 0.1624$$

$$= 0.9980 \text{ บิต}$$

เมื่อเทียบกับวิธีการ C4.5 แบบพื้นฐาน ที่ใช้เทคนิค SMOTE กับชุดข้อมูลดังตารางที่ 2-2 จะสามารถคำนวณค่าเอนโทรปีก่อนแบ่งได้ดังนี้

จากสมการ Entropy ก่อนแบ่ง $= -\left[\left(\frac{S_1}{S}\right)\log_2\left(\frac{S_1}{S}\right)\right] - \left[\left(\frac{S_2+S_3}{S}\right)\log_2\left(\frac{S_2+S_3}{S}\right)\right]$

$$= -\left[\left(\frac{9}{19}\right)\log_2\left(\frac{9}{S19}\right)\right] - \left[\left(\frac{10}{19}\right)\log_2\left(\frac{10}{19}\right)\right]$$

$$= 0.5106 + 0.4874$$

$$= 0.9980 \text{ บิต}$$

<u>เอนโทรปีหลังแบ่ง เมื่อพิจารณาแต่ละคุณลักษณะ</u>

จากตารางที่ 2-2 ซึ่งแสดงตัวอย่างชุดข้อมูลการตัดสินใจออกไปตีกอล์ฟ ที่ทำการ SMOTE ข้อมูลแล้ว ซึ่งมี 4 คุณลักษณะ ได้แก่ outlook, temperature, humidity และ windy

พิจารณาคุณลักษณะ outlook โดยคุณลักษณะ outlook มีค่าที่เป็นไปได้ 3 ค่า คือ sunny, overcast และ rainy รวมคุณลักษณะ outlook มีจำนวนสมาชิกทั้งหมด 19 ตัวอย่าง

Sunny มีจำนวนตัวอย่างทั้งหมด 10 ตัวอย่าง โดยแบ่งเป็น

ตัวอย่างกลุ่มมาก $S_1 = 2$

ตัวอย่างกลุ่มน้อยเดิม $S_2 = 3$

ตัวอย่างกลุ่มน้อยที่ถูกสังเคราะห์ขึ้นใหม่ $S_3 = 5$

overcast มีจำนวนตัวอย่างทั้งหมด 4 ตัวอย่าง โดยแบ่งเป็น

ตัวอย่างกลุ่มมาก $S_1 = 4$

ตัวอย่างกลุ่มน้อยเดิม $S_2 = 0$

ตัวอย่างกลุ่มน้อยที่ถูกสังเคราะห์ขึ้นใหม่ $S_3 = 0$

rainy มีจำนวนตัวอย่างทั้งหมด 5 ตัวอย่าง โดยแบ่งเป็น

ตัวอย่างกลุ่มมาก $S_1 = 3$

ตัวอย่างกลุ่มน้อยเดิม $S_2 = 2$

ตัวอย่างกลุ่มน้อยที่ถูกสังเคราะห์ขึ้นใหม่ $S_3 = 0$

สามารถคำนวณค่าตามสมการที่ (8) ได้ดังนี้

$$\text{DWSMOTE(outlook)} = \left(\frac{10}{19}\right) \times \left[\left\{-\left(\frac{2}{10}\right)\log_2\left(\frac{2}{10}\right)\right\} - \left\{(2\times 0.1670)\left(\left(\frac{3}{10}\right)\log_2\left(\frac{3}{10}\right)\right)\right\} - \left\{(0.1670)\left(\left(\frac{5}{10}\right)\log_2\left(\frac{5}{10}\right)\right)\right\}\right] + \left(\frac{4}{19}\right) \times \left[\left\{-\left(\frac{4}{4}\right)\log_2\left(\frac{4}{4}\right)\right\} - 0 - 0\right] + \left(\frac{5}{19}\right) \times \left[\left\{-\left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right)\right\} - \left\{(2\times 0.5)\left(\left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right)\right)\right\} - 0\right]$$

$$= 0.3800 + 0 + 0.2555$$

$$= 0.6355 \text{ บิต}$$

$$\text{Gain (outlook)} = 0.9980 - 0.6355$$

$$= 0.3625 \text{ บิต}$$

$$\text{SplitInfo (outlook)} = \left[-\left(\frac{10}{19}\right)\log_2\left(\frac{10}{19}\right)\right] + \left[-\left(\frac{4}{19}\right)\log_2\left(\frac{4}{19}\right)\right] + \left[-\left(\frac{5}{19}\right)\log_2\left(\frac{5}{19}\right)\right]$$

$$= 0.4874 + 0.4732 + 0.5068$$

$$= 1.4675 \text{ บิต}$$

$$\text{Gain Ratio (outlook)} = 0.3625/1.4675$$

$$= 0.2470 \text{ บิต}$$

เอนโทรปีหลังแบ่งของคุณลักษณะที่เหลือ คือ temperature, humidity และ windy ได้แก่

DWSMOTE (temperature) = 0.8228 บิต

DWSMOTE (humidity) = 0.7304 บิต

DWSMOTE (windy) = 0.8310 บิต

ค่า Split Information ของแต่ละคุณลักษณะ

SplitInfo (temperature) = 1.5090 บิต

SplitInfo (humidity) = 0.9495 บิต

SplitInfo (windy) = 0.9819 บิต

ค่า Gain Ratio แต่ละคุณลักษณะ คือ

Gain Ratio (temperature) = 0.1161 บิต

Gain Ratio (humidity) = 0.2819 บิต

Gain Ratio (windy) = 0.1701 บิต

เมื่อได้ค่า Gain Ratio ของแต่ละคุณลักษณะแล้วข้างต้น จะพบว่า ค่า Gain Ratio ของคุณลักษณะ humidity มีค่ามากที่สุด จึงเลือก humidity มาเป็นโหนดรากหรือโหนดเริ่มต้น หลังจากนั้นทำการคำนวณตามสมการข้างต้นเพื่อหาโหนดต่อไปกับคุณลักษณะที่เหลือจนครบทุกคุณลักษณะหรือได้ต้นไม้ที่สมบูรณ์

เนื่องจากวิธีการ DWSMOTE เป็นการปรับค่าน้ำหนัก คือ ปรับค่าแอลฟา (α) เพื่อให้ค่าน้ำหนักที่ต่างกันในข้อมูลกลุ่มน้อยที่มีอยู่เดิมกับข้อมูลกลุ่มน้อยที่ถูกสังเคราะห์ขึ้นใหม่ เมื่อกำหนดค่า α = 7 แทนค่าลงในสมการ (6) (7) และ (8) เพื่อหาค่าเอนโทรปีก่อนแบ่ง เอนโทรปีหลังแบ่งด้วยคุณลักษณะต่างๆ จะทำให้จำแนกข้อมูลกลุ่มน้อยได้ดีขึ้น ซึ่งค่าเอนโทรปีก่อนแบ่ง เอนโทรปีหลังแบ่งด้วยคุณลักษณะต่างๆ มีค่าดังนี้

ค่าเอนโทรปีก่อนแบ่ง จะมีค่าเท่ากับ 1.8103

ค่าเอนโทรปีหลังแบ่งด้วยคุณลักษณะต่างๆ คือ

DWSMOTE (outlook) = 0.8645

DWSMOTE (temperature) = 1.0471

DWSMOTE (humidity) = 1.0584

DWSMOTE (windy) = 1.1522

ค่า Gain Ratio ของแต่ละคุณลักษณะ มีค่าดังนี้

Gain Ratio (outlook) = 0.6445

Gain Ratio (temperature) = 0.5057

Gain Ratio (humidity) = 0.7919

Gain Ratio (windy) = 0.6702

เมื่อพิจารณาจากค่า Gain Ratio ของแต่ละคุณลักษณะ เมื่อทำการคำนวณด้วยค่า α = 7 ซึ่งเป็นการให้ค่าน้ำหนักที่ต่างกันบนข้อมูลกลุ่มน้อยแล้วจะพบว่า ค่า Gain Ratio ของคุณลักษณะ Humidity จะมีค่ามากที่สุดเช่นกัน โดยเท่ากับเมื่อกำหนดค่า α = 2 แล้วทำการคำนวณเพื่อหาโหนดในลำดับถัดไปตามลำดับจนครบทุกคุณลักษณะ หรือจนข้อมูลเป็นกลุ่มเดียวกัน

บทที่ 4

ผลการทดลองและการวิเคราะห์ผลการทดลอง

ในบทนี้แสดงผลการทดลองที่ได้ทำการจำแนกข้อมูลตามวิธีการที่ได้นำเสนอไว้ จากการที่นำวิธีการที่นำเสนอมาทำการทดลองและทดสอบแบบไขว้ข้ามสิบกลุ่ม โดยกำหนดนัยสำคัญทางสถิติของค่าความแม่นยำทั้งหมดที่ระดับ 0.05 และ 0.1 ตามลำดับ เปรียบเทียบกับวิธีการ SMOTE ที่ใช้อัลกอริทึม C4.5 และทำการวิเคราะห์ผลการทดลองด้วยค่าความแม่นยำทั้งหมด ค่าความแม่นยำในข้อมูลกลุ่มมาก และค่าความแม่นยำในข้อมูลกลุ่มน้อย

เนื่องจากงานวิจัยฉบับนี้สนใจในชุดข้อมูลไม่สมดุล เพื่อต้องการเพิ่มประสิทธิภาพในการจำแนกข้อมูลกลุ่มน้อยให้ดีขึ้น จึงได้ทำการทดลองเปรียบเทียบเฉพาะค่าความระลึก ค่าความเที่ยง และค่าเอฟเมเชอร์ ในข้อมูลกลุ่มน้อย ข้อมูลกลุ่มมาก และข้อมูลทั้งหมด

4.1 ชุดข้อมูลที่ใช้ในการทดลอง

งานวิจัยฉบับนี้ใช้ชุดข้อมูล 16 ชุด ในการทดลอง โดยคัดเลือกข้อมูลจาก UCI Machine Learning Repository [3] ซึ่งเป็นข้อมูลต้นฉบับ ทำการเลือกและรวมข้อมูลให้มีแค่ 2 กลุ่ม แล้วนำข้อมูลมาทำการสังเคราะห์ข้อมูลกลุ่มน้อยขึ้นใหม่ด้วยวิธีการ SMOTE ที่ 100% มีรายละเอียดดังนี้

ตารางที่ 4-1 แสดงรายละเอียดของข้อมูลที่ไม่สมดุลที่ใช้ในงานวิจัย

ชื่อชุดข้อมูล	ชุดข้อมูล	จำนวนคุณลักษณะ (Attribute)	จำนวนข้อมูลกลุ่มมาก	จำนวนข้อมูลกลุ่มน้อย	จำนวนตัวอย่างทั้งหมด
Breast-cancer	original	9	201	85	286
	SMOTE	9	201	170	371
Breast-cancer-W	original	9	458	241	699
	SMOTE	9	458	482	940
Credit-g	original	20	700	300	1000
	SMOTE	20	700	600	1300
Diabetes	original	8	500	268	768
	SMOTE	8	500	536	1036
Ecoli	original	7	301	35	336
	SMOTE	7	301	70	371

ชื่อชุดข้อมูล	ชุดข้อมูล	จำนวนคุณลักษณะ (Attribute)	จำนวนข้อมูลกลุ่มมาก	จำนวนข้อมูลกลุ่มน้อย	จำนวนตัวอย่างทั้งหมด
Flags	original	28	177	17	194
	SMOTE	28	177	34	211
Glasses	original	9	197	17	214
	SMOTE	9	197	34	231
Haberman	original	3	225	81	306
	SMOTE	3	225	162	387
Hepatitis	original	19	123	32	155
	SMOTE	19	123	64	187
Ionosphere	original	34	225	126	351
	SMOTE	34	225	252	477
Liver	original	6	200	145	345
	SMOTE	6	200	190	390
Post-operative	original	8	66	24	90
	SMOTE	8	66	48	114
Tic-tac-toe	original	9	626	332	958
	SMOTE	9	626	664	1290
Splice-ei	original	60	2423	767	3190
	SMOTE	60	2423	1534	3957
Splice-ie	original	60	2422	768	3190
	SMOTE	60	2422	1536	3958
Vehicle	original	18	647	199	846
	SMOTE	18	647	398	1045

จากตารางที่ 4-1 แสดงรายละเอียดข้อมูล จะแสดงข้อมูลต้นฉบับที่นำมาใช้ในการทดลองและข้อมูลต้นฉบับที่ได้นำไปทำการ SMOTE เพื่อสังเคราะห์ข้อมูลกลุ่มน้อยเพิ่มขึ้น โดยเพิ่มขึ้น 100% จากข้อมูลกลุ่มน้อยเดิม ซึ่งรายละเอียดของชุดข้อมูลที่ใช้ในการทดลองโดยเปรียบเทียบเป็นเปอร์เซ็นต์ของแต่ละกลุ่มตัวอย่าง มีดังนี้

1) ชุดข้อมูล Breast-cancer เป็นชุดข้อมูลที่แสดงถึงการวินิจฉัยของแพทย์ว่าสามารถเกิดโรคมะเร็งเต้านมได้อีกหรือไม่

ตารางที่ 4-2 จำนวนข้อมูลกลุ่มน้อยและจำนวนข้อมูลกลุ่มมากของชุดข้อมูล Breast-cancer

ชุดข้อมูล	จำนวนตัวอย่าง (original)	เปอร์เซ็นต์ของจำนวนตัวอย่าง	จำนวนตัวอย่าง (SMOTE)	เปอร์เซ็นต์ของจำนวนตัวอย่าง
ข้อมูลกลุ่มน้อย	85	29.72	170	45.82
ข้อมูลกลุ่มมาก	201	70.28	201	54.18
ทั้งหมด	286	100	371	100

จากชุดข้อมูลต้นฉบับที่มีจำนวนตัวอย่างทั้งหมด 286 ตัวอย่าง โดยแบ่งเป็นตัวอย่างกลุ่มน้อย 85 ตัวอย่าง และตัวอย่างกลุ่มมาก 201 ตัวอย่าง เมื่อนำชุดข้อมูลดังกล่าวไปทำการ SMOTE โดยกำหนดให้เพิ่มจำนวนตัวอย่างกลุ่มน้อย 100% จะพบว่าจำนวนตัวอย่างกลุ่มน้อยมีจำนวนเพิ่มขึ้น 1 เท่า จาก 85 ตัวอย่าง เพิ่มขึ้นเป็น 170 ตัวอย่าง ซึ่งเมื่อคิดเทียบเป็นเปอร์เซ็นต์กับจำนวนตัวอย่างทั้งหมดแล้ว จะทำให้มีจำนวนใกล้เคียงกับจำนวนตัวอย่างกลุ่มมาก โดยจำนวนตัวอย่างกลุ่มน้อยที่เมื่อทำการ SMOTE ขึ้นมาแล้วรวมกับของเดิมด้วยจะคิดเป็น 45.82% จำนวนตัวอย่างกลุ่มมากคิดเป็น 54.18%

2) ชุดข้อมูล Breast-cancer-w เป็นชุดข้อมูลที่แสดงถึงการวินิจฉัยของแพทย์ว่าเป็นโรคมะเร็งเต้านมชนิดร้ายแรงหรือไม่

ตารางที่ 4-3 จำนวนข้อมูลกลุ่มน้อยและจำนวนข้อมูลกลุ่มมากของชุดข้อมูล Breast-cancer-w

ชุดข้อมูล	จำนวนตัวอย่าง (original)	เปอร์เซ็นต์ของจำนวนตัวอย่าง	จำนวนตัวอย่าง (SMOTE)	เปอร์เซ็นต์ของจำนวนตัวอย่าง
ข้อมูลกลุ่มน้อย	241	34.48	482	51.28
ข้อมูลกลุ่มมาก	458	65.52	458	48.72
ทั้งหมด	699	100	940	100

จากชุดข้อมูลต้นฉบับที่มีจำนวนตัวอย่างทั้งหมด 699 ตัวอย่าง โดยแบ่งเป็นตัวอย่างกลุ่มน้อย 241 ตัวอย่าง และตัวอย่างกลุ่มมาก 458 ตัวอย่าง เมื่อนำชุดข้อมูลดังกล่าวไปทำการ SMOTE โดยกำหนดให้เพิ่มจำนวนตัวอย่างกลุ่มน้อย 100% จะพบว่าจำนวนตัวอย่างกลุ่มน้อยมีจำนวนเพิ่มขึ้น 1 เท่า จาก 241 ตัวอย่าง เพิ่มขึ้นเป็น 482 ตัวอย่าง ซึ่งเมื่อคิดเทียบเป็นเปอร์เซ็นต์กับจำนวนตัวอย่างทั้งหมดแล้ว จะทำให้มีจำนวนมากกว่าจำนวนตัวอย่างกลุ่มมาก ซึ่งได้จำนวนใกล้เคียงกัน โดยจำนวนตัวอย่างกลุ่มน้อยที่เมื่อทำการ SMOTE ขึ้นมาแล้วรวมกับของเดิมด้วยจะคิดเป็น 51.28% จำนวนตัวอย่างกลุ่มมากคิดเป็น 48.72%

3) ชุดข้อมูล Credit-g เป็นชุดข้อมูลที่แสดงถึงการวินิจฉัยของแพทย์ว่าเป็นโรคมะเร็งเต้านมชนิดร้ายแรงหรือไม่

ตารางที่ 4-4 จำนวนข้อมูลกลุ่มน้อยและจำนวนข้อมูลกลุ่มมากของชุดข้อมูล Credit-g

ชุดข้อมูล	จำนวนตัวอย่าง (original)	เปอร์เซ็นต์ของจำนวนตัวอย่าง	จำนวนตัวอย่าง (SMOTE)	เปอร์เซ็นต์ของจำนวนตัวอย่าง
ข้อมูลกลุ่มน้อย	300	30.00	600	46.15
ข้อมูลกลุ่มมาก	700	70.00	700	53.85
ทั้งหมด	1,000	100	1,300	100

จากชุดข้อมูลต้นฉบับที่มีจำนวนตัวอย่างทั้งหมด 1,000 ตัวอย่าง โดยแบ่งเป็นตัวอย่างกลุ่มน้อย 300 ตัวอย่าง และตัวอย่างกลุ่มมาก 700 ตัวอย่าง เมื่อนำชุดข้อมูลดังกล่าวไปทำการ SMOTE โดยกำหนดให้เพิ่มจำนวนตัวอย่างกลุ่มน้อย 100% จะพบว่าจำนวนตัวอย่างกลุ่มน้อยมีจำนวนเพิ่มขึ้น 1 เท่า จาก 300 ตัวอย่าง เพิ่มขึ้นเป็น 600 ตัวอย่าง ซึ่งเมื่อคิดเทียบเป็นเปอร์เซ็นต์กับจำนวนตัวอย่างทั้งหมดแล้ว จะทำให้มีจำนวนใกล้เคียงกับจำนวนตัวอย่างกลุ่มมาก โดยจำนวนตัวอย่างกลุ่มน้อยที่เมื่อทำการ SMOTE ขึ้นมาแล้วรวมกับของเดิมด้วยจะคิดเป็น 46.15% จำนวนตัวอย่างกลุ่มมากคิดเป็น 53.85%

4) ชุดข้อมูล Diabetes เป็นชุดข้อมูลที่แสดงถึงการวินิจฉัยของแพทย์ว่าเป็นโรคเบาหวาน

ตารางที่ 4-5 จำนวนข้อมูลกลุ่มน้อยและจำนวนข้อมูลกลุ่มมากของชุดข้อมูล Diabetes

ชุดข้อมูล	จำนวนตัวอย่าง (original)	เปอร์เซ็นต์ของจำนวนตัวอย่าง	จำนวนตัวอย่าง (SMOTE)	เปอร์เซ็นต์ของจำนวนตัวอย่าง
ข้อมูลกลุ่มน้อย	268	34.90	536	51.74
ข้อมูลกลุ่มมาก	500	65.10	500	48.26
ทั้งหมด	768	100	1,036	100

จากชุดข้อมูลต้นฉบับที่มีจำนวนตัวอย่างทั้งหมด 768 ตัวอย่าง โดยแบ่งเป็นตัวอย่างกลุ่มน้อย 268 ตัวอย่าง และตัวอย่างกลุ่มมาก 500 ตัวอย่าง เมื่อนำชุดข้อมูลดังกล่าวไปทำการ SMOTE โดยกำหนดให้เพิ่มจำนวนตัวอย่างกลุ่มน้อย 100% จะพบว่าจำนวนตัวอย่างกลุ่มน้อยมีจำนวนเพิ่มขึ้น 1 เท่า จาก 268 ตัวอย่าง เพิ่มขึ้นเป็น 536 ตัวอย่าง ซึ่งเมื่อคิดเทียบเป็นเปอร์เซ็นต์กับจำนวนตัวอย่างทั้งหมดแล้ว จะทำให้มีจำนวนมากกว่าจำนวนตัวอย่างกลุ่มมาก ซึ่งจำนวนใกล้เคียงกัน โดยจำนวนตัวอย่างกลุ่มน้อยที่เมื่อทำการ SMOTE ขึ้นมาแล้วรวมกับของเดิมด้วยจะคิดเป็น 51.75% จำนวนตัวอย่างกลุ่มมากคิดเป็น 48.26%

5) ชุดข้อมูล Ecoli เป็นชุดข้อมูลที่แสดงถึงการจำแนกโปรตีนที่มีอยู่ในร่างกาย

ตารางที่ 4-6 จำนวนข้อมูลกลุ่มน้อยและจำนวนข้อมูลกลุ่มมากของชุดข้อมูล Ecoli

ชุดข้อมูล	จำนวนตัวอย่าง (original)	เปอร์เซ็นต์ของจำนวนตัวอย่าง	จำนวนตัวอย่าง (SMOTE)	เปอร์เซ็นต์ของจำนวนตัวอย่าง
ข้อมูลกลุ่มน้อย	35	10.42	70	18.87
ข้อมูลกลุ่มมาก	301	89.58	301	81.13
ทั้งหมด	336	100	371	100

จากชุดข้อมูลต้นฉบับที่มีจำนวนตัวอย่างทั้งหมด 336 ตัวอย่าง โดยแบ่งเป็นตัวอย่างกลุ่มน้อย 35 ตัวอย่าง และตัวอย่างกลุ่มมาก 301 ตัวอย่าง เมื่อนำชุดข้อมูลดังกล่าวไปทำการ SMOTE โดยกำหนดให้เพิ่มจำนวนตัวอย่างกลุ่มน้อย 100% จะพบว่าจำนวนตัวอย่างกลุ่มน้อยมีจำนวนเพิ่มขึ้น 1 เท่า จาก 35 ตัวอย่าง เพิ่มขึ้นเป็น 70 ตัวอย่าง ซึ่งเมื่อคิดเทียบเป็นเปอร์เซ็นต์กับจำนวนตัวอย่างทั้งหมดแล้ว จะทำให้มีจำนวนเพิ่มขึ้นแต่ยังไม่ใกล้เคียงกับจำนวนตัวอย่างกลุ่มมาก โดยจำนวนตัวอย่างกลุ่มน้อยที่เมื่อทำการ SMOTE ขึ้นมาแล้วรวมกับจำนวนตัวอย่างกลุ่มน้อยที่มีอยู่เดิมด้วยจะคิดเป็น 18.87% จำนวนตัวอย่างกลุ่มมากคิดเป็น 81.13%

6) ชุดข้อมูล Flags เป็นชุดข้อมูลที่แสดงถึงการจำแนกลักษณะของธง

ตารางที่ 4-7 จำนวนข้อมูลกลุ่มน้อยและจำนวนข้อมูลกลุ่มมากของชุดข้อมูล Flags

ชุดข้อมูล	จำนวนตัวอย่าง (original)	เปอร์เซ็นต์ของจำนวนตัวอย่าง	จำนวนตัวอย่าง (SMOTE)	เปอร์เซ็นต์ของจำนวนตัวอย่าง
ข้อมูลกลุ่มน้อย	17	8.76	34	16.11
ข้อมูลกลุ่มมาก	177	91.24	177	83.89
ทั้งหมด	194	100	211	100

จากชุดข้อมูลต้นฉบับที่มีจำนวนตัวอย่างทั้งหมด 194 ตัวอย่าง โดยแบ่งเป็นตัวอย่างกลุ่มน้อย 17 ตัวอย่าง และตัวอย่างกลุ่มมาก 177 ตัวอย่าง เมื่อนำชุดข้อมูลดังกล่าวไปทำการ SMOTE โดยกำหนดให้เพิ่มจำนวนตัวอย่างกลุ่มน้อย 100% จะพบว่าจำนวนตัวอย่างกลุ่มน้อยมีจำนวนเพิ่มขึ้น 1 เท่า จาก 17 ตัวอย่าง เพิ่มขึ้นเป็น 34 ตัวอย่าง ซึ่งเมื่อคิดเทียบเป็นเปอร์เซ็นต์กับจำนวนตัวอย่างทั้งหมดแล้ว จะทำให้มีจำนวนเพิ่มขึ้นแต่ยังไม่ใกล้เคียงกับจำนวนตัวอย่างกลุ่มมาก โดยจำนวนตัวอย่างกลุ่มน้อยที่เมื่อทำการ SMOTE ขึ้นมาแล้วรวมกับของเดิมด้วยจะคิดเป็น 16.11% จำนวนตัวอย่างกลุ่มมากคิดเป็น 83.89%

7) ชุดข้อมูล Glasses เป็นชุดข้อมูลที่แสดงถึงการจำแนกประเภทของกระจก

ตารางที่ 4-8 จำนวนข้อมูลกลุ่มน้อยและจำนวนข้อมูลกลุ่มมากของชุดข้อมูล Glasses

ชุดข้อมูล	จำนวนตัวอย่าง (original)	เปอร์เซ็นต์ของจำนวนตัวอย่าง	จำนวนตัวอย่าง (SMOTE)	เปอร์เซ็นต์ของจำนวนตัวอย่าง
ข้อมูลกลุ่มน้อย	17	7.94	34	14.72
ข้อมูลกลุ่มมาก	197	92.06	197	85.28
ทั้งหมด	214	100	231	100

จากชุดข้อมูลต้นฉบับที่มีจำนวนตัวอย่างทั้งหมด 214 ตัวอย่าง โดยแบ่งเป็นตัวอย่างกลุ่มน้อย 17 ตัวอย่าง และตัวอย่างกลุ่มมาก 197 ตัวอย่าง เมื่อนำชุดข้อมูลดังกล่าวไปทำการ SMOTE โดยกำหนดให้เพิ่มจำนวนตัวอย่างกลุ่มน้อย 100% จะพบว่าจำนวนตัวอย่างกลุ่มน้อยมีจำนวนเพิ่มขึ้น 1 เท่า จาก 17 ตัวอย่าง เพิ่มขึ้นเป็น 34 ตัวอย่าง ซึ่งเมื่อคิดเทียบเป็นเปอร์เซ็นต์กับจำนวนตัวอย่างทั้งหมดแล้ว จะทำให้มีจำนวนเพิ่มขึ้นแต่ยังไม่ใกล้เคียงกับจำนวนตัวอย่างกลุ่มมาก โดยจำนวนตัวอย่างกลุ่มน้อยที่เมื่อทำการ SMOTE ขึ้นมาแล้วรวมกับของเดิมด้วยจะคิดเป็น 14.72% จำนวนตัวอย่างกลุ่มมากคิดเป็น 85.28%

8) ชุดข้อมูล Haberman เป็นชุดข้อมูลที่แสดงถึงการจำแนกประเภทของการรอดชีวิตของผู้ป่วยโรคมะเร็งหลังจากการเข้ารับการรักษาโรงพยาบาลแห่งหนึ่ง

ตารางที่ 4-9 จำนวนข้อมูลกลุ่มน้อยและจำนวนข้อมูลกลุ่มมากของชุดข้อมูล Haberman

ชุดข้อมูล	จำนวนตัวอย่าง (original)	เปอร์เซ็นต์ของจำนวนตัวอย่าง	จำนวนตัวอย่าง (SMOTE)	เปอร์เซ็นต์ของจำนวนตัวอย่าง
ข้อมูลกลุ่มน้อย	81	26.47	162	41.86
ข้อมูลกลุ่มมาก	225	73.53	225	58.14
ทั้งหมด	306	100	387	100

จากชุดข้อมูลต้นฉบับที่มีจำนวนตัวอย่างทั้งหมด 306 ตัวอย่าง โดยแบ่งเป็นตัวอย่างกลุ่มน้อย 81 ตัวอย่าง และตัวอย่างกลุ่มมาก 225 ตัวอย่าง เมื่อนำชุดข้อมูลดังกล่าวไปทำการ SMOTE โดยกำหนดให้เพิ่มจำนวนตัวอย่างกลุ่มน้อย 100% จะพบว่าจำนวนตัวอย่างกลุ่มน้อยมีจำนวนเพิ่มขึ้น 1 เท่า จาก 17 ตัวอย่าง เพิ่มขึ้นเป็น 162 ตัวอย่าง ซึ่งเมื่อคิดเทียบเป็นเปอร์เซ็นต์กับจำนวนตัวอย่างทั้งหมดแล้ว จะทำให้มีจำนวนเพิ่มขึ้นใกล้เคียงกับจำนวนตัวอย่างกลุ่มมาก โดยจำนวนตัวอย่างกลุ่มน้อยที่เมื่อทำการ SMOTE ขึ้นมาแล้วรวมกับของเดิมด้วยจะคิดเป็น 41.86% จำนวนตัวอย่างกลุ่มมากคิดเป็น 58.14%

9) ชุดข้อมูล Hepatitis เป็นชุดข้อมูลที่แสดงถึงการจำแนกประเภทของการรอดชีวิตของผู้ป่วยโรคตับอักเสบว่ายังรอดชีวิตอยู่หรือไม่

ตารางที่ 4-10 จำนวนข้อมูลกลุ่มน้อยและจำนวนข้อมูลกลุ่มมากของชุดข้อมูล Hepatitis

ชุดข้อมูล	จำนวนตัวอย่าง (original)	เปอร์เซ็นต์ของจำนวนตัวอย่าง	จำนวนตัวอย่าง (SMOTE)	เปอร์เซ็นต์ของจำนวนตัวอย่าง
ข้อมูลกลุ่มน้อย	32	20.65	64	34.22
ข้อมูลกลุ่มมาก	123	79.35	123	65.78
ทั้งหมด	155	100	187	100

จากชุดข้อมูลต้นฉบับที่มีจำนวนตัวอย่างทั้งหมด 155 ตัวอย่าง โดยแบ่งเป็นตัวอย่างกลุ่มน้อย 32 ตัวอย่าง และตัวอย่างกลุ่มมาก 123 ตัวอย่าง เมื่อนำชุดข้อมูลดังกล่าวไปทำการ SMOTE โดยกำหนดให้เพิ่มจำนวนตัวอย่างกลุ่มน้อย 100% จะพบว่าจำนวนตัวอย่างกลุ่มน้อยมีจำนวนเพิ่มขึ้น 1 เท่า จาก 32 ตัวอย่าง เพิ่มขึ้นเป็น 64 ตัวอย่าง ซึ่งเมื่อคิดเทียบเป็นเปอร์เซ็นต์กับจำนวนตัวอย่างทั้งหมดแล้ว จะทำให้มีจำนวนเพิ่มขึ้นแต่ยังน้อยกว่าจำนวนตัวอย่างกลุ่มมาก โดยจำนวนตัวอย่างกลุ่มน้อยที่เมื่อทำการ SMOTE ขึ้นมาแล้วรวมกับของเดิมด้วยจะคิดเป็น 34.22% จำนวนตัวอย่างกลุ่มมากคิดเป็น 65.78%

10) ชุดข้อมูล Ionosphere เป็นชุดข้อมูลที่แสดงถึงการจำแนกประเภทของการรับสัญญาณเรดาร์ในชั้นบรรยากาศว่ามีสัญญาณดีหรือไม่

ตารางที่ 4-11 จำนวนข้อมูลกลุ่มน้อยและจำนวนข้อมูลกลุ่มมากของชุดข้อมูล Ionosphere

ชุดข้อมูล	จำนวนตัวอย่าง (original)	เปอร์เซ็นต์ของจำนวนตัวอย่าง	จำนวนตัวอย่าง (SMOTE)	เปอร์เซ็นต์ของจำนวนตัวอย่าง
ข้อมูลกลุ่มน้อย	126	35.90	252	52.83
ข้อมูลกลุ่มมาก	225	64.10	225	47.17
ทั้งหมด	351	100	477	100

จากชุดข้อมูลต้นฉบับที่มีจำนวนตัวอย่างทั้งหมด 351 ตัวอย่าง โดยแบ่งเป็นตัวอย่างกลุ่มน้อย 126 ตัวอย่าง และตัวอย่างกลุ่มมาก 225 ตัวอย่าง เมื่อนำชุดข้อมูลดังกล่าวไปทำการ SMOTE โดยกำหนดให้เพิ่มจำนวนตัวอย่างกลุ่มน้อย 100% จะพบว่าจำนวนตัวอย่างกลุ่มน้อยมีจำนวนเพิ่มขึ้น 1 เท่า จาก 126 ตัวอย่าง เพิ่มขึ้นเป็น 252 ตัวอย่าง ซึ่งเมื่อคิดเทียบเป็นเปอร์เซ็นต์กับจำนวนตัวอย่างทั้งหมดแล้ว จะทำให้มีจำนวนเพิ่มขึ้นมีจำนวนมากกว่าจำนวนตัวอย่างกลุ่มมาก แต่จำนวนใกล้เคียงกัน โดยจำนวนตัวอย่างกลุ่มน้อยที่เมื่อทำการ SMOTE ขึ้นมาแล้วรวมกับของเดิมด้วยจะคิดเป็น 52.83% จำนวนตัวอย่างกลุ่มมากคิดเป็น 47.17%

11) ชุดข้อมูล Liver เป็นชุดข้อมูลที่แสดงถึงผลการทดสอบความผิดปกติของตับของกลุ่มทดสอบที่เป็นเพศชายที่ดื่มแอลกอฮอล์

ตารางที่ 4-12 จำนวนข้อมูลกลุ่มน้อยและจำนวนข้อมูลกลุ่มมากของชุดข้อมูล Liver

ชุดข้อมูล	จำนวนตัวอย่าง (original)	เปอร์เซ็นต์ของจำนวนตัวอย่าง	จำนวนตัวอย่าง (SMOTE)	เปอร์เซ็นต์ของจำนวนตัวอย่าง
ข้อมูลกลุ่มน้อย	145	42.03	290	59.18
ข้อมูลกลุ่มมาก	200	57.97	200	40.82
ทั้งหมด	345	100	490	100

จากชุดข้อมูลต้นฉบับที่มีจำนวนตัวอย่างทั้งหมด 345 ตัวอย่าง โดยแบ่งเป็นตัวอย่างกลุ่มน้อย 145 ตัวอย่าง และตัวอย่างกลุ่มมาก 200 ตัวอย่าง เมื่อนำชุดข้อมูลดังกล่าวไปทำการ SMOTE โดยกำหนดให้เพิ่มจำนวนตัวอย่างกลุ่มน้อย 100% จะพบว่าจำนวนตัวอย่างกลุ่มน้อยมีจำนวนเพิ่มขึ้น 1 เท่า จาก 145 ตัวอย่าง เพิ่มขึ้นเป็น 290 ตัวอย่าง ซึ่งเมื่อคิดเทียบเป็นเปอร์เซ็นต์กับจำนวนตัวอย่างทั้งหมดแล้ว จะทำให้มีจำนวนเพิ่มขึ้นจนมีจำนวนมากกว่าจำนวนตัวอย่างกลุ่มมาก แต่จำนวนใกล้เคียงกัน โดยจำนวนตัวอย่างกลุ่มน้อยที่เมื่อทำการ SMOTE ขึ้นมาแล้วรวมกับของเดิมด้วยจะคิดเป็น 59.18% จำนวนตัวอย่างกลุ่มมากคิดเป็น 40.82%

12) ชุดข้อมูล Post-operative เป็นชุดข้อมูลที่แสดงถึงข้อมูลของผู้ป่วยหลังการผ่าตัด

ตารางที่ 4-13 จำนวนข้อมูลกลุ่มน้อยและจำนวนข้อมูลกลุ่มมากของชุดข้อมูล Post-operative

ชุดข้อมูล	จำนวนตัวอย่าง (original)	เปอร์เซ็นต์ของจำนวนตัวอย่าง	จำนวนตัวอย่าง (SMOTE)	เปอร์เซ็นต์ของจำนวนตัวอย่าง
ข้อมูลกลุ่มน้อย	24	26.67	48	42.11
ข้อมูลกลุ่มมาก	66	73.33	66	57.89
ทั้งหมด	90	100	114	100

จากชุดข้อมูลต้นฉบับที่มีจำนวนตัวอย่างทั้งหมด 90 ตัวอย่าง โดยแบ่งเป็นตัวอย่างกลุ่มน้อย 24 ตัวอย่าง และตัวอย่างกลุ่มมาก 66 ตัวอย่าง เมื่อนำชุดข้อมูลดังกล่าวไปทำการ SMOTE โดยกำหนดให้เพิ่มจำนวนตัวอย่างกลุ่มน้อย 100% จะพบว่าจำนวนตัวอย่างกลุ่มน้อยมีจำนวนเพิ่มขึ้น 1 เท่า จาก 24 ตัวอย่าง เพิ่มขึ้นเป็น 48 ตัวอย่าง ซึ่งเมื่อคิดเทียบเป็นเปอร์เซ็นต์กับจำนวนตัวอย่างทั้งหมดแล้ว จะทำให้มีจำนวนเพิ่มขึ้นจนมีจำนวนใกล้เคียงกับจำนวนตัวอย่างกลุ่มมาก โดยจำนวนตัวอย่างกลุ่มน้อยที่เมื่อทำการ SMOTE ขึ้นมาแล้วรวมกับของเดิมด้วยจะคิดเป็น 42.11% จำนวนตัวอย่างกลุ่มมากคิดเป็น 57.89%

13) ชุดข้อมูล tic-tac-toe เป็นชุดข้อมูลที่แสดงถึงข้อมูลในการเล่นเกมส์โอ-เอ็กซ์

ตารางที่ 4-14 จำนวนข้อมูลกลุ่มน้อยและจำนวนข้อมูลกลุ่มมากของชุดข้อมูล tic-tac-toe

ชุดข้อมูล	จำนวนตัวอย่าง (original)	เปอร์เซ็นต์ของจำนวนตัวอย่าง	จำนวนตัวอย่าง (SMOTE)	เปอร์เซ็นต์ของจำนวนตัวอย่าง
ข้อมูลกลุ่มน้อย	332	34.66	664	51.47
ข้อมูลกลุ่มมาก	626	65.34	626	48.53
ทั้งหมด	958	100	1290	100

จากชุดข้อมูลต้นฉบับที่มีจำนวนตัวอย่างทั้งหมด 958 ตัวอย่าง โดยแบ่งเป็นตัวอย่างกลุ่มน้อย 332 ตัวอย่าง และตัวอย่างกลุ่มมาก 626 ตัวอย่าง เมื่อนำชุดข้อมูลดังกล่าวไปทำการ SMOTE โดยกำหนดให้เพิ่มจำนวนตัวอย่างกลุ่มน้อย 100% จะพบว่าจำนวนตัวอย่างกลุ่มน้อยมีจำนวนเพิ่มขึ้น 1 เท่า จาก 332 ตัวอย่าง เพิ่มขึ้นเป็น 664 ตัวอย่าง ซึ่งเมื่อคิดเทียบเป็นเปอร์เซ็นต์กับจำนวนตัวอย่างทั้งหมดแล้ว จะทำให้มีจำนวนเพิ่มขึ้นจนมีจำนวนมากกว่าจำนวนตัวอย่างกลุ่มมาก แต่จำนวนใกล้เคียงกันเกือบเท่ากัน โดยจำนวนตัวอย่างกลุ่มน้อยที่เมื่อทำการ SMOTE ขึ้นมาแล้วรวมกับของเดิมด้วยจะคิดเป็น 51.47% จำนวนตัวอย่างกลุ่มมากคิดเป็น 48.53%

14) ชุดข้อมูล Splice-ei เป็นชุดข้อมูลที่แสดงถึงข้อมูลลักษณะทางพันธุกรรมของมนุษย์ ที่พิจารณาในด้านของผู้บริจาคหรือผู้ให้

ตารางที่ 4-15 จำนวนข้อมูลกลุ่มน้อยและจำนวนข้อมูลกลุ่มมากของชุดข้อมูล splice-ei

ชุดข้อมูล	จำนวนตัวอย่าง (original)	เปอร์เซ็นต์ของจำนวนตัวอย่าง	จำนวนตัวอย่าง (SMOTE)	เปอร์เซ็นต์ของจำนวนตัวอย่าง
ข้อมูลกลุ่มน้อย	767	24.04	1,534	38.77
ข้อมูลกลุ่มมาก	2,423	75.96	2,423	61.23
ทั้งหมด	3,190	100	3,957	100

จากชุดข้อมูลต้นฉบับที่มีจำนวนตัวอย่างทั้งหมด 3,190 ตัวอย่าง โดยแบ่งเป็นตัวอย่างกลุ่มน้อย 767 ตัวอย่าง และตัวอย่างกลุ่มมาก 2,423 ตัวอย่าง เมื่อนำชุดข้อมูลดังกล่าวไปทำการ SMOTE โดยกำหนดให้เพิ่มจำนวนตัวอย่างกลุ่มน้อย 100% จะพบว่าจำนวนตัวอย่างกลุ่มน้อยมีจำนวนเพิ่มขึ้น 1 เท่า จาก 767 ตัวอย่าง เพิ่มขึ้นเป็น 1,534 ตัวอย่าง ซึ่งเมื่อคิดเทียบเป็นเปอร์เซ็นต์กับจำนวนตัวอย่างทั้งหมดแล้ว จะทำให้มีจำนวนเพิ่มขึ้นจนมีจำนวนใกล้เคียงเกือบเท่ากับจำนวนตัวอย่างกลุ่มมาก แต่ก็ยังมีจำนวนน้อยกว่าจำนวนข้อมูลกลุ่มมาก โดยจำนวนตัวอย่างกลุ่มน้อยที่เมื่อทำการ SMOTE ขึ้นมาแล้วรวมกับของเดิมด้วยจะคิดเป็น 38.77% จำนวนตัวอย่างกลุ่มมากคิดเป็น 61.23%

15) ชุดข้อมูล Splice-ie เป็นชุดข้อมูลที่แสดงถึงข้อมูลลักษณะทางพันธุกรรมของมนุษย์ ที่พิจารณาในด้านของผู้รับ

ตารางที่ 4-16 จำนวนข้อมูลกลุ่มน้อยและจำนวนข้อมูลกลุ่มมากของชุดข้อมูล splice-ie

ชุดข้อมูล	จำนวนตัวอย่าง (original)	เปอร์เซ็นต์ของจำนวนตัวอย่าง	จำนวนตัวอย่าง (SMOTE)	เปอร์เซ็นต์ของจำนวนตัวอย่าง
ข้อมูลกลุ่มน้อย	768	24.08	1,536	38.81
ข้อมูลกลุ่มมาก	2,422	75.92	2,422	61.19
ทั้งหมด	3,190	100	3,958	100

จากชุดข้อมูลต้นฉบับที่มีจำนวนตัวอย่างทั้งหมด 3,190 ตัวอย่าง โดยแบ่งเป็นตัวอย่างกลุ่มน้อย 768 ตัวอย่าง และตัวอย่างกลุ่มมาก 2,422 ตัวอย่าง เมื่อนำชุดข้อมูลดังกล่าวไปทำการ SMOTE โดยกำหนดให้เพิ่มจำนวนตัวอย่างกลุ่มน้อย 100% จะพบว่าจำนวนตัวอย่างกลุ่มน้อยมีจำนวนเพิ่มขึ้น 1 เท่า จาก 768 ตัวอย่าง เพิ่มขึ้นเป็น 1,536 ตัวอย่าง ซึ่งเมื่อคิดเทียบเป็นเปอร์เซ็นต์กับจำนวนตัวอย่างทั้งหมดแล้ว จะทำให้มีจำนวนเพิ่มขึ้นจนมีจำนวนใกล้เคียงเกือบเท่ากับจำนวนตัวอย่างกลุ่มมาก แต่ก็ยังมีจำนวนน้อยกว่าจำนวนข้อมูลกลุ่มมาก โดยจำนวนตัวอย่างกลุ่มน้อยที่เมื่อทำการ SMOTE ขึ้นมาแล้วรวมกับของเดิมด้วยจะคิดเป็น 38.81% จำนวนตัวอย่างกลุ่มมากคิดเป็น 61.19%

16) ชุดข้อมูล Vehicle เป็นชุดข้อมูลที่แสดงถึงข้อมูลการจำแนกประเภทของยานพาหนะจากภาพของรถในมุมต่างๆ

ตารางที่ 4-17 จำนวนข้อมูลกลุ่มน้อยและจำนวนข้อมูลกลุ่มมากของชุดข้อมูล vehicle

ชุดข้อมูล	จำนวนตัวอย่าง (original)	เปอร์เซ็นต์ของจำนวนตัวอย่าง	จำนวนตัวอย่าง (SMOTE)	เปอร์เซ็นต์ของจำนวนตัวอย่าง
ข้อมูลกลุ่มน้อย	199	23.52	398	38.09
ข้อมูลกลุ่มมาก	647	76.48	647	61.91
ทั้งหมด	846	100	1,045	100

จากชุดข้อมูลต้นฉบับที่มีจำนวนตัวอย่างทั้งหมด 846 ตัวอย่าง โดยแบ่งเป็นตัวอย่างกลุ่มน้อย 199 ตัวอย่าง และตัวอย่างกลุ่มมาก 647ตัวอย่าง เมื่อนำชุดข้อมูลดังกล่าวไปทำการ SMOTE โดยกำหนดให้เพิ่มจำนวนตัวอย่างกลุ่มน้อย 100% จะพบว่าจำนวนตัวอย่างกลุ่มน้อยมีจำนวนเพิ่มขึ้น 1 เท่า จาก 199 ตัวอย่าง เพิ่มขึ้นเป็น 398 ตัวอย่าง ซึ่งเมื่อคิดเทียบเป็นเปอร์เซ็นต์กับจำนวนตัวอย่างทั้งหมดแล้ว จะทำให้มีจำนวนเพิ่มขึ้นจนมีจำนวนใกล้เคียงกับจำนวนตัวอย่างกลุ่มมาก แต่ก็ยังมีจำนวนน้อยกว่าจำนวนข้อมูลกลุ่มมาก โดยจำนวนตัวอย่างกลุ่มน้อยที่เมื่อทำการ SMOTE ขึ้นมาแล้วรวมกับของเดิมด้วยจะคิดเป็น 38.09% จำนวนตัวอย่างกลุ่มมากคิดเป็น 61.91%

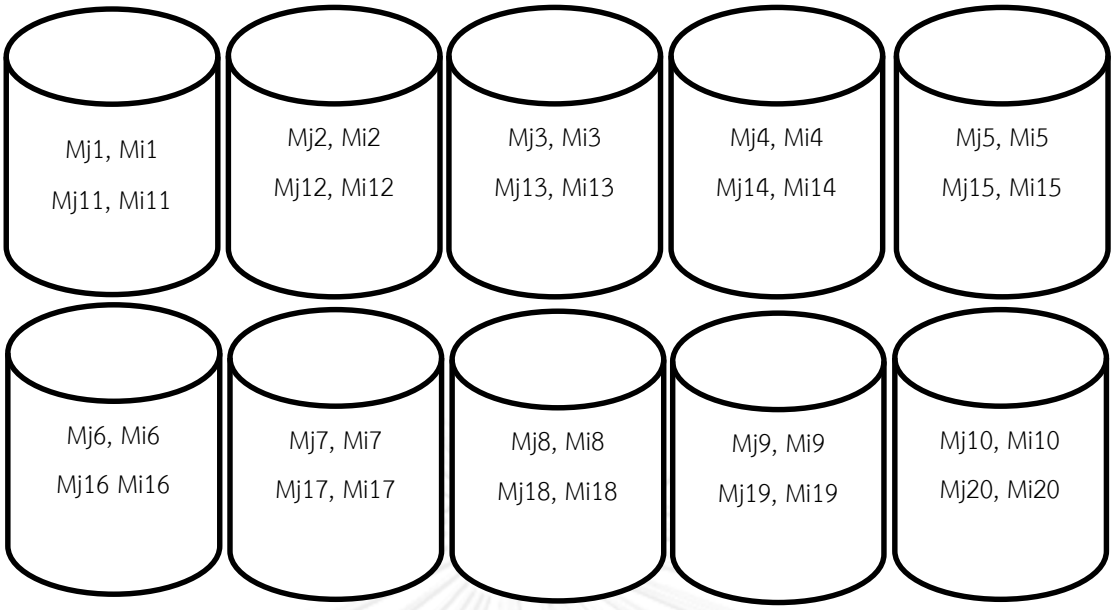
4.2 การเตรียมข้อมูลก่อนทำการทดลอง

ในการเตรียมข้อมูลก่อนทำการทดลองจะนำข้อมูลที่ได้คัดเลือกมาทำการจัดเตรียมข้อมูล เพื่อให้ได้ข้อมูลที่สมบูรณ์จากกรณีของข้อมูลที่มีค่าหายไป (Missing value) หรืออาจมีข้อมูลประเภทตัวเลขอยู่ในชุดข้อมูล จึงมีการทำการกรองข้อมูล โดยจะใช้โปรแกรม WEKA เวอร์ชั่น 3.7.10 [8] ส่วนขั้นตอนในการเพิ่มตัวอย่างกลุ่มน้อยโดยการสังเคราะห์ขึ้นมาใหม่นั้น หรือวิธีการ SMOTE นั้น จัดเป็นขั้นตอนเตรียมข้อมูลก่อนทำการทดลองเช่นกัน จะเป็นการสังเคราะห์ข้อมูลกลุ่มน้อยขึ้นมาตามจำนวนที่ผู้ใช้ต้องการหรือกำหนดขึ้น ซึ่งสามารถใช้ฟังก์ชั่น Filter ในการสังเคราะห์ข้อมูลกลุ่มน้อยขึ้นดังนี้ filters>supervised>instances>SMOTE และสามารถปรับเปลี่ยนพารามิเตอร์ต่างๆ ได้แก่ class value จะเป็นการระบุว่าข้อมูลกลุ่มไหนที่จะเป็นข้อมูลกลุ่มน้อย (minority class) ถ้าระบุเป็น 0 (ศูนย์) ระบบจะตั้งค่าให้เองอัตโนมัติ, Nearest Neighbors จะเป็นการพิจารณาจำนวนข้อมูลที่อยู่ใกล้กับข้อมูลกลุ่มน้อยว่ามีจำนวนกี่ตัวที่จะใช้พิจารณาในการสังเคราะห์ข้อมูลกลุ่มน้อยเพิ่มขึ้น เช่น กำหนดให้เท่ากับ 5 หมายความว่า เลือกข้อมูลกลุ่มน้อยที่อยู่ใกล้เคียงกันจำนวน 5 ตัวอย่าง แล้วทำการสังเคราะห์ข้อมูลกลุ่มน้อยขึ้นใหม่ 1 ตัวอย่าง, percentage เป็นการกำหนดจำนวนเปอร์เซ็นต์ที่ต้องการสังเคราะห์ข้อมูลกลุ่มน้อยเพิ่มขึ้น ว่าต้องการสังเคราะห์ข้อมูลกลุ่มน้อยเพิ่มขึ้นใหม่เป็นจำนวนเท่าไรจากข้อมูลน้อยที่มีอยู่เดิม เช่น ถ้ากำหนด 100% หมายความว่า ต้องการสังเคราะห์ข้อมูลกลุ่มน้อยเพิ่มขึ้นอีก 1 เท่าจากจำนวนกลุ่มน้อยเดิมที่มีอยู่

ในการทดลองได้กำหนดค่าต่างๆ ไว้ ดังนี้ กำหนด Nearest Neighbors เท่ากับ 5 และ percentage เท่ากับ 100% ซึ่งเป็นค่าที่ถูกกำหนดไว้ในโปรแกรมเพื่อเรียกใช้ตอนเริ่มต้นโปรแกรม

4.3 การทดสอบแบบไขว้ข้าม 10 กลุ่ม

ทำการทดสอบแบบ 10 Cross-validation Folds เพื่อวิเคราะห์ความถูกต้องในการจำแนกแบบจำลองที่ออกแบบไว้ วิธีการคือจะทำการแบ่งข้อมูลทั้งหมดออกเป็น 10 ส่วน 9 ส่วนแรกจะเป็นชุดข้อมูลสำหรับสอน และ 1 ส่วนที่เหลือจะใช้เป็นชุดข้อมูลสำหรับทดสอบ ทำการทดลองซ้ำเช่นนี้ 10 ครั้ง เพื่อให้ทุกกลุ่มได้เป็นทั้งชุดข้อมูลสำหรับสอน และชุดข้อมูลสำหรับทดสอบ ส่วนการจัดชุดข้อมูลลงในแต่ละกลุ่มนั้น จะทำการกระจายตัวอย่างทั้งตัวอย่างกลุ่มมาก และตัวอย่างกลุ่มน้อยให้กระจายอยู่ในทุกๆ กลุ่ม เพื่อให้มีการกระจายตัวของข้อมูล และทุกๆ กลุ่มมีทั้งตัวอย่างกลุ่มน้อยและตัวอย่างกลุ่มมาก เช่น ตัวอย่างกลุ่มมากตัวที่ 1 ใส่ลง folds ที่ 1, ตัวอย่างกลุ่มมากตัวที่ 2 ใส่ลง folds ที่ 2 ทำเช่นนี้จนครบทุก folds และในส่วนของตัวอย่างกลุ่มน้อยก็ทำเช่นนี้เหมือนกัน จนครบตัวอย่างทุกตัวในชุดข้อมูล



ภาพที่ 4-1: แสดงวิธีการจัดตัวอย่างข้อมูลกลุ่มมากและตัวอย่างข้อมูลกลุ่มน้อยลงในแต่ละกลุ่ม

จากรูป $Mj_1,Mi_2,Mi_3,...,Mi_n$ คือตัวอย่างกลุ่มมากตัวที่ 1-n ในชุดข้อมูล $Mi_1,Mi_2,Mi_3,...,Mi_n$ คือตัวอย่างกลุ่มน้อยตัวที่ 1-n ในชุดข้อมูล

เนื่องจากถ้าไม่กำหนดตามวิธีการข้างต้น เมื่อทำการทดสอบแบบไขว้ข้ามสิบกลุ่ม การกำหนดกลุ่มตัวอย่างจะถูกเลือกแบบสุ่มเพื่อใส่ไปในแต่ละกลุ่ม ทำให้บางกลุ่มอาจจะมีตัวอย่างกลุ่มมากอยู่จำนวนมากกว่าตัวอย่างกลุ่มน้อย หรือมีจำนวนตัวอย่างกลุ่มน้อยเดิม แต่ไม่มีตัวอย่างกลุ่มน้อยที่ถูกสังเคราะห์ใหม่อยู่เลย ซึ่งวิธีการแบ่งกลุ่มตามวิธีการเบื้องต้น จะทำให้ได้จำนวนตัวอย่างทั้งตัวอย่างกลุ่มมาก ตัวอย่างกลุ่มน้อย และตัวอย่างกลุ่มน้อยที่ถูกสังเคราะห์ใหม่กระจายอยู่ในทุกๆ กลุ่ม ซึ่งจะทำให้ตัวอย่างทุกๆ กลุ่มได้ถูกใช้เป็นทั้งชุดข้อมูลสำหรับฝึกและชุดข้อมูลสำหรับทดสอบ

4.4 การวัดผลการทดลอง

ตัววัดผลการทดลองในงานวิจัยฉบับนี้ ได้แก่ ค่าความแม่นยำในการทำนายผลตัวอย่างทั้งหมด (Total accuracy) ค่าความแม่นยำในการทำนายผลตัวอย่างกลุ่มมาก (Accuracy of Majority class) ค่าความแม่นยำในการทำนายผลตัวอย่างกลุ่มน้อย (Accuracy of Minority class) ค่าความระลึก (Recall) ค่าความเที่ยง (Precision) และค่าเอฟเมเชอร์ (F-measure) ซึ่งเป็นค่าที่ได้จากการคำนวณจากตารางคอนฟิวชันเมตริกซ์

ตารางที่ 4-18 ตารางคอนฟิวชั่นเมทริกซ์ (Confusion Matrix)

	ผลการทำนายตัวอย่างกลุ่มมาก	**ผลการทำนายตัวอย่างกลุ่มน้อย**
ค่าจริงกลุ่มมาก	TP : True Positive	FN : False Negative
ค่าจริงกลุ่มน้อย	FP : False Positive	TN : True Negative

จากตารางคอนฟิวชั่นเมทริกซ์ เป็นตารางที่แสดงผลของโปรแกรมในการทำนายโดยเปรียบเทียบกับผลลัพธ์จริง โดยค่า TP คือ ค่าที่เป็นจริงอยู่ในกลุ่มมาก และถูกทำนายว่าอยู่ในกลุ่มมาก ค่า FP คือ ค่าที่เป็นจริงอยู่ในกลุ่มมาก แต่ถูกทำนายว่าอยู่ในกลุ่มน้อย ค่า FP คือ ค่าที่เป็นจริงอยู่ในกลุ่มน้อย แต่ถูกทำนายว่าอยู่ในกลุ่มบวก และค่า TN คือ ค่าที่เป็นจริงอยู่ในกลุ่มน้อย และถูกทำนายว่าอยู่ในกลุ่มน้อย

4.4.1 ค่าความแม่นยำในการทำนายผลตัวอย่างทั้งหมด (Total accuracy)

เป็นค่าความแม่นยำทั้งหมดของข้อมูล คำนวณได้จาก

$$\text{Total accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \tag{11}$$

4.4.2 ค่าความแม่นยำในการทำนายผลตัวอย่างกลุ่มมาก (Accuracy of Majority class)

เป็นค่าความแม่นยำในข้อมูลกลุ่มมาก คำนวณได้จาก

$$\text{Accuracy of Majority class} = \frac{TN}{TN + FP} \tag{12}$$

4.4.3 ค่าความแม่นยำในการทำนายผลตัวอย่างกลุ่มน้อย (Accuracy of Minority class)

เป็นค่าความแม่นยำในข้อมูลกลุ่มน้อย คำนวณได้จาก

$$\text{Accuracy of Minority class} = \frac{TP}{TP + FN} \tag{13}$$

4.4.4 ค่าความระลึก (Recall) เพื่อทดสอบประสิทธิภาพความแม่นยำในการทำนายผลสามารถคำนวณได้จาก

$$\text{Recall} = \frac{\text{จำนวนข้อมูลที่ทำนายได้ถูกต้องในกลุ่มนั้น}}{\text{จำนวนข้อมูลที่มีอยู่ทั้งหมดในกลุ่มนั้น}} \tag{14}$$

เช่น ถ้าเป็นข้อมูลกลุ่มน้อย จะได้

$$\text{Recall} = \frac{TP}{TP + FN} \tag{15}$$

4.4.5 ค่าความแม่นยำ (Precision) คือ

$$\text{Precision} = \frac{\text{จำนวนข้อมูลที่ทำนายได้ถูกต้องในกลุ่มนั้น}}{\text{จำนวนข้อมูลทั้งหมดที่ถูกทายว่าอยู่ในกลุ่มนั้น}} \tag{16}$$

เช่น ถ้าเป็นข้อมูลกลุ่มมาก จะได้

$$\text{Precision} = \frac{TN}{TN + FN}$$

4.4.6 ค่าเอฟเมเชอร์ (F-measure) คือ ค่าที่แสดงถึงประสิทธิภาพโดยรวม คำนวณได้จาก

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (17)$$

4.4.7 ค่านัยสำคัญทางสถิติ

ทำการทดสอบสมมติฐานทางสถิติโดยกำหนดระดับความมีนัยสำคัญทางสถิติที่ระดับ 0.05 และ 0.1 ตามลำดับ ซึ่งทำการทดสอบแบบ One Tail Paired t-test เพื่อเปรียบเทียบความแตกต่างอย่างมีนัยสำคัญทางสถิติ

4.5 ผลการทดลอง

4.5.1 ค่าความแม่นยำ

จากผลการทดลองกับชุดข้อมูลทั้งหมด 16 ชุดข้อมูล ได้ผลค่าความแม่นยำในแต่ละกลุ่ม ค่าที่ได้ ได้จากการเลือกค่าแอลฟา (α) ที่มีผลในการจำแนกตัวอย่างแต่ละกลุ่มดีที่สุด ดังนี้

ตารางที่ 4-19 ค่าความแม่นยำของข้อมูลในแต่ละกลุ่มข้อมูลที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม

ชุดข้อมูล	ความแม่นยำทั้งหมด		ความแม่นยำกลุ่มมาก		ความแม่นยำกลุ่มน้อย	
	SMOTE	DWSMOTE	SMOTE	DWSMOTE	SMOTE	DWSMOTE
breast-cancer	72.24	72.24 (α = 2.7)	69.41	69.41 (α = 2.7)	74.63	74.63 (α = 2.7)
breast-cancer-w	**96.17**	95.96 (α = 3)	**96.27**	95.85 (α = 3)	96.07	96.07 (α = 3)
credit-g	71.08	**71.77** (α = 4)	**68.67**	67.83 (α = 4)	73.14	**75.14**** (α = 4)
diabetes	**77.32**	77.12 (α = 8)	**79.66**	78.92 (α = 8)	74.80	**75.20** (α = 8)
ecoli	90.03	**90.56*** (α = 7)	78.57	78.57 (α = 7)	92.70	**93.35*** (α = 7)
flags	88.10	**90.51*** (α = 4)	64.17	**70.00** (α = 4)	92.61	**94.35*** (α = 4)

ชุดข้อมูล	ความแม่นยำทั้งหมด		ความแม่นยำกลุ่มมาก		ความแม่นยำกลุ่มน้อย	
	SMOTE	DWSMOTE	SMOTE	DWSMOTE	SMOTE	DWSMOTE
glasses	83.98	**85.71**** (α = 6)	32.35	**35.29** (α = 6)	92.89	**94.42**** (α = 6)
haberman	72.73	**73.23** (α = 5)	65.26	**66.43** (α = 5)	78.12	78.12 (α = 5)
hepatitis	80.70	**83.39** (α = 12)	75.00	**81.43**** (α = 12)	83.72	**84.49** (α = 12)
ionosphere	88.49	**90.56*** (α = 6)	90.54	**90.85** (α = 6)	86.15	**90.22**** (α = 6)
liver	70.41	**73.47*** (α = 2.8)	84.83	**85.17** (α = 2.8)	49.50	**56.50*** (α = 2.8)
post-operative	68.80	**69.39** (α = 8)	50.50	50.50 (α = 8)	81.90	**82.86** (α = 8)
spice-ei	96.74	**96.79** (α = 3.1)	96.48	**96.54** (α = 3.1)	96.91	**96.95** (α = 3.1)
spice-ie	94.95	**95.20** (α = 2.3)	94.86	**94.92** (α = 2.3)	95.00	**95.05** (α = 2.3)
tic-tac-toe	88.76	**89.54** (α = 18)	91.27	**91.87** (α = 18)	86.10	**87.07** (α = 18)
vehicle	95.31	**95.60** (α = 4)	94.98	**95.99**** (α = 4)	**95.52**	95.36 (α = 4)

* หมายถึงนัยสำคัญทางสถิติของค่าความแม่นยำทั้งหมดที่ระดับ 0.1

** หมายถึงนัยสำคัญทางสถิติของค่าความแม่นยำทั้งหมดที่ระดับ 0.05

ผลจากการจำแนกข้อมูล ทั้งหมด 16 ชุดข้อมูล ด้วยวิธี SMOTE ที่ใช้อัลกอริทึม C4.5 กับวิธีการ DWSMOTE ความแม่นยำทั้งหมด วิธี DWSMOTE ดีกว่า วิธี SMOTE จำนวน 13 ชุดข้อมูล โดยมากกว่าอย่างมีนัยสำคัญ 5 ชุดข้อมูล ความแม่นยำในกลุ่มมาก วิธี DWSMOTE ดีกว่า วิธี SMOTE จำนวน 10 ชุดข้อมูล โดยมากกว่าอย่างมีนัยสำคัญ 2 ชุดข้อมูล และความแม่นยำในกลุ่มน้อย วิธี DWSMOTE ดีกว่า วิธี SMOTE จำนวน 12 ชุดข้อมูล โดยมากกว่าอย่างมีนัยสำคัญ 6 ชุดข้อมูล

ตารางที่ 4-20 แสดงค่าความแม่นยำของข้อมูลในแต่ละกลุ่มข้อมูลด้วยการทดสอบแบบไขว้ข้ามสิบกลุ่มโดยแสดงในแต่ละกลุ่มที่ทดสอบของชุดข้อมูล breast-cancer เมื่อกำหนดค่า α = 2.7

กลุ่มที่	ความแม่นยำทั้งหมด		ความแม่นยำกลุ่มมาก		ความแม่นยำกลุ่มน้อย	
	SMOTE	DWSMOTE	SMOTE	DWSMOTE	SMOTE	DWSMOTE
1	73.68	73.68	52.94	52.94	90.48	90.48
2	86.49	83.78	88.24	82.35	85.00	85.00
3	62.16	64.86	58.82	64.71	65.00	65.00
4	75.68	75.68	64.71	64.71	85.00	85.00
5	64.86	64.86	52.94	52.94	75.00	75.00
6	78.38	75.68	88.24	88.24	70.00	65.00
7	70.27	70.27	88.24	88.24	55.00	55.00
8	72.97	75.68	58.82	58.82	85.00	90.00
9	75.68	75.68	82.35	82.35	70.00	70.00
10	62.16	62.16	58.82	58.82	65.00	65.00
ค่าเฉลี่ย	72.23	72.23	69.41	69.41	74.55	74.55

ตารางที่ 4-21 แสดงค่าความแม่นยำของข้อมูลในแต่ละกลุ่มข้อมูลด้วยการทดสอบแบบไขว้ข้ามสิบกลุ่มโดยแสดงในแต่ละกลุ่มที่ทดสอบของชุดข้อมูล breast-cancer-w เมื่อกำหนดค่า α = 3

กลุ่มที่	ความแม่นยำทั้งหมด		ความแม่นยำกลุ่มมาก		ความแม่นยำกลุ่มน้อย	
	SMOTE	DWSMOTE	SMOTE	DWSMOTE	SMOTE	DWSMOTE
1	97.89	97.89	97.96	97.96	97.83	97.83
2	93.68	91.58	93.88	89.80	93.48	93.48
3	95.74	95.74	95.83	95.83	95.65	95.65
4	97.87	97.87	97.92	97.92	97.83	97.83
5	95.74	95.74	95.83	95.83	95.65	95.65
6	92.55	92.55	89.58	89.58	95.65	95.65
7	96.81	96.81	97.92	97.92	95.65	95.65
8	94.68	94.68	95.83	95.83	93.48	93.48
9	97.85	97.85	97.92	97.92	97.78	97.78
10	98.92	98.92	100.00	100.00	97.78	97.78
ค่าเฉลี่ย	**96.18**	95.97	**96.27**	95.86	96.08	96.08

ตารางที่ 4-22 แสดงค่าความแม่นยำของข้อมูลในแต่ละกลุ่มข้อมูลด้วยการทดสอบแบบไขว้ข้ามสิบกลุ่มโดยแสดงในแต่ละกลุ่มที่ทดสอบของชุดข้อมูล credit-g เมื่อกำหนดค่า α = 3

กลุ่มที่	ความแม่นยำทั้งหมด		ความแม่นยำกลุ่มมาก		ความแม่นยำกลุ่มน้อย	
	SMOTE	DWSMOTE	SMOTE	DWSMOTE	SMOTE	DWSMOTE
1	65.38	65.38	58.33	56.67	71.43	72.86
2	75.38	74.62	68.33	66.67	81.43	81.43
3	70.00	69.23	71.67	70.00	68.57	68.57
4	63.08	63.85	55.00	56.67	70.00	70.00
5	68.46	70.77	61.67	63.33	74.29	77.14
6	73.08	69.23	81.67	68.33	65.71	70.00
7	67.69	72.31	66.67	70.00	68.57	74.29
8	75.38	76.15	68.33	70.00	81.43	81.43
9	83.08	80.00	86.67	80.00	80.00	80.00
10	69.23	69.23	68.33	68.33	70.00	70.00
ค่าเฉลี่ย	71.08	71.08	**68.67**	67.00	73.14	**74.57****

ตารางที่ 4-23 แสดงค่าความแม่นยำของข้อมูลในแต่ละกลุ่มข้อมูลด้วยการทดสอบแบบไขว้ข้ามสิบกลุ่มโดยแสดงในแต่ละกลุ่มที่ทดสอบของชุดข้อมูล diabetes เมื่อกำหนดค่า α = 8

กลุ่มที่	ความแม่นยำทั้งหมด		ความแม่นยำกลุ่มมาก		ความแม่นยำกลุ่มน้อย	
	SMOTE	DWSMOTE	SMOTE	DWSMOTE	SMOTE	DWSMOTE
1	81.73	83.65	81.48	79.63	82.00	88.00
2	76.92	78.85	83.33	85.19	70.00	72.00
3	66.35	63.46	66.67	66.67	66.00	60.00
4	73.08	73.08	74.07	70.37	72.00	76.00
5	76.92	74.04	75.93	68.52	78.00	80.00
6	80.77	82.69	88.89	87.04	72.00	78.00
7	84.47	85.44	84.91	86.79	84.00	84.00
8	75.73	75.73	81.13	83.02	70.00	68.00
9	77.67	78.64	81.13	83.02	74.00	74.00
10	79.61	75.73	79.25	79.25	80.00	72.00
ค่าเฉลี่ย	**77.32**	77.13	**79.68**	78.95	74.80	**75.20**

ตารางที่ 4-24 แสดงค่าความแม่นยำของข้อมูลในแต่ละกลุ่มข้อมูลด้วยการทดสอบแบบไขว้ข้ามสิบกลุ่มโดยแสดงในแต่ละกลุ่มที่ทดสอบของชุดข้อมูล ecoli เมื่อกำหนดค่า α = 7

กลุ่มที่	ความแม่นยำทั้งหมด		ความแม่นยำกลุ่มมาก		ความแม่นยำกลุ่มน้อย	
	SMOTE	DWSMOTE	SMOTE	DWSMOTE	SMOTE	DWSMOTE
1	89.47	92.11	85.71	85.71	90.32	93.55
2	81.08	83.78	85.71	85.71	80.00	83.33
3	89.19	89.19	57.14	57.14	96.67	96.67
4	89.19	89.19	71.43	71.43	93.33	93.33
5	94.59	94.59	85.71	85.71	96.67	96.67
6	94.59	94.59	71.43	71.43	100.00	100.00
7	86.49	86.49	71.43	71.43	90.00	90.00
8	86.49	86.49	85.71	85.71	86.67	86.67
9	89.19	89.19	71.43	71.43	93.33	93.33
10	100.00	100.00	100.00	100.00	100.00	100.00
ค่าเฉลี่ย	90.03	**90.56***	78.57	78.57	92.70	**93.35***

ตารางที่ 4-25 แสดงค่าความแม่นยำของข้อมูลในแต่ละกลุ่มข้อมูลด้วยการทดสอบแบบไขว้ข้ามสิบกลุ่มโดยแสดงในแต่ละกลุ่มที่ทดสอบของชุดข้อมูล flags เมื่อกำหนดค่า α = 4

กลุ่มที่	ความแม่นยำทั้งหมด		ความแม่นยำกลุ่มมาก		ความแม่นยำกลุ่มน้อย	
	SMOTE	DWSMOTE	SMOTE	DWSMOTE	SMOTE	DWSMOTE
1	81.82	90.91	50.00	50.00	88.89	100.00
2	95.45	95.45	75.00	75.00	100.00	100.00
3	90.91	86.36	100.00	100.00	88.89	83.33
4	86.36	90.91	50.00	75.00	94.44	94.44
5	90.48	85.71	100.00	66.67	88.89	88.89
6	90.48	90.48	66.67	66.67	94.44	94.44
7	90.48	95.24	33.33	66.67	100.00	100.00
8	75.00	80.00	66.67	66.67	76.47	82.35
9	85.00	95.00	33.33	66.67	94.12	100.00
10	95.00	95.00	66.67	66.67	100.00	100.00
ค่าเฉลี่ย	88.10	**90.51***	64.17	**70.00**	92.61	**94.35***

ตารางที่ 4-26 แสดงค่าความแม่นยำของข้อมูลในแต่ละกลุ่มข้อมูลด้วยการทดสอบแบบไขว้ข้ามสิบกลุ่มโดยแสดงในแต่ละกลุ่มที่ทดสอบของชุดข้อมูล glasses เมื่อกำหนดค่า α = 6

กลุ่มที่	ความแม่นยำทั้งหมด		ความแม่นยำกลุ่มมาก		ความแม่นยำกลุ่มน้อย	
	SMOTE	DWSMOTE	SMOTE	DWSMOTE	SMOTE	DWSMOTE
1	87.50	87.50	25.00	25.00	100.00	100.00
2	83.33	91.67	50.00	75.00	90.00	95.00
3	91.67	91.67	50.00	50.00	100.00	100.00
4	91.67	91.67	50.00	50.00	100.00	100.00
5	82.61	82.61	33.33	33.33	90.00	90.00
6	82.61	86.96	0.00	0.00	95.00	100.00
7	78.26	82.61	0.00	0.00	90.00	95.00
8	77.27	77.27	33.33	33.33	84.21	84.21
9	95.45	95.45	66.67	66.67	100.00	100.00
10	68.18	68.18	0.00	0.00	78.95	78.95
ค่าเฉลี่ย	83.86	**85.56****	30.83	**33.33**	92.82	**94.32****

ตารางที่ 4-27 แสดงค่าความแม่นยำของข้อมูลในแต่ละกลุ่มข้อมูลด้วยการทดสอบแบบไขว้ข้ามสิบกลุ่มโดยแสดงในแต่ละกลุ่มที่ทดสอบของชุดข้อมูล haberman เมื่อกำหนดค่า α = 5

กลุ่มที่	ความแม่นยำทั้งหมด		ความแม่นยำกลุ่มมาก		ความแม่นยำกลุ่มน้อย	
	SMOTE	DWSMOTE	SMOTE	DWSMOTE	SMOTE	DWSMOTE
1	80.00	85.00	76.47	88.24	82.61	82.61
2	87.50	87.50	82.35	82.35	91.30	91.30
3	79.49	79.49	68.75	68.75	86.96	86.96
4	84.62	84.62	81.25	81.25	86.96	86.96
5	64.10	64.10	62.50	62.50	65.22	65.22
6	65.79	65.79	62.50	62.50	68.18	68.18
7	71.05	71.05	56.25	56.25	81.82	81.82
8	63.16	63.16	50.00	50.00	72.73	72.73
9	65.79	65.79	56.25	56.25	72.73	72.73
10	65.79	65.79	56.25	56.25	72.73	72.73
ค่าเฉลี่ย	72.73	**73.23**	65.26	**66.43**	78.12	78.12

ตารางที่ 4-28 แสดงค่าความแม่นยำของข้อมูลในแต่ละกลุ่มข้อมูลด้วยการทดสอบแบบไขว้ข้ามสิบกลุ่มโดยแสดงในแต่ละกลุ่มที่ทดสอบของชุดข้อมูล hepatitis เมื่อกำหนดค่า α = 12

กลุ่มที่	ความแม่นยำทั้งหมด		ความแม่นยำกลุ่มมาก		ความแม่นยำกลุ่มน้อย	
	SMOTE	DWSMOTE	SMOTE	DWSMOTE	SMOTE	DWSMOTE
1	85.00	90.00	85.71	85.71	84.62	92.31
2	80.00	80.00	85.71	85.71	76.92	76.92
3	80.00	85.00	57.14	71.43	92.31	92.31
4	84.21	78.95	71.43	71.43	91.67	83.33
5	88.89	83.33	100.00	100.00	83.33	75.00
6	61.11	83.33	66.67	100.00	58.33	75.00
7	83.33	88.89	66.67	83.33	91.67	91.67
8	77.78	83.33	66.67	66.67	83.33	91.67
9	88.89	83.33	83.33	83.33	91.67	83.33
10	77.78	77.78	66.67	66.67	83.33	83.33
ค่าเฉลี่ย	80.70	**83.39**	75.00	**81.43****	83.72	**84.49**

ตารางที่ 4-29 แสดงค่าความแม่นยำของข้อมูลในแต่ละกลุ่มข้อมูลด้วยการทดสอบแบบไขว้ข้ามสิบกลุ่มโดยแสดงในแต่ละกลุ่มที่ทดสอบของชุดข้อมูล ionosphere เมื่อกำหนดค่า α = 6

กลุ่มที่	ความแม่นยำทั้งหมด		ความแม่นยำกลุ่มมาก		ความแม่นยำกลุ่มน้อย	
	SMOTE	DWSMOTE	SMOTE	DWSMOTE	SMOTE	DWSMOTE
1	87.76	100.00	84.62	100.00	91.30	100.00
2	81.63	87.76	80.77	88.46	82.61	86.96
3	87.50	87.50	88.00	84.00	86.96	91.30
4	93.75	87.50	100.00	88.00	86.96	86.96
5	89.58	87.50	88.00	88.00	91.30	86.96
6	85.11	85.11	92.00	92.00	77.27	77.27
7	91.49	89.36	96.00	88.00	86.36	90.91
8	89.36	97.87	88.00	96.00	90.91	100.00
9	91.49	93.62	100.00	92.00	81.82	95.45
10	87.23	89.36	88.00	92.00	86.36	86.36
ค่าเฉลี่ย	88.49	**90.56***	90.54	**90.85**	86.19	**90.22****

ตารางที่ 4-30 แสดงค่าความแม่นยำของข้อมูลในแต่ละกลุ่มข้อมูลด้วยการทดสอบแบบไขว้ข้ามสิบกลุ่มโดยแสดงในแต่ละกลุ่มที่ทดสอบของชุดข้อมูล liver เมื่อกำหนดค่า α = 2.8

กลุ่มที่	ความแม่นยำทั้งหมด		ความแม่นยำกลุ่มมาก		ความแม่นยำกลุ่มน้อย	
	SMOTE	DWSMOTE	SMOTE	DWSMOTE	SMOTE	DWSMOTE
1	65.31	65.31	86.21	86.21	35.00	35.00
2	67.35	71.43	72.41	79.31	60.00	60.00
3	79.59	79.59	89.66	89.66	65.00	65.00
4	61.22	77.55	93.10	89.66	15.00	60.00
5	73.47	71.43	79.31	79.31	65.00	60.00
6	67.35	67.35	89.66	89.66	35.00	35.00
7	65.31	77.55	82.76	82.76	40.00	70.00
8	71.43	71.43	86.21	86.21	50.00	50.00
9	83.67	83.67	93.10	93.10	70.00	70.00
10	69.39	69.39	75.86	75.86	60.00	60.00
ค่าเฉลี่ย	70.41	**73.47***	84.83	**85.17**	49.50	**56.50***

ตารางที่ 4-31 แสดงค่าความแม่นยำของข้อมูลในแต่ละกลุ่มข้อมูลด้วยการทดสอบแบบไขว้ข้ามสิบกลุ่มโดยแสดงในแต่ละกลุ่มที่ทดสอบของชุดข้อมูล post-operative เมื่อกำหนดค่า α = 8

กลุ่มที่	ความแม่นยำทั้งหมด		ความแม่นยำกลุ่มมาก		ความแม่นยำกลุ่มน้อย	
	SMOTE	DWSMOTE	SMOTE	DWSMOTE	SMOTE	DWSMOTE
1	66.67	66.67	40.00	40.00	85.71	85.71
2	75.00	66.67	60.00	40.00	85.71	85.71
3	50.00	50.00	20.00	20.00	71.43	71.43
4	58.33	66.67	40.00	40.00	71.43	85.71
5	75.00	83.33	60.00	60.00	85.71	100.00
6	66.67	83.33	40.00	60.00	85.71	100.00
7	72.73	63.64	80.00	80.00	66.67	50.00
8	63.64	63.64	40.00	40.00	83.33	83.33
9	80.00	70.00	50.00	50.00	100.00	83.33
10	80.00	80.00	75.00	75.00	83.33	83.33
ค่าเฉลี่ย	68.80	**69.39**	50.50	50.50	81.90	**82.86**

ตารางที่ 4-32 แสดงค่าความแม่นยำของข้อมูลในแต่ละกลุ่มข้อมูลด้วยการทดสอบแบบไขว้ข้ามสิบกลุ่มโดยแสดงในแต่ละกลุ่มที่ทดสอบของชุดข้อมูล splice-ei เมื่อกำหนดค่า α = 3.1

กลุ่มที่	ความแม่นยำทั้งหมด		ความแม่นยำกลุ่มมาก		ความแม่นยำกลุ่มน้อย	
	SMOTE	DWSMOTE	SMOTE	DWSMOTE	SMOTE	DWSMOTE
1	96.73	96.73	96.75	97.40	96.71	96.30
2	96.98	96.98	97.40	97.40	96.71	96.71
3	95.72	95.72	96.10	95.45	95.47	95.88
4	97.47	98.23	98.70	98.70	96.69	97.93
5	97.72	97.72	97.39	97.39	97.93	97.11
6	94.43	94.43	90.85	90.85	96.69	96.69
7	96.46	96.46	98.04	98.04	95.45	95.45
8	96.46	96.46	94.77	94.77	97.52	97.52
9	97.47	97.72	96.73	96.73	97.93	98.35
10	97.97	97.97	98.04	98.69	97.93	97.52
ค่าเฉลี่ย	96.74	**96.79**	96.48	**96.54**	96.91	**96.95**

ตารางที่ 4-33 แสดงค่าความแม่นยำของข้อมูลในแต่ละกลุ่มข้อมูลด้วยการทดสอบแบบไขว้ข้ามสิบกลุ่มโดยแสดงในแต่ละกลุ่มที่ทดสอบของชุดข้อมูล splice-ie เมื่อกำหนดค่า α = 2.3

กลุ่มที่	ความแม่นยำทั้งหมด		ความแม่นยำกลุ่มมาก		ความแม่นยำกลุ่มน้อย	
	SMOTE	DWSMOTE	SMOTE	DWSMOTE	SMOTE	DWSMOTE
1	93.70	93.70	93.51	93.51	93.83	93.83
2	95.47	95.21	94.81	94.81	95.88	95.47
3	94.95	95.45	96.10	96.75	94.21	94.63
4	95.45	95.45	94.81	94.81	95.87	95.87
5	97.47	97.47	98.05	98.05	97.11	97.11
6	94.70	94.44	93.51	92.86	95.45	95.45
7	93.42	93.92	92.81	93.46	93.80	94.21
8	94.94	94.94	92.81	92.81	96.28	96.28
9	95.44	95.44	96.08	96.08	95.04	95.04
10	93.92	93.92	96.08	96.08	92.56	92.56
ค่าเฉลี่ย	94.95	**95.00**	94.86	**94.92**	95.00	**95.05**

ตารางที่ 4-34 แสดงค่าความแม่นยำของข้อมูลในแต่ละกลุ่มข้อมูลด้วยการทดสอบแบบไขว้ข้ามสิบกลุ่มโดยแสดงในแต่ละกลุ่มที่ทดสอบของชุดข้อมูล tic-tac-toe เมื่อกำหนดค่า α = 18

กลุ่มที่	ความแม่นยำทั้งหมด		ความแม่นยำกลุ่มมาก		ความแม่นยำกลุ่มน้อย	
	SMOTE	DWSMOTE	SMOTE	DWSMOTE	SMOTE	DWSMOTE
1	88.46	90.77	91.04	94.03	85.71	87.30
2	85.38	84.62	92.54	89.55	77.78	79.37
3	91.54	91.54	91.04	92.54	92.06	90.48
4	90.77	90.00	89.55	91.04	92.06	88.89
5	90.70	88.37	92.42	92.42	88.89	84.13
6	89.15	90.70	93.94	93.94	84.13	87.30
7	88.28	86.72	90.91	89.39	85.48	83.87
8	91.41	92.97	93.94	89.39	88.71	96.77
9	85.94	89.06	84.85	90.91	87.10	87.10
10	85.94	90.63	92.42	95.45	79.03	85.48
ค่าเฉลี่ย	88.76	**89.54**	91.27	**91.87**	86.10	**87.07**

ตารางที่ 4-35 แสดงค่าความแม่นยำของข้อมูลในแต่ละกลุ่มข้อมูลด้วยการทดสอบแบบไขว้ข้ามสิบกลุ่มโดยแสดงในแต่ละกลุ่มที่ทดสอบของชุดข้อมูล vehicle เมื่อกำหนดค่า α = 4

กลุ่มที่	ความแม่นยำทั้งหมด		ความแม่นยำกลุ่มมาก		ความแม่นยำกลุ่มน้อย	
	SMOTE	DWSMOTE	SMOTE	DWSMOTE	SMOTE	DWSMOTE
1	97.14	96.19	95.00	97.50	98.46	95.38
2	97.14	98.10	95.00	97.50	98.46	98.46
3	94.29	94.29	95.00	95.00	93.85	93.85
4	95.24	97.14	97.50	97.50	93.85	96.92
5	95.24	95.24	95.00	95.00	95.38	95.38
6	92.38	92.38	90.00	90.00	93.85	93.85
7	95.24	96.19	95.00	97.50	95.38	95.38
8	94.23	94.23	95.00	95.00	93.75	93.75
9	94.17	95.15	94.87	97.44	93.75	93.75
10	98.06	97.09	97.44	97.44	98.44	96.88
ค่าเฉลี่ย	95.31	**95.60**	94.98	**95.99****	95.52	**95.36**

ตารางที่ 4-36 สรุปผลค่าความระลึก (Recall) ค่าความแม่นยำ (Precision) และค่าเอฟเมเชอร์ (F-measure) ในข้อมูลกลุ่มน้อยของแต่ละชุดข้อมูลที่ใช้ทดสอบ ด้วยการทดสอบแบบไขว้ข้ามสิบกลุ่ม

ชุดข้อมูล	อัลกอริทึม	ค่าความระลึก	ค่าความแม่นยำ	ค่าเอฟเมเชอร์
breast-cancer	SMOTE	0.75	0.74	0.74
	DWSMOTE (α= 2.7)	0.75	0.74	0.74
breast-cancer-w	SMOTE	0.96	0.96	0.96
	DWSMOTE (α= 3)	0.96	0.96	0.96
credit-g	SMOTE	0.73	0.73	0.73
	DWSMOTE (α= 4)	**0.75**	0.73	**0.74**
diabetes	SMOTE	0.75	0.77	0.76
	DWSMOTE (α= 8)	0.75	0.77	0.76
ecoli	SMOTE	0.93	0.95	0.94
	DWSMOTE (α= 7)	0.93	0.95	0.94
flags	SMOTE	0.93	0.93	0.93
	DWSMOTE (α= 4)	**0.94**	**0.94**	**0.94**
glasses	SMOTE	0.93	0.89	0.91
	DWSMOTE (α= 6)	**0.94**	0.89	**0.92**
haberman	SMOTE	0.78	0.76	0.77
	DWSMOTE (α= 5)	0.78	**0.77**	0.77
hepatitis	SMOTE	0.84	0.87	0.85
	DWSMOTE (α= 12)	**0.85**	**0.90**	**0.87**
ionosphere	SMOTE	0.86	0.89	0.88
	DWSMOTE (α= 6)	**0.90**	**0.90**	**0.90**
liver	SMOTE	0.50	0.69	0.58
	DWSMOTE (α= 2.8)	**0.57**	**0.72**	**0.63**
post-operative	SMOTE	0.82	0.69	0.75
	DWSMOTE (α= 8)	**0.83**	**0.70**	**0.76**
spice-ei	SMOTE	0.97	0.98	0.97
	DWSMOTE (α= 3.1)	0.97	0.98	0.97
spice-ie	SMOTE	0.95	0.97	0.96
	DWSMOTE (α= 2.3)	0.95	0.97	0.96
tic-tac-toe	SMOTE	0.86	0.90	0.88
	DWSMOTE (α= 18)	**0.87**	**0.91**	**0.89**

ชุดข้อมูล	อัลกอริทึม	ค่าความระลึก	ค่าความแม่นยำ	ค่าเอฟเมเซอร์
vehicle	SMOTE	**0.96**	0.97	0.96
	DWSMOTE (α= 4)	0.95	0.97	0.96

เนื่องจากงานวิจัยฉบับนี้ให้ความสนใจชุดข้อมูลไม่สมดุล และเพื่อต้องการเพิ่มประสิทธิภาพในการจำแนกข้อมูลกลุ่มน้อยได้ดีขึ้น ซึ่งจากผลการทดลอง วิธีการที่นำเสนอให้ค่าความระลึกในข้อมูลกลุ่มน้อยได้ดีกว่า วิธี C4.5 (SMOTE) จำนวน 8 ชุดข้อมูล และประสิทธิภาพโดยรวม วิธีการที่นำเสนอดีกว่า C4.5 (SMOTE) จำนวน 8 ชุดข้อมูล จากชุดข้อมูลที่ใช้ทดสอบทั้งหมด 16 ชุดข้อมูล

ตารางที่ 4-37 สรุปผลค่าความระลึก (Recall) ค่าความแม่นยำ (Precision) และค่าเอฟเมเซอร์ (F-measure) ในข้อมูลกลุ่มมากของแต่ละชุดข้อมูลที่ใช้ทดสอบ ด้วยการทดสอบแบบไขว้ข้ามสิบกลุ่ม

ชุดข้อมูล	อัลกอริทึม	ค่าความระลึก	ค่าความแม่นยำ	ค่าเอฟเมเซอร์
breast-cancer	SMOTE	0.69	0.70	0.70
	DWSMOTE (α= 2.7)	0.69	0.70	0.70
breast-cancer-w	SMOTE	0.96	0.96	0.96
	DWSMOTE (α= 3)	0.96	0.96	0.96
credit-g	SMOTE	**0.69**	0.69	**0.69**
	DWSMOTE (α= 4)	0.67	0.69	0.68
diabetes	SMOTE	**0.80**	0.77	0.78
	DWSMOTE (α= 8)	0.79	0.77	0.78
ecoli	SMOTE	0.79	0.71	0.75
	DWSMOTE (α= 7)	0.79	**0.73**	**0.76**
flags	SMOTE	0.65	0.63	0.64
	DWSMOTE (α= 4)	**0.71**	**0.71**	**0.71**
glasses	SMOTE	0.32	0.44	0.37
	DWSMOTE (α= 6)	**0.35**	**0.52**	**0.42**
haberman	SMOTE	0.65	0.68	0.67
	DWSMOTE (α= 5)	**0.67**	**0.69**	**0.68**
hepatitis	SMOTE	0.75	0.71	0.73
	DWSMOTE (α= 12)	**0.81**	**0.73**	**0.77**
ionosphere	SMOTE	0.90	0.88	0.89
	DWSMOTE (α= 6)	**0.91**	**0.91**	**0.91**
liver	SMOTE	0.85	0.71	0.77
	DWSMOTE (α= 2.8)	0.85	**0.74**	**0.79**

ชุดข้อมูล	อัลกอริทึม	ค่าความระลึก	ค่าความแม่นยำ	ค่าเอฟเมเซอร์
post-operative	SMOTE	0.50	0.67	0.57
	DWSMOTE (α= 8)	0.50	**0.69**	**0.58**
spice-ei	SMOTE	0.96	0.95	0.96
	DWSMOTE (α= 3.1)	**0.97**	0.95	0.96
spice-ie	SMOTE	0.95	0.92	0.94
	DWSMOTE (α= 2.3)	0.95	0.92	0.94
tic-tac-toe	SMOTE	0.91	0.87	0.89
	DWSMOTE (α= 18)	**0.92**	**0.88**	**0.90**
vehicle	SMOTE	0.95	0.93	0.94
	DWSMOTE (α= 4)	**0.96**	0.93	0.94

ตารางที่ 4-38 สรุปผลค่าความระลึก (Recall) ค่าความแม่นยำ (Precision) และค่าเอฟเมเซอร์ (F-measure) ของกลุ่มข้อมูลทั้งหมดแต่ละชุดข้อมูลที่ใช้ทดสอบ ด้วยการทดสอบแบบไขว้ข้ามสิบกลุ่ม

ชุดข้อมูล	อัลกอริทึม	ค่าความระลึก	ค่าความแม่นยำ	ค่าเอฟเมเซอร์
breast-cancer	SMOTE	0.72	0.72	0.72
	DWSMOTE (α= 2.7)	0.72	0.72	0.72
breast-cancer-w	SMOTE	0.96	0.96	0.96
	DWSMOTE (α= 3)	0.96	0.96	0.96
credit-g	SMOTE	0.71	0.71	0.71
	DWSMOTE (α= 4)	0.71	0.71	0.71
diabetes	SMOTE	0.77	0.77	0.77
	DWSMOTE (α= 8)	0.77	0.77	0.77
ecoli	SMOTE	0.90	0.90	0.90
	DWSMOTE (α= 7)	**0.91**	**0.91**	**0.91**
flags	SMOTE	0.88	0.88	0.88
	DWSMOTE (α= 4)	**0.91**	**0.91**	**0.91**
glasses	SMOTE	0.84	0.84	0.84
	DWSMOTE (α= 6)	**0.86**	**0.86**	**0.86**
haberman	SMOTE	0.73	0.73	0.73
	DWSMOTE (α= 5)	0.73	0.73	0.73

ชุดข้อมูล	อัลกอริทึม	ค่าความระลึก	ค่าความแม่นยำ	ค่าเอฟเมเซอร์
hepatitis	SMOTE	0.81	0.81	0.81
	DWSMOTE (α= 12)	**0.83**	**0.83**	**0.83**
ionosphere	SMOTE	0.88	0.88	0.88
	DWSMOTE (α= 6)	**0.91**	**0.91**	**0.91**
liver	SMOTE	0.70	0.70	0.70
	DWSMOTE (α= 2.8)	**0.73**	**0.73**	**0.73**
post-operative	SMOTE	0.68	0.68	0.68
	DWSMOTE (α= 8)	**0.69**	**0.69**	**0.69**
spice-ei	SMOTE	0.97	0.97	0.97
	DWSMOTE (α= 3.1)	0.97	0.97	0.97
spice-ie	SMOTE	0.95	0.95	0.95
	DWSMOTE (α= 2.3)	0.95	0.95	0.95
tic-tac-toe	SMOTE	0.89	0.89	0.89
	DWSMOTE (α= 18)	**0.90**	**0.90**	**0.90**
vehicle	SMOTE	0.95	0.95	0.95
	DWSMOTE (α= 4)	**0.96**	**0.96**	**0.96**

ตารางที่ 4-39 แสดงการเปรียบเทียบประสิทธิภาพโดยรวมของวิธีการ C4.5 (SMOTE) กับ DWSMOTE โดยใช้ค่าเอฟเมเซอร์ และค่านัยสำคัญทางสถิติที่ระดับ 0.1

C4.5 (SMOTE) VS DWSMOTE	ค่าเอฟเมเซอร์ (F-measure)	นัยสำคัญทางสถิติ
DWSMOTE ชนะ C4.5 (SMOTE)	13	5
C4.5 (SMOTE) ชนะ DWSMOTE	3	0

จากตารางที่ 4-39 แสดงให้เห็นว่าวิธีการ DWSMOTE ที่ได้ออกแบบ มีประสิทธิภาพโดยรวมดีกว่าวิธี C4.5 (SMOTE) จำนวน 13 ชุดข้อมูล จากชุดข้อมูลทดสอบทั้งหมด 16 ชุดข้อมูล และเมื่อพิจารณาร่วมกับค่านัยสำคัญทางสถิติที่ระดับ 0.1 จะพบว่า วิธีการที่นำเสนอดีกว่า C4.5 (SMOTE)

จากผลการทดลองเบื้องต้นเลือกค่าแอลฟา ที่ให้ผลในการจำแนกชุดข้อมูลได้ดีที่สุด เมื่อเทียบกับวิธีการ C4.5 โดยใช้ชุดข้อมูลชุดเดียวกันทำการทดลอง คือ ชุดข้อมูลที่ได้ทำการสุ่มเพิ่มตัวอย่างกลุ่มน้อยขึ้นเป็นจำนวน 1 เท่าของจำนวนตัวอย่างกลุ่มน้อยเดิม ซึ่งผลที่ได้ในการจำแนก แสดงค่าแอลฟาที่ระดับต่างกัน 5 ค่า มีดังนี้

ตารางที่ 4-40 ค่าความแม่นยำของข้อมูลในแต่ละกลุ่มข้อมูลที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า

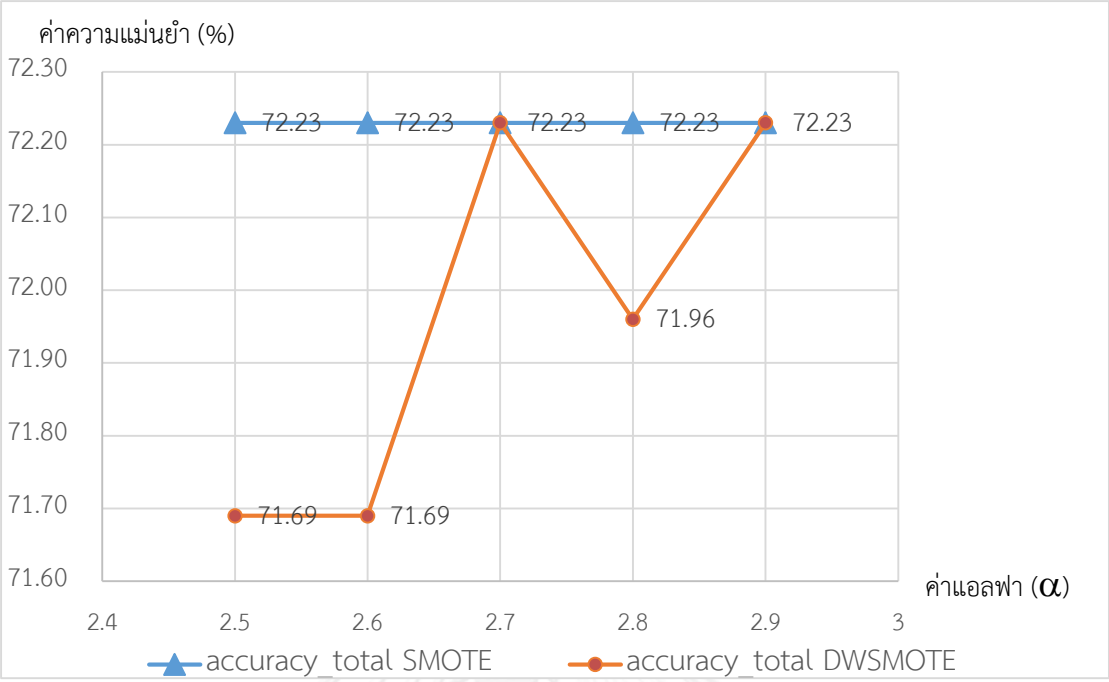
ชุดข้อมูล	ค่า α	ความแม่นยำทั้งหมด		ความแม่นยำกลุ่มมาก		ความแม่นยำกลุ่มน้อย	
		SMOTE	DWSMOTE	SMOTE	DWSMOTE	SMOTE	DWSMOTE
breast-cancer	2.5	**72.23**	71.69	**69.41**	68.82	**74.55**	74.05
	2.6	**72.23**	71.69	**69.41**	68.82	**74.55**	74.05
	2.7	**72.23**	**72.23**	**69.41**	**69.41**	**74.55**	**74.55**
	2.8	**72.23**	71.96	**69.41**	**69.41**	**74.55**	74.05
	2.9	**72.23**	72.23	69.41	**70.00**	**74.55**	74.05
ค่าเฉลี่ย	**2.7**	**72.23**	71.96	**69.41**	69.29	**74.55**	74.15
breast-cancer-w	2.8	**96.18**	95.86	**96.27**	95.65	**96.08**	**96.08**
	2.9	**96.18**	95.86	**96.27**	95.65	**96.08**	**96.08**
	3	**96.18**	95.97	**96.27**	95.86	**96.08**	**96.08**
	3.1	**96.18**	95.97	**96.27**	95.86	**96.08**	**96.08**
	3.2	**96.18**	95.97	**96.27**	95.86	**96.08**	**96.08**
ค่าเฉลี่ย	**3**	**96.18**	95.93	**96.27**	95.78	**96.08**	**96.08**
credit-g	3.8	71.08	**71.31**	**68.67**	67.83	73.14	**74.29**
	3.9	71.08	**71.62**	**68.67**	67.83	73.14	**74.86**
	4	71.08	**71.77**	**68.67**	67.83	73.14	**75.41**
	4.1	71.08	**71.31**	**68.67**	67.33	73.14	**74.71**
	4.2	71.08	**71.31**	**68.67**	67.50	73.14	**74.57**
ค่าเฉลี่ย	**4**	71.08	**71.46**	**68.67**	67.66	73.14	**74.77**
diabetes	6	**77.32**	76.55	**79.68**	78.76	**74.80**	74.20
	7	**77.32**	76.65	**79.68**	78.57	**74.80**	74.60
	8	**77.32**	77.13	**79.68**	78.95	74.80	**75.20**
	9	**77.32**	77.03	**79.68**	78.95	74.80	**75.00**
	10	**77.32**	77.03	**79.68**	78.95	74.80	**75.00**
ค่าเฉลี่ย	**8**	**77.32**	76.88	**79.68**	78.84	**74.80**	**74.80**

ชุดข้อมูล	ค่า α	ความแม่นยำทั้งหมด		ความแม่นยำกลุ่มมาก		ความแม่นยำกลุ่มน้อย	
		SMOTE	DWSMOTE	SMOTE	DWSMOTE	SMOTE	DWSMOTE
ecoli	5	90.03	90.03	78.57	78.57	92.70	92.70
	6	90.03	90.03	78.57	78.57	92.70	92.70
	7	90.03	**90.56**	78.57	78.57	92.70	**93.35**
	8	90.03	**90.56**	78.57	78.57	92.70	**93.35**
	9	90.03	**90.56**	78.57	78.57	92.70	**93.35**
ค่าเฉลี่ย	**7**	90.03	**90.35**	**78.57**	**78.57**	92.70	**93.09**
flags	3	88.10	**88.58**	64.17	64.17	92.61	**93.20**
	4	88.10	**90.51**	64.17	**70.00**	92.61	**94.35**
	5	88.10	**91.10**	64.17	**70.00**	92.61	**95.00**
	6	88.10	**91.10**	64.17	**70.00**	92.61	**95.00**
	7	88.10	**90.64**	64.17	**67.50**	92.61	**95.00**
ค่าเฉลี่ย	**5**	88.10	**90.39**	64.17	**68.33**	92.61	94.51
glasses	6	83.86	**85.56**	30.83	**33.33**	92.82	**94.32**
	7	83.86	**85.56**	30.83	**33.33**	92.82	**94.32**
	8	83.86	**86.01**	30.83	**36.67**	92.82	**94.32**
	9	83.86	**86.01**	30.83	**36.67**	92.82	**94.32**
	10	83.86	**86.01**	30.83	**36.67**	92.82	**94.32**
ค่าเฉลี่ย	**8**	83.86	**85.83**	30.83	**35.33**	92.82	**94.32**
haberman	4	72.73	72.73	65.26	65.26	78.12	78.12
	5	72.73	**73.23**	65.26	**66.43**	78.12	78.12
	6	72.73	**73.23**	65.26	**66.43**	78.12	78.12
	7	72.73	**73.23**	65.26	**66.43**	78.12	78.12
	8	72.73	**73.23**	65.26	**66.43**	78.12	78.12
ค่าเฉลี่ย	**6**	72.73	**73.13**	65.26	**66.20**	**78.12**	**78.12**

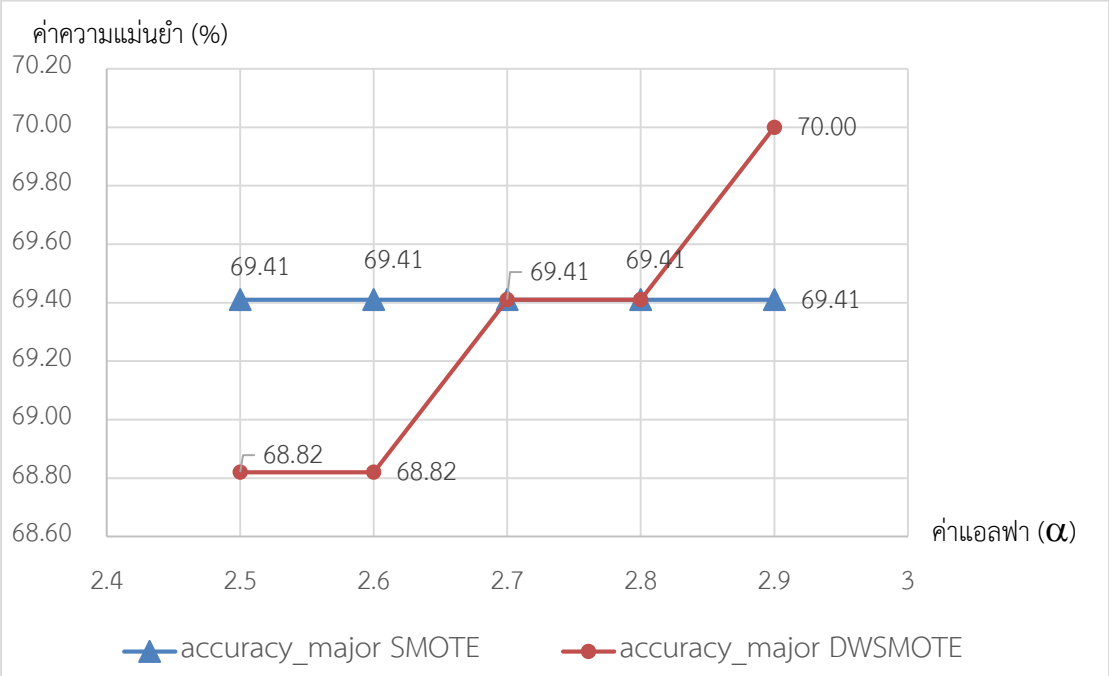
ชุดข้อมูล	ค่า α	ความแม่นยำทั้งหมด		ความแม่นยำกลุ่มมาก		ความแม่นยำกลุ่มน้อย	
		SMOTE	DWSMOTE	SMOTE	DWSMOTE	SMOTE	DWSMOTE
hepatitis	10	80.70	72.73	75.00	65.26	83.72	83.72
	11	80.70	**82.89**	75.00	**81.43**	83.72	83.72
	12	80.70	**83.39**	75.00	**81.43**	83.72	84.49
	13	80.70	**82.84**	75.00	**81.43**	83.72	83.65
	14	80.70	**81.84**	75.00	**81.43**	83.72	82.12
ค่าเฉลี่ย	**12**	80.70	**80.74**	75.00	**78.20**	**83.72**	83.54
ionosphere	4	88.49	**89.28**	90.54	**91.25**	86.19	**87.08**
	5	88.49	**89.91**	90.54	**91.25**	86.19	**88.40**
	6	88.49	**90.56**	90.54	**90.85**	86.19	**90.22**
	7	88.49	**90.56**	90.54	**90.85**	86.19	**90.22**
	8	88.49	**90.78**	90.54	**90.46**	86.19	**91.13**
ค่าเฉลี่ย	**6**	88.49	**90.22**	90.54	**90.93**	86.19	**89.41**
liver	2.6	70.41	**71.84**	84.83	**85.17**	49.50	**52.50**
	2.7	70.41	**73.27**	84.83	84.83	49.50	**56.50**
	2.8	70.41	**73.47**	84.83	**85.17**	49.50	**56.50**
	2.9	70.41	**72.86**	**84.83**	84.48	49.50	**56.00**
	3	70.41	**73.06**	**84.83**	83.45	49.50	**58.00**
ค่าเฉลี่ย	**2.8**	70.41	**72.90**	**84.83**	84.62	49.50	**55.90**
post-operative	6	**68.80**	67.73	50.50	50.50	**81.90**	80.00
	7	**68.80**	68.56	50.50	50.50	**81.90**	81.43
	8	68.80	**69.39**	50.50	50.50	81.90	**82.86**
	9	68.80	**69.39**	50.50	50.50	81.90	**82.86**
	10	**68.80**	68.48	50.50	48.50	81.90	**82.86**
ค่าเฉลี่ย	**8**	**68.80**	68.71	**50.50**	50.10	81.90	**82.00**

ชุดข้อมูล	ค่า α	ความแม่นยำทั้งหมด		ความแม่นยำกลุ่มมาก		ความแม่นยำกลุ่มน้อย	
		SMOTE	DWSMOTE	SMOTE	DWSMOTE	SMOTE	DWSMOTE
tic-tac-toe	16	88.76	**88.84**	91.27	**91.42**	86.10	**86.11**
	17	88.76	**88.84**	91.27	**91.42**	86.10	**86.11**
	18	88.76	**89.54**	91.27	**91.87**	86.10	**87.07**
	19	88.76	**89.15**	91.27	**91.42**	86.10	**86.75**
	20	88.76	**89.15**	91.27	**91.42**	86.10	**86.75**
ค่าเฉลี่ย	**18**	88.76	**89.10**	91.27	**91.51**	86.10	**86.56**
splice-ei	2.9	96.74	**96.79**	96.48	96.48	96.91	**96.99**
	3	96.74	**96.79**	96.48	96.48	96.91	**96.99**
	3.1	96.74	**96.79**	96.48	**96.54**	96.91	**96.95**
	3.2	96.74	**96.77**	96.48	96.48	96.91	**96.95**
	3.3	96.74	**96.77**	96.48	96.48	96.91	**96.95**
ค่าเฉลี่ย	**3.1**	96.74	**96.78**	96.48	**96.49**	96.91	**96.97**
splice-ie	2.1	94.95	94.95	94.86	94.86	95.00	95.00
	2.2	94.95	**94.97**	94.86	**94.92**	95.00	95.00
	2.3	94.95	**95.00**	94.86	**94.92**	95.00	**95.05**
	2.4	94.95	94.95	94.86	94.86	95.00	95.00
	2.5	**94.95**	94.87	**94.86**	94.66	95.00	95.00
ค่าเฉลี่ย	**2.3**	**94.95**	**94.95**	**94.86**	94.84	95.00	**95.01**
vehicle	3.8	95.31	**95.50**	94.98	**95.73**	**95.52**	95.36
	3.9	95.31	**95.60**	94.98	**95.99**	**95.52**	95.36
	4	95.31	**95.60**	94.98	**95.99**	**95.52**	95.36
	4.1	95.31	**95.50**	94.98	**95.74**	**95.52**	95.36
	4.2	95.31	**95.50**	94.98	**95.74**	**95.52**	95.36
ค่าเฉลี่ย	**4**	95.31	**95.54**	94.98	**95.84**	**95.52**	95.36

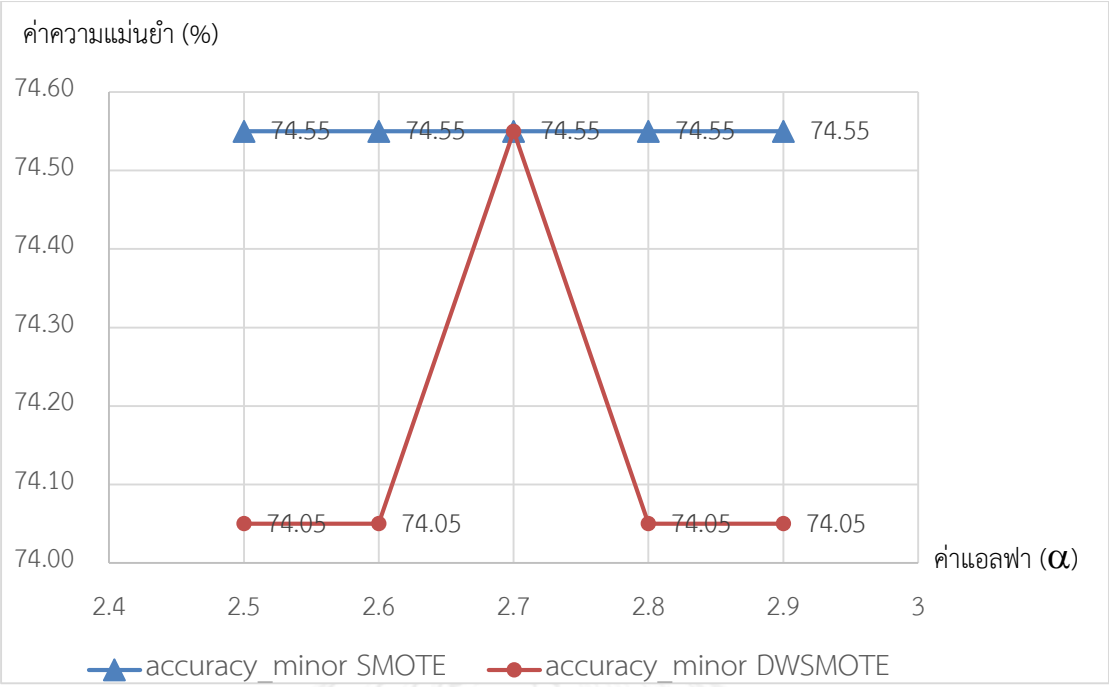
จากตารางข้างต้นแสดงค่าแอลฟาที่แตกต่างกัน 5 ค่า ที่ใช้ในทดลองกับแต่ละชุดข้อมูล ซึ่งแต่ละชุดข้อมูลจะมีค่าแอลฟาที่ทำให้ผลในการจำแนกแต่ละกลุ่มตัวอย่างมีค่าใกล้เคียงหรือดีกว่าวิธีการพื้นฐาน C4.5 (SMOTE) เมื่อเทียบกับวิธีการ DWSMOTE โดยสามารถแสดงผลการจำแนกแต่ละชุดข้อมูล ดังนี้



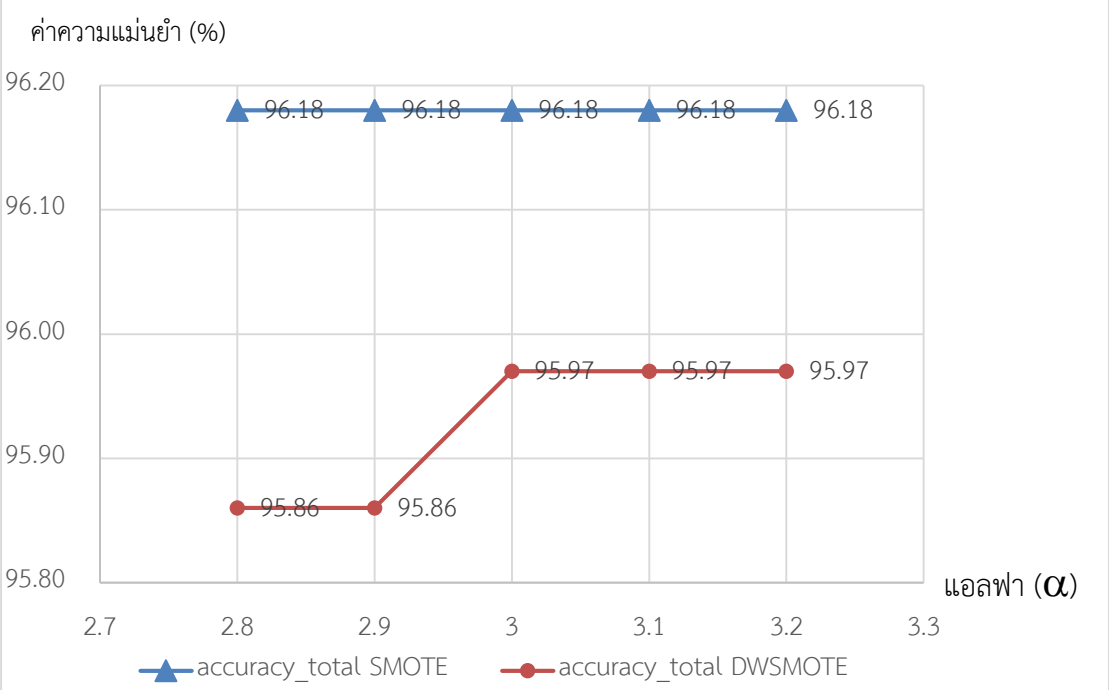
ภาพที่ 4-2 ค่าความแม่นยำของข้อมูลในตัวอย่างทั้งหมดที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Breast-cancer



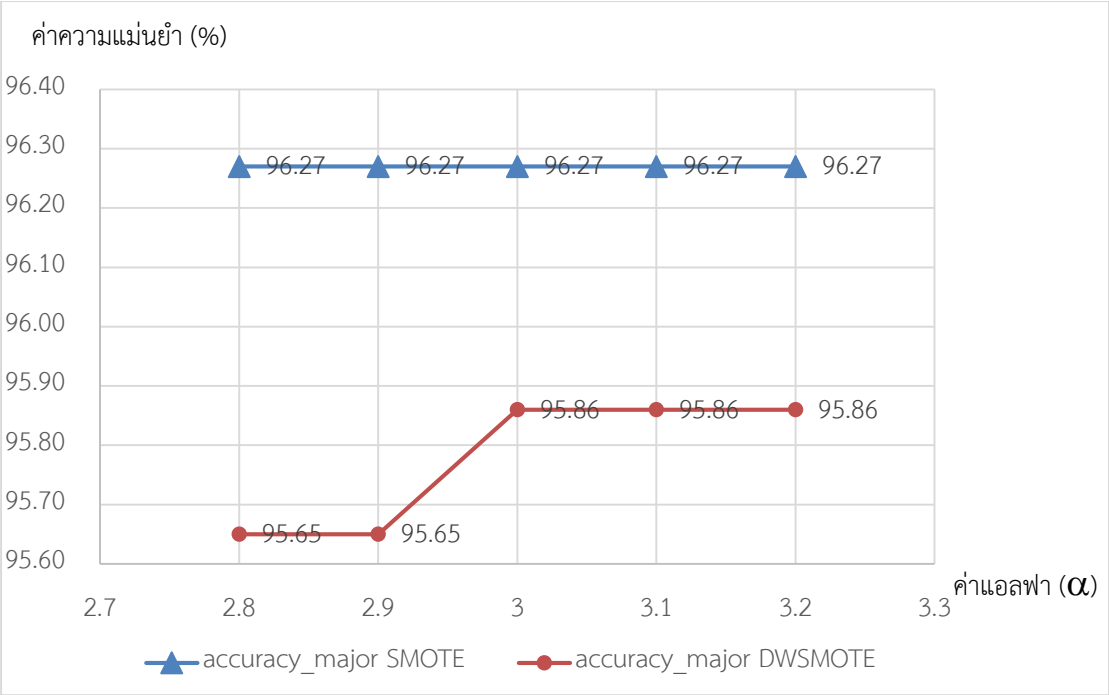
ภาพที่ 4-3 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มมากที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Breast-cancer



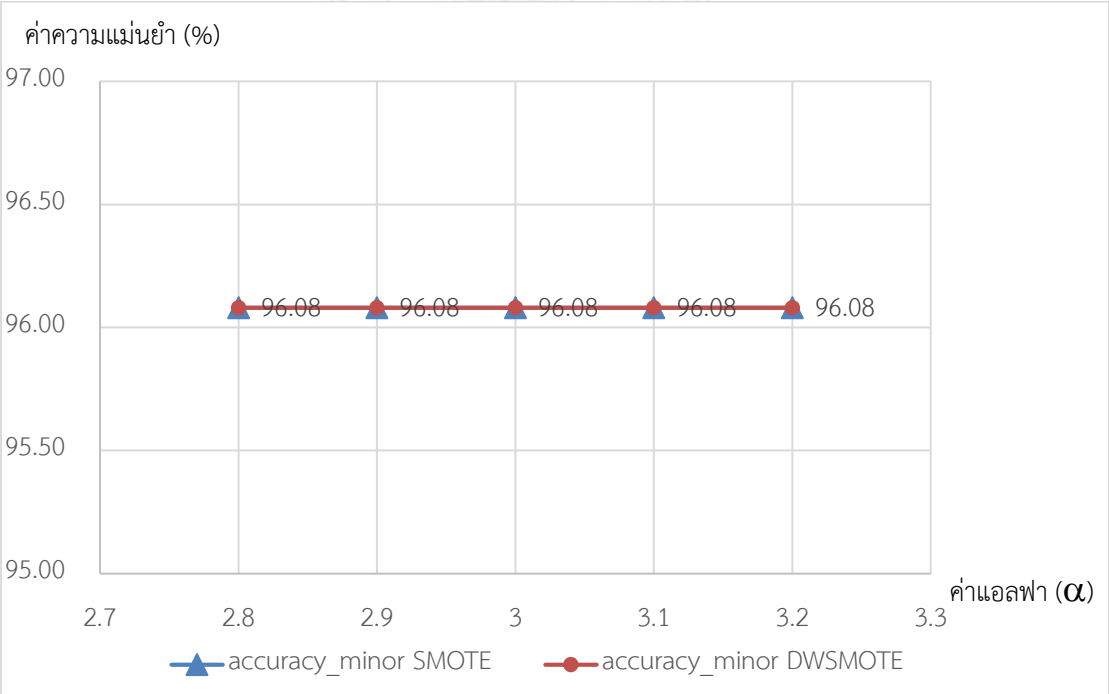
ภาพที่ 4-4 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มน้อยที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Breast-cancer



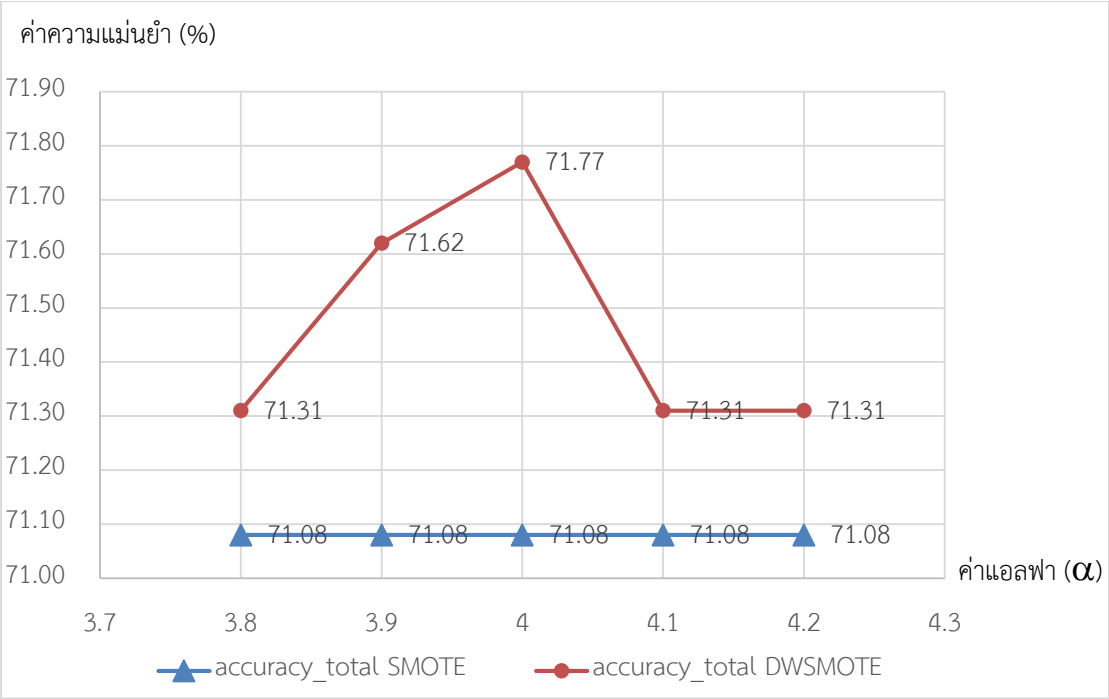
ภาพที่ 4-5 ค่าความแม่นยำของข้อมูลในตัวอย่างทั้งหมดที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Breast-cancer-w



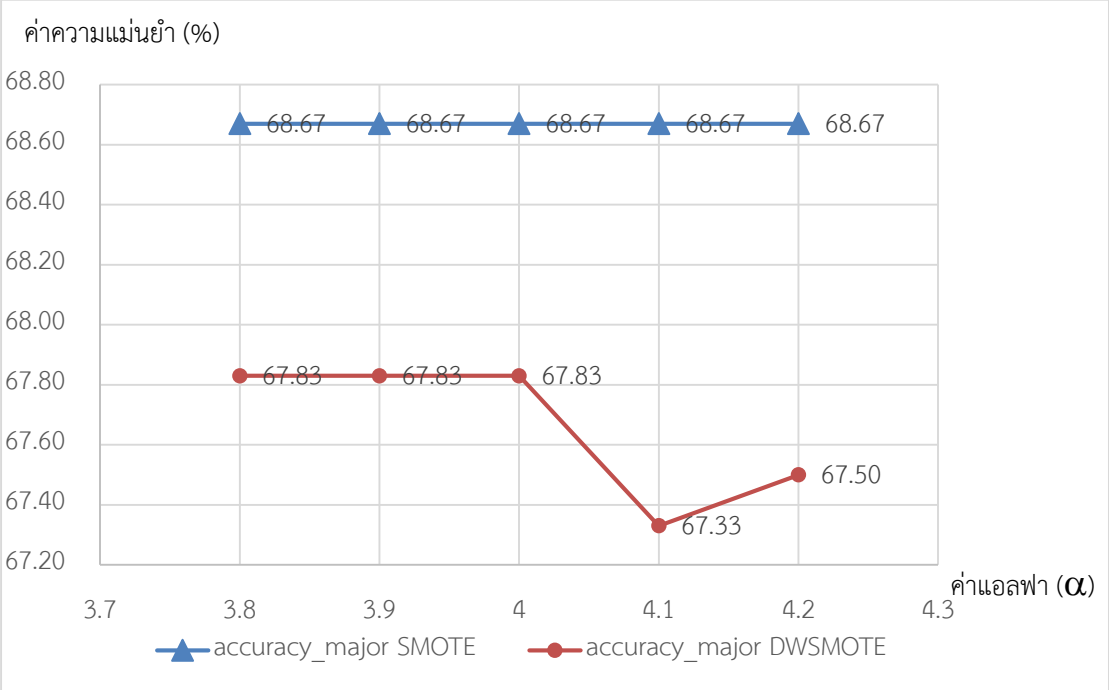
ภาพที่ 4-6 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มมากที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Breast-cancer-w



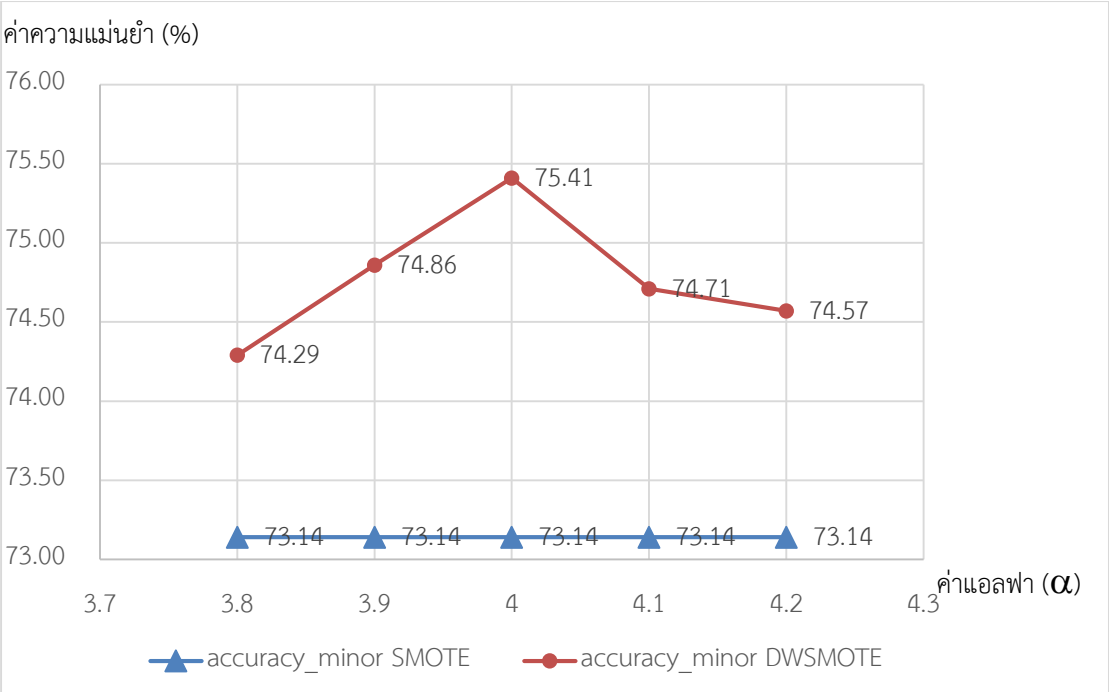
ภาพที่ 4-7 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มน้อยที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Breast-cancer-w



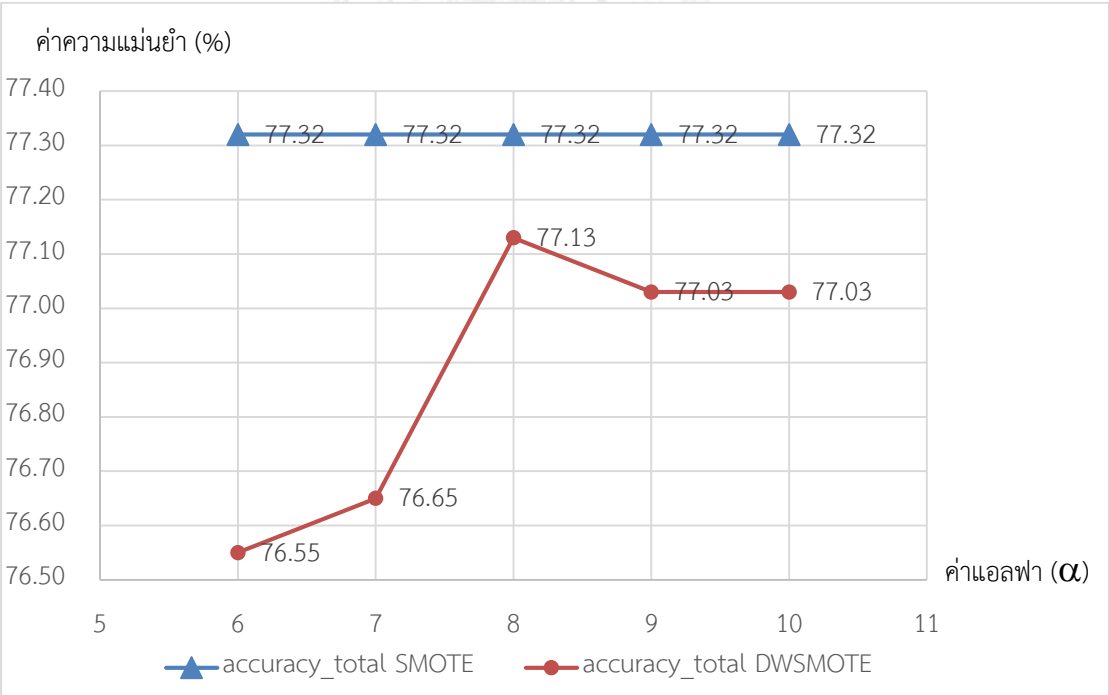
ภาพที่ 4-8 ค่าความแม่นยำของข้อมูลในตัวอย่างทั้งหมดที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Credit-g



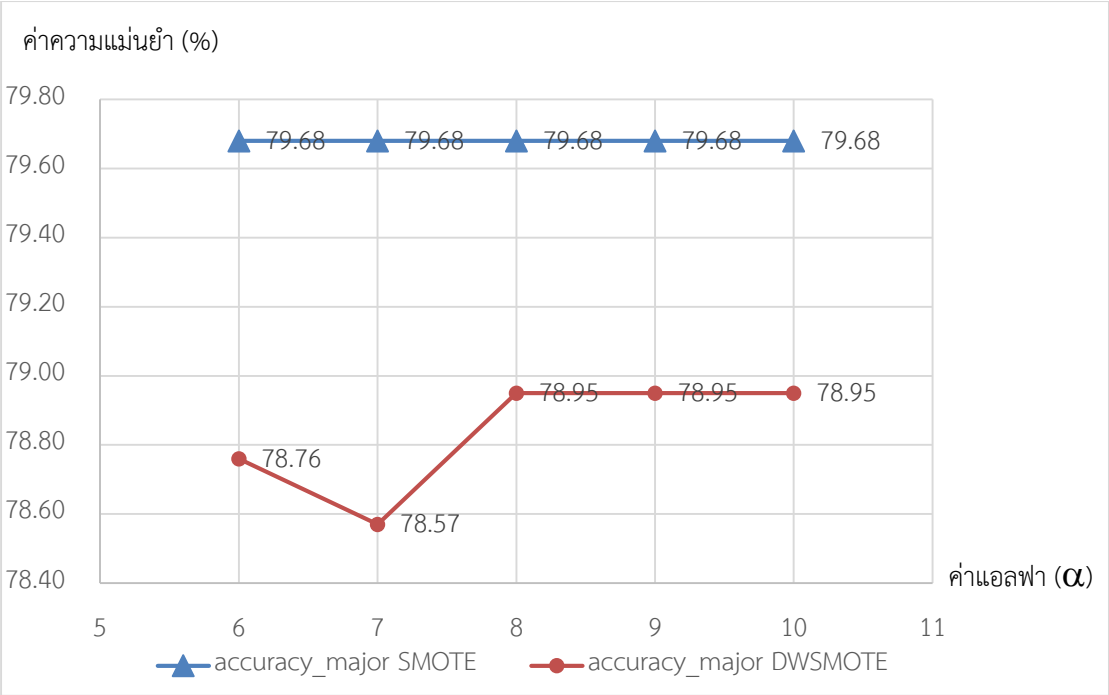
ภาพที่ 4-9 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มมากที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Credit-g



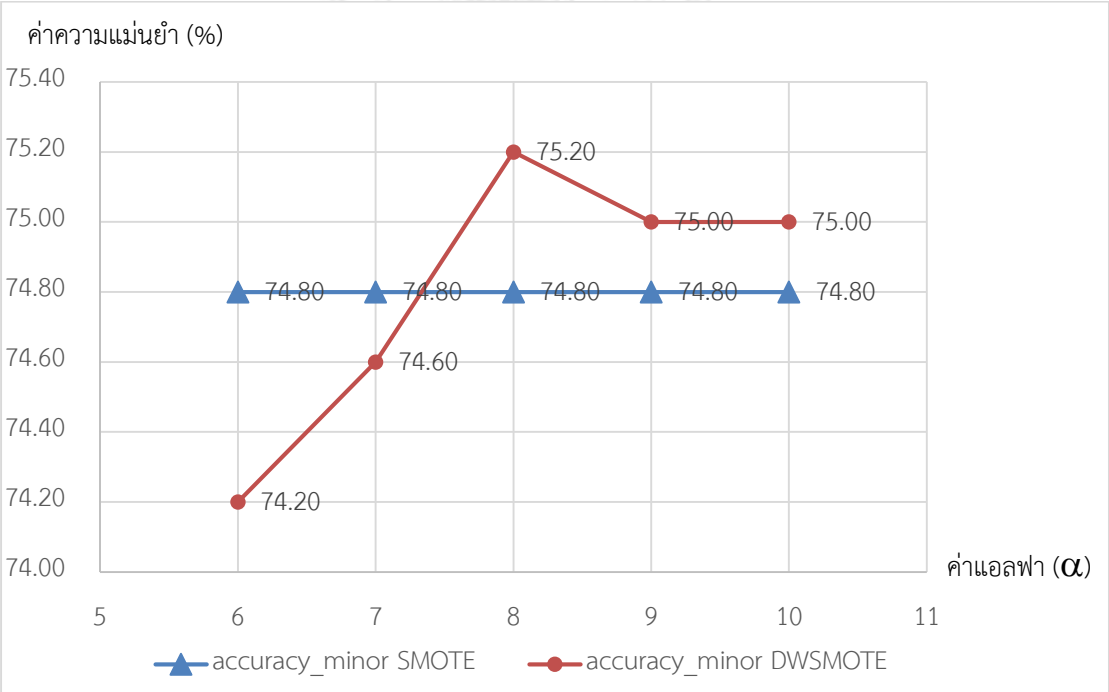
ภาพที่ 4-10 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มน้อยที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Credit-g



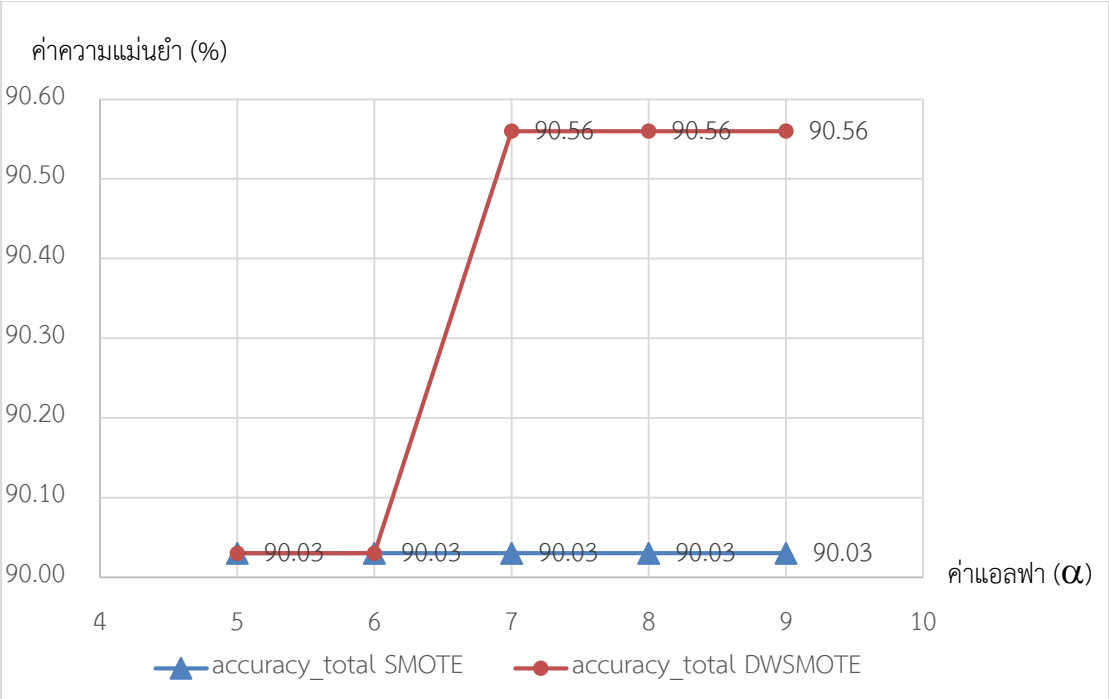
ภาพที่ 4-11 ค่าความแม่นยำของข้อมูลในตัวอย่างทั้งหมดที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Diabetes



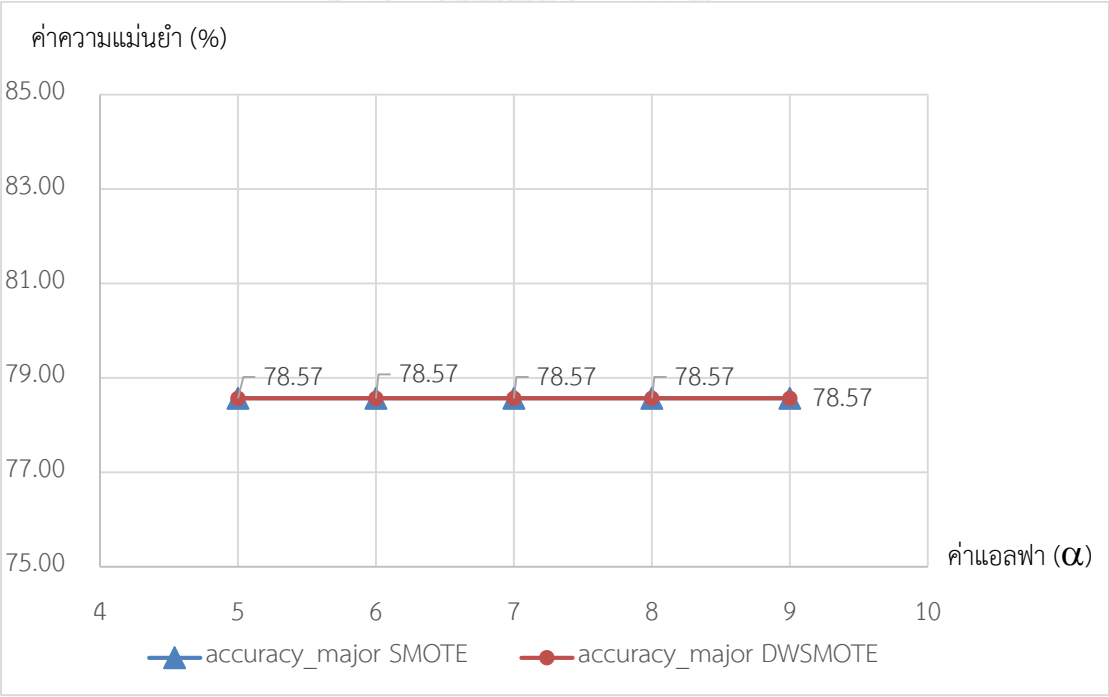
ภาพที่ 4-12 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มมากที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Diabetes



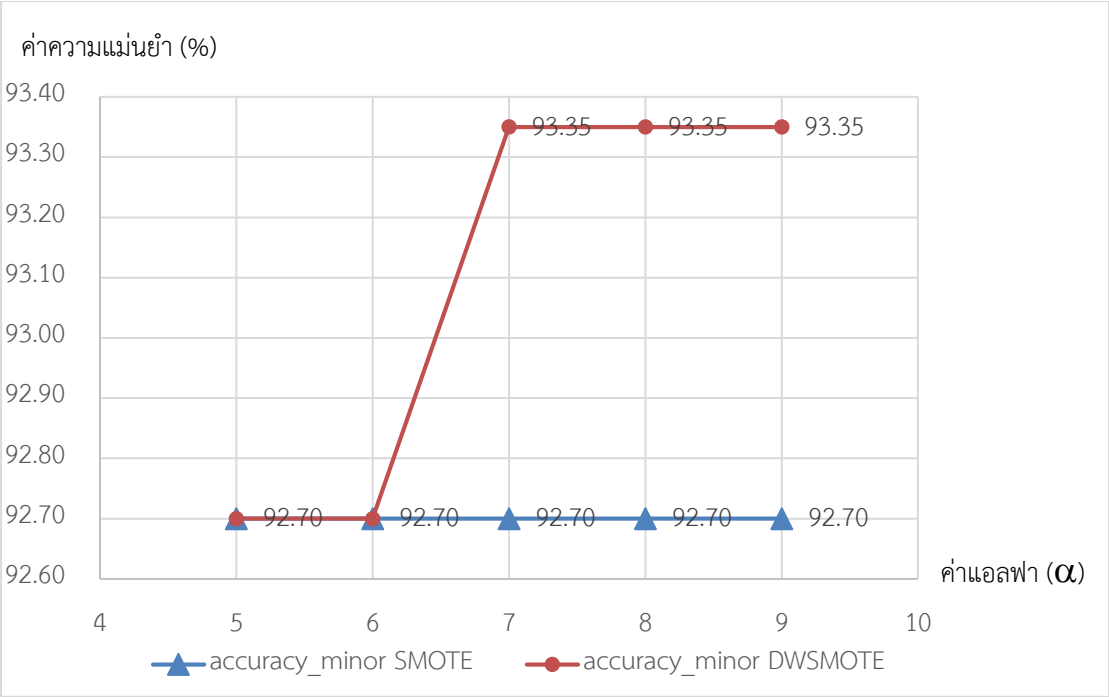
ภาพที่ 4-13 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มน้อยที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Diabetes



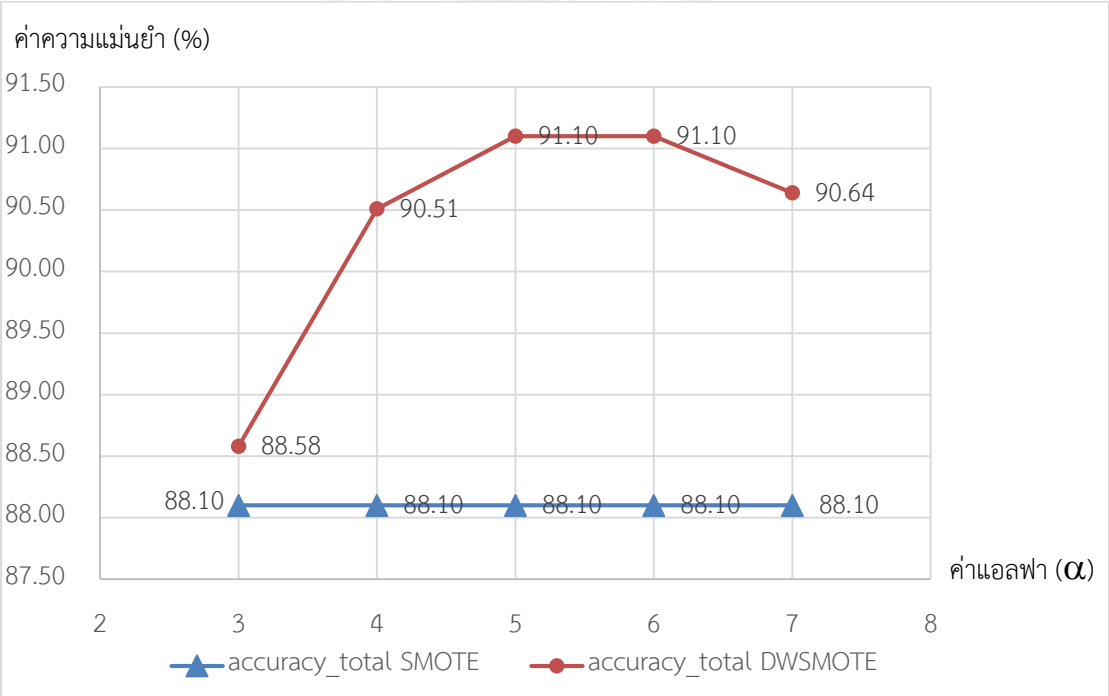
ภาพที่ 4-14 ค่าความแม่นยำของข้อมูลในตัวอย่างทั้งหมดที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Ecoli



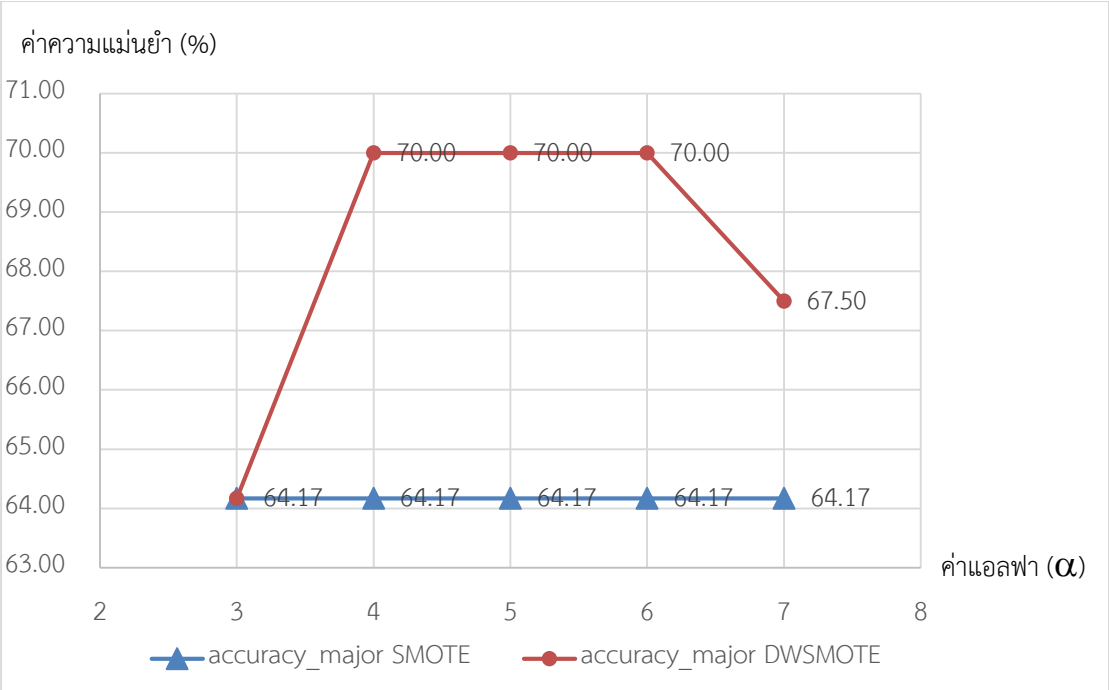
ภาพที่ 4-15 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มมากที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Ecoli



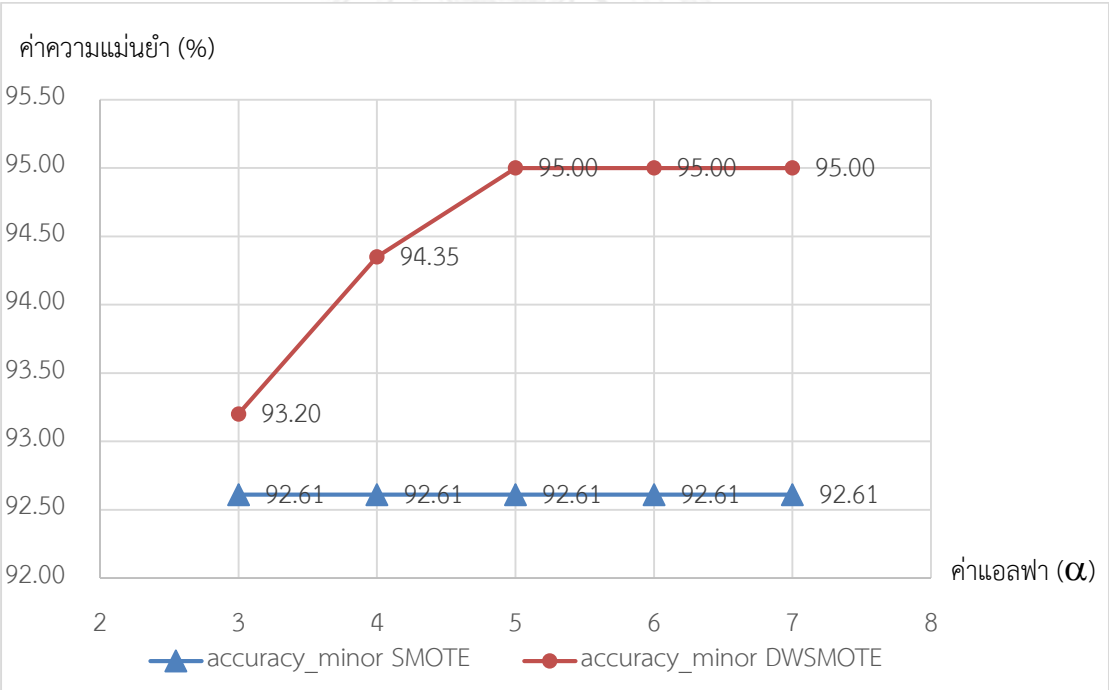
ภาพที่ 4-16 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มน้อยที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Ecoli



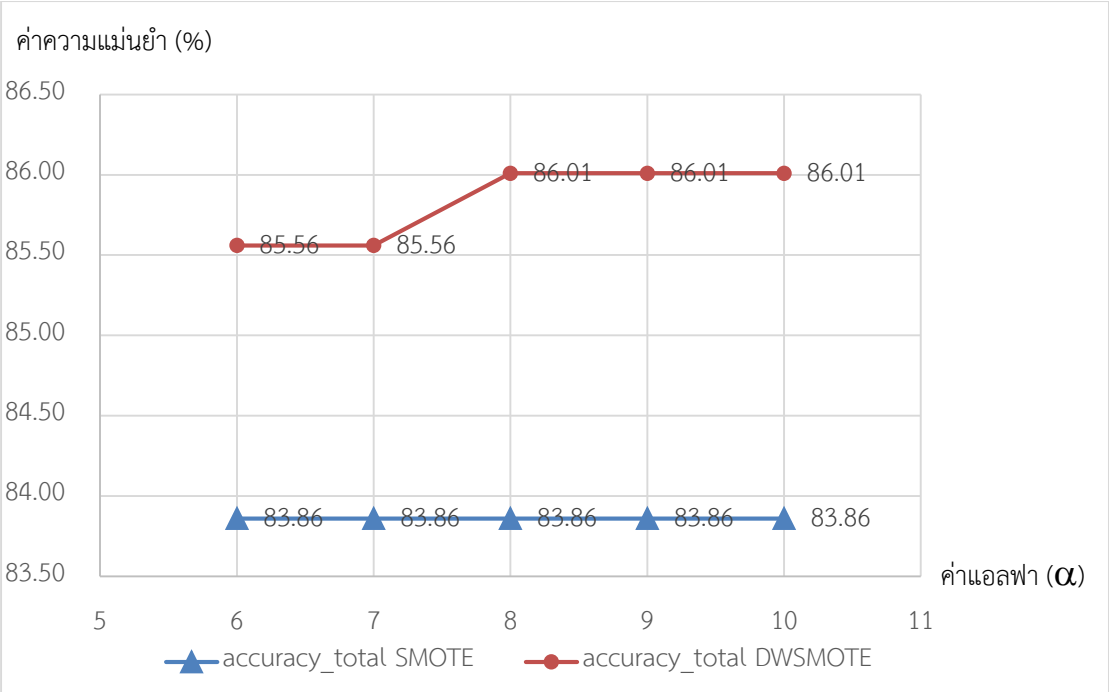
ภาพที่ 4-17 ค่าความแม่นยำของข้อมูลในตัวอย่างทั้งหมดที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Flags



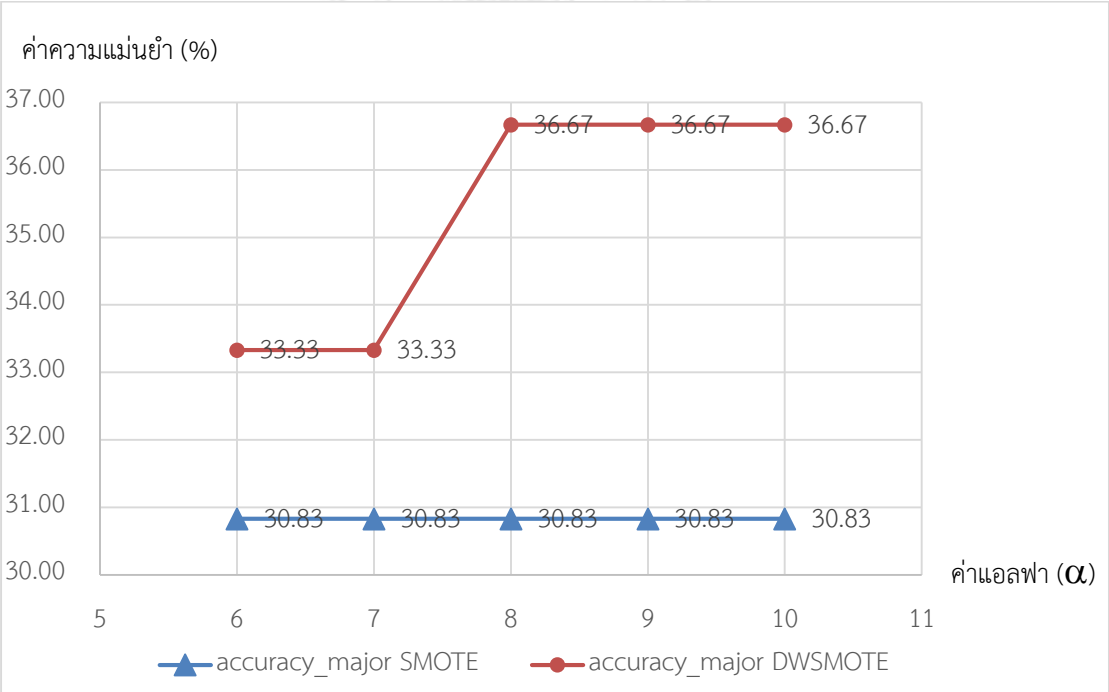
ภาพที่ 4-18 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มมากที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Flags



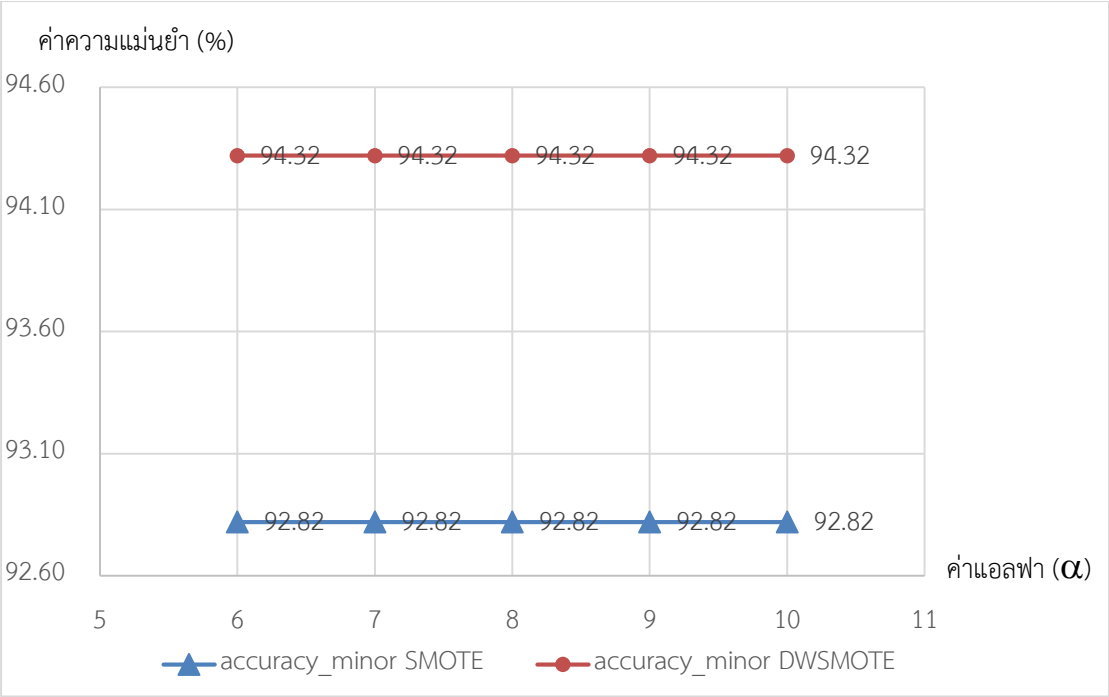
ภาพที่ 4-19 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มน้อยที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Flags



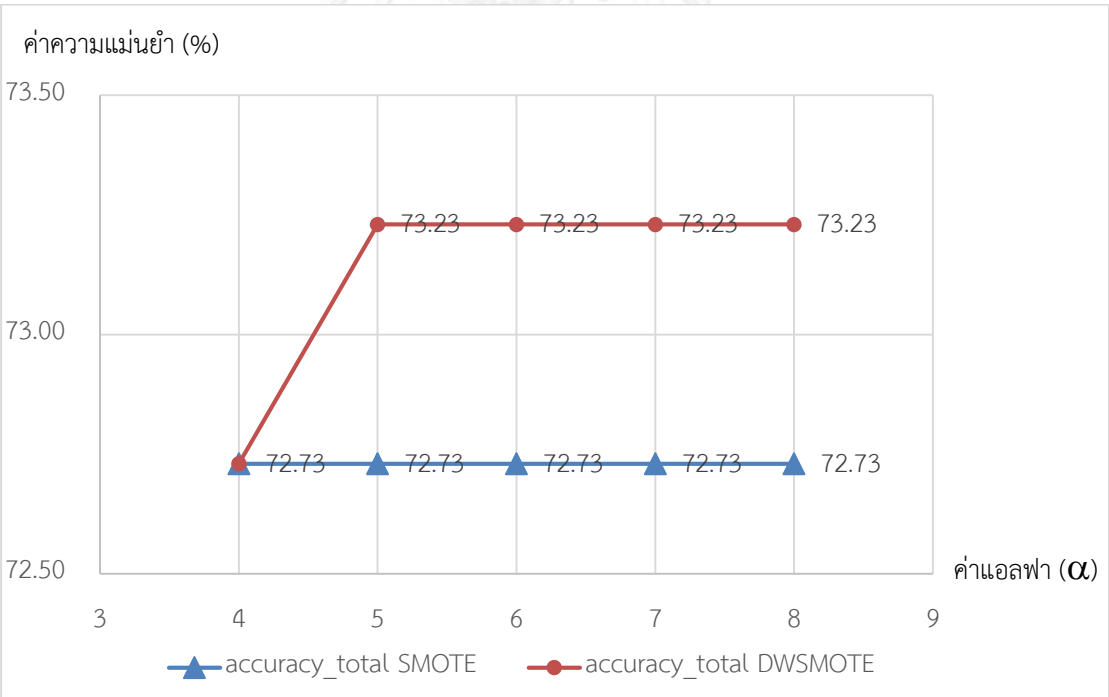
ภาพที่ 4-20 ค่าความแม่นยำของข้อมูลในตัวอย่างทั้งหมดที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Glasses



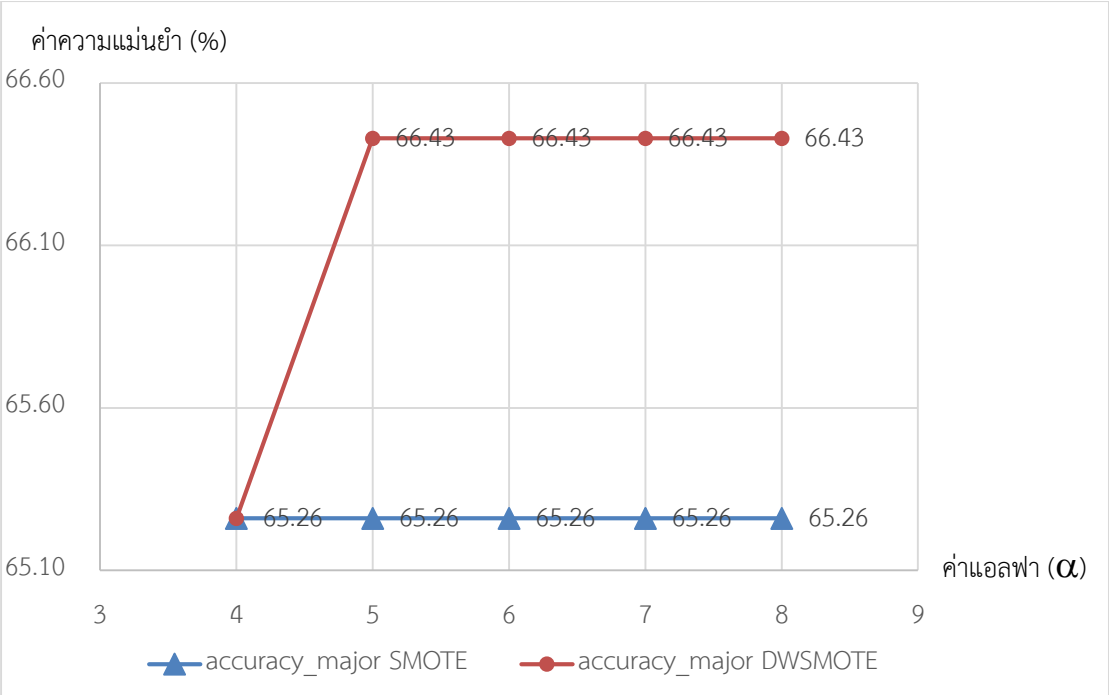
ภาพที่ 4-21 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มมากที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Glasses



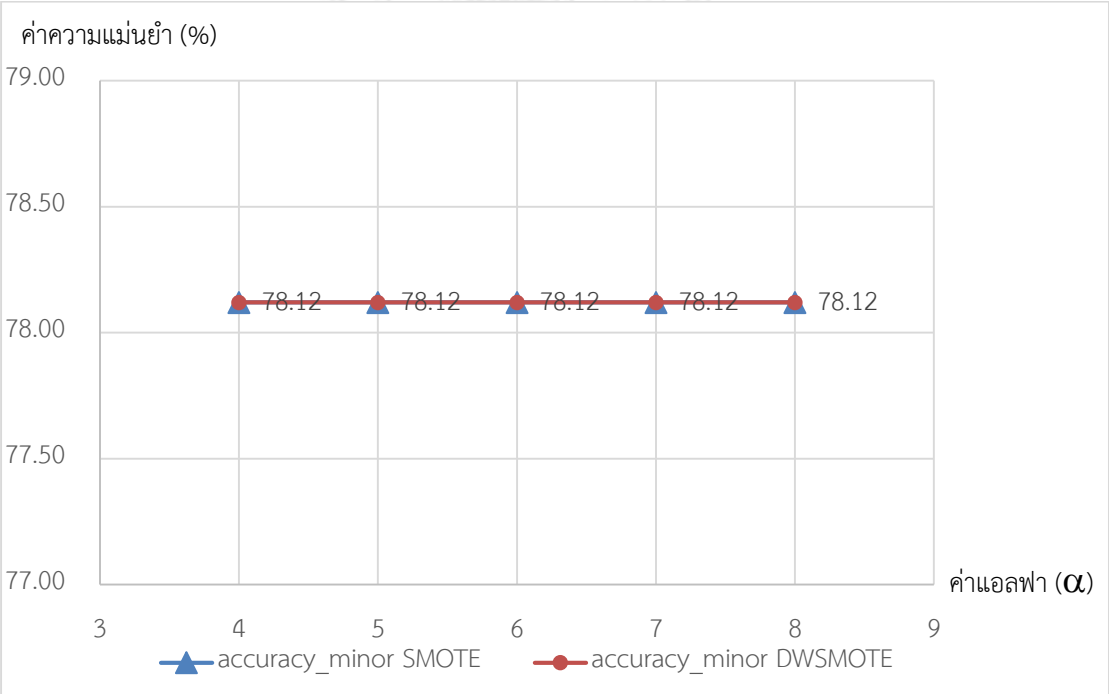
ภาพที่ 4-22 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มน้อยที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Glasses



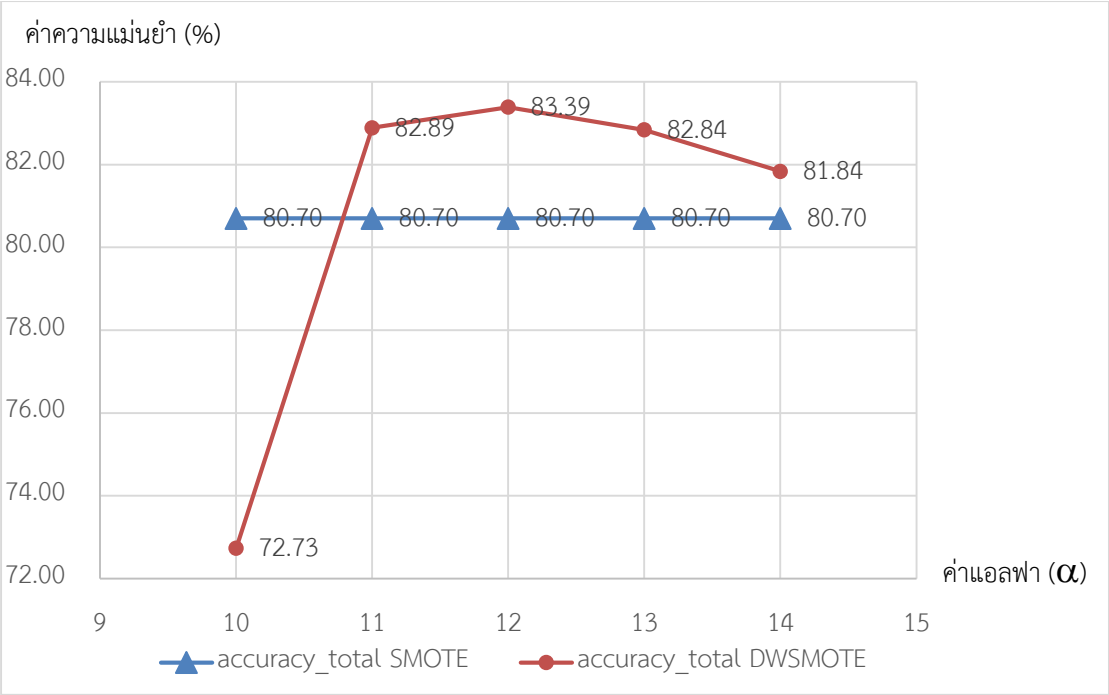
ภาพที่ 4-23 ค่าความแม่นยำของข้อมูลในตัวอย่างทั้งหมดที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Haberman



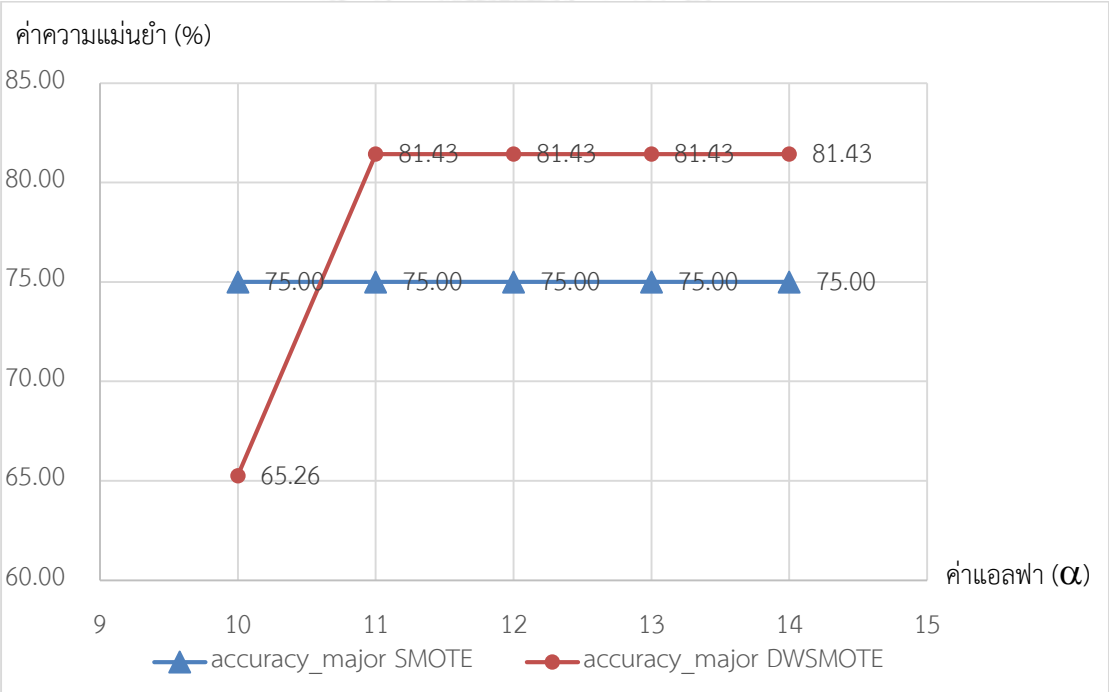
ภาพที่ 4-24 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มมากที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Haberman



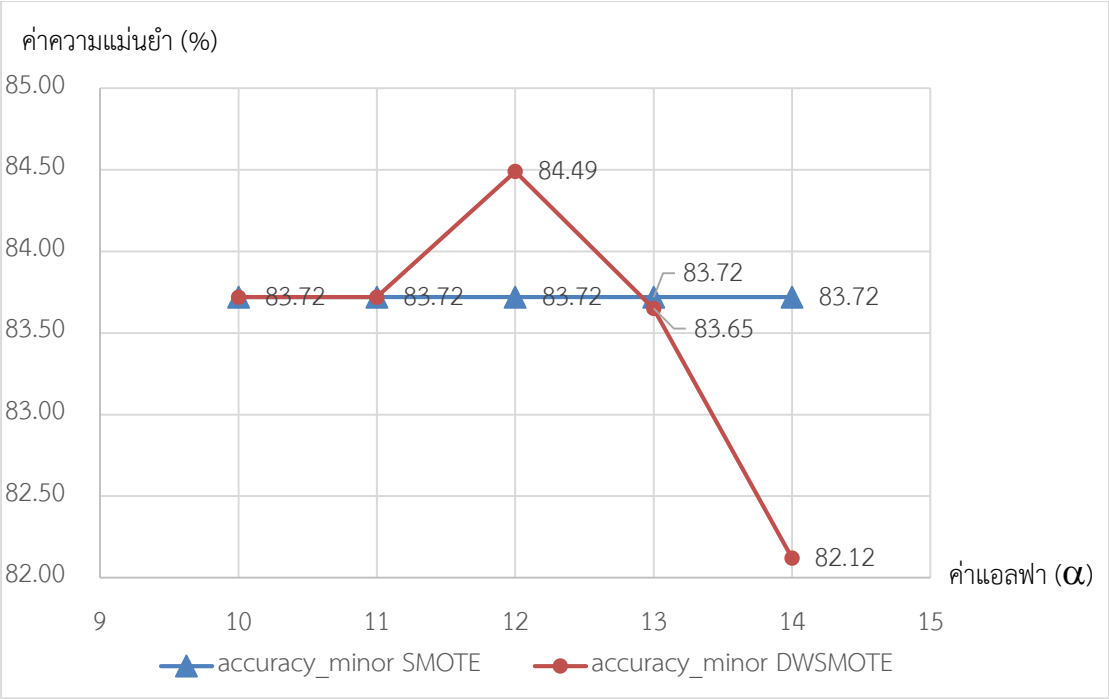
ภาพที่ 4-25 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มน้อยที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Haberman



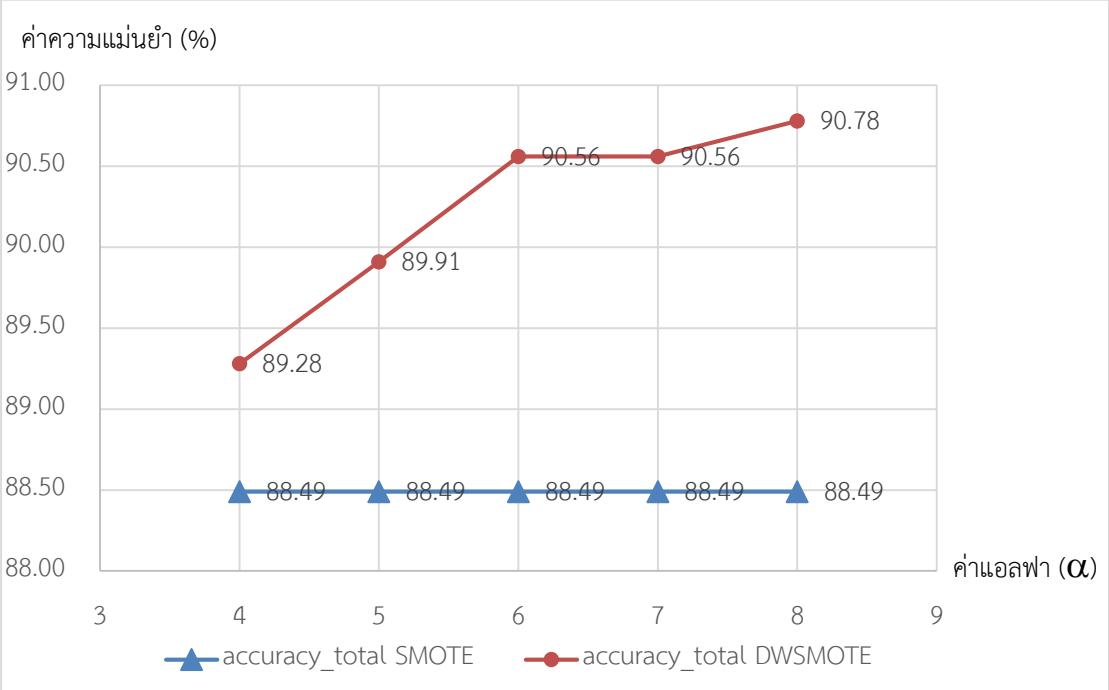
ภาพที่ 4-26 ค่าความแม่นยำของข้อมูลในตัวอย่างทั้งหมดที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Hepatitis



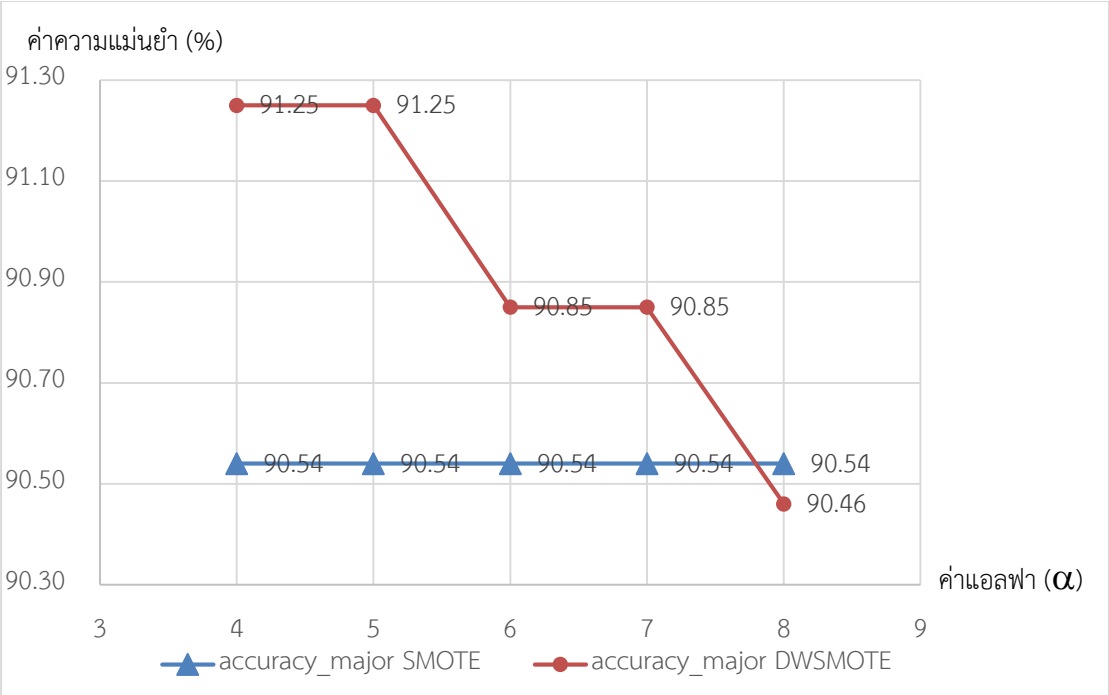
ภาพที่ 4-27 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มมากที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Hepatitis



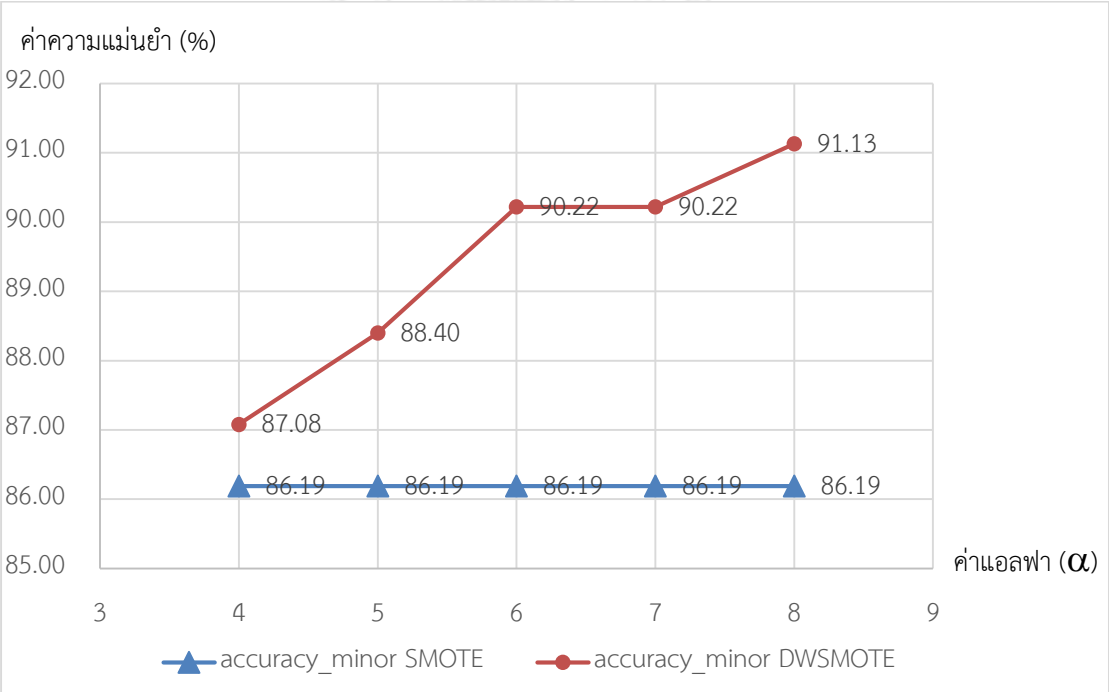
ภาพที่ 4-28 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มน้อยที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Hepatitis



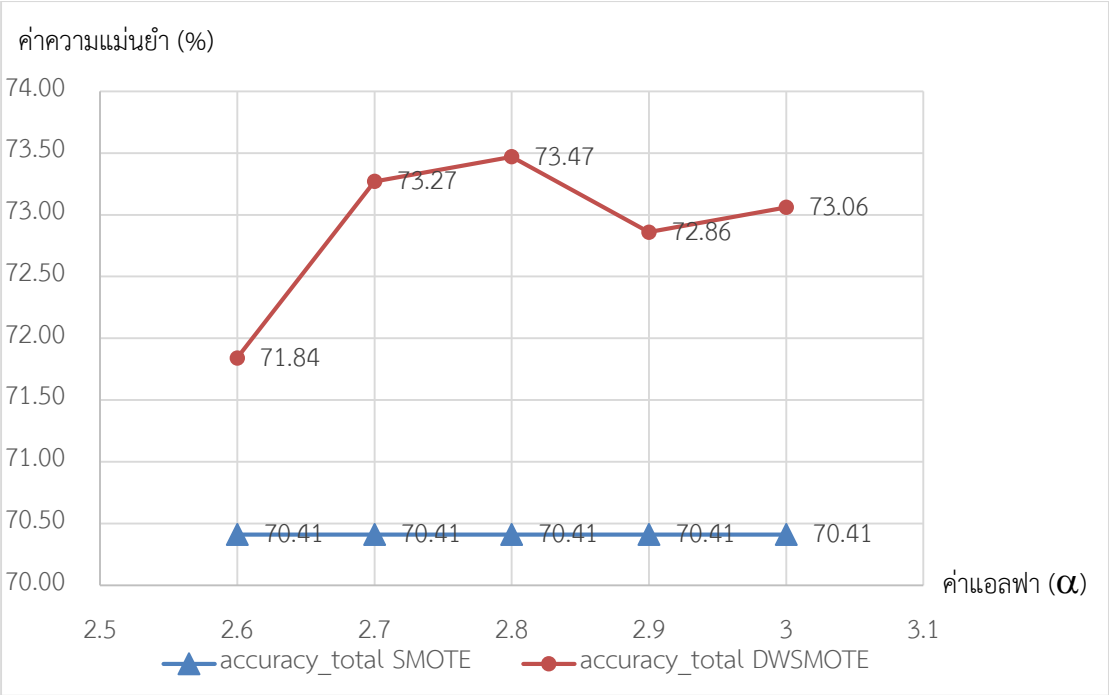
ภาพที่ 4-29 ค่าความแม่นยำของข้อมูลในตัวอย่างทั้งหมดที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Ionosphere



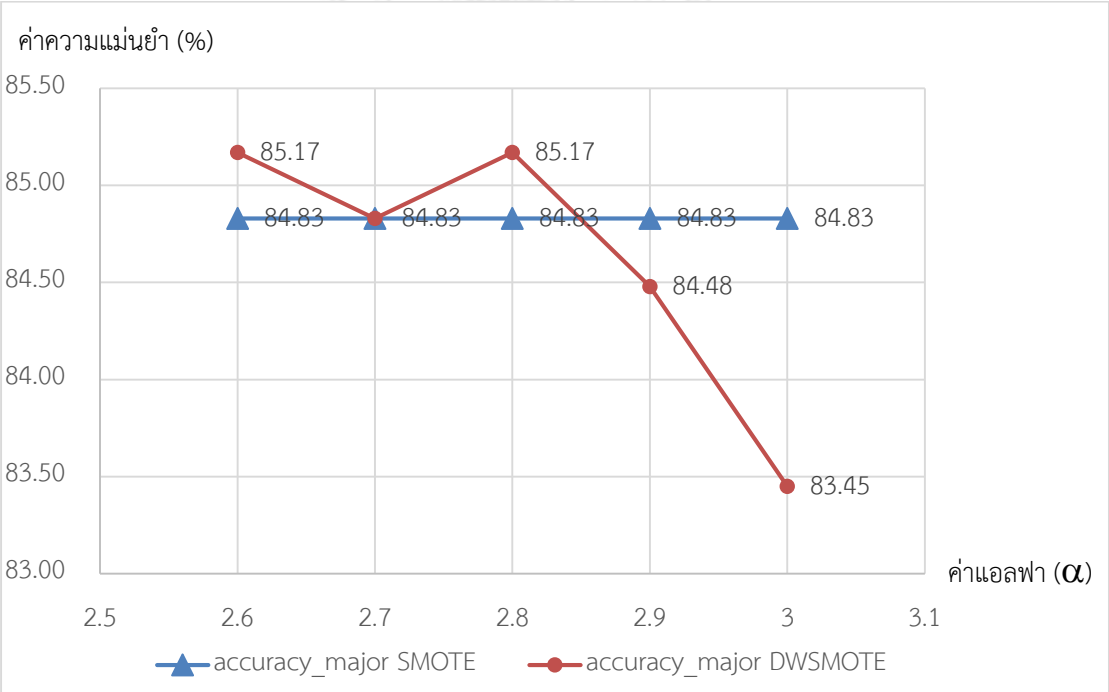
ภาพที่ 4-30 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มมากที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Ionosphere



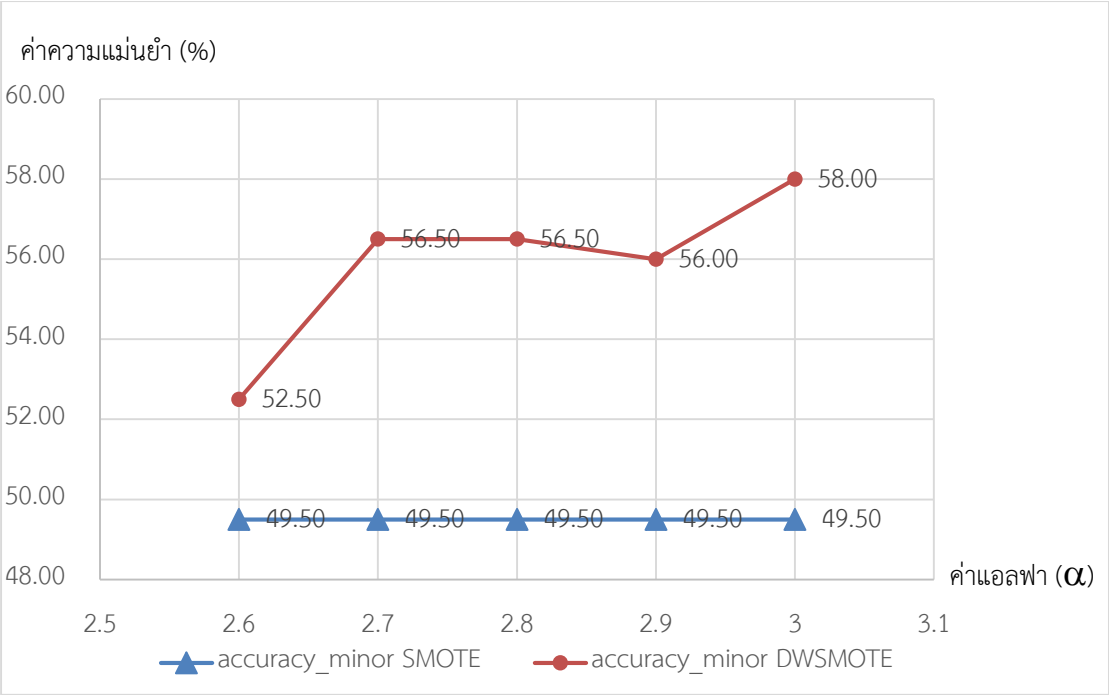
ภาพที่ 4-31 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มน้อยที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Ionosphere



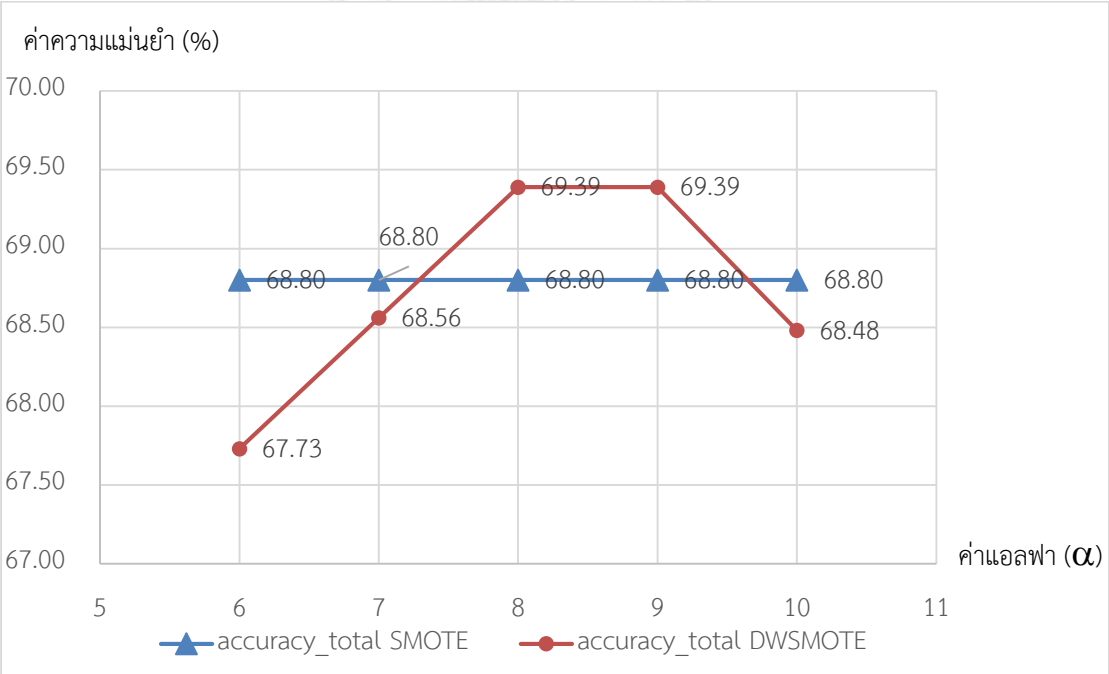
ภาพที่ 4-32 ค่าความแม่นยำของข้อมูลในตัวอย่างทั้งหมดที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Liver



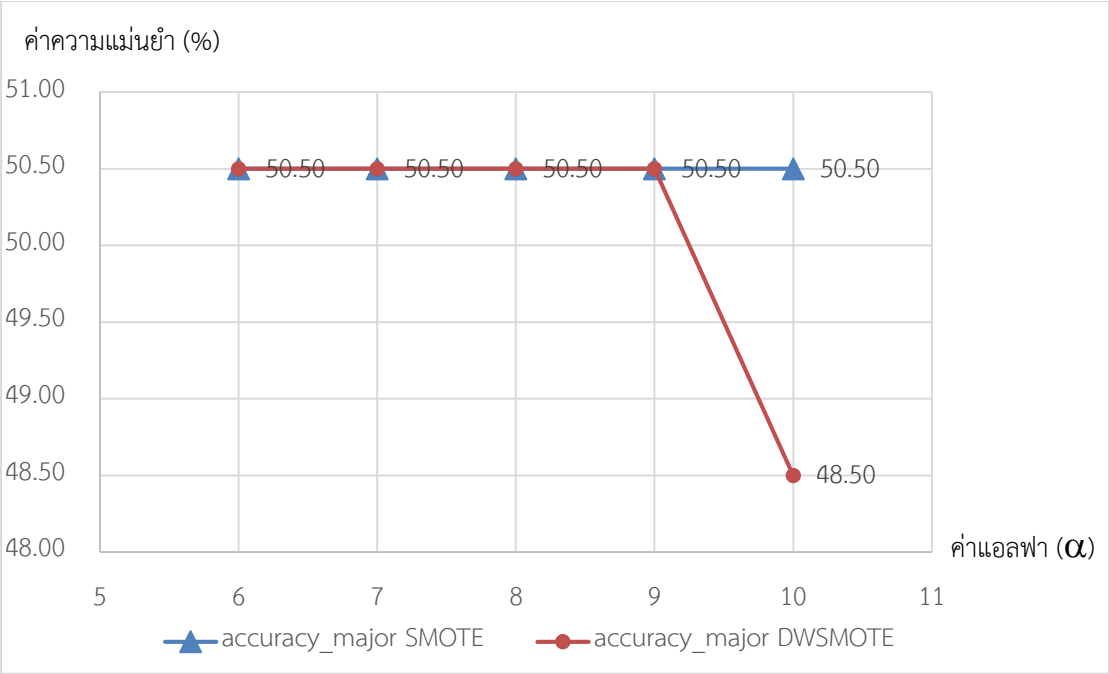
ภาพที่ 4-33 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มมากที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Liver



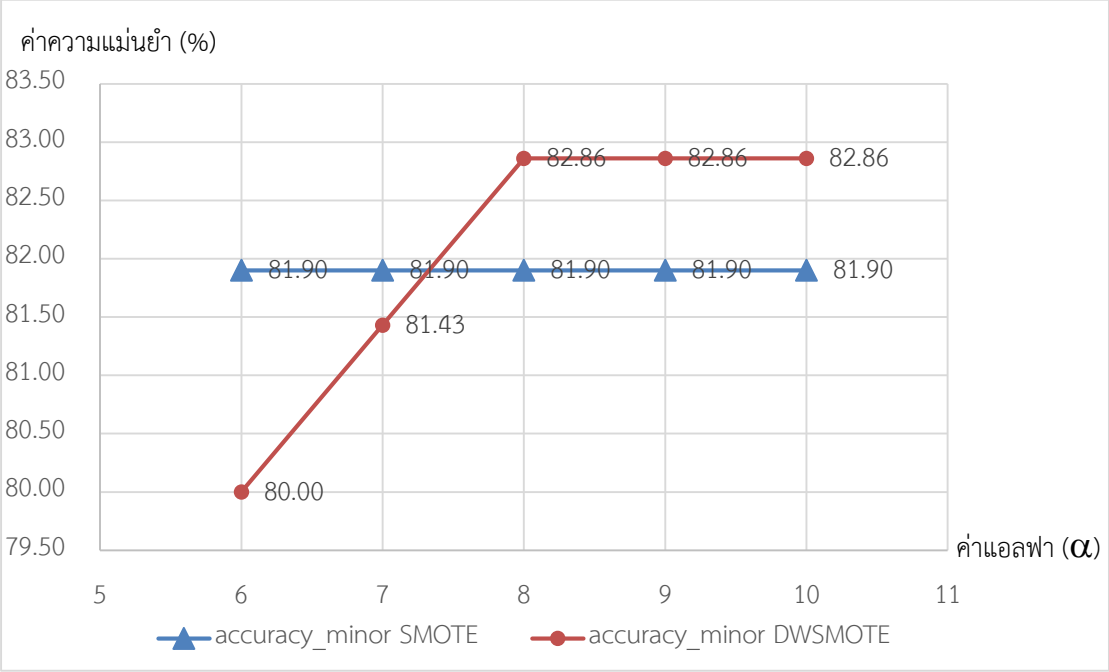
ภาพที่ 4-34 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มน้อยที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Liver



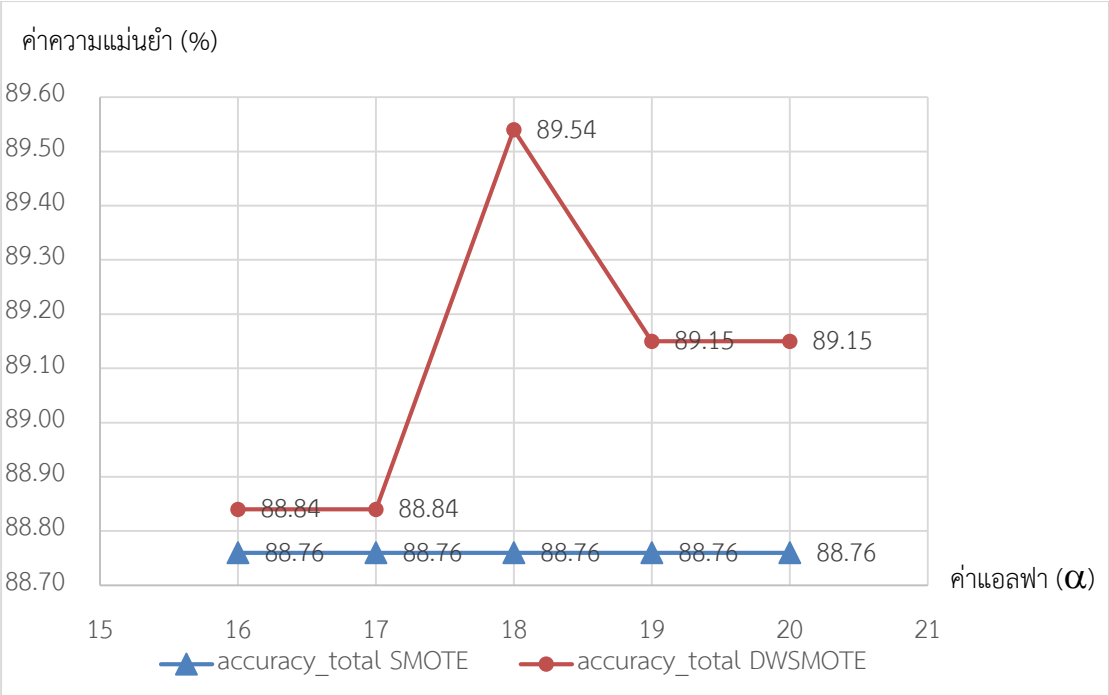
ภาพที่ 4-35 ค่าความแม่นยำของข้อมูลในตัวอย่างทั้งหมดที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Post-operative



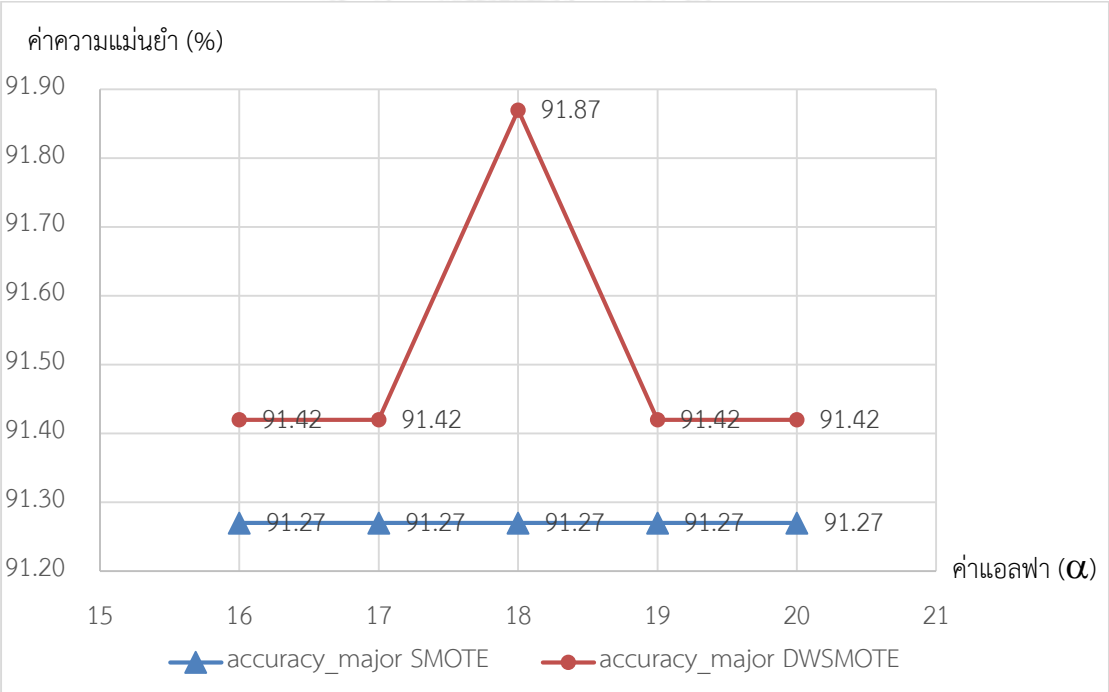
ภาพที่ 4-36 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มมากที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Post-operative



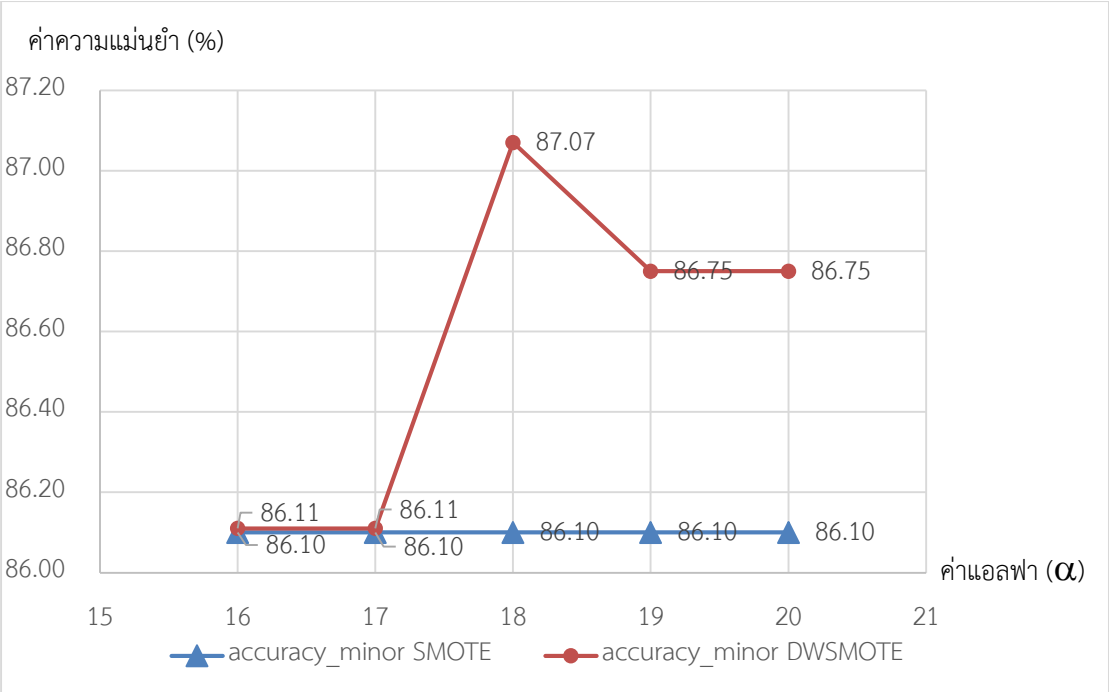
ภาพที่ 4-37 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มน้อยที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Post-operative



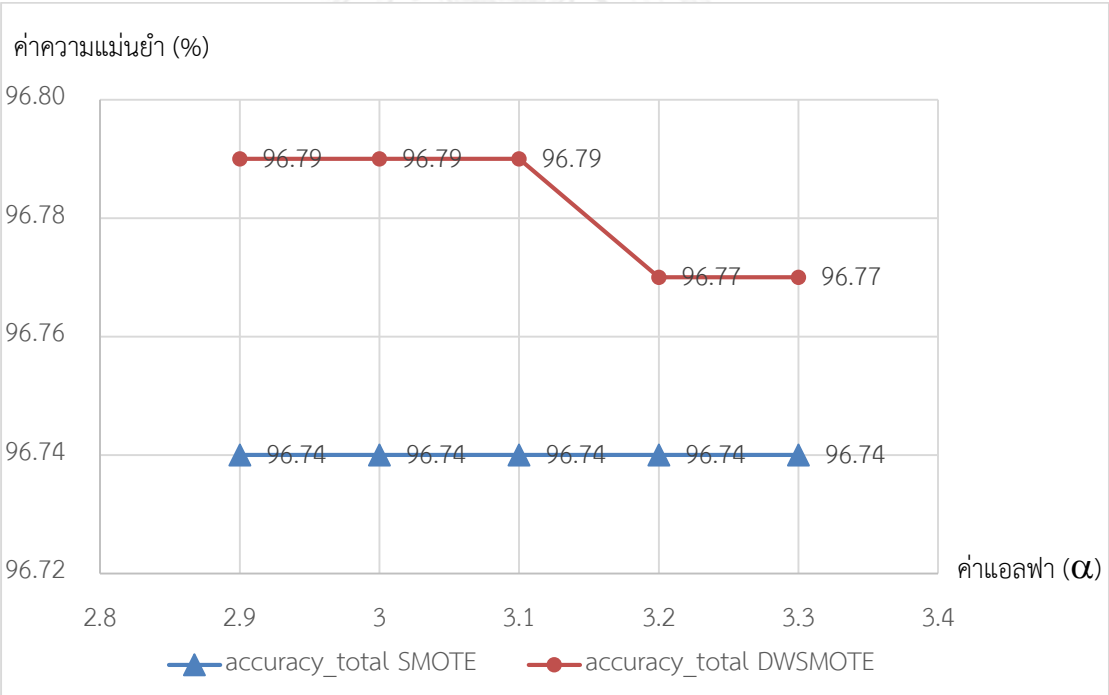
ภาพที่ 4-38 ค่าความแม่นยำของข้อมูลในตัวอย่างทั้งหมดที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Tic-tac-toe



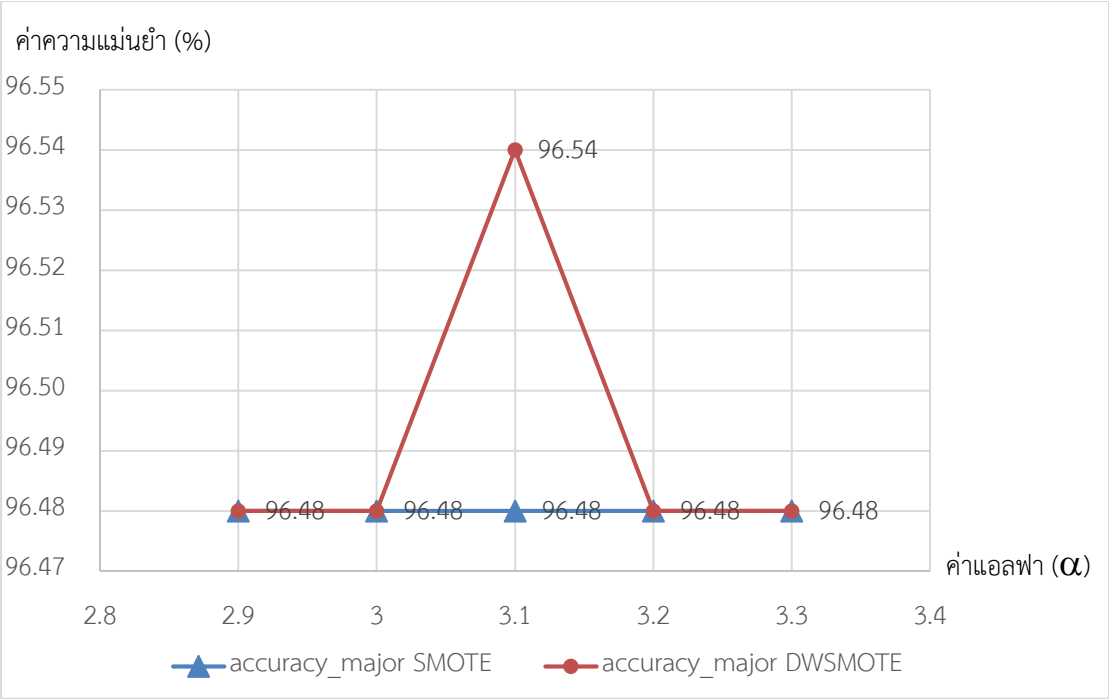
ภาพที่ 4-39 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มมากที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Tic-tac-toe



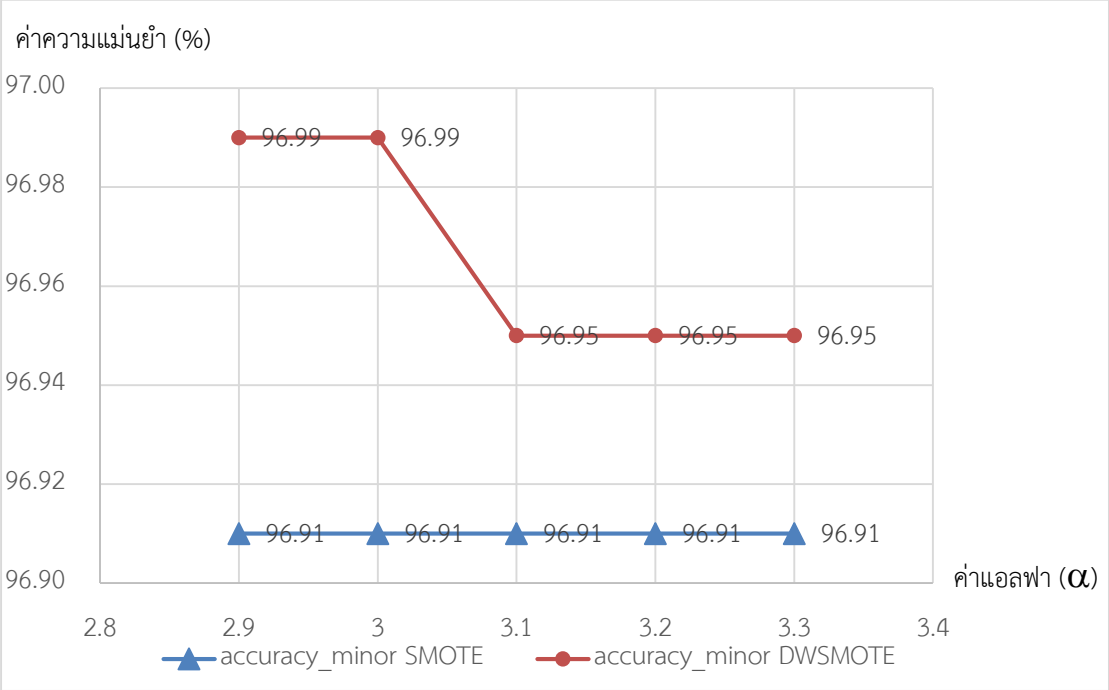
ภาพที่ 4-40 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มน้อยที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Tic-tac-toe



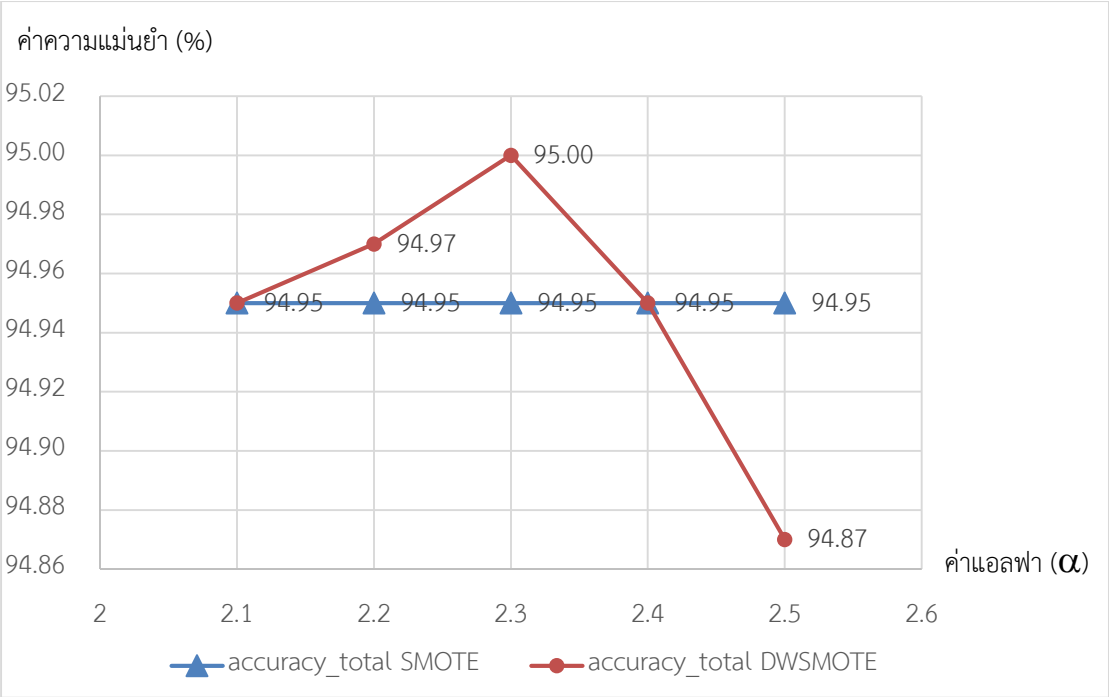
ภาพที่ 4-41 ค่าความแม่นยำของข้อมูลในตัวอย่างทั้งหมดที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Splice-ei



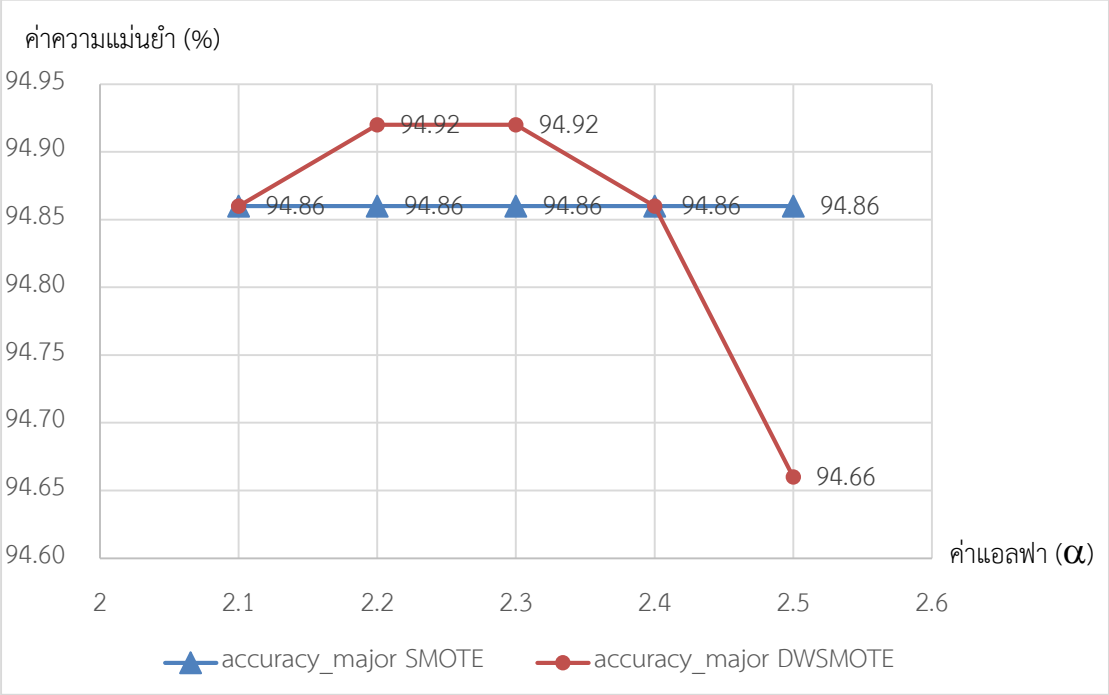
ภาพที่ 4-42 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มมากที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Splice-ei



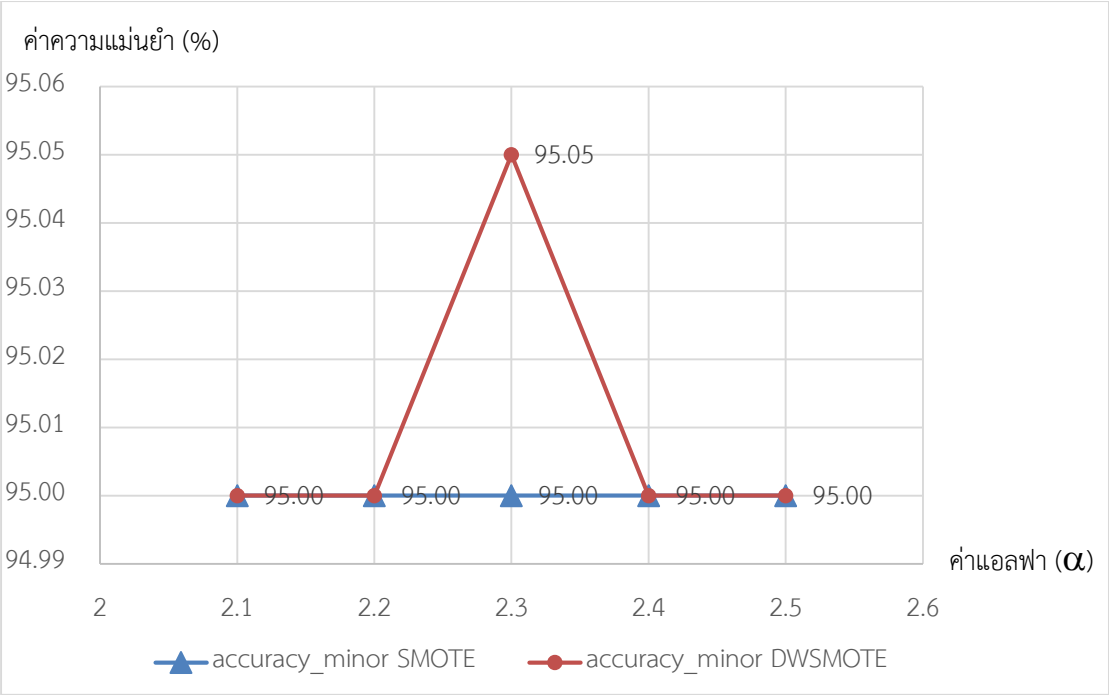
ภาพที่ 4-43 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มน้อยที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Splice-ei



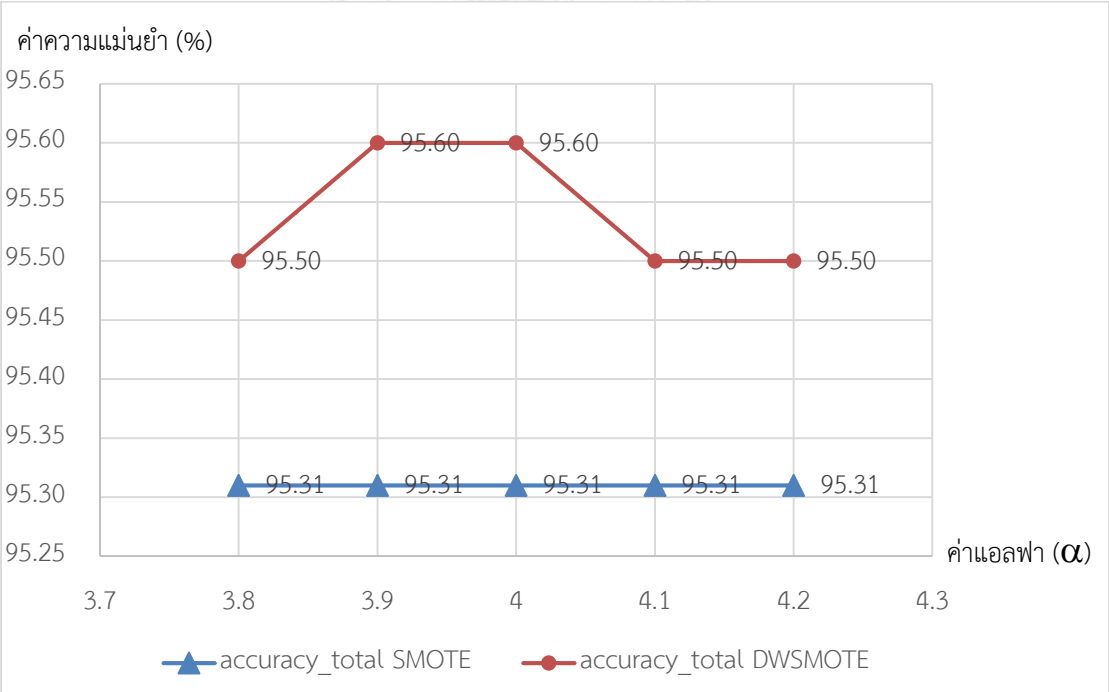
ภาพที่ 4-44 ค่าความแม่นยำของข้อมูลในตัวอย่างทั้งหมดที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Splice-ie



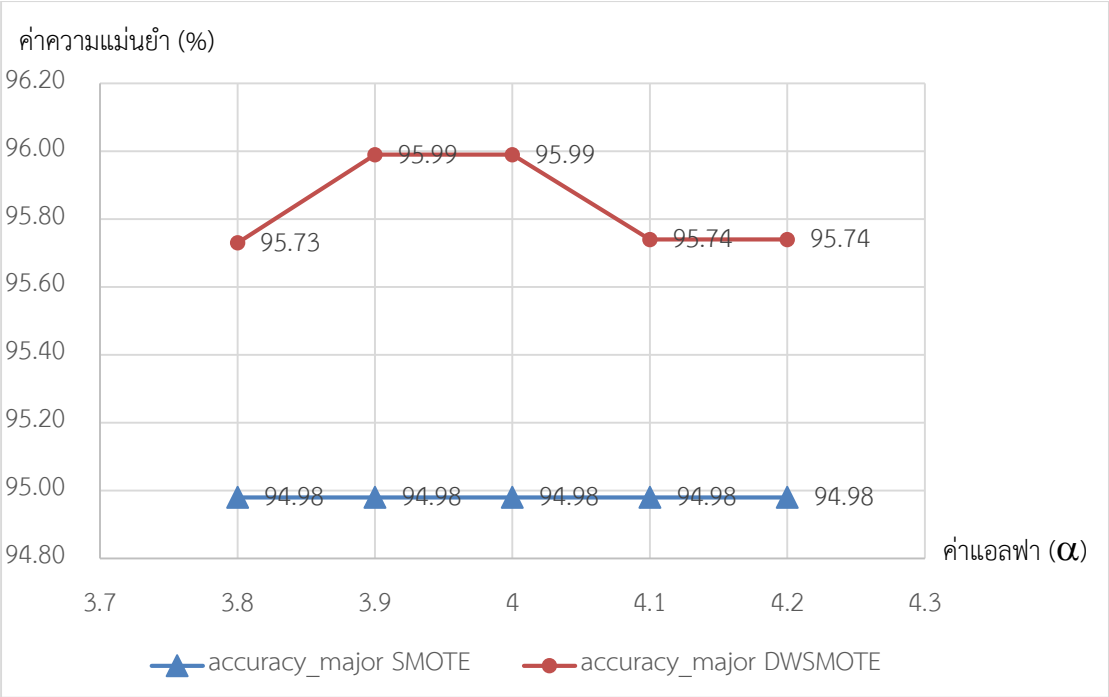
ภาพที่ 4-45 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มมากที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Splice-ie



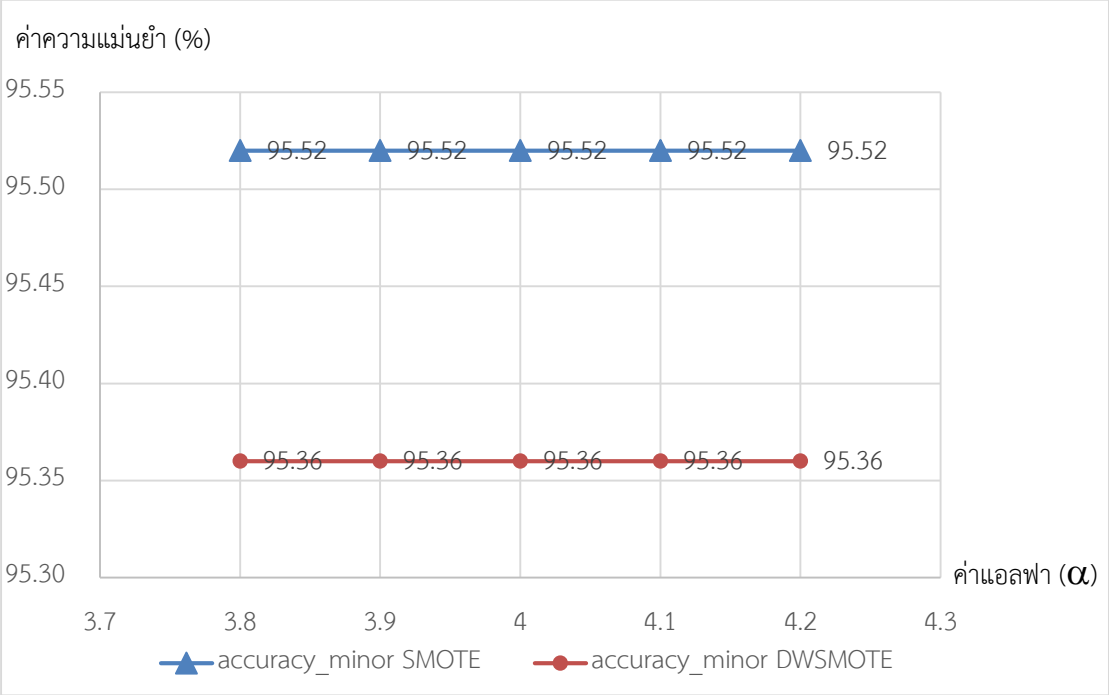
ภาพที่ 4-46 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มน้อยที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Splice-ie



ภาพที่ 4-47 ค่าความแม่นยำของข้อมูลในตัวอย่างทั้งหมดที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Vehicle



ภาพที่ 4-48 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มมากที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Vehicle



ภาพที่ 4-49 ค่าความแม่นยำของข้อมูลในตัวอย่างกลุ่มน้อยที่ได้จากผลการทดลองด้วยค่าเฉลี่ยการทดสอบไขว้ข้ามสิบกลุ่ม โดยใช้ค่าแอลฟาที่ต่างกัน 5 ค่า ของชุดข้อมูล Vehicle

บทที่ 5

สรุปผลการวิจัย และข้อเสนอแนะ

5.1 สรุปผลการวิจัย

จากวิธีการที่ได้นำเสนอในเบื้องต้น และจากการทดลอง มาจากปัญหาข้อมูลไม่สมดุล โดยเป็นปัญหาที่เกิดขึ้นโดยทั่วไป และพบอยู่ในชีวิตประจำวัน เนื่องจากข้อมูลจะถูกแบ่งออกเป็น 2 กลุ่ม คือ ข้อมูลกลุ่มมาก และข้อมูลกลุ่มน้อย เมื่อนำข้อมูลมาทำการจำแนกประเภทจะพบว่า เกิดความโน้มเอียงไปทางข้อมูลกลุ่มมาก วิธีการที่นำเสนอจึงต้องการแก้ปัญหาการจำแนกข้อมูลที่ไม่สมดุลที่มีลักษณะสองกลุ่ม ให้ความสนใจกับข้อมูลกลุ่มน้อยในชุดข้อมูล เพื่อให้จำแนกข้อมูลกลุ่มน้อยได้ดีขึ้น เป็นการนำข้อมูลที่ได้ทำการสังเคราะห์ข้อมูลกลุ่มน้อยเพิ่มขึ้นมาจนมีจำนวนใกล้เคียงกับข้อมูลกลุ่มมาก จัดเป็นขั้นตอนการเตรียมข้อมูลก่อนการจำแนก เรียกว่า SMOTE แล้วนำชุดข้อมูลที่ผ่านการทำ SMOTE มาแล้ว มาทำการจำแนกประเภท ตามวิธีการที่ออกแบบไว้ โดยใช้อัลกอริทึม C4.5 เป็นพื้นฐาน และปรับการหาค่าเอนโทรปีใหม่ ใช้น้ำหนักต่างกันบนข้อมูลสังเคราะห์ จากผลการทดลองแสดงให้เห็นว่าการใช้ค่าน้ำหนักที่ต่างกันบนข้อมูลกลุ่มน้อย ที่ระดับต่างๆ หรือจำนวนเท่าที่แตกต่างกัน สำหรับชุดข้อมูลทดสอบแต่ละชุดไม่เท่ากัน เพราะเมื่อทำการเพิ่มจำนวนตัวอย่างกลุ่มน้อยที่ 100% แล้ว ชุดข้อมูลทดสอบบางชุด จำนวนตัวอย่างกลุ่มน้อยอาจยังมีจำนวนน้อยอยู่เมื่อเทียบกับจำนวนตัวอย่างกลุ่มมาก ทำให้ผลในการจำแนกตัวอย่างกลุ่มน้อยยังไม่ดี แต่บางชุดข้อมูล เมื่อมีการเพิ่มข้อมูลกลุ่มน้อยจากการสังเคราะห์ข้อมูลกลุ่มน้อยขึ้นมาแล้ว ทำให้มีจำนวนกลุ่มน้อยใกล้เคียงกับจำนวนข้อมูลกลุ่มมาก หรือในบางตัวอย่างจำนวนกลุ่มน้อยมีจำนวนมากกว่าจำนวนตัวอย่างกลุ่มมาก ทำให้การจำแนกข้อมูลกลุ่มน้อยสามารถทำให้ดีขึ้น เมื่อเทียบกับวิธีที่ใช้ C4.5 (SMOTE) แสดงให้เห็นว่า เมื่อทำการสังเคราะห์ข้อมูลกลุ่มน้อยเพิ่มขึ้นจากวิธีการ SMOTE ที่ 100% หรือเพิ่มขึ้นเป็น 1 เท่าของข้อมูลกลุ่มน้อยเดิม และมีการให้ค่าน้ำหนักกับข้อมูลกลุ่มน้อยที่ต่างกันในการจำแนก จะทำให้การจำแนกข้อมูลกลุ่มน้อยได้ดีขึ้น แต่อาจจะส่งผลให้การจำแนกตัวอย่างกลุ่มมากมีประสิทธิภาพที่ลดลงได้ในบางชุดข้อมูลที่ใช้ทดลอง และเมื่อเปรียบเทียบระดับความมีนัยสำคัญทางสถิติที่ระดับ 0.1 กับค่าเอฟเมเชอร์ จากการทดลองจะพบว่าวิธีการที่นำเสนอสามารถทำการจำแนกตัวอย่างกลุ่มน้อยได้ดีกว่าวิธี C4.5 แบบพื้นฐานที่ใช้ข้อมูลที่ได้จากการ SMOTE และมีประสิทธิภาพโดยรวมดีกว่า

เมื่อพิจารณาถึงลักษณะของข้อมูลในชุดข้อมูลทั้ง 16 ชุดข้อมูล จะพบว่าชุดข้อมูลส่วนใหญ่จะมีลักษณะเป็นข้อมูลที่มีค่าเป็นตัวเลข (numeric) และค่าโนมินัล (nominal) และยังมีบางชุดข้อมูลที่มีลักษณะเป็นเชิงเส้น (linear), แบบต่อเนื่อง (continuous), ข้อมูลที่ไม่ไม่ใช่ตัวเลข (non-numeric) ซึ่งส่งผลต่อประสิทธิภาพในการจำแนกประเภทข้อมูล เนื่องจากบางเทคนิคจะทำงานได้ดีเฉพาะข้อมูล

ที่เป็นโนมินัลเท่านั้น และการแก้ไขข้อมูลที่ผิดพลาด (Missing Value) โดยเพิ่มข้อมูลที่ขาดหายไป (Replace Missing Value) อาจส่งผลต่อประสิทธิภาพของการจำแนก เนื่องจากข้อมูลบางส่วนที่หายไป อาจเกิดจากความผิดพลาดในการกรอกข้อมูล ทำให้เมื่อนำมาทำการจำแนกข้อมูลอาจไม่ได้ข้อมูลที่แท้จริง

ตารางที่ 4-41 จำนวนข้อมูลที่สังเคราะห์ขึ้นในตัวอย่างกลุ่มน้อย เมื่อเทียบกับตัวอย่างกลุ่มมากกับผลในการจำแนกตัวอย่างในชุดข้อมูล โดยพิจารณาจากค่าแอลฟาที่ให้ผลการจำแนกที่ดีที่สุด

ชุดข้อมูล	ความถูกต้องในการจำแนกข้อมูล						ลักษณะข้อมูล	ข้อมูลที่ผิดพลาด	จำนวนตัวอย่าง	
	ทั้งหมด		กลุ่มมาก		กลุ่มน้อย					
	SMOTE	DWSMOTE	SMOTE	DWSMOTE	SMOTE	DWSMOTE			Major	Minor
Breast-cancer	✓	✓	✓	✓	✓	✓	linear, nominal	Missing value	201	170
Breast-cancer-w	✓		✓		✓	✓	numeric	Missing value	458	482
Credit-g		✓	✓			✓	numeric, nominal	Missing value	700	600
Diabetes	✓			✓		✓	numeric, nominal	no	500	536
Ecoli		✓	✓	✓		✓	numeric, nominal	no	301	70
Flags		✓		✓		✓	numeric, nominal	no	177	34
Glass		✓		✓		✓	numeric, nominal	no	197	34
Haberman		✓		✓	✓	✓	numeric, nominal	no	225	162
Hepatitis		✓		✓		✓	continuous, nominal	Missing value	123	64
Ionosphere		✓		✓		✓	continuous	no	225	252
Liver		✓		✓		✓	numeric, nominal	no	200	290
Post-operative		✓	✓	✓		✓	nominal	Missing value	66	48
Tic-tac-toe		✓		✓		✓	nominal	no	626	664

ชุดข้อมูล	ความถูกต้องในการจำแนกข้อมูล						ลักษณะข้อมูล	ข้อมูลที่ผิดพลาด	จำนวนตัวอย่าง	
	ทั้งหมด		กลุ่มมาก		กลุ่มน้อย					
	SMOTE	DWSMOTE	SMOTE	DWSMOTE	SMOTE	DWSMOTE			Major	Minor
Splice-ei		✓		✓		✓	non-numeric, nominal	no	2423	1524
Splice-ie		✓		✓		✓	non-numeric, nominal	no	2422	1536
Vehicle		✓		✓	✓		numeric	no	647	398

หมายเหตุ ✓ คือ วิธีการที่สามารถจำแนกข้อมูลได้ดีกว่า เมื่อพิจารณาค่าแอลฟาที่ทำให้ค่าความถูกต้องในการจำแนกข้อมูลที่ดีที่สุด

จากตารางข้างต้นเมื่อพิจารณาถึงจำนวนตัวอย่างกลุ่มน้อยที่สังเคราะห์เพิ่มขึ้น ถ้าจำนวนตัวอย่างกลุ่มน้อยทั้งหมด มีจำนวนมากกว่าจำนวนตัวอย่างกลุ่มมากจะส่งผลให้จำแนกประเภทกลุ่มตัวอย่างได้ดีขึ้นเมื่อเทียบกับวิธีการ C4.5 แบบพื้นฐาน แต่ในบางชุดข้อมูลที่มีจำนวนตัวอย่างกลุ่มน้อยที่สังเคราะห์เพิ่มขึ้นแล้วรวมกับจำนวนตัวอย่างกลุ่มน้อยเดิม ยังมีจำนวนน้อยกว่าจำนวนตัวอย่างกลุ่มมาก หรือบางชุดข้อมูลมีจำนวนใกล้เคียงกัน จะให้ผลในการจำแนกดีขึ้นในตัวอย่างกลุ่มน้อย แต่ในกลุ่มตัวอย่างมากหรือ ตัวอย่างทั้งหมด อาจจะยังทำได้ไม่ดีเท่ากับวิธีการแบบพื้นฐาน คือ C4.5 เนื่องจากบางชุดข้อมูลมีข้อมูลที่สูญหาย ซึ่งจะส่งผลต่อข้อมูลที่ใช้ในการจำแนก

โดยค่าแอลฟาที่ทำให้การจำแนกข้อมูลในชุดข้อมูลได้ดีที่สุด ได้จากการทดลอง กำหนดค่าแอลฟาที่แตกต่างกันไปเรื่อยๆ จนพบค่าแอลฟาที่ทำให้ประสิทธิภาพในการจำแนกข้อมูลในชุดข้อมูลที่ใช้ในการทดลองให้ผลได้ใกล้เคียง หรือดีกว่าวิธีการพื้นฐาน คือ C4.5

5.2 ข้อเสนอแนะ

1. งานวิจัยฉบับนี้เป็นงานวิจัยที่ทำการจำแนกข้อมูลที่มีลักษณะเป็น 2 กลุ่ม ถ้ามีการพัฒนาเพื่อให้สามารถทำการจำแนกข้อมูลที่มีลักษณะหลายกลุ่มจะสามารถประยุกต์ใช้งานในด้านอื่นๆ และจะเป็นประโยชน์ในหลายๆ ด้าน

2. ถ้าเพิ่มในส่วนการของตัดเล็มของต้นไม้ตัดสินใจ (pruning) เข้าไป อาจจะทำให้มีความแม่นยำในการจำแนก หรือสามารถจำแนกได้ดีกว่าวิธีที่ได้ออกแบบไว้

3. กรณีที่ทำการเพิ่มตัวอย่างกลุ่มน้อยโดยการสังเคราะห์ขึ้นมาใหม่จากการทำ SMOTE ถ้ากำหนดให้เพิ่มขึ้น 100% ของจำนวนกลุ่มน้อยที่มีอยู่ เมื่อเพิ่มจำนวนตัวอย่างกลุ่มน้อยขึ้นมาแล้ว

อาจจะยังมีจำนวนไม่ใกล้เคียงกับจำนวนกลุ่มมาก ทำให้ผลในการจำแนกอาจจะยังทำได้ไม่ดี เนื่องจากยังมีความแตกต่างระหว่างจำนวนข้อมูลกลุ่มมากกับจำนวนข้อมูลกลุ่มน้อยอยู่ อาจส่งผลให้การจำแนกอาจเอนเอียงไปทางกลุ่มมากได้ หรือจำแนกแล้วไม่ส่งผลอะไรในกลุ่มน้อย ในอนาคต ถ้ามีการกำหนดค่าในการสังเคราะห์เพิ่มตัวอย่างกลุ่มน้อยที่ระดับเหมาะสมจนเท่ากับตัวอย่างกลุ่มมากในชุดข้อมูล อาจจะส่งผลให้การจำแนกทำได้ดีขึ้น

4. ส่วนของการจำแนก อาจจะต้องมีการพัฒนาโดยทำการหาค่าน้ำหนักที่เหมาะสม เมื่อเทียบเป็นสัดส่วนระหว่างข้อมูลกลุ่มมากและข้อมูลกลุ่มน้อย ซึ่งในงานวิจัยฉบับนี้ยังเป็นการหาค่าที่ดีที่สุด โดยการทดลอง

รายการอ้างอิง

1. แต้วิจิตร, ส., *What is big data?*, in *Data Mining Innovation*. 2014.
2. กิจศิริกุล, บ., รายงานการวิจัยฉบับสมบูรณ์ โครงการวิจัย โครงการวิจัยร่วมภาครัฐและเอกชน ปีงบประมาณ *2545*, in โครงการย่อยที่ *7* อัลกอริทึมการทำเหมืองข้อมูล. 2546. p. 19-27, 42-44, 82-89.
3. Merz., C.L.B.a.C.J., *UCI repository of machine learning databases*. 1998: Department of Information and Computer Science.
4. เชี่ยวสกุลวัฒนา, ส. and ส. สินธุภิญโญ. การจำแนกต้นไม้ตัดสินใจสำหรับชุดข้อมูลไม่สมดุลโดยใช้น้ำหนักต่างกันบนข้อมูลสังเคราะห์. in การประชุมวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ *10 (NCCIT 2014)*. 2014. ณ Angsana laguna จ.ภูเก็ต ประเทศไทย.
5. Quinlan, J.R., *Induction of decision trees, Machine Learning*. 1986.
6. Shannon, C.E., *A Mathematical Theory of Communication.* Bell System Technical Journal, 1948. **27**.
7. Quinlan, J.R., *Induction of decision trees, Machine Learning.* 1986: p. 81-106.
8. Fletcher, D. and P. Reutemann, *Weka 3: Data Mining Software in Java*. 1999, University of Waikato.
9. Chawla, N.V., et al., *SMOTE: Synthetic Minority Over-Sampling Technique.* Journal of Artificial Intelligent Research 2002: p. 321-357.
10. Bunkhumpornpat, C., K. Sinapiromsaran, and C. Lursinsap, *Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-sampling Technique for Handling the Class Imbalanced Problem*, in *Lecture Notes in Computer Science* 2009. p. 475-482.
11. Songwattanasiri, P. and K. Sinapiromsaran, *Smoute: Synthetic Minority Over-sampling and Under-sampling Techniques for class imbalanced problem,.* Annual International Conference on computer Science Education Innovation&Technology (CSEIT) 2010, 2010: p. 78-83.
12. ทรงวัฒนศิริ, ป., เทคนิคการสุ่มเพิ่มตัวอย่างข้างน้อยสังเคราะห์และเทคนิคการสุ่มลดตัวอย่างข้างมากสำหรับปัญหาความไม่สมดุลระหว่างกลุ่ม, in ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ 2553, จุฬาลงกรณ์มหาวิทยาลัย.
13. Thach, N., H.,, P. Rojanavasu, and O. Pinngern, *Cost-sensitive XCS Classifier System Addressing Imbalance Problems.* Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 2008: p. 132-136.

14. Udomthanapong, S., K. Tamee, and O. Pinngern, *Using Accuracy-Based Learning Classifier Systems for Imbalance Datasets.* Proceedings of ECTI-CON 2008, 2008: p. 21-24.
15. Lenca, P., et al., *A Comparison of Different Off-Centered Entropies to Deal with Class Imbalance for Decision Trees.* PAKDD'08 Proceedings of the 12th Pacific-Asia conference on Advances in knowledge discovery and data mining, 2008: p. 634-643.
16. กฤษดาวาณิชย์, อ., ตัวปรับยืดเกาะในต้นไม้ตัดสินใจสำหรับเซตข้อมูลไม่สมดุล, in ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ 2553, จุฬาลงกรณ์มหาวิทยาลัย: จุฬาลงกรณ์มหาวิทยาลัย.
17. Boonchuay, K., K. Sinapiromsaran, and C. Lursinsap, *Minority Split and Gain Ratio for a Class Imbalance.* Eight International Conference on Fuzzy Systems and Knowledge Discovery (FSKD) 2011 2011: p. 2060-3064.

ประวัติผู้เขียนวิทยานิพนธ์

ชื่อ-สกุล นายสุรพงษ์ เชี่ยวสกุลวัฒนา

วัน เดือน ปี เกิด 12 มกราคม พ.ศ. 2521 จังหวัดสุพรรณบุรี

ประวัติการศึกษา ปริญญาตรี วิทยาศาสตรบัณฑิต ภาควิชาเทคโนโลยีชนบท คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์ ศูนย์รังสิต ปีการศึกษา 2539

เข้าศึกษาที่คณะวิศวกรรมศาสตร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ สาขาวิทยาศาสตร์คอมพิวเตอร์ ปีการศึกษา 2553