บทความวิจัย (Research Article)

การปรับความไม่สมดุลของข้อมูลด้วยการจำแนก 5 วิธี

Adjusting the Imbalanced Data with 5 Classification Methods

อัจฉรา แผ้วบาง และสายชล สินสมบูรณ์ทอง*

ภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ถนนฉลองกรุง เขตลาดกระบัง กรุงเทพมหานคร 10520

Achara Phaeobang and Saichon Sinsomboonthong*

Department of Statistics, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang,

Chalongkrung Road, Ladkrabang, Bangkok 10520

Received: November 15, 2019; Accepted: November 23, 2019

บทคัดย่อ

งานวิจัยนี้มีนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการปรับข้อมูลที่ไม่สมดุล 4 วิธี คือ วิธีการสุ่มเกิน วิธีการสุ่มเกินโดยเทคนิค SMOTE วิธีการสุ่มลด และวิธีการสุ่มผสมผสาน โดยวิธีการจำแนก 5 วิธี คือ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีโครงข่ายประสาทเทียม วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีฐานกฎ และวิธีลาดลงสโตแคสติก ว่าวิธีใดมีประสิทธิภาพในการจำแนกดีที่สุด โดยพิจารณาจากค่าความถูกต้อง ค่าความไว ค่าความจำเพาะ ค่าคลาดเคลื่อนกำลังสองเฉลี่ย และค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย โดยแบ่งข้อมูลในอัตราส่วน 70, 20 และ 10 ตามลำดับ ในข้อมูลส่วนที่ 1 ข้อมูลเรียนรู้ นำไปสร้างตัวแบบ ร้อยละ 70 ข้อมูลส่วนที่ 2 ข้อมูลตรวจสอบความถูกต้อง นำข้อมูลไปประเมินความผิดพลาดของตัวแบบ ร้อยละ 20 และข้อมูลส่วนที่ 3 ข้อมูลทดสอบ นำไปทดสอบตัวแบบ ร้อยละ 10 โดยการกำหนดตัวสร้างเลขสุ่มเทียม เป็น 10, 20, 30, 40 และ 50 มีข้อมูลที่ไม่สมดุลในการศึกษา 3 ชุด คือ ชุดข้อมูลเคมีบำบัดมะเร็งลำไส้ใหญ่ระยะ B/C ชุดข้อมูลโรคที่มีความผิดปกติของโปรตีน และชุดข้อมูลการรักษาอาการปวดศีรษะขั้นรุนแรง โดยใช้โปรแกรม WEKA การเปรียบเทียบข้อมูลทั้ง 3 ชุด คือ ข้อมูลเคมีบำบัดมะเร็งลำไส้ใหญ่ระยะ B/C ชุดข้อมูลโรคที่มีความผิดปกติของโปรตีน และชุดข้อมูลการรักษาอาการปวดศีรษะขั้นรุนแรง วิธีที่มีประสิทธิภาพสูงสุดคือวิธีฐานกฎโดยการปรับความไม่สมดุลด้วยวิธีการสุ่มเกินเทคนิค SMOTE

คำสำคัญ : ความไม่สมดุลของข้อมูล; วิธีเพื่อนบ้านใกล้สุด k ตัว; วิธีโครงข่ายประสาทเทียม; วิธีซัพพอร์ตเวกเตอร์แมชชีน; วิธีฐานกฎ; วิธีลาดลงสโตแคสติก

Abstract

We compared the imbalanced data of four methods; i.e. over sampling, synthetic minority over sampling technique, under sampling, and hybrid methods, using five classification methods; i.e. k-nearest neighbor, artificial neural network, support vector machine, rule-based, and stochastic gradient descent. Metrics were accuracy, sensitivity, specificity, mean square error and mean absolute error.

**Corresponding author: saichon.ss49@gmail.com* DOI: 10.14456/tjst.2020.46

The data sets were chemotherapy for stage B/C colon cancer, monoclonal gammopathy and treatment of migraine headaches. Each of these data sets was divided into three proportions in the ratio of 70:20:10 using the data part 1. Training data are used to create a model 70 percentages; the data part 2. Validation data are used to evaluate an error a model 20 percentages, and the data part 3, testing data are used to test a model 10 percentages using the random seed 10, 20, 30, 40, and 50 by WEKA program. When we compared the chemotherapy for stage B/C colon cancer data set, the monoclonal gammopathy data sets, and the treatment of migraine headaches data sets, the best method was the ruled-based in imbalanced data adapting the synthetic minority over sampling technique.

Keywords: imbalanced data; k-nearest neighbor; artificial neural network; support vector machine; rule-based; stochastic gradient descent

1. คำนำ

ปัจจุบันมีปัญหาการแบ่งข้อมูลที่กำลังได้รับความสนใจ คือ ปัญหาการแบ่งกลุ่มข้อมูลที่ไม่สมดุล (imbalanced dataset) ซึ่งเกิดจากการที่มีข้อมูล 2 กลุ่ม หรือมากกว่า 2 กลุ่ม โดยที่ข้อมูลกลุ่มส่วนมาก (majority) จะมีข้อมูลจำนวนมากกว่า ขณะเดียวกันข้อมูลกลุ่มส่วนน้อย (minority) จะมีข้อมูลจำนวนน้อยกว่า ทั้งนี้เนื่องจากโดยธรรมชาติของความเป็นจริง การที่จะกำหนดให้ขนาดของข้อมูลในกลุ่มส่วนมากและกลุ่มส่วนน้อยมีขนาดเท่า ๆ กัน เพื่อการสอนหรือการจัดกลุ่มข้อมูลนั้นเป็นเรื่องยากหรืออาจจะเป็นไปไม่ได้ ดังนั้นจึงเป็นปัญหาที่ท้าทายและมีความยากมากสำหรับการหาขั้นตอนวิธี (algorithm) ที่เหมาะสมสำหรับการแบ่งกลุ่มข้อมูลที่ไม่สมดุล ทั้งนี้เนื่องจากถ้านำข้อมูลทั้งสองชุดเข้าสู่ขั้นตอนวิธีพร้อมกันทั้งหมด จะทำให้ผลการแบ่งกลุ่มข้อมูลเกิดความผิดพลาด กล่าวคือ ข้อมูลที่อยู่ในกลุ่มส่วนน้อยจะถูกจัดให้ไปอยู่ในกลุ่มส่วนมากทั้งหมด ซึ่งจะนำไปสู่ปัญหาที่เรียกว่า ปัญหาการแบ่งกลุ่มข้อมูลผิดกลุ่ม (misclassification) (เบญจภรณ์ และคณะ, 2557)

มีการศึกษาวิธีการสุ่มเกินโดยใช้เทคนิค synthetic minority over- sampling technique (SMOTE) ในหลายงานวิจัย เช่น ในการจัดการปัญหาเกี่ยวกับกลุ่มไม่สมดุลในข้อมูลทางการแพทย์ โดยใช้เทคนิค FURIA14, decision tree ปรับความสมดุลของข้อมูลด้วยวิธี SMOTE พบว่าวิธีต้นไม้ตัดสินใจ (decision tree) มีประสิทธิภาพในการทำนายที่ดีกว่าวิธีอื่น โดยมีค่าความถูกต้องร้อยละ 85.78 ค่าความไวร้อยละ 84.21 และค่าความจำเพาะร้อยละ 87.34 หลังจากปรับความไม่สมดุลพบว่าค่าความถูกต้องเพิ่มขึ้นร้อยละ 6.19 ค่าความไวเพิ่มขึ้นร้อยละ 64.21 และค่าความจำเพาะลดลงร้อยละ 2.42 (Rahman and Davis, 2013) ส่วนการศึกษาตัวแบบการทำนายผลการรักษาผู้ป่วยมะเร็งปากมดลูกด้วยวิธีโครงข่ายประสาทเทียม โดยนำเสนอตัวแบบเพื่อทำนายผลการรักษาผู้ป่วยมะเร็งปากมดลูกที่เข้ารับการรักษาด้วยวิธีการฉายรังสีโดยการประยุกต์ใช้วิธีโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับด้วยการนำวิธีการเรียนรู้แบบมีค่าใช้จ่าย (cost-sensitive learning, CSL) และวิธีการสุ่มเกินโดยใช้เทคนิค SMOTE เมื่อเปรียบเทียบประสิทธิภาพการทำนาย พบว่าวิธีโครงข่ายประสาทเทียมที่มีการแก้ปัญหาความไม่สมดุลของข้อมูลด้วยวิธีการสุ่มเกินโดยใช้เทคนิค SMOTE มีประสิทธิภาพการทำนายด้วยค่าความถูกต้อร้อยละง 81.71 ค่าความไวร้อยละ 94.47 และค่าความจำเพาะร้อยละ 55.47 สูงกว่าวิธีการถดถอยลอจิสติกที่มีการ

แก้ปัญหาความไม่สมดุลของข้อมูลด้วยวิธีการเรียนรู้แบบมีค่าใช้จ่าย ซึ่งมีค่าความถูกต้องร้อยละ 81.00 ค่าความไวร้อยละ 84.52 และค่าความจำเพาะร้อยละ 30.66 (เชาวนันท์ และคณะ, 2556) และการศึกษาเกี่ยวกับการรักษาซ้ำของผู้ป่วยจิตเภท เรื่อง การพัฒนาตัวแบบเพื่อการพยากรณ์การรักษาซ้ำของผู้ป่วยโรคจิตเภทโดยเทคนิคเหมืองข้อมูล พบว่าข้อมูลการรักษาของผู้ป่วยโรคจิตเภททางการแพทย์มีข้อมูลที่มีความผิดปกติและมีความไม่สมดุลของข้อมูล ซึ่งผู้วิจัยได้ใช้วิธีการกรองเอาค่าผิดปกติออกด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนแล้วใช้การสุ่มเกินโดยใช้เทคนิค SMOTE และวิธีการสุ่มลด (under sampling) ในการแก้ปัญหา แล้วนำข้อมูลที่สมดุลแล้วมาจำแนก พบว่าวิธีการกรองแล้วใช้วิธีการสุ่มเกินโดยใช้เทคนิค SMOTE สามารถเพิ่มประสิทธิภาพให้ตัวแบบเพิ่มขึ้น โดยค่าความถูกต้องเพิ่มขึ้นเฉลี่ยร้อยละ 46.36 ค่าความไวเพิ่มขึ้นเฉลี่ยร้อยละ 20.05 และค่าความจำเพาะเพิ่มขึ้นร้อยละ 32.69 ซึ่งมากกว่าวิธีการกรองแล้วใช้วิธีการสุ่มลด (วีระยุทธ และคณะ, 2557)

นอกจากนี้ได้มีการศึกษาวิธีการสุ่มผสมผสาน (hybrid method) อีกด้วย โดยศึกษาเกี่ยวกับวิธีการแบ่งกลุ่มข้อมูลที่ไม่สมดุลเรื่องวิธีการที่เหมาะสมสำหรับการแบ่งกลุ่มข้อมูลที่ไม่สมดุลสูง โดยได้นำเสนอวิธีการเลือกคุณลักษณะสำหรับข้อมูลไม่สมดุลที่มีจำนวนมิติข้อมูลที่สูง เพื่อลดจำนวนมิติข้อมูลที่ซ้ำซ้อนและเพิ่มประสิทธิภาพของการจำแนกข้อมูลด้วยวิธี SMOTE และวิธีแบบผสมผสาน (hybrid method) จากนั้นค้นหาคุณลักษณะย่อยด้วยขั้นตอนวิธีทางพันธุกรรมร่วมกับโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ เปรียบเทียบประสิทธิภาพของวิธีการที่นำเสนอโดยใช้ค่าความถูกต้อง ค่าเฉลี่ยเรขาคณิต และความถูกต้องของการจำแนกข้อมูลกลุ่มส่วนน้อย รวมถึงเวลาที่ใช้ในการประมวลผล ทดสอบกับ 3 ตัวจำแนกข้อมูล ได้แก่ โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ โครงข่ายฟังก์ชันรัศมีฐานและซัพพอร์ตเวกเตอร์แมชชีน ผลปรากฏว่าตัวจำแนกข้อมูลทั้ง 3 วิธี ให้ค่าความถูกต้อง ค่าเฉลี่ยเรขาคณิตโดยส่วนมากของการจำแนกชุดข้อมูลของวิธีการแบบผสมผสานดีกว่าข้อมูลเดิมที่ไม่ได้ปรับปรุง (เบญจภรณ์ และคณะ, 2557) และมีการศึกษาวิธีการสุ่มเกิน (Over Sampling) วิธีการสุ่มเกินโดยเทคนิค SMOTE วิธีการสุ่มลดและวิธีการสุ่มผสมผสานในการเปรียบเทียบประสิทธิภาพในการทำนายผลการปรับความไม่สมดุลของข้อมูลในการจำแนกด้วยเทคนิคการทำเหมืองข้อมูล 4 วิธี คือ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม และวิธีซัพพอร์ตเวกเตอร์แมชชีน โดยพิจารณาจากค่าความถูกต้อง ค่าความไว ค่าความจำเพาะ และค่าคลาดเคลื่อนกำลังสองเฉลี่ย พบว่าชุดข้อมูลการรับรู้ทางหูของเด็กที่มีภาวะน้ำคั่งในหูชั้นกลางหรือหูน้ำหนวก วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีซัพพอร์ตเวกเตอร์แมชชีน โดยการปรับความไม่สมดุลด้วยวิธีการสุ่มเกินเทคนิค SMOTE ส่วนชุดข้อมูลยอดคงเหลือในบัตรเครดิตของลูกค้า วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีเพื่อนบ้านใกล้สุด k ตัว โดยการปรับความไม่สมดุลด้วยวิธีการสุ่มเกินเทคนิค SMOTE และชุดข้อมูลคุณภาพไวน์แดง วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีโครงข่ายประสาทเทียม โดยการปรับความไม่สมดุลด้วยวิธีการสุ่มเกิน (พัชรียา และคณะ, 2561)

ดังนั้นผู้วิจัยจึงให้ความสนใจการปรับความไม่สมดุลของข้อมูลด้วยวิธีต่าง ๆ 4 วิธี คือ วิธีการสุ่มเกิน วิธีการสุ่มเกินโดยใช้เทคนิค SMOTE วิธีการสุ่มลด และวิธีการสุ่มผสมผสาน แล้วจำแนก (classification) ด้วยวิธีต่าง ๆ 5 วิธี คือ วิธีเพื่อนบ้านใกล้สุด k ตัว (k nearest neighbor) วิธีโครงข่ายประสาทเทียม (artificial neural network) วิธีซัพพอร์ตเวกเตอร์แมชชีน (support vector machine)

เนื่องจากมีประสิทธิภาพสูงสุดตามงานวิจัยของ พัชรียา และคณะ (2561) ส่วนวิธีฐานกฎ (rule based) และวิธีลาดลงสโตแคสติก (stochastic gradient) ยังไม่มีการศึกษามากนัก ผู้วิจัยจึงสนใจเลือก 2 วิธี นี้มาศึกษาด้วย เพื่อต้องการเปรียบเทียบประสิทธิภาพในการจำแนกทั้ง 5 วิธี ว่าวิธีใดมีประสิทธิภาพและเหมาะสมกับรูปแบบของชุดข้อมูล โดยใช้การทดสอบเปรียบเทียบประสิทธิภาพด้วยค่าความถูกต้อง (accuracy) ค่าความไว (sensitivity) ค่าความจำเพาะ (specificity) ที่มีค่าสูงที่สุด ส่วนค่าคลาดเคลื่อนกำลังสองเฉลี่ย (mean square error, MSE) และค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย (mean absolute error, MAE) ที่มีค่าต่ำที่สุด

2. วิธีการวิจัย

2.1 เครื่องมือที่ใช้ในการวิจัย

โปรแกรมที่ใช้ในการวิจัย คือ WEKA (Waikato environment for knowledge analysis) เวอร์ชั่น 3.9.2

2.2 การเก็บรวบรวมข้อมูล การปรับข้อมูลให้มีความสมดุล การแบ่งข้อมูล การศึกษาขั้นตอนวิธี และการเปรียบเทียบประสิทธิภาพของวิธีการจำแนก

2.2.1 การเก็บรวบรวมข้อมูล ศึกษาข้อมูลที่ไม่สมดุลจากเว็บไซต์ Vincentarelbundock จำนวน 3 ชุด คือ

(1) เคมีบำบัดมะเร็งลำไส้ใหญ่ระยะ B/C (chemotherapy for stage B/C colon cancer) มีจำนวนข้อมูลทั้งหมด 1,858 ค่า พบค่าในกลุ่มส่วนมาก 1,760 ค่า (ร้อยละ 94.72 และค่าในกลุ่มส่วนน้อย 98 ค่า (ร้อยละ 5.27) (Laurie *et al*., 1994)

(2) โรคที่มีความผิดปกติของโปรตีน (monoclonal gammopathy) มีจำนวนข้อมูลทั้งหมด 1,384 ค่า พบค่าในกลุ่มส่วนมาก 1,245 ค่า (ร้อยละ 89.96) และค่าในกลุ่มส่วนน้อย 139 ค่า (ร้อยละ 10.04) (Kyle *et al*., 1994)

(3) การรักษาอาการปวดศีรษะขั้นรุนแรง (treatment of migraine headaches) มีจำนวนข้อมูลทั้งหมด 3,491 ค่า พบค่าในกลุ่มส่วนมาก 2,666 ค่า (ร้อยละ 76.37) และค่าในกลุ่มส่วนน้อย 825 ค่า (ร้อยละ 23.63) (Kostecki *et al*., 1999)

2.2.2 การปรับข้อมูลให้มีความสมดุล

(1) วิธีการสุ่มเกิน (over sampling) คือ การสุ่มข้อมูลในกลุ่มส่วนน้อยเพื่อสร้างข้อมูลใหม่ของกลุ่มส่วนน้อยให้มีจำนวนเพิ่มมากขึ้นให้ใกล้เคียงหรือเท่ากับจำนวนในกลุ่มส่วนมาก โดยใช้วิธีการสุ่มกลุ่มตัวอย่างอย่างง่าย (simple random sampling) การศึกษาของ กีระชาติ (2559) เกี่ยวกับชุดข้อมูลผู้ป่วยเป็นเนื้อร้ายได้ใช้การปรับความไม่สุมดุล 4 วิธี คือ วิธีการสุ่มเกิน วิธีการสุ่มเกินโดยเทคนิค SMOTE วิธีการสุ่มลด วิธีการสุ่มผสมผสาน วิธีที่ให้ผลการปรับดีที่สุด คือ วิธีการสุ่มเกินโดยเทคนิค SMOTE

(2) วิธีการสุ่มเกินโดยเทคนิค SMOTE คือ การสุ่มสร้างข้อมูลจากกลุ่มส่วนน้อยตามจำนวนที่กำหนด โดยการวัดระยะห่างจากจุดข้อมูลตัวอย่างไปยังจุดข้อมูลใกล้เคียง แล้วสุ่มสร้างข้อมูลสังเคราะห์ขึ้นให้ใกล้เคียงกับกลุ่มส่วนมาก การศึกษาของ ภรันยา (2559) เกี่ยวกับสัดส่วนการใช้อินเทอร์เน็ตสูงที่สุดในเยาชนอายุ 15-24 ปี โดยปรับความสมดุลของข้อมูลด้วยวิธีการสุ่มเกินเทคนิค SMOTE แล้วใช้การจำแนกด้วยวิธีต้นไม้ตัดสินใจ โดยใช้ขั้นตอนวิธี J48, ID3, LMT, CART และ random forest พบว่าขั้นตอนวิธี random forest ดีกว่าขั้นตอนวิธี J48, ID3, LMT และ CART

(3) วิธีการสุ่มลด (under sampling) คือ การปรับข้อมูลให้มีความสมดุลด้วยวิธีการสุ่มลดจำนวนข้อมูลจากกลุ่มส่วนมากลงเพื่อทำให้จำนวนข้อมูลระหว่างกลุ่มส่วนมากและกลุ่มส่วน

น้อยมีจำนวนใกล้เคียงกันมากขึ้น (กีระชาติ, 2559)

(4) วิธีการสุ่มผสมผสาน (hybrid method) คือ การนำวิธีสุ่มเกินและวิธีสุ่มลดมาทำงานร่วมกัน โดยวิธีนี้จะเป็นการสุ่มลดจำนวนข้อมูลจากกลุ่มส่วนมากและสุ่มเพิ่มข้อมูลในกลุ่มส่วนน้อย ให้จำนวนข้อมูลจากทั้งสองกลุ่มมีจำนวนใกล้เคียงกันหรือเท่ากัน (กีระชาติ, 2559)

2.2.3 การแบ่งข้อมูล

แบ่งชุดข้อมูลโดยโปรแกรม WEKA เวอร์ชั่น 3.9.2 สุ่ม 5 รอบ โดยการกำหนดตัวสร้างเลขสุ่มเทียมเป็น 10, 20, 30, 40 และ 50 ในอัตราส่วน 70:20:10 ส่วนที่ 1 ข้อมูลเรียนรู้ (training data) นำไปสร้างตัวแบบร้อยละ 70 ข้อมูลส่วนที่ 2 ข้อมูลตรวจสอบความถูกต้อง (validation data) นำไปประเมินความผิดพลาดของตัวแบบร้อยละ 20 และข้อมูลส่วนที่ 3 ข้อมูลทดสอบ (testing data) นำไปทดสอบตัวแบบร้อยละ 10 (พนิดา และคณะ, 2560)

2.2.4 การศึกษาขั้นตอนวิธี

(1) วิธีเพื่อนบ้านใกล้สุด k ตัว (K-nearest neighbor) เป็นวิธีการที่ได้รับความนิยมอย่างมาก เนื่องจากเป็นวิธีการที่ง่ายและมีประสิทธิภาพ ซึ่งสามารถนำไปประยุกต์ใช้กับงานอย่างหลากหลาย เช่น งานด้านการจำแนก รวมถึงงานด้านการแทนที่ข้อมูลที่สูญหาย ใช้ขั้นตอนวิธี IBK ซึ่งมีวิธีการดำเนินการดังนี้ (Troyanskaya, 2001)

(1.1) กำหนดค่า k เพื่อใช้พิจารณาสมาชิกที่อยู่ใกล้กันมากที่สุด เช่น k = 3 คือ จะพิจารณาเฉพาะข้อมูล 3 ตัวแรก ที่อยู่ใกล้กับจุดที่ต้องการทำนาย

(1.2) คำนวณระยะห่างระหว่างข้อมูลตัวอย่างที่สนใจกับข้อมูลอื่น ๆ ทุกตัวด้วยระยะห่างยูคลิเดียน (Euclidian distance) จากสมการ

$$D_{\text{Euclidian}}(x_i, y_i) = \sqrt{\sum_{k=1}^{n}(x_i, y_i)^2} \quad (1)$$

โดยที่ $D_{\text{Euclidian}}(x_i, y_i)$ คือ ระยะห่างระหว่างตัวอย่าง x_i กับตัวอย่าง y_i; k คือ คุณลักษณะทั้งหมดของตัวอย่าง

(1.3) เลือกค่าข้อมูลที่มีค่าระยะห่างน้อยที่สุด k ตัว เพื่อนำมาพิจารณาหาคำตอบ

(2) วิธีโครงข่ายประสาทเทียม (artificial neural network) โครงข่ายประสาทเทียมใช้ขั้นตอนวิธีชนิดเพอร์เซปตรอนหลายชั้น โดยกำหนดค่าอัตราการเรียนรู้เป็น 0.1 ค่าโมเมนตัมเป็น 0.9 จำนวนรอบการสอน 20,000 รอบ การวิจัยครั้งนี้ใช้ขั้นตอนวิธีของวิธีโครงข่ายประสาทเทียมชนิดเพอร์เซปตรอนหลายชั้นที่มีชั้นซ่อน 1 ชั้น การเชื่อมโยงกันระหว่างเซลล์ประสาทโดยทั่วไปนิยมใช้การเชื่อมโยงแบบแพร่ย้อนกลับ (back-propagation) ซึ่งเป็นขั้นตอนที่ใช้สอนโครงข่ายประสาทเทียมแบบเพอร์เซปตรอนหลายชั้น โดยแบบจำลองโครงข่ายประสาทเทียมมีการเชื่อมโยงกันเป็นโครงข่ายแบบเป็นชั้น ๆ โครงข่ายชนิดนี้มีการเชื่อมโยงกัน 3 ชั้น ประกอบด้วยชั้นข้อมูลเข้า (input layer) ถัดมาเป็นชั้นซ่อน (hidden layer) และชั้นสุดท้าย คือ ชั้นข้อมูลออก (output layer) (Berson and Smith, 1997) โดยส่วนประกอบที่ถูกบรรจุอยู่ในเซลล์ประสาทแต่ละตัวประกอบด้วย 2 ฟังก์ชันย่อย คือ ฟังก์ชันผลรวม (summation function) และฟังก์ชันกระตุ้น (activation function)

(2.1) ฟังก์ชันผลรวม ทำหน้าที่ในการคำนวณผลรวมของข้อมูลที่ได้จากชั้นข้อมูลเข้า คำนวณจากสมการ $g = \sum_{i=1}^{z} x_i w_i + \beta$ (2) (Hagan *et al*., 1996) โดยกำหนดให้ตัวแปร x คือ ค่าข้อมูลเข้าตัวที่ i; ตัวแปร w คือ ค่าน้ำหนักถ่วงของข้อมูลเข้าตัวที่ i; ตัวแปร g คือ ข้อมูลออกจากฟังก์ชันผลรวม; ตัวแปร z คือ จำนวนเซลล์ประสาทของข้อมูลเข้า; ตัวแปร β คือ ค่าความเอนเอียง (bias)

(2.2) ฟังก์ชันกระตุ้น ทำหน้าที่ปรับเปลี่ยนค่าของข้อมูลที่ได้จากฟังก์ชันผลรวมให้อยู่ในช่วงที่ต้องการ ฟังก์ชันกระตุ้นที่นิยม ได้แก่ ฟังก์ชันเชิงเส้น (linear function) ฟังก์ชันซิกมอยด์ (sigmoid function) และฟังก์ชันไฮเปอร์โบลิกแทนเจนต์ (hyperbolic tangent function) (ธนาวุฒิ, 2552)

(2.3) วิธีซัพพอร์ตเวกเตอร์แมชชีน เป็นวิธีที่ใช้ในการแก้ปัญหาด้านการรู้จำรูปแบบข้อมูล โดยอาศัยหลักการของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการสอนให้ระบบเรียนรู้ โดยเน้นไปยังเส้นแบ่งแยกกลุ่มข้อมูลได้ดีที่สุด (optimal separating hyperplane) สำหรับรากฐานเดิมของซัพพอร์ตเวกเตอร์แมชชีนถูกนำมาใช้กับข้อมูลที่เป็นเชิงเส้น แต่ในความเป็นจริงแล้วข้อมูลที่นำมาใช้ในระบบการสอนให้ระบบเรียนรู้ส่วนใหญ่มักเป็นข้อมูลแบบไม่เป็นเชิงเส้น ซึ่งสามารถแก้ปัญหาดังกล่าวด้วยเคอร์เนลที่เป็นการเปลี่ยนแปลงมิติของข้อมูลให้สูงขึ้น เพื่อช่วยในการเรียงตัวของข้อมูลใหม่ โดยงานวิจัยนี้ใช้ขั้นตอนวิธี SMO ชนิดโพลิโนเมียลเคอร์เนล (polynomial Kernel) (สุรเดช และคณะ, 2554)

(2.4) วิธีฐานกฎ (rules-based) ใช้ชุดลำดับของกฎมาสร้างรูปแบบการแยกประเภทข้อมูล โดยส่วนใหญ่แล้วจะใช้กฎที่เป็น if ... then ซึ่งเป็นกฎอย่างง่าย ใช้ขั้นตอนวิธี decision table เป็นเครื่องมือที่ใช้แสดงเงื่อนไขการตัดสินใจและเลือกการทำงานหรือกระทำกิจกรรมภายใต้เหตุการณ์ของเงื่อนไขที่ระบุ วิธีการตัดสินใจแบบ decision table เป็นตาราง 2 มิติ วิธีฐานกฎเป็นวิธีหนึ่งที่นิยมใช้เช่นเดียวกับวิธีต้นไม้ตัดสินใจ ข้อกำหนด (antecedent) หรือเงื่อนไข (precondition) ของวิธีฐานกฎเป็นการทดสอบคล้ายกับการทดสอบที่โหนดของวิธีต้นไม้ตัดสินใจ แต่ผลของการทดสอบ (consequent) หรือผลลัพธ์ (cnclusion) ที่ได้จะให้คำตอบ (class) ที่ใช้กับตัวอย่างที่อยู่ภายใต้กฎนั้น หรือบางครั้งก็จะให้ค่าการแจกแจงความน่าจะเป็นของคำตอบต่าง ๆ (Murti and Mahantappa, 2012)

(2.5) วิธีลาดลงสโตแคสติก (stochastic gradient descent method) เป็นวิธีที่มีประสิทธิภาพในการจำแนกการเรียนรู้ของตัวจำแนกเชิงเส้น ภายใต้ฟังก์ชันการสูญเสียโค้งนูน ได้แก่ ซัพพอร์ตเวกเตอร์แมชชีน และการถดถอยลอจิสติก (logistic regression) ใช้การเรียนรู้ตัวแบบเชิงเส้นต่าง ๆ ได้แก่ ซัพพอร์ตเวกเตอร์แมชชีนที่มีคำตอบเป็นทวิภาค การถดถอยลอจิสติกที่มีคำตอบเป็นทวิภาค และการถดถอยเชิงเส้น แทนที่ข้อมูลสูญหายและแปลงคุณลักษณะเชิงกลุ่มให้เป็นทวิภาค แปลงคุณลักษณะต่าง ๆ ให้อยู่ในรูปปรกติ (normalization) ดังนั้นค่าสัมประสิทธิ์ของผลลัพธ์เป็นข้อมูลที่อยู่ในรูปปรกติ (Nektarios, 2013) สำหรับคุณลักษณะที่มีคำตอบเป็นนามบัญญัติจะใช้ฟังก์ชันสูญเสียไฮนด์ (hinge loss function) หรือฟังก์ชันการสูญเสียล็อก (log loss function) ส่วนคุณลักษณะที่มีคำตอบเป็นเชิงตัวเลขจะใช้ฟังก์ชันการสูญเสียกำลังสอง (squared loss function) ฟังก์ชันการสูญเสียเอฟซิลอน (epsilon-insensitive loss function) หรือฟังก์ชันการสูญเสียฮูเบอร์ (Huber loss function) หลังจากนั้นนำข้อมูลที่แบ่งออกเป็น 3 ส่วน มาวิเคราะห์โดยใช้โปรแกรม WEKA ซึ่งวิเคราะห์จากวิธีการจำแนก 5 วิธี ข้างต้น

2.2.5 การเปรียบเทียบประสิทธิภาพของวิธีการจำแนก

นำผลการวิเคราะห์ของแต่ละวิธีทั้ง 5 วิธี มาเปรียบเทียบประสิทธิภาพโดยพิจารณาจากเมทริกซ์ความสับสน (confusion matrix) ซึ่งเป็นรูปแบบตารางที่เฉพาะเจาะจงที่นำผลลัพธ์จากการทำนายมาใส่ในตารางเมทริกซ์ความสับสน จะช่วยให้ง่ายต่อการมองเห็นค่าทำนายของขั้นตอนวิธีดังตารางที่ 1

ตารางที่ 1 เมทริกซ์ความสับสน

		ผลลัพธ์จากสมการหรือการทดสอบ	
		คำตอบเป็นบวก	คำตอบที่เป็นลบ
ผลลัพธ์ที่เกิดขึ้นจริง	คำตอบเป็นบวก	TP (true positive)	FN (false negative)
	คำตอบเป็นลบ	FP (false positive)	TN (true negative)

บวกจริง (true positive, TP) คือ จำนวนข้อมูลที่จำแนกถูกว่าเป็นบวก ซึ่งค่าที่แท้จริงเป็นบวก; ลบจริง (true negative, TN) คือ จำนวนข้อมูลที่จำแนกถูกว่าเป็นลบ ซึ่งค่าที่แท้จริงเป็นลบ; บวกเท็จ (false positive, FP) คือ จำนวนข้อมูลที่จำแนกผิดว่าเป็นบวก ซึ่งค่าที่แท้จริงเป็นลบ; ลบเท็จ (false negative, FN) คือ จำนวนข้อมูลที่จำแนกผิดว่าเป็นลบ ซึ่งค่าที่แท้จริงเป็นบวก

(1) ค่าความถูกต้อง (accuracy) ในการทำนาย คือ การแสดงการวัดที่ได้มีความถูกต้องในรูปอัตราส่วนโดยคิดเป็นร้อยละ (สุรวัชร และสายชล, 2560) โดย ค่าความถูกต้อง = (จำนวนข้อมูลที่จำแนกถูกว่าคำตอบเป็นบวกและลบ x 100 %) ÷ จำนวนข้อมูลทั้งหมด

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\ \%$$

(2) ค่าความไว (sensitivity หรือ true positive rate, TPR) คือ สัดส่วนของผลบวกที่เป็นจริงสำหรับภาวะนั้น ๆ (กีระชาติ, 2559)

(3) ค่าความจำเพาะ (specificity, SPC หรือ true negative rate, TNR) คือ สัดส่วนของผลลบที่เป็นจริงสำหรับภาวะนั้น ๆ (กีระชาติ, 2559)

(4) ค่าคลาดเคลื่อนกำลังสองเฉลี่ย (mean square error, MSE) เป็นมาตรวัดการประเมินค่าได้ดี เนื่องจากค่าคลาดเคลื่อนกำลังสองเฉลี่ยประกอบด้วยความเอนเอียงและความแปรปรวน (พนิดา และคณะ, 2560)

$$\text{Mean Square Error} = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}$$

โดยที่ y_i แทนค่าจริง; $\hat{y}_i$ แทนค่าทำนาย; n แทนจำนวนข้อมูลของกลุ่มตัวอย่าง

(5) ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย (mean absolute error, MAE) คือ ค่าวัดความถูกต้องของการทำนายที่วัดจากค่าความคลาดเคลื่อนโดยไม่คำนึงถึงทิศทางของความคลาดเคลื่อน MAE มีหน่วยวัดหน่วยเดียวกับค่าสังเกต (สายชล, 2560)

$$\text{Mean Absolute Error} = \frac{\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|}{n}$$

3. ผลการวิจัย

3.1 ผลการเปรียบเทียบประสิทธิภาพในการทำนายผลของวิธีการจำแนก

งานวิจัยครั้งนี้ใช้การจำแนกโดยใช้วิธีการทำเหมืองข้อมูล โดยนำชุดข้อมูลที่ค้นคว้าจำนวน 3 ชุด มาวิเคราะห์ข้อมูล ซึ่งจะสุ่มแบ่งข้อมูลออกเป็น 3 ส่วน คือ ส่วนที่ 1 ข้อมูลเรียนรู้ นำไปสร้างตัวแบบร้อยละ 70 ข้อมูลส่วนที่ 2 ข้อมูลตรวจสอบความถูกต้อง นำไปประเมินความผิดพลาดของตัวแบบร้อยละ 20 และข้อมูลส่วนที่ 3 ข้อมูลทดสอบ นำไปทดสอบตัวแบบร้อยละ 10 และผู้วิจัยได้นำมาเปรียบเทียบประสิทธิภาพในการทำนายผลของวิธีการจำแนกโดยพิจารณาจากค่าความถูกต้อง ค่าความไว ค่าความจำเพาะ ค่าคลาดเคลื่อนกำลังสองเฉลี่ย และค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย ซึ่งวิธีที่ใช้ในการทดสอบครั้งนี้มี 5 วิธี คือ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีโครงข่ายประสาทเทียม วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีฐานกฎ แลวิธีลาดลงสโตแคสติก ผลการวิเคราะห์ข้อมูลจากตารางที่ 2 ซึ่งเป็นกรณีที่มีจำนวนข้อมูลในกลุ่มส่วนน้อยไม่เกินร้อยละ 5 พบว่าวิธีฐานกฎโดยการปรับความไม่สมดุลด้วยวิธีการสุ่มเกินเทคนิค SMOTE ให้ค่าความถูกต้องสูงสุด คือ ร้อยละ 92.5862 ค่าความไวสูงสุด 0.920 และค่าความจำเพาะสูงสุด 0.931 ส่วน

วิธีโครงข่ายประสาทเทียมโดยการปรับความไม่สมดุลด้วยวิธีการสุ่มเกินเทคนิค SMOTE ให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.0840 ส่วนวิธีลาดลงสโตแคสติกโดยการปรับความไม่สมดุลด้วยวิธีการสุ่มเกินเทคนิค SMOTE ให้ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยต่ำสุด คือ 0.1017 ส่วนตารางที่ 3 ซึ่งเป็นกรณีที่มีจำนวนข้อมูลในกลุ่มส่วนน้อยไม่เกินร้อยละ 10 พบว่าวิธีฐานกฎโดยการปรับความไม่สมดุลด้วยวิธีการสุ่มเกินเทคนิค SMOTE ให้ค่าความถูกต้องสูงสุด คือ ร้อยละ 91.7269 ค่าความไวสูงสุด คือ 0.916 และค่าความจำเพาะสูงสุด คือ 0.921 และค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.0669 ส่วนวิธีเพื่อนบ้านใกล้สุด k ตัว โดยการปรับความไม่สมดุลด้วยวิธีการสุ่มเกินเทคนิค SMOTE ให้ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยต่ำสุด คือ 0.1230 และจากตารางที่ 4 ซึ่งเป็นกรณีที่มีจำนวนข้อมูลในกลุ่มส่วนน้อยไม่เกินร้อยละ 25 พบว่าวิธีฐานกฎโดยการปรับความไม่สมดุลด้วยวิธีการสุ่มเกินเทคนิค SMOTE ให้ค่าความถูกต้องสูงสุด คือ ร้อยละ 80.0377 ค่าความไวสูงสุด 0.795 ค่าความจำเพาะสูงสุด 0.817 และค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.1449 ส่วนวิธีเพื่อนบ้านใกล้สุด k ตัว โดยการปรับความไม่สมดุลด้วยวิธีการสุ่มเกินเทคนิค SMOTE ให้ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยต่ำสุด คือ 0.2260

ตารางที่ 2 ผลการเปรียบเทียบประสิทธิภาพของวิธีการจำแนกสำหรับข้อมูลเคมีบำบัดมะเร็งลำไส้ใหญ่ระยะ B/C (ข้อมูลชุดที่ 1)

วิธีการปรับความไม่สมดุล	ตัวสร้างเลขสุ่มเทียม	ค่าความถูกต้อง	ค่าความไว	ค่าความจำเพาะ	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย	ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย
การจำแนกด้วยวิธีเพื่อนบ้านใกล้สุด k ตัว						
วิธีการสุ่มเกิน	10, 20, 30, 40, 50	76.8523, 76.4205, 74.7159, 79.5455, 76.9886	0.843, 0.881, 0.833, 0.904, 0.701	0.683, 0.635, 0.657, 0.684, 0.840	0.2325, 0.2349, 0.2439, 0.2052, 0.2161	0.2575, 0.2521, 0.2651, 0.2211, 0.2277
	ค่าเฉลี่ย	76.9046	0.832	0.700	0.2265	0.2447
วิธีการสุ่มเกินเทคนิค SMOTE	10, 20, 30, 40, 50	87.6437, 88.2184, 86.4943, 88.2184, 89.0805	0.895, 0.889, 0.878, 0.860, 0.891	0.856, 0.876, 0.849, 0.905, 0.890	0.1171, 0.1147, 0.1321, 0.1162, 0.1088	0.1231, 0.1207, 0.1359, 0.1207, 0.1125
	ค่าเฉลี่ย	87.9311	0.883	0.875	0.1178	0.1226
วิธีการสุ่มลด	10, 20, 30, 40, 50	54.5455, 81.8182, 68.1818, 40.9091, 54.5455	0.583, 0.778, 0.818,0.533, 0.455	0.500, 0.846, 0.545, 0.143, 0.636	0.4264, 0.1674, 0.3018, 0.5324, 0.4149	0.4797, 0.2113, 0.3552, 0.5824, 0.4589
	ค่าเฉลี่ย	60	0.633	0.534	0.3686	0.4175
วิธีการสุ่มผสมผสาน	10, 20, 30, 40, 50	74.7312, 70.9677, 73.1183, 69.3548, 72.0430	0.684, 0.687, 0.852, 0.667, 0.678	0.813, 0.736, 0.564, 0.722, 0.758	0.2433, 0.2846, 0.2613, 0.2928, 0.2595	0.2509, 0.2925, 0.2748, 0.3032, 0.2776
	ค่าเฉลี่ย	72.0430	0.714	0.719	0.2683	0.2798

ตารางที่ 2 (ต่อ)

วิธีการปรับความไม่สมดุล	ตัวสร้างเลขสุ่มเทียม	ค่าความถูกต้อง	ค่าความไว	ค่าความจำเพาะ	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย	ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย
การจำแนกด้วยวิธีโครงข่ายประสาทเทียม						
วิธีการสุ่มเกิน	10, 20, 30, 40, 50	55.9659, 54.5455, 56.5341, 53.1250, 51.9886	0.476, 0.773, 0.767, 0.494, 0.571	0.634, 0.293, 0.355, 0.569, 0.469	0.2619, 0.2546, 0.2552, 0.2681, 0.2633	0.4775, 0.4675, 0.4684, 0.4806, 0.4846
	ค่าเฉลี่ย	54.4318	0.616	0.464	0.2606	0.4757
วิธีการสุ่มเกินเทคนิค SMOTE	10, 20, 30, 40, 50	92.5287, 91.9540, 90.8046, 90.8046, 87.6437	0.950, 0.895, 0.884, 0.883, 0.846	0.898, 0.944, 0.937, 0.935, 0.908	0.0695, 0.0748, 0.0839, 0.0819, 0.1101	0.1123, 0.1216, 0.1494, 0.1434, 0.1903
	ค่าเฉลี่ย	90.7471	0.892	0.924	0.0840	0.1434
วิธีการสุ่มลด	10, 20, 30, 40, 50	31.8182, 68.1818, 54.5455, 31.8182, 22.7273	0.167, 0.667, 0.636, 0.400, 0.182	0.500, 0.692, 0.455, 0.143, 0.273	0.4585, 0.2829, 0.3035, 0.3754, 0.3950	0.6210, 0.3769, 0.4543, 0.6229, 0.7264
	ค่าเฉลี่ย	41.8182	0.410	0.413	0.3630	0.5603
วิธีการผสมผสาน	10, 20, 30, 40, 50	55.9140, 58.6022, 51.0753, 55.3763, 54.3011	0.800, 0.586, 0.639, 0.563, 0.655	0.308, 0.586, 0.333, 0.544, 0.444	0.2793, 0.2578, 0.2748, 0.2597, 0.2590	0.4733, 0.4561, 0.4963, 0.4713, 0.4701
	ค่าเฉลี่ย	55.0538	0.649	0.443	0.2661	0.4734
การจำแนกด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน						
วิธีการสุ่มเกิน	10, 20, 30, 40, 50	55.3977, 57.1023, 50.2841, 55.1136, 48.0114	0.307, 0.665, 0.639, 0.725, 0.362	0.774, 0.467, 0.360, 0.374, 0.600	0.4460, 0.4290, 0.4972, 0.4489, 0.5198	0.4460, 0.4290, 0.4972, 0.4489, 0.5199
	ค่าเฉลี่ย	53.1818	0.540	0.515	0.4682	0.4682
วิธีการสุ่มเกินเทคนิค SMOTE	10, 20, 30, 40, 50	90.2299, 89.0805, 87.6437, 88.2184, 89.0805	0.950, 0.830, 0.825, 0.860, 0.891	0.850, 0.949, 0.937, 0.905, 0.890	0.0977, 0.1091, 0.1235, 0.1162, 0.1088	0.0977, 0.1092, 0.1236, 0.1207, 0.1125
	ค่าเฉลี่ย	88.8506	0.871	0.906	0.1111	0.1127
วิธีการสุ่มลด	10, 20, 30, 40, 50	36.3636, 54.5455, 54.5455, 63.6364, 13.6364	0.583, 0.444, 0.545, 0.933, 0.091	0.100, 0.615, 0.545, 0, 0.182	0.6363, 0.4545, 0.4545, 0.3636, 0.8636	0.6364, 0.4545, 0.4545,0.3636, 0.8636
	ค่าเฉลี่ย	44.5455	0.519	0.288	0.5545	0.5545
วิธีการสุ่มผสมผสาน	10, 20, 30, 40, 50	58.0645, 62.9032, 50.5376, 51.6129, 58.6022	0.747, 0.646, 0.806, 0.563, 0.552	0.407, 0.609, 0.090, 0.467, 0.616	0.4194, 0.3710, 0.4946, 0.4839, 0.4140	0.4194, 0.3710, 0.4946, 0.4839, 0.4140
	ค่าเฉลี่ย	56.3441	0.663	0.438	0.4366	0.4366

ตารางที่ 2 (ต่อ)

วิธีการปรับความไม่สมดุล	ตัวสร้างเลขสุ่มเทียม	ค่าความถูกต้อง	ค่าความไว	ค่าความจำเพาะ	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย	ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย
การจำแนกด้วยวิธีฐานกฎ						
วิธีการสุ่มเกิน	10, 20, 30, 40, 50	55.9659, 53.1250, 53.6932, 58.5227, 51.1364	0.367, 0.616, 0.622, 0.494, 0.480	0.731, 0.437, 0.448, 0.678, 0.543	0.2493, 0.2545, 0.2531, 0.2341, 0.2629	0.4794, 0.4942, 0.4852, 0.4629, 0.5059
	ค่าเฉลี่ย	54.4886	0.516	0.567	0.2507	0.4855
วิธีการสุ่มเกินเทคนิค SMOTE	10, 20, 30, 40, 50	93.6782, 91.9540, 93.1034, 92.5287, 91.6667	0.950, 0.889, 0.926, 0.911, 0.926	0.922, 0.949, 0.937, 0.941, 0.908	0.0626, 0.0751, 0.1100, 0.0935, 0.1064	0.1400, 0.1630, 0.2415, 0.2008, 0.2316
	ค่าเฉลี่ย	92.5862	0.920	0.931	0.0895	0.1954
วิธีการสุ่มลด	10, 20, 30, 40, 50	50, 68.1818, 40.9091, 63.6364, 36.3636	0.667, 0.556, 0.364, 0.933, 0.455	0.300, 0.769, 0.455, 0, 0.273	0.2670, 0.2143, 0.3025, 0.2616, 0.2769	0.5024, 0.4194, 0.5287, 0.4468, 0.5180
	ค่าเฉลี่ย	51.8182	0.595	0.359	0.2645	0.4831
วิธีการสุ่มผสมผสาน	10, 20, 30, 40, 50	57.5269, 61.8280, 54.8387, 48.9247, 58.6022	0.779, 0.717, 0.815, 0.531, 0.667	0.363, 0.506, 0.179, 0.444, 0.515	0.2537, 0.2429, 0.2656, 0.2607, 0.2390	0.4856, 0.4693, 0.4976, 0.4963, 0.4641
	ค่าเฉลี่ย	56.3441	0.702	0.401	0.2524	0.4826
การจำแนกด้วยวิธีลาดลงสโตแคสติก						
วิธีการสุ่มเกิน	10, 20, 30, 40, 50	57.1023, 54.8295, 49.7159, 55.1136, 48.0114	0.410, 0.600, 0.633, 0.708, 0.379	0.715, 0.491, 0.355, 0.391, 0.583	0.4290, 0.4517, 0.5028, 0.4489, 0.5198	0.4290, 0.4517, 0.5028, 0.4489, 0.5199
	ค่าเฉลี่ย	52.9545	0.546	0.507	0.4705	0.4705
วิธีการสุ่มเกินเทคนิค SMOTE	10, 20, 30, 40, 50	91.9540, 91.0920, 89.3678, 89.9425, 86.7816	0.950, 0.877, 0.857, 0.866, 0.829	0.886, 0.944, 0.937, 0.935, 0.908	0.0805, 0.0891, 0.1063, 0.1005, 0.1322	0.0805, 0.0891, 0.1063, 0.1006, 0.1322
	ค่าเฉลี่ย	89.8275	0.876	0.922	0.1017	0.1017
วิธีการสุ่มลด	10, 20, 30, 40, 50	31.8182, 59.0909, 50, 54.5455, 27.2727	0.333, 0.667, 0.636, 0.733, 0.182	0.300, 0.583, 0.364, 0.143, 0.364	0.6818, 0.4091, 0.5000, 0.4545, 0.7273	0.6818, 0.4091, 0.5000, 0.4545, 0.7273
	ค่าเฉลี่ย	44.5455	0.510	0.351	0.5545	0.5545
วิธีการสุ่มผสมผสาน	10, 20, 30, 40, 50	55.9140, 61.8280, 50.5376, 55.3763, 55.9140	0.684, 0.646, 0.796, 0.583, 0.586	0.429, 0.586, 0.103, 0.522, 0.535	0.2496, 0.3817, 0.4946, 0.4462, 0.4409	0.4409, 0.3817, 0.4946, 0.4462, 0.4409
	ค่าเฉลี่ย	55.9140	0.659	0.435	0.4026	0.4409

ตารางที่ 3 ผลการเปรียบเทียบประสิทธิภาพของวิธีการจำแนกสำหรับข้อมูลโรคที่มีความผิดปกติของโปรตีน (ข้อมูลชุดที่ 2)

วิธีการปรับความไม่สมดุล	ตัวสร้างเลขสุ่มเทียม	ค่าความถูกต้อง	ค่าความไว	ค่าความจำเพาะ	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย	ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย
การจำแนกด้วยวิธีเพื่อนบ้านใกล้สุด k ตัว						
วิธีการสุ่มเกิน	10, 20, 30, 40, 50	71.3710, 75, 63.3065, 72.9839, 70.5645	0.669, 0.722, 0.464, 0.717, 0.643	0.758, 0.783, 0.780, 0.745, 0.773	0.2841, 0.2450, 0.3641, 0.2680, 0.2920	0.2881, 0.2501, 0.3680, 0.2720, 0.2962
	ค่าเฉลี่ย	70.6452	0.643	0.768	0.2907	0.2949
วิธีการสุ่มเกินเทคนิค SMOTE	10, 20, 30, 40, 50	89.5582, 89.1566, 89.1566, 81.9277, 90.3614	0.913, 0.936, 0.861, 0.867, 0.870	0.881, 0.847, 0.921, 0.769, 0.933	0.1034, 0.1075, 0.1075, 0.1791, 0.0955	0.1079, 0.1119, 0.1119, 0.1835, 0.1000
	ค่าเฉลี่ย	88.0321	0.889	0.870	0.1186	0.1230
วิธีการสุ่มลด	10, 20, 30, 40, 50	87.5000, 40.6250, 50, 81.2500, 46.8750	0.923, 0.471, 0.667, 0.833, 0.556	0.824, 0.333, 0.353, 0.750, 0.357	0.1180, 0.5562, 0.4685, 0.1764, 0.4977	0.1495, 0.5876, 0.5001, 0.2079, 0.5292
	ค่าเฉลี่ย	61.2500	0.690	0.523	0.3634	0.3949
วิธีการสุ่มผสมผสาน	10, 20, 30, 40, 50	66.1871, 48.2014, 64.7482, 68.3453, 71.2230	0.671, 0.475, 0.614, 0.735, 0.736	0.651, 0.487, 0.681, 0.634, 0.687	0.3336, 0.5101, 0.3525, 0.3119, 0.2838	0.3406, 0.5176, 0.3525, 0.3195, 0.2908
	ค่าเฉลี่ย	63.7410	0.646	0.628	0.3584	0.3642
การจำแนกด้วยวิธีโครงข่ายประสาทเทียม						
วิธีการสุ่มเกิน	10, 20, 30, 40, 50	66.1280, 63.7097, 66.5323, 62.5000, 66.9355	0.613, 0.767, 0.414, 0.572, 0.574	0.710, 0.487, 0.886, 0.691, 0.773	0.2408, 0.2547, 0.2326, 0.2452, 0.2342	0.4176, 0.4309, 0.4140, 0.4262, 0.4096
	ค่าเฉลี่ย	65.1611	0.588	0.709	0.2415	0.4197
วิธีการสุ่มเกินเทคนิค SMOTE	10, 20, 30, 40, 50	87.5502, 92.3695, 87.9518, 80.7229, 87.5502	0.896, 0.904, 0.820, 0.867, 0.809	0.858, 0.944, 0.937, 0.744, 0.933	0.1075, 0.0683, 0.1062, 0.1603, 0.1071	0.1836, 0.1050, 0.1752, 0.2712, 0.1885
	ค่าเฉลี่ย	87.2289	0.859	0.883	0.1099	0.1847
วิธีการสุ่มลด	10, 20, 30, 40, 50	68.7500, 56.2500, 68.7500, 65.6250, 46.8750	0.615, 0.647, 0.800, 0.667, 0.611	0.737, 0.467, 0.588, 0.625, 0.286	0.2656, 0.2901, 0.2497, 0.2258, 0.3294	0.3382, 0.4797, 0.3748, 0.3082, 0.5225
	ค่าเฉลี่ย	61.2500	0.668	0.541	0.2721	0.4047
วิธีการผสมผสาน	10, 20, 30, 40, 50	63.3094, 53.2374, 60.4317, 66.1871, 64.0288	0.487, 0.443, 0.671, 0.559, 0.556	0.810, 0.603, 0.536, 0.761, 0.731	0.2480, 0.2566, 0.2583, 0.2458, 0.2454	0.4113, 0.4635, 0.4382, 0.3975, 0.4263
	ค่าเฉลี่ย	61.43888	0.543	0.688	0.25078	0.42736

ตารางที่ 3 (ต่อ)

วิธีการปรับความไม่สมดุล	ตัวสร้างเลขสุ่มเทียม	ค่าความถูกต้อง	ค่าความไว	ค่าความจำเพาะ	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย	ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย
การจำแนกด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน						
วิธีการสุ่มเกิน	10, 20, 30, 40, 50	69.3548, 68.5484, 64.1129, 66.1290, 70.1613	0.782, 0.797, 0.543, 0.833, 0.744	0.605, 0.557, 0.727, 0.445, 0.655	0.3065, 0.3145, 0.3589, 0.3387, 0.2944	0.3065, 0.3145, 0.3589, 0.3387, 0.2984
	ค่าเฉลี่ย	67.6613	0.740	0.598	0.3226	0.3234
วิธีการสุ่มเกินเทคนิค SMOTE	10, 20, 30, 40, 50	82.3293, 84.3373, 84.7390, 81.5261, 84.7390	0.913, 0.936, 0.738, 0.898, 0.757	0.746, 0.750, 0.953, 0.727, 0.925	0.1767, 0.1566, 0.1526, 0.1847, 0.1526	0.1767, 0.1566, 0.1526, 0.1847, 0.1526
	ค่าเฉลี่ย	83.5341	0.848	0.820	0.1646	0.1646
วิธีการสุ่มลด	10, 20, 30, 40, 50	68.7500, 53.1250, 50, 75, 65.6250	0.385, 0.765, 0.133, 1, 0.722	0.895, 0.267, 0.824, 0, 0.571	0.3125, 0.4688, 0.5000, 0.2500, 0.3437	0.3125, 0.4688, 0.5000, 0.2500, 0.3438
	ค่าเฉลี่ย	62.5000	0.601	0.511	0.3750	0.3750
วิธีการสุ่มผสมผสาน	10, 20, 30, 40, 50	69.0647, 52.5180, 61.1511, 70.5036, 60.4317	0.776, 0.148, 0.600, 0.588, 0.653	0.587, 0.821, 0.623, 0.817, 0.552	0.3094, 0.4749, 0.3885, 0.2950, 0.3956	0.3094, 0.4748, 0.3885, 0.2950, 0.3957
	ค่าเฉลี่ย	62.7338	0.553	0.680	0.3727	0.3727
การจำแนกด้วยวิธีฐานกฎ						
วิธีการสุ่มเกิน	10, 20, 30, 40, 50	67.3387, 62.0968, 62.9032, 66.5323, 69.7581	0.742, 0.932, 0.353, 0.913, 0.589	0.605, 0.261, 0.871, 0.355, 0.815	0.2171, 0.2270, 0.2330, 0.2224, 0.2068	0.4197, 0.4380, 0.4406, 0.4382, 0.4107
	ค่าเฉลี่ย	65.7258	0.706	0.581	0.2212	0.4294
วิธีการสุ่มเกินเทคนิค SMOTE	10, 20, 30, 40, 50	91.1647, 91.1647, 94.7791, 89.5582, 91.9679	0.913, 0.896, 0.984, 0.820, 0.965	0.910, 0.927, 0.913, 0.975, 0.881	0.0626, 0.0706, 0.0491, 0.0848, 0.0675	0.1554, 0.1398, 0.1071, 0.1715, 0.1365
	ค่าเฉลี่ย	91.7269	0.916	0.921	0.0669	0.1421
วิธีการสุ่มลด	10, 20, 30, 40, 50	84.3750, 53.1250, 46.8750, 75, 46.8750	0.769, 0.941, 0.200, 1, 0.611	0.895, 0.067, 0.706, 1, 0.286	0.1387, 0.2673, 0.2693, 0.1918, 0.3035	0.2805, 0.4931, 0.5120, 0.3875, 0.5321
	ค่าเฉลี่ย	61.2500	0.704	0.391	0.2341	0.4410
วิธีการสุ่มผสมผสาน	10, 20, 30, 40, 50	53.9568, 58.2734, 65.4676, 69.0647, 59.7122	0.368, 0.377, 0.471, 0.485, 0.639	0.746, 0.744, 0.841, 0.887, 0.552	0.2267, 0.2345, 0.2175, 0.2112, 0.2441	0.4399, 0.4654, 0.4323, 0.4077, 0.4529
	ค่าเฉลี่ย	61.2949	0.468	0.754	0.2268	0.4396

ตารางที่ 3 (ต่อ)

วิธีการปรับความไม่สมดุล	ตัวสร้างเลขสุ่มเทียม	ค่าความถูกต้อง	ค่าความไว	ค่าความจำเพาะ	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย	ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย
การจำแนกด้วยวิธีลาดลงสโตแคสติก						
วิธีการสุ่มเกิน	10, 20, 30, 40, 50	67.7419, 68.1452, 62.9032, 68.5484, 70.9677	0.774, 0.782, 0.543, 0.804, 0.729	0.581, 0.565, 0.705, 0.536, 0.689	0.3226, 0.3185, 0.3710, 0.3145, 0.2903	0.3226, 0.3185, 0.3710, 0.3145, 0.2903
	ค่าเฉลี่ย	67.6613	0.726	0.615	0.3234	0.3234
วิธีการสุ่มเกินเทคนิค SMOTE	10, 20, 30, 40, 50	84.3373, 88.7550, 88.3534, 81.1245, 85.5422	0.913, 0.928, 0.828, 0.875, 0.783	0.784, 0.847, 0.937, 0.744, 0.918	0.1566, 0.1124, 0.1165, 0.1888, 0.1466	0.1566, 0.1124, 0.1165, 0.1888, 0.1466
	ค่าเฉลี่ย	85.62248	0.865	0.846	0.1442	0.1442
วิธีการสุ่มลด	10, 20, 30, 40, 50	75, 56.2500, 62.5000, 68.7500, 62.5000	0.769, 0.706, 0.600, 0.917, 0.722	0.737, 0.400, 0.647, 0, 0.500	0.2500, 0.4374, 0.3750, 0.3125, 0.3750	0.2500, 0.4375, 0.4375, 0.3125, 0.3750
	ค่าเฉลี่ย	65	0.743	0.457	0.3500	0.3500
วิธีการสุ่มผสมผสาน	10, 20, 30, 40, 50	66.9065, 58.2734, 64.7482, 69.7842, 60.4317	0.724, 0.361, 0.614, 0.574, 0.653	0.603, 0.756, 0.681, 0.817, 0.552	0.3310, 0.4173, 0.3525, 0.3022, 0.3956	0.3309, 0.4173, 0.3525, 0.3022, 0.3957
	ค่าเฉลี่ย	64.0288	0.585	0.682	0.3597	0.3597

ตารางที่ 4 ผลการเปรียบเทียบประสิทธิภาพของวิธีการจำแนกสำหรับข้อมูลการรักษาอาการปวดศีรษะขั้นรุนแรง (ข้อมูลชุดที่ 3)

วิธีการปรับความไม่สมดุล	ตัวสร้างเลขสุ่มเทียม	ค่าความถูกต้อง	ค่าความไว	ค่าความจำเพาะ	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย	ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย
การจำแนกด้วยวิธีเพื่อนบ้านใกล้สุด k ตัว						
วิธีการสุ่มเกิน	10, 20, 30, 40, 50	73.4463, 78.3427, 77.5895, 78.1544, 73.6347	0.760, 0.775, 0.805, 0.761, 0.771	0.712, 0.792, 0.750, 0.805, 0.705	0.2645, 0.2158, 0.2233, 0.2176, 0.2631	0.2665, 0.2177, 0.2252, 0.2196, 0.2655
	ค่าเฉลี่ย	76.2335	0.774	0.753	0.2368	0.2389
วิธีการสุ่มเกินเทคนิค SMOTE	10, 20, 30, 40, 50	77.9661, 79.0960, 77.2128, 76.0829, 77.2128	0.783, 0.785, 0.766, 0.746, 0.773	0.777, 0.797, 0.778, 0.775, 0.771	0.2193, 0.2082, 0.2269, 0.2381, 0.2269	0.2215, 0.2103, 0.2290, 0.2403, 0.2290
	ค่าเฉลี่ย	77.5141	0.771	0.780	0.2239	0.2260
วิธีการสุ่มลด	10, 20, 30, 40, 50	73.1183, 64.5161, 73.1183, 69.3548, 72.0430	0.733, 0.616, 0.852, 0.667, 0.678	0.729, 0.664, 0.564, 0.722, 0.758	0.2656, 0.3507, 0.2613, 0.2928, 0.2595	0.2715, 0.3566, 0.2748, 0.3032, 0.7760
	ค่าเฉลี่ย	70.4301	0.709	0.687	0.2860	0.3964
วิธีการสุ่มผสมผสาน	10, 20, 30, 40, 50	72.4928, 78.7966, 71.9198, 73.6390, 76.2178	0.688, 0.800, 0.696, 0.698, 0.777	0.760, 0.777, 0.744, 0.778, 0.745	0.2712, 0.2108, 0.2791, 0.2620, 0.2371	0.2750, 0.2138, 0.2822, 0.2651, 0.2408
	ค่าเฉลี่ย	74.6132	0.732	0.761	0.2520	0.2554

ตารางที่ 4 (ต่อ)

วิธีการปรับความไม่สมดุล	ตัวสร้างเลขสุ่มเทียม	ค่าความถูกต้อง	ค่าความไว	ค่าความจำเพาะ	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย	ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย
การจำแนกด้วยวิธีโครงข่ายประสาทเทียม						
วิธีการสุ่มเกิน	10, 20, 30, 40, 50	69.3032, 75.5179, 70.2448, 72.6930, 73.2580	0.878, 0.650, 0.825, 0.536, 0.846	0.533, 0.838, 0.593, 0.940, 0.629	0.1927, 0.1836, 0.2025, 0.1835, 0.1911	0.3645, 0.3407, 0.3688, 0.3464, 0.3472
	ค่าเฉลี่ย	72.2034	0.747	0.707	0.1907	0.3535
วิธีการสุ่มเกินเทคนิค SMOTE	10, 20, 30, 40, 50	79.6610, 80.2260, 79.6610, 78.7194, 78.7194	0.814, 0.875, 0.751, 0.688, 0.809	0.780, 0.729, 0.841, 0.882, 0.767	0.1649, 0.1624, 0.1628, 0.1684, 0.1701	0.2996, 0.2898, 0.3015, 0.2998, 0.3009
	ค่าเฉลี่ย	79.3974	0.787	0.8000	0.1657	0.2983
วิธีการสุ่มลด	10, 20, 30, 40, 50	72.0430, 64.5161, 51.0753, 55.3763, 54.3011	0.505, 0.699, 0.639, 0.563, 0.655	0.976, 0.611, 0.333, 0.544, 0.444	0.1804, 0.2271, 0.2748, 0.2597, 0.2590	0.3446, 0.3656, 0.4963, 0.4713, 0.4701
	ค่าเฉลี่ย	59.4624	0.612	0.582	0.2402	0.4296
วิธีการผสมผสาน	10, 20, 30, 40, 50	69.3410, 64.4699, 65.6160, 71.6332, 67.9083	0.476, 0.824, 0.492, 0.484, 0.842	0.899, 0.484, 0.833, 0.970, 0.497	0.2137, 0.2374, 0.2337, 0.2183, 0.2335	0.3821, 0.4079, 0.4673, 0.3453, 0.3869
	ค่าเฉลี่ย	67.7937	0.624	0.737	0.2273	0.3979
การจำแนกด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน						
วิธีการสุ่มเกิน	10, 20, 30, 40, 50	68.3616, 73.2580, 70.0565, 67.6083, 73.2580	0.764, 0.679, 0.733, 0.614, 0.759	0.614, 0.788, 0.671, 0.745, 0.709	0.3164, 0.2674, 0.2994, 0.3239, 0.2674	0.3164, 0.2674, 0.2994, 0.3239, 0.2674
	ค่าเฉลี่ย	70.5085	0.710	0.705	0.2949	0.2949
วิธีการสุ่มเกินเทคนิค SMOTE	10, 20, 30, 40, 50	74.3879, 79.0960, 73.8230, 72.3164, 75.1412	0.814, 0.887, 0.621, 0.581, 0.805	0.678, 0.695, 0.852, 0.860, 0.702	0.2561, 0.2090, 0.2618, 0.2768, 0.2486	0.2561, 0.2090, 0.2618, 0.2768, 0.2486
	ค่าเฉลี่ย	74.9529	0.742	0.757	0.2505	0.2505
วิธีการสุ่มลด	10, 20, 30, 40, 50	70.9677, 65.0538, 50.5376, 51.6129, 58.6022	0.634, 0.356, 0.806, 0.563, 0.552	0.800, 0.841, 0.090, 0.467, 0.616	0.2903, 0.3495, 0.4946, 0.4839, 0.4140	0.2903, 0.3495, 0.4946, 0.4839, 0.4140
	ค่าเฉลี่ย	59.3548	0.582	0.563	0.4065	0.4065
วิธีการสุ่มผสมผสาน	10, 20, 30, 40, 50	68.7679, 66.7622, 60.7450, 65.0430, 67.6218	0.553, 0.588, 0.547, 0.566, 0.832	0.816, 0.739, 0.673, 0.743, 0.503	0.3124, 0.3324, 0.3925, 0.3495, 0.3238	0.3123, 0.3324, 0.3926, 0.3496, 0.3238
	ค่าเฉลี่ย	65.7880	0.617	0.695	0.3421	0.3421

ตารางที่ 4 (ต่อ)

วิธีการปรับความไม่สมดุล	ตัวสร้างเลขสุ่มเทียม	ค่าความถูกต้อง	ค่าความไว	ค่าความจำเพาะ	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย	ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย
การจำแนกด้วยวิธีฐานกฎ						
วิธีการสุ่มเกิน	10, 20, 30, 40, 50	68.9266, 73.0697, 70.2448, 71.3748, 71.7514	0.915, 0.793, 0.837, 0.629, 0.656	0.495, 0.665, 0.582, 0.809, 0.773	0.1824, 0.1711, 0.1833, 0.1665, 0.1606	0.3721, 0.3401, 0.3706, 0.3288, 0.3215
	ค่าเฉลี่ย	71.0735	0.766	0.665	0.1728	0.3466
วิธีการสุ่มเกินเทคนิค SMOTE	10, 20, 30, 40, 50	79.4727, 80.9793, 82.2976, 77.4011, 83..0508	0.810, 0.834, 0.793, 0.704, 0.836	0.780, 0.786, 0.852, 0.841, 0.825	0.1487, 0.1401, 0.1370, 0.1636, 0.1351	0.3032, 0.2768, 0.2822, 0.3226, 0.2820
	ค่าเฉลี่ย	80.0377	0.795	0.817	0.1449	0.2934
วิธีการสุ่มลด	10, 20, 30, 40, 50	73.1183, 67.7419, 54.8387, 48.9247, 58.6022	0.584, 0.671, 0.815, 0.531, 0.667	0.906, 0.681, 0.179, 0.444, 0.515	0.1852, 0.4187, 0.2656, 0.2607, 0.2390	0.3685, 0.3565, 0.4976, 0.4963, 0.4641
	ค่าเฉลี่ย	60.6452	0.654	0.545	0.2738	0.4366
วิธีการสุ่มผสมผสาน	10, 20, 30, 40, 50	70.7736, 66.4756, 65.9026, 71.6332, 71.9198	0.506, 0.891, 0.459, 0.582, 0.967	0.899, 0.462, 0.875, 0.862, 0.442	0.1884, 0.1871, 0.1997, 0.1865, 0.1911	0.3718, 0.3796, 0.4030, 0.3638, 0.3757
	ค่าเฉลี่ย	69.3410	0.681	0.708	0.1906	0.3788
การจำแนกด้วยวิธีลาดลงสโตแคสติก						
วิธีการสุ่มเกิน	10, 20, 30, 40, 50	68.5499, 73.6347, 70.4331, 66.6667, 73.6347	0.756, 0.664, 0.745, 0.604, 0.775	0.625, 0.812, 0.668, 0.737, 0.701	0.3145, 0.2637, 0.2957, 0.3334, 0.2637	0.3145, 0.2637, 0.2957, 0.3333, 0.2637
	ค่าเฉลี่ย	70.5838	0.709	0.709	0.2942	0.2942
วิธีการสุ่มเกินเทคนิค SMOTE	10, 20, 30, 40, 50	76.4595, 80.4143, 78.3427, 72.693, 77.7778	0.810, 0.891, 0.686, 0.585, 0.836	0.722, 0.718, 0.878, 0.863, 0.724	0.2354, 0.1959, 0.2166, 0.2731, 0.2222	0.2354, 0.1959, 0.2166, 0.2731, 0.2222
	ค่าเฉลี่ย	77.1375	0.762	0.781	0.2286	0.2286
วิธีการสุ่มลด	10, 20, 30, 40, 50	70.4301, 63.4409, 50.5376, 55.3763, 55.9140	0.634, 0.384, 0.796, 0.583, 0.586	0.788, 0.796, 0.103, 0.522, 0.535	0.2957, 0.3655, 0.0106, 0.4462, 0.4409	0.2957, 0.3656, 0.4946, 0.4462, 0.4409
	ค่าเฉลี่ย	59.1398	0.597	0.549	0.3118	0.4086
วิธีการสุ่มผสมผสาน	10, 20, 30, 40, 50	67.3352, 68.4814, 61.6046, 67.3352, 68.4814	0.547, 0.618, 0.525, 0.588, 0.832	0.793, 0.745, 0.714, 0.766, 0.521	0.3266, 0.3152, 0.3839, 0.3266, 0.3152	0.3266, 0.3152, 0.3840, 0.3266, 0.3152
	ค่าเฉลี่ย	66.6476	0.622	0.708	0.3335	0.3335

4. สรุปผลการวิจัย ข้อเสนอแนะ และการนำไปใช้ประโยชน์

4.1 สรุปผลการวิจัย

งานวิจัยนี้ได้ศึกษาประสิทธิภาพในการปรับข้อมูลไม่สมดุล 4 วิธี ด้วยวิธีการจำแนก 5 วิธี คือ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีโครงข่ายประสาทเทียม วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีฐานกฎ และวิธีลาดลงสโตแคสติก ว่าวิธีการจำแนกข้อมูลดังกล่าวเมื่อปรับข้อมูลไม่สมดุล 4 วิธี คือ วิธีการสุ่มเกิน วิธีการสุ่มเกินโดยเทคนิค SMOTE วิธีการสุ่มลด และวิธีการสุ่มผสมผสาน ว่าวิธีใดมีประสิทธิภาพในการจำแนกและการปรับความไม่สมดุลของข้อมูลดีที่สุด โดยพิจารณาจากค่าความถูกต้อง ค่าความไว ค่าความจำเพาะ ค่าคลาดเคลื่อนกำลังสองเฉลี่ย และค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย โดยใช้ข้อมูลไม่สมดุล 3 ชุด

ชุดข้อมูลที่มีจำนวนข้อมูลในกลุ่มส่วนน้อยไม่เกินร้อยละ 5 คือ ชุดข้อมูลเคมีบำบัดมะเร็งลำไส้ใหญ่ระยะ B/C วิธีการจำแนกที่มีประสิทธิภาพสูงสุด คือ วิธีฐานกฎโดยการปรับความไม่สมดุลด้วยวิธีการสุ่มเกินเทคนิค SMOTE ส่วนชุดข้อมูลที่มีจำนวนข้อมูลในกลุ่มส่วนน้อยไม่เกินร้อยละ 10 คือ ชุดข้อมูลชุดข้อมูลโรคที่มีความผิดปกติของโปรตีน วิธีการจำแนกที่มีประสิทธิภาพสูงสุด คือ วิธีฐานกฎโดยการปรับความไม่สมดุลด้วยวิธีการสุ่มเกินเทคนิค SMOTE และชุดข้อมูลที่มีจำนวนข้อมูลในกลุ่มส่วนน้อยไม่เกินร้อยละ 25 คือ ชุดข้อมูลการรักษาอาการปวดศีรษะขั้นรุนแรง วิธีการจำแนกที่มีประสิทธิภาพสูงสุด คือ วิธีฐานกฎโดยการปรับความไม่สมดุลด้วยวิธีการสุ่มเกินเทคนิค SMOTE

4.2 ข้อเสนอแนะ

4.2.1 การปรับความไม่สมดุลของข้อมูลอาจจะเพิ่มจำนวนชุดข้อมูลให้มากขึ้น และหลากหลายสาขาวิชา ได้แก่ แพทยศาสตร์ วิทยาศาสตร์ สารสนเทศศาสตร์ เกษตรศาสตร์ วิศวกรรมศาสตร์ สังคมศาสตร์ เป็นต้น

4.2.2 การปรับความไม่สมดุลของข้อมูลอาจศึกษาวิธีการอื่น ๆ เช่น การเรียนรู้แบบมีค่าใช้จ่าย (cost sensitive learning)

4.2.3 เพื่อให้ได้ข้อสรุปที่มีความสมบูรณ์มากขึ้น สามารถวิเคราะห์ข้อมูลด้วยวิธีการอื่น ๆ เพิ่มเติม ได้แก่ วิธีนาอีฟเบส์ (Naïve-Bayes) วิธีการถดถอยลอจิสติก และวิธีเพอร์เซปตรอนให้คะแนน (voted perceptron) เป็นต้น

4.3 การนำไปใช้ประโยชน์

การปรับความไม่สมดุลของข้อมูลมีประโยชน์ทำให้ผลการแบ่งกลุ่มข้อมูลเกิดความผิดพลาดน้อยลง กล่าวคือ ข้อมูลที่อยู่ในกลุ่มส่วนน้อยไม่ถูกจัดให้ไปอยู่ในกลุ่มส่วนมากทั้งหมด

5. กิตติกรรมประกาศ

ขอขอบคุณ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ที่สนับสนุนการให้ทุนวิจัย เรื่อง การปรับความไม่สมดุลของข้อมูลโดยใช้การจำแนก 5 วิธี

6. รายการอ้างอิง

กีระชาติ สุขสุทธิ์, 2559, การจำแนกข้อมูลไม่สมดุลโดยใช้การปรับปรุงข้อมูลร่วมกับการหาค่าพารามิเตอร์ที่เหมาะสมด้วยขั้นตอนวิธีทางพันธุกรรมที่มีการเริ่มต้นใหม่, วิทยานิพนธ์ปริญญาเอก, มหาวิทยาลัยเทคโนโลยีสุรนารี, นครราชสีมา.

ธนาวุฒิ ประกอบผล, 2552, โครงข่ายประสาทเทียม, ว.มฉก.วิชาการ 12(24): 73-87.

วีระยุทธ มายุศิริ, จารี ทองคำ และวาทินี สุขมาก, 2557, การพัฒนาแบบจำลองเพื่อการพยากรณ์การรักษาซ้ำของผู้ป่วยโรคจิตเภทโดยเทคนิคเหมืองข้อมูล, ว.วิทยาศาสตร์และ

เทคโนโลยี มหาวิทยาลัยมหาสารคาม 10(พิเศษ): 144-153.

พัชรียา ทองพูล, พิมพ์ชนก จำเรือง และรมย์นลิน บุญฤทธิ์, 2561, การเปรียบเทียบประสิทธิภาพในการทำนายผลการปรับความไม่สมดุลของข้อมูลในการจำแนกด้วยเทคนิคการทำเหมืองข้อมูล, ปัญหาพิเศษปริญญาตรี, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง, กรุงเทพฯ.

พนิดา สมบัติมาก, ภัสสร จันทร์หอม, ศุภกร รัศมี และโอพาร รุ่งมณีธรรมคุณ, 2560, การเปรียบเทียบประสิทธิภาพในการจำแนกกลุ่มเมื่อข้อมูลมีค่านอกเกณฑ์ในการทำเหมืองข้อมูล, ปัญหาพิเศษปริญญาตรี, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง, กรุงเทพฯ.

ภรันยา ปาลวิสุทธิ์, 2559, การเพิ่มประสิทธิภาพเทคนิคต้นไม้ตัดสินใจบนชุดข้อมูลที่ไม่สมดุลโดยวิธีการการสุ่มเพิ่มตัวอย่างกลุ่มน้อยสำหรับสำหรับข้อมูลการเป็นโรคอินเตอร์เน็ต, ว.เทคโนโลยีสารสนเทศ 12(1): 54-63.

สายชล สินสมบูรณ์ทอง, 2560, การทำเหมืองข้อมูล เล่ม 1 : การค้นหาความรู้จากข้อมูล, พิมพ์ครั้งที่ 2, จามจุรีโปรดักส์ จำกัด, กรุงเทพฯ.

สุรวัชร ศรีเปารยะ และสายชล สินสมบูรณ์ทอง, 2560, การเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มการเป็นโรคไตเรื้อรัง : กรณีศึกษาโรงพยาบาลแห่งหนึ่งในประเทศอินเดีย, ว.วิทยาศาสตร์และเทคโนโลยี 25(5): 839-853.

สุรเดช บุญลือ, ชฎาพร สุขแจ่ม และศศิธร สนิทผล, 2554, การประยุกต์ใช้ซัพพอร์ทเว็กเตอร์แมชชีนในการทำนายการอยู่รอดของผู้ป่วยมะเร็งเต้านม, ศรีนครินทรวิโรฒวิชาการ ครั้งที่ 5, มหาวิทยาลัยศรีนครินทรวิโรฒ, กรุงเทพฯ.

เชาวนันท์ โสโท, พุธษดี ศิริแสงตระกูล และวรชัย ตั้งวรพงศ์ชัย, 2556, แบบจำลองการทำนายผลการรักษาผู้ป่วยมะเร็งปากมดลูกด้วยโครงข่ายประสาทเทียม, ว.วิจัย มข. 13(1): 39-50.

เบญจภรณ์ จันทรกองกุล, สุวรรณา รัศมีขวัญ, สุนิสา ริมเจริญ, ภูสิต กุลเกษม, กฤษณะ ชินสาร, อัณณ์นุพันธ์ รอดทุกข์, ปิยนุช วรบุตร และจรรยา อ้นปันส์, 2557, วิธีการที่เหมาะสมสำหรับการแบ่งกลุ่มข้อมูลที่ไม่สมดุลสูง, แหล่งที่มา : http://digital_collect .lib.buu.ac.th/dcms/files/2559_047.pdf, 10 พฤศจิกายน 2561.

Berson, A. and Smith, S.J., 1997, Data Warehousing, Data Mining and OLAP, McGraw-Hll, Inc., New York.

Hagan, M., Demuth, H. and Beale, M., 1996, Neural Network Design, Martin T. Hagan, Oklahoma.

Kostecki, T., Monette, G. and Wong, P., 1999, Treatment of Migraine Headaches, Available Source: https://vincentarelbundock.github.io/Rdataset/doc/carData/KosteckiDillon.html, February 7, 2019.

Kyle, R., Therneau, T., Rajkumar, V., Larson, D., Plevak, M. and Melton, L., 1994, Monoclonal Gammopathy, Available Source: https://vincentarelbundock.github.io/Rdataset/doc/survival/mgus.2.html, January 15, 2019.

Laurie, J., Moertel, C. and Lin, D., 1994, Chemotherapy for Stage B/C Colon Cancer, Available Source: https://vincentarelbundock.github.io/Rdataset/doc/survival/colon.html, January 15, 2019.

Murti, S. and Mahantappa, M., 2012, Using Rule

Based Classifiers for the Predictive Analysis of Breast Cancer Recurrence, Available Source: https://archive.ics.uci.edu/ml/data sets/pima+indians+diabetes, February 15, 2019.

Nektarios, T.G., 2013, Weka Classify Summary, Athens University of Economics and Business, Available Source: https://www. academia.edu/5167325/Weka_Classifiers_ Summary, January 10, 2019.

Rahman, M.M. and Davis, D.N., 2013, Addressing the class imbalance problem in medical datasets, Int. J. Mach. Learn. Comput. 3: 224-228.

Troyanskaya, O., 2001, Missing Value Estimation methods for DNA microarrays, Bionforma tics 17: 520-525.