Research Article

Received: August 27, 2018; Accepted: September 19, 2018

การเปรียบเทียบประสิทธิภาพในการทำนายผลความไม่สมดุลของข้อมูลในการจำแนกด้วยเทคนิคการทำเหมืองข้อมูล

An Efficiency Comparison in Prediction of Imbalanced Data Classification with Data Mining Techniques

สายชล สินสมบูรณ์ทอง*

ภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ถนนฉลองกรุง เขตลาดกระบัง กรุงเทพมหานคร 10520

Saichon Sinsomboonthong*

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

Chalongkrung Road, Ladkrabang, Bangkok 10520

บทคัดย่อ

การศึกษานี้เป็นการเปรียบเทียบประสิทธิภาพในการทำนายผลความไม่สมดุลของข้อมูลในการจำแนกด้วยเทคนิคการทำเหมืองข้อมูล วิธีการจำแนกที่นำมาเปรียบเทียบมี 7 วิธี ได้แก่ วิธีเพื่อนบ้านใกล้สุด k ตัว โดยใช้อัลกอริทึมชนิด IBk วิธีต้นไม้ตัดสินใจโดยใช้อัลกอริทึมชนิด J48 วิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึมชนิดเพอร์เซปตรอนแบบหลายชั้น วิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึม SMO ชนิดโพลิโนเมียลเคอร์เนล วิธีฐานกฎโดยใช้อัลกอริทึม decision table วิธีการถดถอยลอจิสติกทวิภาค และวิธีนาอีฟเบส์ การเปรียบเทียบประสิทธิภาพของวิธีการจำแนกจะพิจารณาจากค่าความถูกต้อง ค่าความไว ค่าความจำเพาะ ระยะเวลาในการประมวลผล และค่าคลาดเคลื่อนกำลังสองเฉลี่ย โดยใช้ชุดข้อมูล fertility, vertibral column และ diabetes ผลการศึกษาพบว่าชุดข้อมูล fertility วิธีการถดถอยลอจิสติกทวิภาคที่ random seed = 10, 20 และ 30 มีค่าความถูกต้อง ค่าความไว ค่าความจำเพาะ และค่าคลาดเคลื่อนกำลังสองเฉลี่ยดีที่สุด คือ ร้อยละ 100, 1.0000, 1.0000 และ 0.00000 ตามลำดับ ส่วนชุดข้อมูล vertibral volumn วิธีเพื่อนบ้านใกล้สุด k ตัว ที่ random seed = 10, 20 และ 30 มีค่าความถูกต้อง ค่าความไว ค่าความจำเพาะ และค่าคลาดเคลื่อนกำลังสองเฉลี่ยดีที่สุด คือ ร้อยละ 100, 1.0000, 1.0000 และ 0.00024 ตามลำดับ และชุดข้อมูล diabetes วิธีเพื่อนบ้านใกล้สุด k ตัว ที่ random seed = 10, 20 และ 30 มีค่าความถูกต้อง ค่าความไว ค่าความจำเพาะ และค่าคลาดเคลื่อนกำลังสองเฉลี่ยดีที่สุด คือ ร้อยละ 100, 1.0000, 1.0000 และ 0.00004 ตามลำดับ และเมื่อพิจารณาข้อมูลทั้ง 3 ชุด ร่วมกัน พบว่าวิธีที่ดีที่สุดในการทำนายผล คือ วิธีเพื่อนบ้านใกล้สุด k ตัว

***ผู้รับผิดชอบบทความ :** saichon.ss49@gmail.com **doi:** 10.14456/tstj.2020.31

คำสำคัญ : วิธีเพื่อนบ้านใกล้สุด k ตัว; วิธีต้นไม้ตัดสินใจ; วิธีโครงข่ายประสาทเทียม; วิธีซัพพอร์ตเวกเตอร์แมชชีน; วิธีฐานกฎ; วิธีการถดถอยลอจิสติกทวิภาค; วิธีนาอีฟเบส์

Abstract

In this study, an efficiency comparison in prediction of imbalanced data classification with data mining techniques was compared. The seven classification methods were the following: (1) k-nearest neighbor method using IBk algorithm; (2) decision tree method using J48 algorithm; (3) neural network method using multilayer perceptron algorithm; (4) support vector machine method using polynomial kernel; (5) rule-based method using decision table algorithm; (6) binary logistic regression method; and (7) naïve Bayes method. The following efficiency comparison of classification were employed: accuracy, sensitivity, specificity, time and mean square error (MSE) using fertility, vertibral volumn and diabetes data set. The important results are as follows. The binary logistic regression method using random seed = 10, 20 and 30 showed the best accuracy, sensitivity, specificity, and MSE at 100 %, 1.0000, 1.0000 and 0.00000 respectively for fertility data set. The k-nearest neighbor method using random seed = 10, 20 and 30 showed the best accuracy, sensitivity, specificity, and MSE at 100 %, 1.0000, 1.0000 and 0.00024 respectively for vertebral column data set. The k-nearest neighbor method using random seed = 10, 20 and 30 showed the best accuracy, sensitivity, specificity, and MSE at 100 %, 1.0000, 1.0000 and 0.00004 respectively for diabetes data set. In the three data sets, the k-nearest neighbor method offered the best prediction method.

Keywords: k-nearest neighbor; decision tree; neural network; support vector machine; rule-based; binary logistic regression; naïve-Bayes

1. บทนำ

ความไม่สมดุลของข้อมูล (imbalanced data) หมายถึงแถว (instance) ที่เป็นบวกมีจำนวนมากกว่าแถวที่เป็นลบ แถวที่เป็นลบบ่อยครั้งแสดงคำตอบปกติเป็นส่วนใหญ่ และแถวที่เป็นบวกแสดงคำตอบไม่ปกติเป็นส่วนน้อย [1] ปัญหาข้อมูลไม่สมดุล (class imbalance) เป็นปัญหาที่เกิดขึ้นกับข้อมูลทุก ๆ ประเภท ข้อมูลที่เกิดขึ้นโดยทั่วไปแบ่งเป็น 2 ประเภท คือ ข้อมูลกลุ่มน้อย (minority class) เป็นข้อมูลที่มีอยู่จำนวนน้อยในชุดข้อมูล ส่วนอีกประเภทหนึ่ง คือ ข้อมูลกลุ่มมาก (majority class) เป็นกลุ่มข้อมูลที่มีอยู่จำนวนมากในชุดข้อมูล จากสาเหตุของกรณีที่มีข้อมูลที่ไม่สมดุลเกิดขึ้น ผลลัพธ์ที่ได้เมื่อจำแนกข้อมูล คือ ข้อมูลจะจัดจำแนกไปอยู่ในกลุ่มข้อมูลกลุ่มมาก [2]

ความไม่สมดุลข้อมูลนั้น มีนักวิจัยหลายท่านศึกษาเกี่ยวกับเรื่องนี้ ปี ค.ศ. 2004 Akbani และคณะศึกษาพบว่าวิธีซัพพอร์ตเวกเตอร์แมชชีนไม่ได้แย่สำหรับคำตอบไม่สมดุลจำนวนปานกลางเมื่อเปรียบ

เทียบเทคนิคการเรียนรู้เครื่องจักรอื่น ๆ แต่ไม่ดีสำหรับคำตอบที่ไม่สมดุลอย่างมากในสถานการณ์หลัง วิธีซัพพอร์ตเวกเตอร์แมชชีนจะจำแนกแถวทั้งหมดเป็นคำตอบส่วนใหญ่ วิธีการสุ่มตัวอย่างใช้เพื่อแก้ปัญหาความไม่สมดุล ซึ่งประกอบด้วยคำตอบส่วนใหญ่ที่อยู่ภายใต้การสุ่มตัวอย่าง และ/หรือ คำตอบส่วนน้อยที่ไม่ได้อยู่ภายใต้การสุ่มตัวอย่างเพื่อสร้างข้อมูลที่ไม่สมดุลให้ลดน้อยลง ปี ค.ศ. 2009 Chen [4] เปรียบเทียบวิธีการจำแนก 3 วิธี คือ วิธีนาอีฟเบส์ วิธีต้นไม้ตัดสินใจ และวิธีโครงข่ายประสาทเทียมสำหรับข้อมูลที่ทำให้สมดุลใหม่ การทดลองพบว่าเทคนิคการสุ่มข้อมูลอีกครั้งช่วยเพิ่มประสิทธิภาพการจำแนกสำหรับวิธีต้นไม้ตัดสินใจและวิธีโครงข่ายประสาทเทียม แต่ไม่ได้ช่วยเพิ่มประสิทธิภาพการจำแนกสำหรับวิธีนาอีฟเบส์ นอกจากนี้ในปี ค.ศ. 2013 Sobran และคณะ [5] เปรียบเทียบวิธีนาอีฟเบส์มาตรฐานกับวิธีเพื่อนบ้านใกล้สุด k ตัว วิธีซัพพอร์ตเวกเตอร์แมชชีน และวิธีโครงข่ายประสาทเทียมดัดแปลง โดยเปรียบเทียบกับข้อมูล UCI 3 ชุด คือ Herbaman's Survival, German Credit และ Pima Indian โดยใช้เมทริกซ์ G-mean พบว่าวิธีนาอีฟเบส์ไม่ได้มีประสิทธิภาพดีกว่าวิธีการอื่น ๆ และปี ค.ศ. 2015 Zhang และคณะ [6] วิเคราะห์การจำแนกข้อมูลไม่สมดุลของข้อมูล UCI ซึ่งมีอัตราส่วน ขนาด และความซับซ้อนที่ไม่สมดุลที่ต่างกัน การทดลองประกอบด้วยการเปรียบเทียบผลการจำแนกของวิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีนาอีฟเบส์ และวิธีต้นไม้ตัดสินใจ C4.5 เพื่อหาข้อดีและข้อเสีย โดยการหาค่าความไว (sensitivity) ค่าความจำเพาะ (specificity) ค่า G-mean และระยะเวลาในการประมวลผล พบว่าวิธีซัพพอร์ตเวกเตอร์แมชชีนดีกว่าวิธีนาอีฟเบส์และวิธีต้นไม้ตัดสินใจ C4.5 ในเทอมของค่าความไวหรือค่าความจำเพาะสำหรับข้อมูลทั้งหมด และมีความถูกต้องมากกว่าในเทอมของค่า G-mean เมื่อข้อมูลมีขนาดใหญ่

เนื่องจากข้อมูลไม่มีความสมดุลกันในคำตอบเป็นส่วนใหญ่และคำตอบเป็นส่วนน้อย ผู้วิจัยจึงให้ความสนใจในการเปรียบเทียบประสิทธิภาพในการทำนายผลความไม่สมดุลของข้อมูลในการจำแนกด้วยเทคนิคการทำเหมืองข้อมูล ซึ่งการเปรียบเทียบประสิทธิภาพในความไม่สมดุลกันในคำตอบของข้อมูลในการจำแนกนั้นขึ้นอยู่กับปัจจัยหลายอย่าง เช่น จำนวนแถว จำนวนแถวที่มีคำตอบเป็นส่วนใหญ่ (ข้อมูลกลุ่มมาก) จำนวนแถวที่มีคำตอบเป็นส่วนน้อย (ข้อมูลกลุ่มน้อย) จำนวนคุณลักษณะ (คอลัมน์) ปี และสาขา เป็นต้น โดยผู้วิจัยสนใจศึกษาการเปรียบเทียบประสิทธิภาพในการทำนายผลความไม่สมดุลของข้อมูลในการจำแนกทั้ง 7 วิธี คือ วิธีเพื่อนบ้านใกล้สุด k ตัว (k-nearest neighbor) วิธีต้นไม้ตัดสินใจ (decision tree) วิธีโครงข่ายประสาทเทียม (artificial neural network) วิธีซัพพอร์ตเวกเตอร์แมชชีน (support vector machine) วิธีฐานกฎ (rule-based) วิธีการถดถอยลอจิสติกทวิภาค (binary logistic regression) และวิธีนาอีฟเบส์ (naïve Bayes) เพื่อหาวิธีที่มีประสิทธิภาพในการทำนายผลความไม่สมดุลของข้อมูลในการจำแนกที่เหมาะสมต่อไป

2. วิธีการวิจัย

2.1 เครื่องมือที่ใช้ในการวิจัย

เครื่องมือที่ใช้ในการวิจัยครั้งนี้ใช้โปรแกรม WEKA (Waikato Environment for Knowledge Analysis) เวอร์ชั่น 3.9 ซึ่งเป็นโปรแกรมที่สามารถดาวน์โหลดได้จากเว็บไซต์ ซึ่งอยู่ภายใต้การควบคุมของ GPL license ซึ่งโปรแกรม WEKA ได้ถูกพัฒนามาจากภาษาจาวาทั้งหมด ซึ่งเป็นที่นิยมในการใช้งานด้านการทำเหมืองข้อมูล

2.2 การเก็บรวบรวมข้อมูล การแปลงข้อมูล และการแบ่งข้อมูล

2.2.1 การเก็บรวบรวมข้อมูล เก็บรวบรวมข้อมูลในสาขาการแพทย์และชีวการแพทย์จากฐานข้อมูล UCI ในเวปไซต์ที่ได้มีการรวบรวมไว้แล้ว จำนวน 3 ชุด ได้แก่ ชุดข้อมูล fertility ขนาดตัวอย่างที่ใช้ คือ 100 หน่วยตัวอย่าง ชุดข้อมูล vertebral column ขนาดตัวอย่างที่ใช้ คือ 310 หน่วยตัวอย่าง และชุดข้อมูล diabetes ขนาดตัวอย่างที่ใช้ คือ 768 หน่วยตัวอย่าง ดังตารางที่ 1

Table 1 Number of row, number of row with most answers, number of row with least answers, number of attributes, year and major of fertility, vertebral column and diabetes data set

Data	Number of rows	Number of rows with most answers	Number of rows with least answers	Number of attributes	Year	Major
Fertility	100	88	12	10	2013	biomedical
Vertebral column	310	210	100	6	2011	biomedical
Diabetes	768	500	268	8	2012	medical

โดยตารางที่ 1 อธิบายได้ดังนี้ (1) แถว (instance) แปรผันจากจำนวนแถวน้อย คือ fertility มีจำนวนแถว 100 แถว จนถึงจำนวนแถวมาก คือ diabetes มีจำนวนแถว 768 แถว (2) ข้อมูลกลุ่มมาก (majority class) มีจำนวนแถวในคำตอบ (class) เป็นส่วนใหญ่ เช่น ข้อมูล fertility มีคำตอบกลุ่มมากจำนวนถึง 88 ค่า (แถว) (3) ข้อมูลกลุ่มน้อย (minority class) มีจำนวนแถวในคำตอบ (class) เป็นส่วนน้อย เช่น ข้อมูล fertility มีคำตอบกลุ่มน้อยจำนวนเพียง 12 ค่า (แถว) (4) คุณลักษณะ (attribute) แทนจำนวนคุณลักษณะ เช่น ชุดข้อมูล fertility มีคุณลักษณะต่าง ๆ ทั้งหมด 10 คุณลักษณะ (คอลัมน์) (5) ปี (year) ตั้งแต่ปี ค.ศ. 2011, 2012 และ 2013 และ (6) สาขา (area) ประกอบด้วยสาขาชีวการแพทย์ (biomedical) และการแพทย์ (medical)

อัตราส่วนความไม่สมดุลแปรผันจากความไม่สมดุลเล็กน้อย คือ ชุดข้อมูล diabetes 500 : 268 จนถึงความไม่สมดุลมาก คือ ชุดข้อมูล fertility 88 : 12

2.2.2 การแปลงข้อมูล โดยแปลงข้อมูลที่อยู่ในฐานข้อมูล UCI ลงในโปรแกรม Microsoft Excel โดยชุดข้อมูล fertility, vertebral column และ diabetes ไม่ต้องจัดเตรียมข้อมูล เนื่องจากข้อมูลมีความสมบูรณ์ครบถ้วน

2.2.3 การแบ่งข้อมูล สุ่มด้วยโปรแกรม WEKA จำนวน 3 รอบ โดยกำหนดตัวเลขสุ่มเทียม (random seed) เป็น 10, 20 และ 30 แล้วนำข้อมูลทั้งหมดมาแบ่งเป็น 2 ส่วน การตรวจสอบความไม่สมดุลของข้อมูลในการจำแนกเพื่อใช้สำหรับการวิเคราะห์ข้อมูล โดยแบ่งข้อมูลเป็นสัดส่วนดังนี้ [7] (1) ส่วนที่ 1 ข้อมูลชุดฝึกหัด (training data set) เพื่อนำไปสร้างตัวแบบ มีข้อมูลร้อยละ 80 ของข้อมูลทั้งหมด โดยมีชุดข้อมูล fertility จำนวน 80 คน ชุดข้อมูล vertebral column จำนวน 248 คน และชุดข้อมูล diabetes จำนวน 615 คน และ (2) ส่วนที่ 2 ข้อมูลชุดทดสอบ (testing data set) เพื่อนำไปทำนายตัวแบบ มีข้อมูลร้อยละ 20 ของข้อมูลทั้งหมด โดยมีชุดข้อมูล fertility จำนวน 20 คน ชุดข้อมูล vertebral column จำนวน 62 คน และชุดข้อมูล diabetes จำนวน 153 คน

2.3 การวิเคราะห์ข้อมูล

การวิจัยครั้งนี้ได้จัดข้อมูลแต่ละชุดเป็น 2 ส่วน โดยส่วนที่ 1 ข้อมูลชุดฝึกหัด ใช้ข้อมูลร้อยละ 80 ของข้อมูลทั้งหมดในการสร้างตัวแบบ ส่วนที่ 2 ข้อมูลชุดทดสอบ ใช้ข้อมูลร้อยละ 20 ของข้อมูลทั้งหมดในการทำนายตัวแบบ แล้วแปลงไฟล์ข้อมูลให้เป็นนามสกุล *.csv เพื่อใช้วิเคราะห์ประสิทธิภาพการจำแนกข้อมูลในโปรแกรม WEKA ซึ่งเป็นโปรแกรมที่สามารถนำมาทดสอบอัลกอริทึมของวิธีการจำแนก เนื่องจากมีอัลกอริทึมที่ได้ระบุไว้ให้เลือกใช้ในโปรแกรมครบตามที่กำหนด โดยกำหนดวิธีการจำแนกเพื่อนำมาทดสอบดังนี้

2.3.1 วิธีเพื่อนบ้านใกล้สุด k ตัว ใช้อัลกอริทึมชนิด IBk เนื่องจากเป็นฟังก์ชันหลักที่สนใจ ซึ่งเป็นพื้นฐานของอัลกอริทึม 8.1 อัลกอริทึม IBk ยังสามารถกำหนดน้ำหนัก ระยะห่างและทางเลือก (option) เพื่อกำหนดค่า k โดยใช้ cross-validation [8]

2.3.2 วิธีต้นไม้ตัดสินใจ ใช้อัลกอริทึมชนิด J48 ซึ่งพัฒนามาจาก ID3 สามารถใช้ได้กับข้อมูลแบบไม่ต่อเนื่องและแบบต่อเนื่อง ต่างจาก ID3 ที่ใช้ได้เพียงข้อมูลแบบไม่ต่อเนื่องเท่านั้น [9]

2.3.3 วิธีโครงข่ายประสาทเทียม ใช้อัลกอริทึมชนิดเพอร์เซปตรอนหลายชั้น (multilayer perceptron) โดยกำหนดค่าอัตราการเรียนรู้ (learning rate) เป็น 0.1 ค่าโมเมนตัม (momentum) เป็น 0.9 จำนวนรอบการสอน (training time) 20,000 รอบ การวิจัยครั้งนี้ใช้อัลกอริทึมของวิธีโครงข่ายประสาทเทียมชนิดเพอร์เซปตรอนหลายชั้นที่มีชั้นซ่อน (hidden layer) 1 ชั้น แม้ว่าโครงสร้างโครงข่ายประสาทเทียมที่ซับซ้อนสามารถมีชั้นซ่อนมากกว่า 1 ชั้น แต่ในทางปฏิบัตินั้นการกำหนดชั้นซ่อน 1 ชั้น ก็เพียงพอต่อการวิเคราะห์ข้อมูล [10]

2.3.4 วิธีซัพพอร์ตเวกเตอร์แมชชีน ใช้อัลกอริทึม SMO ชนิดโพลิโนเมียลเคอร์เนล (polynomial kernel) อ้างอิงจากงานวิจัยของ วาทินี และคณะ [11] ที่พบว่าวิธีซัพพอร์ตเวกเตอร์แมชชีนที่ใช้อัลกอริทึมชนิดโพลิโนเมียลเคอร์เนลดีที่สุด

2.3.5 วิธีฐานกฎ ใช้ชุดลำดับของกฎมาสร้างรูปแบบการแยกประเภทข้อมูล โดยส่วนใหญ่แล้วจะใช้กฎที่เป็น If ... then ซึ่งเป็นกฎอย่างง่าย ใช้อัลกอริทึม decision table เป็นเครื่องมือที่ใช้แสดงเงื่อนไขการตัดสินใจและเลือกการทำงานหรือกระทำกิจกรรมภายใต้เหตุการณ์ของเงื่อนไขที่ระบุ วิธีการตัดสินใจแบบ decision table จะเป็นตาราง 2 มิติ [12]

2.3.6 วิธีการถดถอยลอจิสติกทวิภาค เป็นการวิเคราะห์การถดถอยแบบหนึ่งโดยที่ตัวแปรตามเป็นตัวแปรเชิงคุณภาพ ส่วนตัวแปรอิสระอาจเป็นตัวแปรเชิงปริมาณหรือเชิงคุณภาพ หรืออาจมีทั้งตัวแปรเชิงปริมาณและตัวแปรเชิงคุณภาพก็ได้ [13]

2.3.7 วิธีนาอีฟเบส์ คือ อัลกอริทึมที่ใช้หลักการของความน่าจะเป็นในการคัดกรองแต่ละคำตอบ (class) โดยมีคำตอบ 2 คำตอบ [14]

การนำผลการวิเคราะห์ข้อมูลมาประเมินผลเพื่อเปรียบเทียบประสิทธิภาพในการทำนายผลความไม่สมดุลของข้อมูลในการจำแนกด้วยเทคนิคการทำเหมืองข้อมูลของวิธีการจำแนก 7 วิธี ว่าวิธีใดมีค่าความถูกต้อง (accuracy) ค่าความไว (sensitivity) ค่าความจำเพาะ (specificity) สูงที่สุด และมีระยะเวลาในการประมวลผล (time) และค่าคลาดเคลื่อนกำลังสองเฉลี่ย (mean square error, MSE) ต่ำที่สุด ซึ่งจะเป็นวิธีที่มีประสิทธิภาพในการทำนายผลดีที่สุด ดังตารางที่ 2

ตารางที่ 2 แสดงเมทริกซ์ความสับสน ซึ่งใช้เป็นค่าพื้นฐานในการวัดประสิทธิภาพในการทำนาย โดยที่ true positive (TP) คือ จำนวนข้อมูลที่จำแนก

Table 2 Confusion matrix

Class of real value	Class of predicted value	
	A	B
A	TP	FN
B	FP	TN

ถูกว่าเป็นชั้น A; true negative (TN) คือ จำนวนข้อมูลที่จำแนกถูกว่าเป็นชั้น B; false positive (FP) คือ จำนวนข้อมูลที่จำแนกผิดว่าเป็นชั้น A แต่ชั้นที่แท้จริงเป็นชั้น B; false negative (FN) คือ จำนวนข้อมูลที่จำแนกผิดว่าเป็นชั้น B แต่ชั้นที่แท้จริงเป็นชั้น A

โดยมีสูตรในการคำนวณค่าต่าง ๆ คือ (1) ค่าความถูกต้อง = จำนวนข้อมูลที่จำแนกถูกว่าเป็นชั้น A และ B ÷ จำนวนข้อมูลทั้งหมด = (TP + TN) ÷ (TP + TN + FP + FN) (2) ค่าความไว = TP ÷ (TP + FN) (3) ค่าความจำเพาะ = TN ÷ (FP + TN) และ (4) ค่าคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) = $\sum_{i=1}^{n}\frac{(y_i-\hat{y}_i)^2}{n} = \sum_{i=1}^{n}\frac{e_i^2}{n}$ โดยที่ y_i แทนค่าที่แท้จริง และ $\hat{y}_i$ แทนค่าที่ทำนายได้

3. ผลการวิจัย

3.1 ผลการเปรียบเทียบประสิทธิภาพในการทำนายผลของวิธีการจำแนก

การเปรียบเทียบประสิทธิภาพในการทำนายผลความไม่สมดุลของข้อมูลในการจำแนกด้วยเทคนิคการทำเหมืองข้อมูล ระหว่างวิธีเพื่อนบ้านใกล้สุด k ตัว วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีฐานกฎ วิธีการถดถอยลอจิสติกทวิภาค และวิธีนาอีฟเบส์ โดยพิจารณาจากค่าความถูกต้อง ค่าความไว ค่าความจำเพาะ ระยะเวลาในการประมวลผล และค่าคลาดเคลื่อนกำลังสองเฉลี่ย ใช้ชุดข้อมูล fertility, vertebral volume และ diabetes ได้ผลดังตารางที่ 3, 4 และ 5

ตารางที่ 3 พบว่าวิธีการจำแนกที่ให้ค่าความถูกต้องสูงสุด คือ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีโครงข่ายประสาทเทียม วิธีการถดถอยลอจิสติกทวิภาค และวิธีนาอีฟเบส์ที่ random seed = 10, 20 และ 30 โดยพบว่า ทุกวิธีการจำแนกให้ค่าความไวสูงสุดเท่ากันที่

Table 3a Accuracy, sensitivity, specificity, processing time and mean square error in prediction of imbalanced data classifications for fertility data set

Classification	Accuracy (percentage)			Sensitivity			Specificity		
	Random seed			Random seed			Random seed		
	10	20	30	10	20	30	10	20	30
k-Nearest neighbor	***100***	***100***	***100***	***1.0000***	***1.0000***	***1.0000***	***1.0000***	***1.0000***	***1.0000***
Decision tree	90	90	95	***1.0000***	***1.0000***	***1.0000***	0.0000	0.0000	0.0000
Artificial neural network	***100***	***100***	***100***	***1.0000***	***1.0000***	***1.0000***	***1.0000***	***1.0000***	***1.0000***
Support vector machine	90	90	95	***1.0000***	***1.0000***	***1.0000***	0.0000	0.0000	0.0000
Rule based	90	90	95	***1.0000***	***1.0000***	***1.0000***	0.0000	0.0000	0.0000
Binary logistic regression	***100***	***100***	***100***	***1.0000***	***1.0000***	***1.0000***	***1.0000***	***1.0000***	***1.0000***
Naïve Bayes	***100***	***100***	***100***	***1.0000***	***1.0000***	***1.0000***	***1.0000***	***1.0000***	***1.0000***

Table 3b Accuracy, sensitivity, specificity, processing time and mean square error in prediction of imbalanced data classifications for fertility data set (ext.)

Classification	Processing time (second)			Mean square error		
	Random seed			Random seed		
	10	20	30	10	20	30
k-Nearest neighbor	0.0200	0.0300	***0.0100***	0.00207	0.00207	0.00207
Decision tree	***0.0100***	***0.0100***	***0.0100***	0.09000	0.09000	0.04748
Artificial neural network	0.0200	***0.0100***	***0.0100***	0.00001	0.00001	0.00001
Support vector machine	0.0200	0.0200	***0.0100***	0.09998	0.09998	0.05000
Rule based	0.0300	0.0300	***0.0100***	0.09181	0.09181	0.04955
Binary logistic regression	***0.0100***	0.0200	0.0200	***0.00000***	***0.00000***	***0.00000***
Naïve Bayes	0.0200	0.0200	***0.0100***	0.00294	0.00018	0.00000

Table 4a Accuracy, sensitivity, specificity, processing time and mean square error in prediction of imbalanced data classifications for vertebral column data set

Classification	Accuracy (percentage)			Sensitivity			Specificity		
	Random seed			Random seed			Random seed		
	10	20	30	10	20	30	10	20	30
k-Nearest neighbor	***100***	***100***	***100***	***1.0000***	***1.0000***	***1.0000***	***1.0000***	***1.0000***	***1.0000***
Decision tree	90.3226	95.1613	95.1613	0.9778	0.9767	***1.0000***	0.7059	0.8947	0.8000
Artificial neural network	95.1613	91.9355	96.7742	***1.0000***	0.8837	0.9787	0.8235	***1.0000***	0.9333
Support vector machine	72.5806	69.3548	75.8065	***1.0000***	***1.0000***	***1.0000***	0.0000	0.0000	0.0000
Rule based	90.3226	91.9355	87.0968	0.9556	0.9302	0.8723	0.7647	0.8947	0.8667
Binary logistic regression	87.0968	83.8710	93.5484	0.9111	0.9070	0.9787	0.7647	0.6842	0.8000
Naïve Bayes	79.0323	85.4839	83.8710	0.7778	0.8140	0.8298	0.8235	0.9474	0.8667

random seed = 10, 20 และ 30 วิธีการจำแนกที่ให้ค่าความ จำเพาะสูงสุด คือ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีโครงข่ายประสาทเทียม วิธีการถดถอยลอจิสติกทวิภาค และวิธีนาอีฟเบส์ที่ random seed = 10, 20 และ 30 วิธีการจำแนกที่ให้ระยะเวลาในการประมวลผลต่ำสุด คือ วิธีต้นไม้ตัดสินใจที่ random seed = 10, 20 และ 30 และวิธีการจำแนกที่ให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ วิธีการถดถอยลอจิสติกทวิภาคที่ random seed = 10, 20 และ 30

ตารางที่ 4 พบว่าวิธีเพื่อนบ้านใกล้สุด k ตัว ที่ random seed = 10, 20 และ 30 ให้ค่าความถูกต้อง ค่าความไวและค่าความจำเพาะสูงสุด และยังให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด วิธีซัพพอร์ตเวกเตอร์แมชชีนที่ random seed = 10, 20 และ 30 ให้ค่าความไวสูงสุด ส่วนวิธีโครงข่ายประสาทเทียมที่ random seed = 10, 20 และ 30 ให้ระยะเวลาในการประมวลผลต่ำสุด

Table 4b Accuracy, sensitivity, specificity, processing time and mean square error in prediction of imbalanced data classifications for vertebral column data set (ext.)

Classification	Processing time (second)			Mean square error		
	Random seed			Random seed		
	10	20	30	10	20	30
k-Nearest neighbor	0.07000	***0.04000***	0.04000	***0.00024***	***0.00024***	***0.00024***
Decision tree	0.06000	***0.04000***	0.04000	0.08579	0.03849	0.04456
Artificial neural network	***0.05000***	***0.04000***	***0.03000***	0.04631	0.06812	0.03226
Support vector machine	0.07000	***0.04000***	0.05000	0.27416	0.30647	0.24197
Rule based	0.06000	0.05000	0.05000	0.08839	0.07081	0.09766
Binary logistic regression	0.06000	***0.04000***	0.05000	0.08916	0.09272	0.05664
Naïve Bayes	***0.05000***	***0.04000***	0.04000	0.18888	0.11621	0.09691

Table 5a Accuracy, sensitivity, specificity, processing time and mean square error in prediction of imbalanced data classifications for diabetes data set

Classification	Accuracy (percentage)			Sensitivity			Specificity		
	Random seed			Random seed			Random seed		
	10	20	30	10	20	30	10	20	30
k-Nearest neighbor	***100***	***100***	***100***	***1.0000***	***1.0000***	***1.0000***	***1.0000***	***1.0000***	***1.0000***
Decision tree	84.3137	94.7712	86.9281	0.9615	0.9412	0.7018	0.7822	0.9510	0.9688
Artificial neural network	81.0458	76.1634	84.3137	0.8077	0.6078	0.7544	0.8119	0.8235	0.8958
Support vector machine	73.2026	79.7386	78.4314	0.3654	0.5490	0.5263	0.9208	0.9216	0.9375
Rule based	79.7386	82.3529	79.7386	0.5962	0.6471	0.5088	0.9010	0.9118	0.9688
Binary logistic regression	74.5098	80.3922	80.3922	0.5000	0.6275	0.6491	0.8713	0.8922	0.8958
Naïve Bayes	77.1242	79.0850	78.4314	0.5385	0.6863	0.6667	0.8911	0.8431	0.8542

ตารางที่ 5 พบว่าวิธีเพื่อนบ้านใกล้สุด k ตัว ที่ random seed = 10, 20 และ 30 ให้ค่าความถูกต้อง ค่าความไว และค่าความจำเพาะสูงสุด วิธีการถดถอยลอจิสติกทวิภาคที่ random seed = 10, 20 และ 30 ใช้ระยะเวลาในการประมวลผลต่ำสุด และวิธีเพื่อนบ้านใกล้สุด k ตัว ที่ random seed = 10, 20 และ 30 ให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด

4. สรุปผลการวิจัยและข้อเสนอแนะ

4.1 สรุปผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพในการทำนายผลความไม่สมดุลของข้อมูลในการจำแนกด้วยเทคนิคการทำเหมืองข้อมูลของชุดข้อมูล fertility, vertebral column และ diabetes โดยใช้วิธีจำแนก 7 วิธี คือ วิธีเพื่อนบ้านใกล้สุด k ตัว

Table 5b Accuracy, sensitivity, specificity, processing time and mean square error in prediction of imbalanced data classifications for diabetes data set (ext.)

Classification	Processing time (second)			Mean square error		
	Random seed			Random seed		
	10	20	30	10	20	30
k-Nearest neighbor	0.1300	0.1200	0.0900	***0.00004***	***0.00004***	***0.00004***
Decision tree	0.1100	0.0700	0.0900	0.09666	0.00082	0.10036
Artificial neural network	0.1200	0.1200	0.0800	0.15016	0.19963	0.13264
Support vector machine	0.1000	0.1100	0.1200	0.26801	0.20259	0.21567
Rule based	0.12000	0.0800	0.0700	0.14033	***0.13148***	0.15952
Binary logistic regression	***0.0000***	***0.0000***	***0.0000***	0.15976	***0.12996***	0.15085
Naïve Bayes	0.0100	0.1100	0.0900	0.15264	0.16040	0.15233

วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีฐานกฎ วิธีการถดถอยลอจิสติกทวิภาค และวิธีนาอีฟเบส์ โดยพิจารณาจากประสิทธิภาพของวิธีการจำแนกค่าความถูกต้อง ค่าความไว ค่าความจำเพาะ ที่มีค่าสูงที่สุด และพิจารณาจากระยะเวลาในการประมวลผลและค่าคลาดเคลื่อนกำลังสองเฉลี่ยที่มีค่าต่ำที่สุด โดยใช้หลักการของการทำเหมืองข้อมูลมาใช้ในการจำแนกข้อมูล ผลการวิจัยพบว่าชุดข้อมูล fertility วิธีการถดถอยลอจิสติกทวิภาคมีประสิทธิภาพในการทำนายผลดีที่สุด 4 ค่า คือ ค่าความถูกต้อง ค่าความไว ค่าความจำเพาะ และค่าคลาดเคลื่อนกำลังสองเฉลี่ย ส่วนชุดข้อมูล vertibral column วิธีเพื่อนบ้านใกล้สุด k ตัว มีประสิทธิภาพในการทำนายผลดีที่สุด 4 ค่า คือ ค่าความถูกต้อง ค่าความไว ค่าความจำเพาะ และค่าคลาดเคลื่อนกำลังสองเฉลี่ย และชุดข้อมูล diabetes วิธีเพื่อนบ้านใกล้สุด k ตัว มีประสิทธิภาพในการทำนายผลดีที่สุด 4 ค่า คือ ค่าความถูกต้อง ค่าความไว ค่าความจำเพาะ และค่าคลาดเคลื่อนกำลังสองเฉลี่ย และถ้าพิจารณาข้อมูลทั้ง 3 ชุด ร่วมกัน วิธีที่ดีที่สุดในการทำนายผล คือ วิธีเพื่อนบ้านใกล้สุด k ตัว

4.2 ข้อเสนอแนะ

4.2.1 เพื่อให้ได้ข้อสรุปของผลการวิเคราะห์ข้อมูลที่มีความครอบคลุมมากขึ้น อาจใช้วิธีการจำแนกโดยใช้อัลกอริทึมประเภทอื่น ๆ เช่น วิธีเพื่อนบ้านใกล้สุด k ตัว อาจใช้อัลกอริทึม KStar และ LWL วิธีต้นไม้ตัดสินใจอาจใช้อัลกอริทึม decision stump, LMT, random forest, random tree และ REP tree วิธีโครงข่ายประสาทเทียมใช้อัลกอริทึมเพอร์เซปตรอนหลายชั้น อาจกำหนดอัตราการเรียนรู้เป็น 0.2 และค่าโมเมนตัมเป็น 0.8 วิธีซัพพอร์ตเวกเตอร์แมชชีนใช้อัลกอริทึม SMO และอาจใช้ฟังก์ชันเคอร์เนลแบบ normalized polykernel, Puk และ RBF kernel วิธีฐานกฎอาจใช้อัลกอริทึม JRip, OneR, PART และ ZeroR วิธีนาอีพเบส์อาจใช้ naïve-Bayes updatable เป็นต้น

4.2.2 ชุดข้อมูลที่นำมาใช้ในงานวิจัยนี้เป็นเพียงส่วนหนึ่งของการทำนายผลความไม่สมดุลเท่านั้น เพื่อให้การเลือกวิธีการจำแนกด้วยเทคนิคการทำเหมืองข้อมูลที่ดี ควรเพิ่มชุดข้อมูลในด้านอื่น ๆ อีก เช่น การ

ศึกษา สังคม ธุรกิจ และการเงินการธนาคาร

5. References

[1] Cao, P., Zhao, D. and Zaiane, O., 2013, An optimized cost-sensitive SVM for imbalanced data learning, Int. Conf. Adv. Knowl. Disc. Comp. Sci. 78: 280-292.

[2] Boonchuay, K., Sinapiromsaran, K. and Lursinsap, C., 2011, Minority split and gain ratio for a class imbalance, Int. Conf. Fuz. Sys. Knowl. Disc. 8: 2060-2064.

[3] Akbani, R., Kwek, S. and Japkowicz, N., 2004, Applying support vector machines to imbalanced datasets. Eur. Conf. Mach. Learn. 32: 39-50.

[4] Chen, Y., 2009, Learning Classifiers from Imbalanced, Only Positive and Unlabelled Data Sets, Project Report for UC San Diego Data Mining Contest, Department of Computer Science, Iowa State University, Iowa, 78 p.

[5] Sobran, N.M.M., Ahmad, A. and Ibrahim, Z., 2013, Classification of imbalanced dataset using conventional Naïve Bayes classifier, Int. Conf. Artif. Intell. Comput. Sci. 10: 35-42.

[6] Zhang, S., Sadaoui, S. and Mouhoub, M., 2015, An empirical analysis of imbalanced data classification, J. Comp. Inform. Sci. 8: 151-162.

[7] Panichkul, P., 2005, Development Data Mining System by Decision Tree, Work System Development Project, Master Thesis, King Montkut's Institute of Technology Ladkrabang, Bangkok, 62 p. (in Thai)

[8] Wu, X. and Kumar, V., 2009, The Top Ten Algorithms in Data Mining, Department of Computer Science and Engineering, University of Minnesota, CRC Press, Minneapolis, 215 p.

[9] Thammasombut, R., 2012, Decision Support System for Selection the Mobile Internet Package Using Decision Tree, Major of Business Computer, Faculty of Business Administration, Rajapruek College, Sakon Nakhon, 77 p. (in Thai)

[10] Berson, A. and Smith, S.J., 1997, Data Warehousing, Data Mining, and OLAP, McGraw-Hill, New York, 612 p.

[11] Nuipian, V., 2010, Comparison of Efficiency and Analysis of Data Classification using Artificial Neural Network, Support Vector Machine, Naïve Bayes and k-Nearest Neighbor, Department of Information Technology, Faculty of Information Technology, King Montkut's University of Technology North Bangkok, Bangkok, 85 p. (in Thai)

[12] Murti, S. and Mahantappa, M., 2012, Using rule based classifiers for the predictive analysis of breast cancer recurrence, J. Inform. Eng. Appl. 2(2): 12-19.

[14] Vanichbuncha, K., 2009, Multivariate Analysis, Thammasan Co., Ltd., Bangkok, 589 p. (in Thai)

[15] Sinsomboonthong, S., 2017, Data Mining 1: Discovering Knowledge in Data, 2nd Ed., Chamchuree Products Co., Ltd., Bangkok, 512 p. (in Thai)