Research Article

Received: May 21, 2018; Accepted: July 6, 2018

การเปรียบเทียบประสิทธิภาพในการจำแนก เมื่อข้อมูลมีค่านอกเกณฑ์ในการทำเหมืองข้อมูล

Performance Comparison of Data Mining's Classification Methods on Data Set with Outliers

พนิดา สมบัติมาก, ภัสสร จันทร์หอม*, ศุภกร รัศมี,
โอฬาร รุ่งมณีธรรมคุณ และสายชล สินสมบูรณ์ทอง
ภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ถนนฉลองกรุง เขตลาดกระบัง กรุงเทพมหานคร 10520

Phanida Sombatmak, Patsorn Janhom*, Suphakorn Ratsamee,
Oran Rungmaneethummakun and Saichon Sinsomboonthong
Department of Statistics, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang,
Chalongkrung Road, Ladkrabang, Bangkok 10520

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพในการจำแนก 5 วิธี คือ วิธีนาอีฟเบส์ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม และวิธีซัพพอร์ตเวกเตอร์แมชชีน โดยพิจารณาจากค่าความถูกต้อง ค่าคลาดเคลื่อนกำลังสองเฉลี่ยและค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย และเพื่อเปรียบเทียบวิธีการสุ่มตัวอย่างระหว่างโปรแกรม SPSS และ WEKA โดยแบ่งข้อมูลเป็นชุดข้อมูลเรียนรู้ ชุดข้อมูลตรวจสอบความถูกต้อง และชุดข้อมูลทดสอบ ในอัตราส่วน 70, 20 และ 10 ตามลำดับ สำหรับการค้นคว้าและศึกษาค่านอกเกณฑ์ได้ใช้ข้อมูลมีข้อมูล 3 ชุด คือ โรคมะเร็งเต้านมของรัฐวิสคอนซิน เป็นชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับต่ำ โรคเบาหวานของชาวพิม่า ประเทศอินเดีย เป็นชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับปานกลาง และการชำระเงินด้วยบัตรเครดิตของลูกค้า เป็นชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับสูง โดยใช้เครื่องมือ Highlight Exceptions ในการตรวจจับค่านอกเกณฑ์ จากการเปรียบเทียบข้อมูลโรคมะเร็งเต้านมของรัฐวิสคอนซิน วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีโครงข่ายประสาทเทียม โดยการสุ่มของโปรแกรม SPSS โรคเบาหวานของชาวพิม่า ประเทศอินเดีย วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีเพื่อนบ้านใกล้สุด k ตัว โดยการสุ่มของโปรแกรม SPSS และ WEKA และการชำระเงินด้วยบัตรเครดิตของลูกค้า วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีเพื่อนบ้านใกล้สุด k ตัว โดยการสุ่มของโปรแกรม SPSS และ WEKA ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับปานกลางและสูงให้ผลการจำแนกที่เหมือนกัน ซึ่งแตกต่างจากชุดข้อมูลที่มีค่านอกเกณฑ์ในระดับที่ต่ำ

***ผู้รับผิดชอบบทความ :** patsorn.mook1995@gmail.com

คำสำคัญ : ค่านอกเกณฑ์; วิธีนาอีฟเบส์; วิธีเพื่อนบ้านใกล้สุด k ตัว; วิธีต้นไม้ตัดสินใจ; วิธีโครงข่ายประสาทเทียม; วิธีซัพพอร์ตเวกเตอร์แมชชีน

Abstract

The objectives of this study were to evaluate and compare the performances of 5 classification methods: Naïve Bayes, k-nearest neighbors, decision tree, artificial neural network, and support vector machine and to compare the sampling methods by SPSS and WEKA. The performance measures were prediction accuracy, mean squared error, and mean absolute deviation. In sampling methods comparison, the data sets used were a data set on the prevalence of breast cancer in Wisconsin, USA, another data set on the prevalence of diabetes in Pima people, India, and another one on Taiwanese customer's payment through credit card. Each of these data sets were divided into three smaller sets: training, validating, and testing sets at a proportion of 70 : 20 : 10. Using Highlight Exceptions add-in to examine outliers. For the prevalence of breast cancer data set, the best classification method was the artificial neural network method in combination with the SPSS sampling method. For both the prevalence of diabetes and payment through credit card data sets, the best classification method was the k-nearest neighbors' method in combination with either SPSS or WEKA sampling method. The data sets that had a moderate to high number of outliers favored the same classification method while the data set that had a low number of outliers did not favor the same classification method as those two mentioned above.

Keywords: outlier; Naïve Bayes; k-nearest neighbors; decision tree; artificial neural network; support vector machine

1. บทนำ

ปัจจุบันข้อมูลที่มีคุณภาพน่าเชื่อถือมีความสำคัญเป็นอย่างมากในการวิเคราะห์ข้อมูล เพื่อการนำข้อมูลไปใช้ประโยชน์ได้สูงสุด บางครั้งจากข้อมูลจะพบว่าข้อมูลมีค่าที่มากเกินไปหรือน้อยเกินไปแฝงอยู่ ซึ่งเรียกว่าค่านอกเกณฑ์ (outlier) เป็นค่าที่อยู่ปลายสุดซึ่งตกอยู่ใกล้กับขีดจำกัดของพิสัยข้อมูล การหาค่านอกเกณฑ์มีความสำคัญ เนื่องจากแสดงค่าความคลาดเคลื่อนในข้อมูล ถ้าเรานำข้อมูลที่มีค่านอกเกณฑ์ไปวิเคราะห์จะส่งผลให้เกิดความคลาดเคลื่อนของผลลัพธ์ข้อมูลที่ได้ การกระจายของข้อมูลและค่าเฉลี่ยของข้อมูลไม่ดี ส่งผลให้ข้อมูลไม่เป็นไปตามข้อกำหนดเบื้องต้น (assumption) ที่กำหนดไว้ในการวิเคราะห์และทำให้ไม่สามารถนำข้อมูลไปใช้ประโยชน์ได้อย่างสูงสุด สำหรับสาเหตุที่ทำให้เกิดค่านอกเกณฑ์ คือ ความคลาดเคลื่อนจากการแปรผันของข้อมูลที่เก็บรวบรวมมา ซึ่งเป็นความคลาดเคลื่อนที่ไม่สามารถควบคุมได้ ความคลาดเคลื่อนที่เกิดจากเครื่องมือที่ใช้วัดมีคุณภาพต่ำทำให้เกิดค่านอกเกณฑ์ ความคลาดเคลื่อนที่เกิดจากการบันทึกข้อมูลจากการปฏิบัติโดยไม่

ตรวจสอบให้ถี่ถ้วน เพื่อให้ได้ผลการวิเคราะห์ที่เชื่อถือได้ การตรวจสอบหาค่านอกเกณฑ์จึงเป็นสิ่งที่สำคัญก่อนการนำข้อมูลไปวิเคราะห์ เพื่อให้ได้ข้อมูลที่มีค่าความคลาดเคลื่อนน้อยที่สุด บางครั้งข้อมูลอาจไม่สามารถใช้งานอย่างเต็มประสิทธิภาพหรือตรงตามความต้องการ เราจึงมีการจำแนกของข้อมูล เพื่อให้ได้วิธีที่มีความเหมาะสมกับข้อมูลที่มีความแตกต่างกันไป ดังนั้นผู้ทำวิจัยจึงต้องเลือกวิธีการจำแนกให้เหมาะสมกับข้อมูล [1]

การศึกษางานวิจัยที่เกี่ยวข้อง [2] โดยศึกษางานวิจัยเกี่ยวกับโรคมะเร็งเต้านม ซึ่งมุ่งเน้นการค้นหาเทคนิคด้านเหมืองข้อมูล สร้างตัวแบบการวิเคราะห์โรคอัตโนมัติ ค้นหาอัลกอริทึมที่เหมาะสมที่สุดสำหรับฐานข้อมูลทางการแพทย์ โดยวัดประสิทธิภาพจากค่าความถูกต้อง พบว่าวิธีต้นไม้ตัดสินใจให้ประสิทธิภาพค่าความถูกต้องดีที่สุด คือ ร้อยละ 75.52 เมื่อเปรียบเทียบกับวิธีซัพพอร์ตเวกเตอร์แมชชีนและวิธีเพื่อนบ้านใกล้สุด k ตัว [3] ศึกษางานวิจัยเกี่ยวกับโรคมะเร็ง นำเทคนิคการทำเหมืองข้อมูลมาประยุกต์ใช้กับการตรวจวิเคราะห์การเกิดโรคมะเร็ง เพื่อนำกฎการจำแนกข้อมูลที่ได้ไปพัฒนาเป็นระบบตรวจวิเคราะห์ปัจจัยที่ส่งผลต่อการเกิดโรคมะเร็ง โดยวัดประสิทธิภาพจากค่าสัมบูรณ์ของความคลาดเคลื่อนเฉลี่ย พบว่าวิธีต้นไม้ตัดสินใจให้ค่าความถูกต้องสูงสุด คือ ร้อยละ 98.63 เมื่อเปรียบเทียบกับวิธีเพื่อนบ้านใกล้สุด k ตัว และวิธีนาอีฟเบส์ ซึ่งงานวิจัยเกี่ยวกับโรคมะเร็งเต้านมให้ผลที่สอดคล้องกัน [4] ศึกษางานวิจัยเกี่ยวกับโรคเบาหวานเป็นโรคที่มีสาเหตุเกิดจากโรคแทรกซ้อนของโรคเบาหวานโดยจำแนกเป็น 2 กลุ่ม คือ ผู้ที่เป็นโรคเบาหวานในจอประสาทตาและไม่เป็นโรคเบาหวานในจอประสาทตา พบว่าวิธีซัพพอร์ตเวกเตอร์แมชชีนให้ค่าความถูกต้อง ร้อยละ 97.61 มากกว่าวิธีโครงข่ายประสาทเทียม [5] ศึกษางานวิจัยเกี่ยวกับโรคเบาหวานนำเทคนิคการทำเหมืองข้อมูลมาเป็นเครื่องมือประเมินการเกิดโรคเบาหวานโดยไม่ต้องอาศัยการตรวจเลือด พบว่าวิธีโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับให้ค่าความถูกต้องมากที่สุด เมื่อเปรียบเทียบกับวิธีโครงข่ายประสาทเทียมแบบธรรมดาและวิธีนาอีฟเบส์ [6] ศึกษางานวิจัยเกี่ยวกับการให้คะแนนสินเชื่อ ใช้เทคนิคการทำเหมืองข้อมูลเพื่อลดความเสี่ยงสำหรับการให้เครดิตสินเชื่อแก่ลูกค้าที่มีความเสี่ยงในการผิดสัญญาหรือขาดการชำระเงินในการหาตัวแบบที่เหมาะสม พบว่าวิธีต้นไม้ตัดสินใจมีค่าความถูกต้องมากที่สุด คือ ร้อยละ 79.48 เมื่อเปรียบเทียบกับวิธีโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับและวิธีซัพพอร์ตเวกเตอร์แมชชีน [7] ศึกษางานวิจัยเกี่ยวกับสินเชื่อ ใช้เทคนิคการทำเหมืองข้อมูลเพื่อเป็นแนวทางการสนับสนุนการตัดสินใจการอนุมัติสินเชื่อของบริษัทได้อย่างมีประสิทธิภาพมากขึ้น ซึ่งมีส่วนช่วยลดปริมาณหนี้สินสูญได้ พบว่าวิธีต้นไม้ตัดสินใจให้ค่าความถูกต้องมากที่สุด คือ ร้อยละ 90.47 มากกว่าวิธีนาอีฟเบส์จะเห็นว่าในเรื่องการให้คะแนนสินเชื่อและบัตรเครดิตมีผลที่สอดคล้องกันว่าวิธีต้นไม้ตัดสินใจให้ผลดีที่สุดเมื่อเทียบกับวิธีอื่น

ดังนั้นบทความฉบับนี้จึงศึกษาการจำแนกด้วยวิธีต่าง ๆ 5 วิธี คือ วิธีนาอีฟเบส์ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม และวิธีซัพพอร์ตเวกเตอร์แมชชีน เพื่อเปรียบเทียบประสิทธิภาพวิธีการจำแนกทั้ง 5 วิธี ว่าวิธีใดมีประสิทธิภาพและเหมาะสมกับรูปแบบของชุดข้อมูล โดยใช้วิธีการสุ่มตัวอย่างระหว่างโปรแกรม SPSS กับ WEKA เพื่อเปรียบเทียบว่าโปรแกรมใดให้ค่าความถูกต้อง รวมถึงการเปรียบเทียบค่าคลาดเคลื่อนกำลังสองเฉลี่ยและค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย

2. วิธีการจำแนก

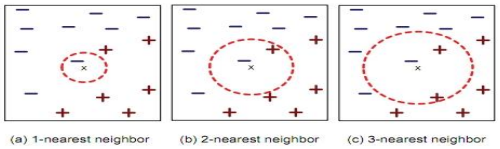
วิธีจำแนกข้อมูลด้วยคุณลักษณะต่าง ๆ ที่ได้มีการกำหนดไว้แล้ว วิธีนี้เหมาะกับการสร้างตัวแบบเพื่อการพยากรณ์ค่าข้อมูล (predictive modeling)

2.1 วิธีนาอีฟเบส์

วิธีนาอีฟเบส์เป็นเครื่องจักรเรียนรู้ที่อาศัยหลักการความน่าจะเป็น (probability) ตามทฤษฎีของเบส์ (Bayes' theorem) ซึ่งมีอัลกอริทึมที่ไม่ซับซ้อนเป็นขั้นตอนวิธีในการจำแนกข้อมูล โดยการเรียนรู้ปัญหาที่เกิดขึ้น เพื่อนำมาสร้างเงื่อนไขการจำแนกข้อมูลใหม่ เป็นการจำแนกข้อมูลโดยใช้ความน่าจะเป็นและคำนวณการแจกแจงความน่าจะเป็นตามสมมติฐานที่ตั้งให้กับข้อมูล จากการคำนวณตัวอย่างใหม่ที่ได้จะถูกนำมาปรับเปลี่ยนการแจกแจง ซึ่งมีผลต่อการเพิ่มหรือลดความน่าจะเป็นของข้อมูล จากข้อมูลเดิมตัวแบบจะถูกปรับเปลี่ยนความน่าจะเป็นใหม่ โดยมีการเพิ่มและลดความน่าจะเป็น โดยผนวกกับข้อมูลเดิมที่มีหลักการของนาอีฟเบส์ใช้การคำนวณหาความน่าจะเป็นซึ่งถูกใช้ในการทำนายผลเป็นวิธีในการแก้ปัญหาแบบการจำแนกที่สามารถคาดการณ์ผลลัพธ์และวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ นาอีฟเบส์เป็นวิธีจำแนกข้อมูลที่มีประสิทธิภาพ มีอัลกอริทึมในการทำงานที่ไม่ซับซ้อน เหมาะกับกรณีของเซตตัวอย่างที่มีจำนวนมากและสมบัติ (attribute) ของตัวอย่างไม่ขึ้นต่อกัน [8]

2.2 วิธีเพื่อนบ้านใกล้สุด k ตัว

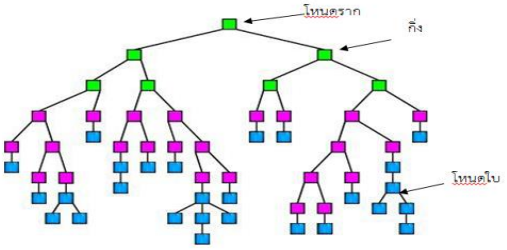
วิธีเพื่อนบ้านใกล้สุด k ตัว เป็นวิธีการที่ได้รับความนิยมในการใช้งานอย่างมาก เนื่องจากเป็นวิธีการที่ง่ายและมีประสิทธิภาพ ซึ่งสามารถนำไปประยุกต์ใช้กับงานอย่างหลากหลาย เช่น งานด้านการจำแนก รวมถึงงานด้านการแทนที่ข้อมูลที่สูญหาย (missing values imputation) บทความใช้อัลกอริทึม IBk [9] ดังรูปที่ 1



รูปที่ 1 ตัวอย่างของเพื่อนบ้านใกล้สุด k ตัว โดย (a) เพื่อนบ้านใกล้สุดโดยพิจารณาจากข้อมูล 1 ตัว, (b) เพื่อนบ้านใกล้สุดโดยพิจารณาจากข้อมูล 2 ตัว และ (c) เพื่อนบ้านใกล้สุดโดยพิจารณาจากข้อมูล 3 ตัว

2.3 วิธีต้นไม้ตัดสินใจ

วิธีต้นไม้ตัดสินใจเป็นการนำข้อมูลมาสร้างตัวแบบการพยากรณ์ในรูปแบบของโครงสร้างต้นไม้ ซึ่งมีการเรียนรู้ข้อมูลแบบมีผู้สอน (supervised learning) สามารถสร้างตัวแบบการจัดกลุ่ม (clustering) ได้จากกลุ่มตัวอย่างของชุดข้อมูลเรียนรู้ (training data set) ได้โดยอัตโนมัติและสามารถพยากรณ์กลุ่มของรายการที่ยังไม่เคยนำมาจัดกลุ่มอีกด้วย บทความนี้ใช้อัลกอริทึม J48 [10] ดังรูปที่ 2

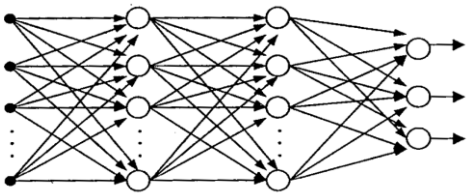


รูปที่ 2 ส่วนประกอบของต้นไม้ตัดสินใจ

2.4 วิธีโครงข่ายประสาทเทียม

โครงข่ายประสาทเทียมเป็นศาสตร์ที่จำลองแบบความสามารถของมนุษย์ด้านการเรียนรู้จดจำ และจำแนกสิ่งต่าง ๆ ซึ่งใช้สมองเป็นส่วนสำคัญ ในการประมวลระบบของโครงข่ายประสาทเทียมนั้นจะเลียนแบบการทำงานของระบบสมอง คือ มีการส่งผ่าน

ข้อมูลระหว่างกันโดยมีการเชื่อมต่อของเซลล์ประสาท (neuron) กันเป็นโครงข่ายร่างแหจำนวนมาก และประมวลผลในลักษณะขนาน (parallel processsing) สาเหตุหลักที่โครงข่ายประสาทเทียมเป็นที่นิยมกันมากขึ้นเนื่องจากมีความยืดหยุ่นในการทำงานสูงและสามารถปรับตัวเองให้ทำงานในสภาพที่เปลี่ยนแปลง อีกทั้งไม่จำเป็นต้องทราบตัวแบบทางคณิตศาสตร์ (mathematical model) ที่แน่นอนของกระบวนการ เพียงแต่ใช้ชุดข้อมูลที่ประกอบด้วยข้อมูลนำเข้า (input data) และข้อมูลเป้าหมาย (target data) ของกระบวนการในจำนวนมากพอที่ใช้ในการสอน (training) โครงข่ายประสาทเทียม บทความนี้ใช้อัลกอริทึมแบบเพอร์เซปตรอนหลายชั้น อัตราการเรียนรู้เป็น 0.1 ค่าโมเมนตัมเป็น 0.9 จำนวนรอบการสอน 20,000 และชั้นซ่อน 1 ชั้น [11] ดังรูปที่ 3



รูปที่ 3 โครงข่ายประสาทเทียมแบบเพอร์เซปตรอนหลายชั้น

2.5 วิธีซัพพอร์ตเวกเตอร์แมชชีน

วิธีซัพพอร์ตเวกเตอร์แมชชีนเป็นสมการที่ใช้ในการจำแนกค่าคุณลักษณะของ 2 กลุ่ม ที่วางตัวอยู่ในพื้นที่คุณลักษณะ (feature space) ออกจากกันโดยจะสร้างเส้นแบ่ง (plane) ที่เป็นเส้นตรงขึ้นมา และเพื่อให้ทราบว่าเส้นตรงที่แบ่ง 2 กลุ่ม ออกจากกันนั้นเส้นตรงใดที่เป็นเส้นที่ดีที่สุด โดยเส้นตรงนั้นจะเพิ่มเส้นขอบ (margin) ออกไปทั้งสองข้าง โดยเส้นขอบที่เพิ่มนั้นจะขนานกับเส้นเดิมเสมอ เส้นขอบที่เพิ่มขึ้นมานี้จะขยายออกไปจนกว่าจะสัมผัสกับค่าของกลุ่มตัวอย่างที่ใกล้ที่สุด บทความนี้ใช้อัลกอริทึม SMO ชนิดโพลิโนเมียลเคอร์เนล [12]

2.6 การเปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่ม

2.6.1 เมทริกซ์ความสับสน (confusion matrix) เป็นรูปแบบตารางที่เฉพาะเจาะจงที่นำผลลัพธ์จากการทำนายมาใส่ในรูปตารางเมทริกซ์ ซึ่งช่วยให้ง่ายต่อการมองเห็นค่าทำนายของอัลกอริทึม ดังรูปที่ 4

	ชั้นของค่าทำนาย	
ชั้นของค่าจริง	A	B
A	TP	FN
B	FP	TN

รูปที่ 4 เมทริกซ์ความสับสน

โดยที่ true positive (TP) คือ จำนวนข้อมูลที่จำแนกถูกว่าเป็นชั้น A, true negative (TN) คือ จำนวนข้อมูลที่จำแนกถูกว่าเป็นชั้น B, false positive (FP) คือ จำนวนข้อมูลที่จำแนกผิดว่าเป็นชั้น A ซึ่งชั้นที่แท้จริงเป็นชั้น B, false negative (FN) คือ จำนวนข้อมูลที่จำแนกผิดว่าเป็นชั้น B ซึ่งชั้นที่แท้จริงเป็นชั้น A, ค่าการทำนาย (prediction) คือ ส่วนที่แสดงผลการทำนายของแต่ละตัวอย่างของชุดข้อมูล และ ค่าการจำแนกได้ถูกต้อง (correctly classified instance) คือ ค่าที่บอกว่าชุดข้อมูลมีอัตราการทำนายถูกต้องและผิดพลาดเท่าไร

2.6.2 ค่าความถูกต้อง (accuracy) คือ การแสดงการวัดที่ได้มีความถูกต้องในรูปอัตราส่วน

Accuracy = จำนวนข้อมูลที่จำแนกถูกว่าเป็นชั้น A และ B ÷ จำนวนข้อมูลทั้งหมด

= (TP + TN) ÷ (TP + TN + FP + FN)

2.6.3 ค่าคลาดเคลื่อนกำลังสองเฉลี่ย เป็นมาตรวัดการประเมินค่าได้ดี เนื่องจากค่าคลาดเคลื่อนกำลังสองเฉลี่ยประกอบด้วยทั้งความเอนเอียงและความแปรปรวน [13]

$$\text{Mean Square Error}(\hat{\theta}) = \text{Var}(\hat{\theta}) + (\text{Bias}(\hat{\theta}))^2$$

$$\text{MSE} = \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{n}$$

โดยที่ y_i แทน ค่าจริงที่ I และ $\hat{y_i}$ แทน ค่าทำนายที่ i

2.6.4 ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย คือ ค่าวัดความถูกต้องของการพยากรณ์ที่วัดจากค่าคลาดเคลื่อนโดยไม่คำนึงถึงทิศทางของความคลาดเคลื่อน MAD มีหน่วยวัดหน่วยเดียวกับค่าสังเกต [13]

$$\text{MAD} = \frac{\sum_{i=1}^{n}|y_i - \hat{y_i}|}{n}$$

โดยที่ y_i แทน ค่าจริงที่ I และ $\hat{y_i}$ แทน ค่าทำนายที่ i

3. วิธีการดำเนินงานวิจัย

3.1 การเก็บรวบรวมข้อมูล

ค้นหาและศึกษาข้อมูลที่มีค่านอกเกณฑ์จากเว็บไซต์ UCI (University of California, Irvine) จากการศึกษาในการหาค่านอกเกณฑ์ พบว่าข้อมูลส่วนมากมีค่านอกเกณฑ์อยู่ระหว่างร้อยละ 0-10 จึงแบ่งข้อมูลเป็น 3 ระดับ คือ ค่านอกเกณฑ์ระดับต่ำ ร้อยละ 0-3.3 ค่านอกเกณฑ์ระดับปานกลาง ร้อยละ 3.4-6.7 ค่านอกเกณฑ์ระดับสูงร้อยละ 6.8-10 โดยได้ข้อมูล 3 ชุด คือ

3.1.1 โรคมะเร็งเต้านมของรัฐวิสคอนซิน (Breast Cancer Wisconsin) จำนวนข้อมูลทั้งหมด 699 ค่า พบค่านอกเกณฑ์จำนวน 13 ค่า คิดเป็นร้อยละ 1.85 เป็นชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่ต่ำ (จากเว็บไซต์ https://archive.ics.uci.edu/ml/datasets/ breast+cancer+wisconsin+(original))

3.1.2 โรคเบาหวานของชาวพิม่า ประเทศอินเดีย (Pima Indians Diabetes) จำนวนข้อมูลทั้งหมด 768 ค่า พบค่านอกเกณฑ์จำนวน 32 ค่า คิดเป็นร้อยละ 4.16 เป็นชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่ปานกลาง (จากเว็บไซต์ https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes)

3.1.3 การชำระเงินด้วยบัตรเครดิตของลูกค้า (Default of Credit Card Clients) จำนวนข้อมูลทั้งหมด 646 ค่า พบค่านอกเกณฑ์จำนวน 62 ค่า คิดเป็นร้อยละ 9.6 เป็นชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับที่สูง (จากเว็บไซต์ https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients)

3.2 ขั้นตอนการดำเนินงานวิจัย

3.2.1 ศึกษาข้อกำหนดเบื้องต้นและวิธีการหาค่านอกเกณฑ์ของข้อมูล

3.2.1.1 ติดตั้ง SQL Server 2014 ในโปรแกรม Excel

3.2.1.2 ติดตั้ง Data Mining Add-in สำหรับโปรแกรม Excel เพื่อสามารถใช้เครื่องมือวิเคราะห์ Data Mining เบื้องต้นในโปรแกรม Excel ได้

3.2.1.3 นำชุดข้อมูลที่รวบรวมได้มาวิเคราะห์ค่านอกเกณฑ์ โดยใช้เครื่องมือ Highlight Exceptions จะได้จำนวนค่านอกเกณฑ์ของข้อมูลแต่ละชุด

3.2.2 วิเคราะห์หาค่านอกเกณฑ์

วิเคราะห์ค่านอกเกณฑ์ จากโปรแกรม Microsoft Excel Data Mining Client สำหรับ Excel Add-in โดยใช้เครื่องมือ Highlight Exceptions ซึ่งเป็นอัลกอริทึมการจัดกลุ่มของ Microsoft รูปแบบการจัดกลุ่มจะตรวจจับกลุ่มของแถวที่มีลักษณะคล้ายกัน

3.2.3 วิธีการแบ่งข้อมูล

3.2.3.1 แบ่งชุดข้อมูลโดยโปรแกรม SPSS ด้วยวิธีการสุ่มตัวอย่างอย่างง่าย (simple

random sampling, SRS) สุ่มจำนวน 3 รอบ โดยการกำหนดตัวเลขสุ่มเทียม (random seed) เป็น 10, 20, 30 ในอัตราส่วน 70 : 20 : 10 ส่วนที่ 1 ข้อมูลเรียนรู้ (training data) นำไปสร้างตัวแบบร้อยละ 70 ข้อมูลส่วนที่ 2 ข้อมูลตรวจสอบ (validation data) นำไปประเมินความผิดพลาดของตัวแบบร้อยละ 20 และข้อมูลส่วนที่ 3 ข้อมูลทดสอบ (testing data) นำไปทดสอบตัวแบบร้อยละ 10

(1) เปิดไฟล์ชุดข้อมูลนามสกุล .CSV เข้าสู่โปรแกรม SPSS

(2) กำหนดค่า Seed → Transform Random Number Generators → ในช่อง Active Generator → เลือก Set active Generator → เลือก SPSS 12 Compatible

(3) ในช่อง Active Generator Initialization → เลือก Set Starting Point → เลือก Fixed Value → ในช่อง Value กำหนดเป็น 10, 20, 30 → OK

(4) เลือก Data → Select Case → Random Sample of Case Sample → Approximately ใส่จำนวนร้อยละที่ต้องการแบ่งข้อมูลชุดแรก (ในงานวิจัยนี้ใส่ ร้อยละ 70) → Continue → OK

(5) เมื่อดูในหน้าต่าง Data View จะเห็นว่าในคอลัมน์ filter_$ จะมีหมายเลข 1 จำนวนร้อยละ 70 ตามที่กำหนดไว้ข้างต้นเพื่อนำไปสร้างตัวแบบและที่เหลือหมายเลข 0 อีก ร้อยละ 30 เพื่อนำไปทดสอบตัวแบบ บันทึกข้อมูลที่แบ่งได้ นั่นคือ ร้อยละ 70 และร้อยละ 30 → นำข้อมูลร้อยละ 30 มาแบ่งอีกรอบเพื่อแบ่งเป็นร้อยละ 20 และร้อยละ 10

(6) เมื่อดูในหน้าต่าง Data View จะเห็นว่าในคอลัมน์ filter_$ จะมีหมายเลข 1 จำนวนร้อยละ 20 และที่เหลือหมายเลข 0 อีกร้อยละ 10 เพื่อนำไปทดสอบตัวแบบ บันทึกข้อมูลที่แบ่งได้ นั่นคือ ร้อยละ 20 และร้อยละ 10

3.2.3.2 แบ่งชุดข้อมูลโดยโปรแกรม WEKA สุ่มจำนวน 3 รอบ โดยการกำหนดตัวเลขสุ่มเทียมเป็น 10, 20, 30 ในอัตราส่วน 70 : 20 : 10 ส่วนที่ 1 ข้อมูลเรียนรู้ นำไปสร้างตัวแบบร้อยละ 70 ข้อมูลส่วนที่ 2 ข้อมูลตรวจสอบ นำไปประเมินความผิดพลาดของตัวแบบร้อยละ 20 และข้อมูลส่วนที่ 3 ข้อมูลทดสอบ นำไปทดสอบตัวแบบร้อยละ 10

(1) เปิดไฟล์ชุดข้อมูลนามสกุล .CSV เข้าสู่โปรแกรม WEKA → กด Explorer

(2) กำหนด Seed และแบ่งข้อมูล → Choose → Filters → Unsupervised → Instance → Resample

(3) คลิก Resample → ต้องการแบ่งข้อมูลร้อยละ 70 → กำหนด Invert Selection เป็น False → No Replacement เป็น True → Random Seed เป็น 10, 20, 30 ตามลำดับ → Sample Size Percent ร้อยละ 70 → Apply → Save เป็นไฟล์ร้อยละ 70 → Undo

(4) คลิก Resample → ต้องการแบ่งข้อมูลร้อยละ 20 → กำหนด Invert Selection เป็น True → No Replacement เป็น True → Sample Size Percent ร้อยละ 70 → Apply → กำหนด Invert Selection เป็น False → No Replacement เป็น True → Sample Size Percent ร้อยละ 66.67 → Apply → Save เป็นไฟล์ร้อยละ 20 → Undo

(5) คลิก Resample → ต้องการแบ่งข้อมูลร้อยละ 10 → กำหนด Invert Selection เป็น True → No Replacement เป็น True → Random Seed เป็น 10, 20, 30 ตามลำดับ → Sample Size Percent ร้อยละ 66.67 → Apply → Save เป็นไฟล์ร้อยละ 10

3.2.4 วิธีการจำแนกด้วยวิธีต่าง ๆ 5 วิธี คือ วิธีนาอีฟเบส์ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม และวิธีซัพพอร์ตเวกเตอร์แมชชีน โดยนำข้อมูลส่วนที่ 3 ข้อมูลทดสอบนำมาทดสอบตัวแบบร้อยละ 10

คลิก Classify → กด Choose เพื่อเลือกวิธีการจำแกที่ต้องการ → ในช่อง Test Options กด Use Training Set → กด Start → ในช่อง Classifier Output จะแสดงผลลัพธ์ ค่าความถูกต้อง ค่าคลาดเคลื่อนกำลังสองเฉลี่ย และค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย

4. ผลการวิจัย

ผลการเปรียบเทียบประสิทธิภาพของวิธีการจำแนกโดยใช้โปรแกรม SPSS และ WEKA

4.1 ชุดข้อมูลโรคมะเร็งเต้านมของรัฐวิสคอนซิน

เป็นชุดข้อมูลที่มีค่านอกเกณฑ์ระดับต่ำ

ตารางที่ 1 วิธีเพื่อนบ้านใกล้สุด k ตัว การสุ่มด้วยโปรแกรม SPSS วิธีต้นไม้ตัดสินใจ การสุ่มด้วยโปรแกรม WEKA วิธีโครงข่ายประสาทเทียม การสุ่มด้วยโปรแกรม SPSS ให้ค่าความถูกต้องสูงสุด คือ ร้อยละ 100.0000 วิธีโครงข่ายประสาทเทียม การสุ่มด้วยโปรแกรม SPSS ให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.0000 และให้ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยต่ำสุด คือ 0.0013

ตารางที่ 2 วิธีเพื่อนบ้านใกล้สุด k ตัว การสุ่มด้วยโปรแกรม SPSS และ WEKA และวิธีโครงข่ายประสาทเทียม การสุ่มด้วยโปรแกรม SPSS ให้ค่าความถูกต้องสูงสุด คือ ร้อยละ 100.0000 วิธีโครงข่ายประสาทเทียม การสุ่มด้วยโปรแกรม SPSS ให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.0000 และให้ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยต่ำสุด คือ 0.0013

ตารางที่ 3 วิธีเพื่อนบ้านใกล้สุด k ตัว การสุ่มด้วยโปรแกรม SPSS และ WEKA วิธีต้นไม้ตัดสินใจ การสุ่มด้วยโปรแกรม WEKA วิธีโครงข่ายประสาทเทียม การสุ่มด้วยโปรแกรม SPSS และ WEKA วิธีซัพพอร์ตเวกเตอร์แมชชีน การสุ่มด้วยโปรแกรม WEKA

ตารางที่ 1 ชุดข้อมูลโรคมะเร็งเต้านมของรัฐวิสคอนซิน โดยการสุ่มตัวอย่าง Random Seed 10

วิธีการจำแนก	โปรแกรม	ค่าความถูกต้อง	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย	ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย
นาอีฟเบส์	SPSS	97.1429	0.0281	0.0284
	WEKA	97.1429	0.0282	0.0285
เพื่อนบ้านใกล้สุด k ตัว	SPSS	100.0000	0.0002	0.0121
	WEKA	100.0000	0.0005	0.0115
ต้นไม้ตัดสินใจ	SPSS	98.5714	0.0114	0.0229
	WEKA	98.5714	0.0135	0.0271
โครงข่ายประสาทเทียม	SPSS	100.0000	0.0000	0.0013
	WEKA	98.5714	0.0136	0.0274
ซัพพอร์ตเวกเตอร์แมชชีน	SPSS	98.5714	0.0143	0.0143
	WEKA	97.1429	0.0286	0.0286

ตารางที่ 2 ชุดข้อมูลโรคมะเร็งเต้านมของรัฐวิสคอนซิน โดยการสุ่มตัวอย่าง Random Seed 20

วิธีการจำแนก	โปรแกรม	ค่าความถูกต้อง	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย	ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย
นาอีฟเบส์	SPSS	95.7143	0.0427	0.0428
	WEKA	94.2857	0.0608	0.0665
เพื่อนบ้านใกล้สุด k ตัว	SPSS	100.0000	0.0002	0.0117
	WEKA	100.0000	0.00016	0.0121
ต้นไม้ตัดสินใจ	SPSS	98.5714	0.0107	0.0214
	WEKA	94.2857	0.0510	0.1020
โครงข่ายประสาทเทียม	SPSS	100.0000	0.0000	0.0013
	WEKA	98.5714	0.0136	0.0272
ซัพพอร์ตเวกเตอร์แมชชีน	SPSS	98.5714	0.0143	0.0143
	WEKA	94.2857	0.0571	0.0571

ตารางที่ 3 ชุดข้อมูลโรคมะเร็งเต้านมของรัฐวิสคอนซิน โดยการสุ่มตัวอย่าง Random Seed 30

วิธีการจำแนก	โปรแกรม	ค่าความถูกต้อง	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย	ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย
นาอีฟเบส์	SPSS	95.7143	0.0430	0.0465
	WEKA	98.5714	0.0143	0.0143
เพื่อนบ้านใกล้สุด k ตัว	SPSS	100.0000	0.0002	0.0133
	WEKA	100.0000	0.0002	0.0119
ต้นไม้ตัดสินใจ	SPSS	98.5714	0.0139	0.0279
	WEKA	100.0000	0.0000	0.0000
โครงข่ายประสาทเทียม	SPSS	100.0000	0.0000	0.0013
	WEKA	100.0000	0.0000	0.0013
ซัพพอร์ตเวกเตอร์แมชชีน	SPSS	98.5714	0.0143	0.0143
	WEKA	100.0000	0.0000	0.0000

ให้ค่าความถูกต้องสูงสุด คือ ร้อยละ 100.0000 วิธีต้นไม้ตัดสินใจและวิธีซัพพอร์ตเวกเตอร์แมชชีน การสุ่มด้วยโปรแกรม WEKA ให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.0000 และให้ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยต่ำสุด คือ 0.0000

4.2 ชุดข้อมูลโรคเบาหวานของชาวพิม่า ประเทศอินเดีย

เป็นชุดข้อมูลที่มีค่านอกเกณฑ์ระดับปานกลาง

ตารางที่ 4 วิธีเพื่อนบ้านใกล้สุด k ตัว การสุ่มด้วยโปรแกรม SPSS และ WEKA ให้ค่าความถูกต้องสูงสุด คือ ร้อยละ 100.0000 ให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.0002 และให้ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยต่ำสุด คือ 0.0127

ตารางที่ 4 ชุดข้อมูลโรคเบาหวานของชาวพิม่า ประเทศอินเดีย โดยการสุ่มตัวอย่าง Random Seed 10

วิธีการจำแนก	โปรแกรม	ค่าความถูกต้อง	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย	ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย
นาอีฟเบส์	SPSS	75.3247	0.1648	0.2930
	WEKA	79.2208	0.1638	0.2409
เพื่อนบ้านใกล้สุด k ตัว	SPSS	100.0000	0.0002	0.0127
	WEKA	100.0000	0.0002	0.0127
ต้นไม้ตัดสินใจ	SPSS	87.0130	0.1041	0.2081
	WEKA	81.8182	0.1482	0.2964
โครงข่ายประสาทเทียม	SPSS	87.0130	0.1176	0.2227
	WEKA	88.3117	0.1002	0.1955
ซัพพอร์ตเวกเตอร์แมชชีน	SPSS	71.4286	0.2857	0.2857
	WEKA	75.3247	0.1648	0.2930

ตารางที่ 5 ชุดข้อมูลโรคเบาหวานของชาวพิม่า ประเทศอินเดีย โดยการสุ่มตัวอย่าง Random Seed 20

วิธีการจำแนก	โปรแกรม	ค่าความถูกต้อง	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย	ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย
นาอีฟเบส์	SPSS	72.7273	0.2031	0.2923
	WEKA	76.6234	0.1736	0.252
เพื่อนบ้านใกล้สุด k ตัว	SPSS	100.0000	0.0002	0.0127
	WEKA	100.0000	0.0002	0.0127
ต้นไม้ตัดสินใจ	SPSS	84.4156	0.1257	0.2514
	WEKA	88.3117	0.0892	0.1784
โครงข่ายประสาทเทียม	SPSS	84.4156	0.1321	0.2554
	WEKA	89.6104	0.0920	0.1751
ซัพพอร์ตเวกเตอร์แมชชีน	SPSS	75.3247	0.2467	0.2468
	WEKA	79.2208	0.2077	0.2078

ตารางที่ 5 วิธีเพื่อนบ้านใกล้สุด k ตัว การสุ่มด้วยโปรแกรม SPSS และ WEKA ให้ค่าความถูกต้องสูงสุด คือ ร้อยละ 100.0000 ให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.0002 และให้ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยต่ำสุด คือ 0.0127

ตารางที่ 6 วิธีเพื่อนบ้านใกล้สุด k ตัว การสุ่มด้วยโปรแกรม SPSS และ WEKA ให้ค่าความถูกต้องสูงสุด คือ ร้อยละ 100.0000 ให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.0002 และให้ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยต่ำสุด คือ 0.0127

ตารางที่ 6 ชุดข้อมูลโรคเบาหวานของชาวพิม่า ประเทศอินเดีย โดยการสุ่มตัวอย่าง Random Seed 30

วิธีการจำแนก	โปรแกรม	ค่าความถูกต้อง	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย	ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย
นาอีฟเบส์	SPSS	74.0260	0.1722	0.2723
	WEKA	81.8182	0.1443	0.2109
เพื่อนบ้านใกล้สุด k ตัว	SPSS	100.0000	0.0002	0.0127
	WEKA	100.0000	0.0002	0.0127
ต้นไม้ตัดสินใจ	SPSS	93.5065	0.0556	0.1112
	WEKA	87.0130	0.1092	0.2185
โครงข่ายประสาทเทียม	SPSS	85.7143	0.1220	0.2456
	WEKA	92.2078	0.0701	0.1387
ซัพพอร์ตเวกเตอร์แมชชีน	SPSS	72.7273	0.2727	0.2727
	WEKA	83.1169	0.1688	0.1688

4.3 ชุดข้อมูลการชำระเงินด้วยบัตรเครดิตของลูกค้า

เป็นข้อมูลที่มีค่านอกเกณฑ์ระดับสูง

ตารางที่ 7 วิธีเพื่อนบ้านใกล้สุด k ตัว การสุ่มด้วยโปรเกรม SPSS และ WEKA ให้ค่าความถูกต้องสูงสุด คือ ร้อยละ 100.000 ให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.0002 และให้ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยต่ำสุด คือ 0.0149

ตารางที่ 8 วิธีเพื่อนบ้านใกล้สุด k ตัว การสุ่มด้วยโปรแกรม SPSS และ WEKA ให้ค่าความถูกต้องสูงสุด คือ ร้อยละ 100.0000 ให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.0002 และให้ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยต่ำสุด คือ 0.0149

ตารางที่ 9 วิธีเพื่อนบ้านใกล้สุด k ตัว การ

ตารางที่ 7 ชุดข้อมูลการชำระเงินด้วยบัตรเครดิตของลูกค้า โดยการสุ่มตัวอย่าง Random Seed 10

วิธีการจำแนก	โปรแกรม	ค่าความถูกต้อง	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย	ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย
นาอีฟเบส์	SPSS	53.8462	0.4444	0.4579
	WEKA	67.6923	0.3085	0.3329
เพื่อนบ้านใกล้สุด k ตัว	SPSS	100.0000	0.0002	0.0149
	WEKA	100.0000	0.0002	0.0149
ต้นไม้ตัดสินใจ	SPSS	96.9231	0.0284	0.0568
	WEKA	95.3846	0.0393	0.0785
โครงข่ายประสาทเทียม	SPSS	92.3077	0.0700	0.1381
	WEKA	93.8462	0.0565	0.1113
ซัพพอร์ตเวกเตอร์แมชชีน	SPSS	81.5385	0.1846	0.1846
	WEKA	84.6154	0.1538	0.1538

ตารางที่ 8 ชุดข้อมูลการชำระเงินด้วยบัตรเครดิตของลูกค้า โดยการสุ่มตัวอย่าง Random Seed 20

วิธีการจำแนก	โปรแกรม	ค่าความถูกต้อง	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย	ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย
นาอีฟเบส์	SPSS	53.8462	0.4331	0.4491
	WEKA	72.3077	0.2236	0.2790
เพื่อนบ้านใกล้สุด k ตัว	SPSS	100.0000	0.0002	0.0149
	WEKA	100.0000	0.0002	0.0149
ต้นไม้ตัดสินใจ	SPSS	96.9231	0.0286	0.0571
	WEKA	89.2308	0.0900	0.1800
โครงข่ายประสาทเทียม	SPSS	87.6923	0.0823	0.1670
	WEKA	87.6923	0.1075	0.2062
ซัพพอร์ตเวกเตอร์แมชชีน	SPSS	75.3846	0.2461	0.2462
	WEKA	81.5385	0.1846	0.1846

ตารางที่ 9 ชุดข้อมูลการชำระเงินด้วยบัตรเครดิตของลูกค้า โดยการสุ่มตัวอย่าง Random Seed 30

วิธีการจำแนก	โปรแกรม	ค่าความถูกต้อง	ค่าคลาดเคลื่อนกำลังสองเฉลี่ย	ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย
นาอีฟเบส์	SPSS	60.0000	0.3714	0.4011
	WEKA	61.5385	0.3049	0.3465
เพื่อนบ้านใกล้สุด k ตัว	SPSS	100.0000	0.0002	0.0149
	WEKA	100.0000	0.0002	0.0149
ต้นไม้ตัดสินใจ	SPSS	89.2308	0.0960	0.1919
	WEKA	98.4615	0.0103	0.0205
โครงข่ายประสาทเทียม	SPSS	95.3846	0.0438	0.0857
	WEKA	96.9231	0.0298	0.0562
ซัพพอร์ตเวกเตอร์แมชชีน	SPSS	86.1538	0.1385	0.1385
	WEKA	80.0000	0.1999	0.2000

สุ่มด้วยโปรแกรม SPSS และ WEKA ให้ค่าความถูกต้องสูงสุด คือ ร้อยละ 100.000 ให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.0002 และให้ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ยต่ำสุด คือ 0.0149

5. สรุปผลการวิจัย

การค้นคว้าและศึกษาในการหาค่านอกเกณฑ์ได้ข้อมูล 3 ชุด โรคมะเร็งเต้านมของรัฐวิสคอนซิน Random Seed 10, 20 วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีโครงข่ายประสาทเทียมของโปรแกรม SPSS มีค่าความถูกต้อง คือ ร้อยละ 100.0000 ค่าคลาดเคลื่อนกำลังสองเฉลี่ย คือ 0.0000 ค่าส่วนเบี่ยงเบนสัมบูรณ์

เฉลี่ย คือ 0.0013 และ Random Seed 30 วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีต้นไม้ตัดสินใจและวิธีซัพพอร์ตเวกเตอร์แมชชีนของโปรแกรม WEKA มีค่าความถูกต้อง คือ ร้อยละ 100.0000 ค่าคลาดเคลื่อนกำลังสองเฉลี่ย คือ 0.0000 ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย คือ 0.0000 โรคเบาหวานของชาวพิม่า ประเทศอินเดีย Random Seed 10, 20, 30 วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีเพื่อนบ้านใกล้สุด k ตัวของโปรแกรม SPSS และ WEKA มีค่าความถูกต้อง คือ ร้อยละ 100.0000 ค่าคลาดเคลื่อนกำลังสองเฉลี่ย คือ 0.0002 ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย คือ 0.0127 และการชำระเงินด้วยบัตรเครดิตของลูกค้า Random Seed 10, 20, 30 วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีเพื่อนบ้านใกล้สุด k ตัวของโปรแกรม SPSS และ WEKA มีค่าความถูกต้อง คือ ร้อยละ 100.0000 ค่าคลาดเคลื่อนกำลังสองเฉลี่ย คือ 0.0002 ค่าส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย คือ 0.0149

ดังนั้นโรคมะเร็งเต้านมของรัฐวิสคอนซิน วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีโครงข่ายประสาทเทียม โดยการสุ่มของโปรแกรม SPSS โรคเบาหวานของชาวพิม่าประเทศอินเดีย วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีเพื่อนบ้านใกล้สุด k ตัว โดยการสุ่มของโปรแกรม SPSS และ WEKA และการชำระเงินด้วยบัตรเครดิตของลูกค้า วิธีที่มีประสิทธิภาพสูงสุด คือ วิธีเพื่อนบ้านใกล้สุด k ตัว โดยการสุ่มของโปรแกรม SPSS และ WEKA ชุดข้อมูลที่มีค่านอกเกณฑ์อยู่ในระดับปานกลางและสูงให้ผลการจำแนกที่เหมือนกัน ซึ่งแตกต่างจากชุดข้อมูลที่มีค่านอกเกณฑ์ในระดับที่ต่ำ

6. ข้อเสนอแนะ

การวิเคราะห์ค่านอกเกณฑ์ ไม่มีหลักการแบ่งที่แน่นอน ผู้วิจัยได้ค้นหาข้อมูลจำนวนหนึ่งมาตรวจสอบค่านอกเกณฑ์ พบว่ามีค่าอยู่ระหว่าง 0-10 จึงแบ่งข้อมูลออกเป็น 3 ช่วง คือ ต่ำ ปานกลาง สูง ซึ่งอาจมีข้อมูลอื่น ๆ ที่มีค่านอกเกณฑ์ไม่ได้อยู่ในช่วงดังกล่าว

7. รายการอ้างอิง

[1] วรพรรณ เจริญขำ, 2556, การตรวจสอบค่านอกเกณฑ์ในตัวอย่างสุ่มจากประชากรปรกติ, วิทยานิพนธ์ปริญญาโท, สถาบันบัณฑิตพัฒนบริหารศาสตร์, กรุงเทพฯ.

[2] นิเวศ จิระวิชิตชัย, 2553, การค้นหาเทคนิคเหมืองข้อมูลเพื่อสร้างโมเดลการวิเคราะห์โรคอัตโนมัติ, มหาวิทยาลัยราชภัฏสวนสุนันทา, กรุงเทพฯ.

[3] Sriwiboon, N., 2016, A comparative efficiency of data mining algorithms for analysis of factors affecting the cancer, SNRU J. Sci. Technol. 8: 344-352.

[4] Priya, R. and Aruna, P., 2012, SVM and neural network based diagnosis of diabetic retinopathy, Int. J. Comp. Appl. 41: 6-12.

[5] กิตติพล วิแสง, สิรภัทร เชี่ยวชาญวัฒนา และคำรณ สุนัติ, 2552, การวิเคราะห์ปัจจัยเสี่ยงของโรคเบาหวาน, การประชุมวิชาการแห่งชาติทางด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 5, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพฯ.

[6] เดช ธรรมศิริ, วาทินี นุ้ยเพียร, ภัทราวุฒิ แสงศิริ, ภรัณยา อำมฤครัตน์, ณรงค์ โพธิ และพยุง มีสัจ, 2552, การให้คะแนนสินเชื่อโดยวิธีการทำเหมืองข้อมูลด้วยเทคนิคซับพอร์ตเวกเตอร์แมชชีน รวมทั้งการเลือกใช้ลักษณะที่เหมาะสมร่วมกับการหาค่าพารามิเตอร์ที่เหมาะสมด้วยวิธีค้นหาแบบกริช, การประชุมวิชาการระดับชาติด้าน

คอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 5, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพฯ.

[7] ทิพย์ธิดา วงศ์พิพันธ์, 2555, การใช้เหมืองข้อมูลช่วยในการตัดสินใจการให้สินเชื่อ, วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยธุรกิจบัณฑิตย์, กรุงเทพฯ.

[8] วรรณสิริ ธุระชน, วรพจน์ สุเมธาวัฒนพงศ์ และณัฐวิภา ส่งสุข, 2557, ระบบการจำแนกพันธุ์ยางพาราโดยใช้ตัวจำแนกนาอีฟเบย์, สาขาวิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏอุดรธานี, อุดรธานี.

[9] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R.B., 2001, Missing values estimation methods for DNA microarrays, Bioinformatics 17: 520-525.

[10] รุจิรา ธรรมสมบัติ, 2554, ระบบสนับสนุนการตัดสินใจในการเลือกใช้แพคเกจอินเทอร์เน็ตมือถือโดยใช้ต้นไม้ตัดสินใจ, สาขาคอมพิวเตอร์ธุรกิจ คณะบริหารธุรกิจ วิทยาลัยราชพฤกษ์, กรุงเทพฯ.

[11] วาทินี นุ้ยเพียร, พยุง มีสัจ และเดช ธรรมศิริ, 2553, การเปรียบเทียบประสิทธิภาพและวิเคราะห์การจำแนกข้อมูลด้วยโครงข่ายประสาทเทียม ซัพพอร์ตเวกเตอร์แมชชีน นาอีฟเบย์ และแครเนียรเรสต์เนเบอร์, การประชุมวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 5, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพฯ.

[12] จิรา แก้วสุวรรณ์, 2549, การตรวจจับและการแก้ไขการวางตัวของภาพโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน, วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพฯ.

[13] สายชล สินสมบูรณ์ทอง, 2558, การทำเหมืองข้อมูล Data Mining, จามจุรี โปรดัก (จำกัด), กรุงเทพฯ.