การเปรียบเทียบประสิทธิภาพในการแทนค่าข้อมูลสูญหาย โดยวิธีการประมาณค่าการถดถอย วิธีการประมาณค่าทดแทนพหุ และวิธีค่าคาดหมายสูงสุด สำหรับการจำแนกในการทำเหมืองข้อมูล

Efficiency Comparison in Replace Missing Value Using Regression Imputation, Multiple Imputation and Expectation Maximization for Classification in Data Mining

ดวงแก้ว หุ่นทอง, ธีริศรา เงินวิลัย* และสายชล สินสมบูรณ์ทอง

ภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ถนนฉลองกรุง เขตลาดกระบัง กรุงเทพมหานคร 10520

Doungkaew Hunthong, Theeridsara Ngernwilai* and Saichon Sinsomboonthong

Department of Statistics, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Chalongkrung Road, Ladkrabang, Bangkok 10520

Received: April 12, 2020; Accepted: May 12, 2020

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อต้องการเปรียบเทียบประสิทธิภาพวิธีการแทนค่าข้อมูลสูญหาย 3 วิธี คือ วิธีการประมาณค่าการถดถอย วิธีการประมาณค่าทดแทนพหุ และวิธีค่าคาดหมายสูงสุด ด้วยการจำแนก 4 วิธี คือ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม และวิธีซัพพอร์ตเวกเตอร์แมชชีน โดยค้นคว้าและศึกษาการแทนค่าข้อมูลสูญหายจากข้อมูล 6 ชุด ชุดข้อมูลทดสอบมีดังนี้ ข้อมูลโรคตับในรัฐอานธรประเทศ ประเทศอินเดีย และข้อมูลการตรวจชิ้นเนื้อในผู้ป่วยมะเร็งเต้านม เป็นข้อมูลที่มีค่าสูญหายต่ำ ข้อมูลการศึกษาระยะยาวของสารภูมิต้านทานโมโนโคลน และข้อมูลการตลาดของธนาคาร เป็นข้อมูลที่มีค่าสูญหายปานกลาง ข้อมูลระดับสินเชื่อครอบครัวเดี่ยว และข้อมูลโรคหลอดเลือดหัวใจของผู้อยู่อาศัยในเมือง Framingham รัฐ Massachusetts เป็นข้อมูลที่มีค่าสูญหายสูง โดยใช้โปรแกรม SPSS ในการแทนค่าข้อมูลสูญหายว่าวิธีใดมีประสิทธิภาพในการจำแนกดีที่สุด โดยพิจารณาจากค่าความถูกต้อง 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 โดยใช้โปรแกรม WEKA พบว่าจำแนก

**Corresponding author: 59051199@kmitl.ac.th*

DOI: 10.14456/tjst.2020.62

ข้อมูลโรคตับในรัฐอานธรประเทศ ประเทศอินเดีย วิธีที่มีประสิทธิภาพสูงสุด คือ การจำแนกด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยวิธีการประมาณค่าการถดถอย วิธีการประมาณค่าทดแทนพหุ และวิธีค่าคาดหมายสูงสุด ชุดข้อมูลการตรวจชิ้นเนื้อในผู้ป่วยมะเร็งเต้านม วิธีที่มีประสิทธิภาพสูงสุด คือ การจำแนกด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยวิธีการแทนค่าสูญหายด้วยวิธีการประมาณค่าการถดถอย และวิธีค่าคาดหมายสูงสุด ข้อมูลการศึกษาระยะยาวของสารภูมิต้านทานโมโนโคลน วิธีที่มีประสิทธิภาพสูงสุด คือ การจำแนกด้วยวิธีโครงข่ายประสาทเทียม โดยวิธีการประมาณค่าทดแทนพหุ ข้อมูลการตลาดของธนาคาร วิธีที่มีประสิทธิภาพสูงสุด คือ การจำแนกด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยวิธีการแทนค่าสูญหายด้วยวิธีค่าคาดหมายสูงสุด ข้อมูลระดับสินเชื่อครอบครัวเดี่ยว วิธีที่มีประสิทธิภาพสูงสุด คือ การจำแนกด้วยวิธีต้นไม้ตัดสินใจโดยวิธีการแทนค่าสูญหายด้วยวิธีการประมาณค่าทดแทนพหุ และข้อมูลโรคหลอดเลือดหัวใจของผู้อยู่อาศัยในเมือง Framingham รัฐ Massachusetts วิธีที่มีประสิทธิภาพสูงสุด คือ การจำแนกด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยวิธีการประมาณค่าการถดถอย วิธีการประมาณค่าทดแทนพหุ และวิธีค่าคาดหมายสูงสุด

คำสำคัญ : ข้อมูลสูญหาย; วิธีการประมาณค่าการถดถอย; วิธีการประมาณค่าทดแทนพหุ; วิธีค่าคาดหมายสูงสุด; วิธีเพื่อนบ้านใกล้สุด k ตัว; วิธีต้นไม้ตัดสินใจ; วิธีโครงข่ายประสาทเทียม; วิธีซัพพอร์ตเวกเตอร์แมชชีน

Abstract

The objective of this research was to compare the efficiencies of three missing value replacement methods, i.e. regression imputation, multiple imputation, and expectation maximization using four classification methods including K-nearest neighbor, decision tree, artificial neural network and support vector machine, on six datasets with some missing values. The tested datasets were the followings: a dataset of liver disease in Andhra Pradesh, India, and a dataset of biopsy data on breast cancer patients, which had the least amount of missing value; a dataset of monoclonal gammopathy data, and a dataset of issued and non-issued credit cards by a bank, which had a moderate amount of missing value; and a dataset of single family loan-level and a dataset of cardiovascular disease in Framingham, Massachusetts, which had the highest amount of missing value. By offered in SPSS software program, the metrics that indicated the efficiency of a classification method were its accuracy, mean squared error and mean absolute error. Each of these data sets was divided into three proportions in the ratio of 70 : 20 : 10. By using the data part 1, training data are used to create a model 70 percentages. For 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 seeds of 10, 20, 30, 40, and 50 by WEKA program. For the classification of the dataset of liver disease in Andhra Pradesh, India, the best method was the support vector machine method by the regression imputation method, multiple imputation method and expectation maximization method. For the classification of the dataset of biopsy data on breast cancer patients, the best method was the support vector machine method by the regression imputation method and expectation maximization

method. For the classification of the dataset of monoclonal gammopathy data, the best method was the artificial neural network method by the multiple imputation method. For the classification of the dataset of issued and non-issued credit cards by a bank, the best method was the support vector machine method by the expectation maximization method. For the classification of the dataset of single-family loan-level, the best method was the decision tree method by the multiple imputation method. For the classification of the dataset of cardiovascular disease in Framingham, Massachusetts, the best method was the support vector machine method by the regression imputation method, multiple imputation method and expectation maximization method.

Keywords: missing value; regression imputation; multiple imputation; expectation maximization; K-nearest neighbor; decision tree; artificial neural network; support vector machine

1. คำนำ

ข้อมูลสูญหาย (missing value) ในการศึกษาวิจัยเป็นเรื่องที่พบเห็นกันโดยทั่วไป อาจเกิดจากขั้นตอนการจัดเก็บข้อมูล ปัญหาจากการป้อนข้อมูล เครื่องมือจัดเก็บ การโอนถ่ายข้อมูลเกิดความผิดพลาด หรือเกิดจากข้อจำกัดของเทคโนโลยีที่ใช้งาน ซึ่งสามารถส่งผลกระทบตั้งแต่ระดับเล็กน้อยไปจนถึงก่อให้เกิดผลกระทบอย่างรุนแรงต่อผลการศึกษา การจัดการเกี่ยวกับข้อมูลสูญหายเป็นสิ่งจำเป็นเพราะผลกระทบจากปัญหานี้สามารถเกิดขึ้นต่อเนื่องเริ่มตั้งแต่การวิเคราะห์ผล ไปจนถึงการสรุปผลและวิจารณ์ผล ปัญหาหนึ่งที่ผู้วิจัยพบบ่อย คือ ข้อมูลที่รวบรวมมาได้นั้นมีคำตอบไม่ครบถ้วนสมบูรณ์ ข้อมูลได้รับคำตอบเพียงบางส่วน ทำให้เกิดข้อมูลสูญหาย ซึ่งส่งผลให้ไม่สามารถใช้ประโยชน์จากข้อมูลชุดนั้นอย่างเต็มที่ ถ้าในด้านธุรกิจอาจทำให้ขาดทุนหรือสูญเสียรายได้ ในส่วนของงานด้านการแพทย์ ถ้าการทำนายเกิดข้อผิดพลาด ไม่ใช่เพียงแต่เสียทรัพย์ แต่อาจทำให้ผู้ป่วยถึงกับเสียชีวิตได้ จึงต้องมีการให้ความสำคัญกับข้อมูลที่สูญหาย โดยมีเทคนิคที่จะช่วยในการทำนายค่าของข้อมูลสูญหาย เพื่อให้การทำเหมืองข้อมูลมีประสิทธิภาพเพิ่มมากยิ่งขึ้น ซึ่งเทคนิคในการจัดการกับข้อมูลสูญหายนั้นมีหลากหลายเทคนิคที่ส่งผลต่อประสิทธิภาพในการทำนายของตัวแบบแตกต่างกันซึ่งนำไปสู่วิธีการหาค่าของข้อมูลสูญหาย (ปิยะภรณ์ และสุคนธ์, 2549)

การศึกษางานวิจัยที่เกี่ยวข้อง ปูเป้ และคณะ (2561) ศึกษาวิธีการแทนค่าสูญหาย 4 วิธี ได้แก่ วิธีค่าเฉลี่ย (mean imputation) วิธีประมาณค่าทดแทนพหุ (multiple imputation, MI) วิธีเพื่อนบ้านใกล้เคียง k ตัว (K-nearest neighbor) และวิธี weight locally linear reconstruction (WLLR) กำหนดขนาดตัวอย่างเท่ากับ 60, 100, 300 และ 500 ร้อยละการสูญหายเท่ากับ 5, 10, 15, 20, 30, 40 และ 50 เกณฑ์การเปรียบเทียบ คือ ค่าประมาณค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ย (estimated mean square error, EMSE) ผลการวิจัยพบว่าที่ทุกร้อยละการสูญหายที่กำหนด เมื่อขนาดตัวอย่างเท่ากับ 60, 100 และ 300 วิธีค่าเฉลี่ยมีประสิทธิภาพดีที่สุด เมื่อขนาดตัวอย่างเท่ากับ 500 วิธีประมาณค่าทดแทนพหุมีประสิทธิ ภาพที่ดีที่สุด นอกจากนี้พบว่าค่า EMSE เพิ่มขึ้น เมื่อร้อยละการสูญหายเพิ่มขึ้น และลดลงเมื่อขนาดตัวอย่างเพิ่มขึ้น

ทัตดา (2554) ศึกษาการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าข้อมูลสูญหายระหว่างวิธีการประมาณค่าข้อมูลสูญหายด้วยค่าถดถอย (RI) กับวิธีการประมาณข้อมูลสูญหายด้วย

ค่าถดถอยแบบสโตแคสติก (SRI) โดยกำหนดขนาดตัวอย่างเท่ากับ 50, 100, 200 และ 300 สุ่มโดยขั้นตอนวิธีของเบ็บบิงตัน (Bebbington algorithm) ซึ่งใช้วิธีการชักตัวอย่างแบบสุ่มเชิงเดียวชนิดไม่คืนที่ (simple random sampling without replacement, SRSWOR) และกำหนดร้อยละของข้อมูลสูญหายของตัวอย่างเท่ากับ 5, 10, 15 และ 20 ของขนาดตัวอย่าง ผลการศึกษาการเปรียบเทียบประสิทธิภาพระหว่างวิธีการประมาณค่าข้อมูลสูญหายนั้นได้วัดจากค่าคลาดเคลื่อนกำลังสองเฉลี่ยของค่าประมาณของข้อมูลสูญหาย (mean square error, MSE) โดยได้ข้อสรุปว่าวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยมีประสิทธิภาพมากกว่าวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยแบบสโตแคสติก (SRI)

Dong และ Peng (2013) ศึกษาการแทนค่าสูญหาย 3 วิธี ได้แก่ วิธีภาวะน่าจะเป็นสูงสุดโดยใช้สารสนเทศ (full information maximum likelihood, FIML) วิธีประมาณค่าทดแทนพหุ (MI) และวิธีค่าคาดหมายสูงสุด (expectation maximization, EM) มีร้อยละของข้อมูลสูญหาย 3 อัตรา คือ 20, 40 และ 60 ใช้เกณฑ์ความคลาดเคลื่อนมาตรฐาน (standard error, SE) ซึ่งการศึกษาทำให้ทราบว่าวิธีค่าคาดหมายสูงสุดแทนค่าสูญหายได้ดีกว่าวิธีการประมาณค่าทดแทนพหุ

Peng และ Zhu (2008) ศึกษาการเปรียบเทียบวิธีการประมาณค่าสูญหาย 2 วิธี ได้แก่ วิธีค่าคาดหมายสูงสุด (EM) และวิธีการประมาณค่าทดแทนพหุ (MI) โดยใช้เกณฑ์ความเอนเอียง ประสิทธิภาพและอัตราการปฏิเสธผิดพลาด ซึ่งงานวิจัยนี้ทำให้ทราบว่าวิธี วิธีการประมาณค่าทดแทนพหุ และประมาณค่าสูญหายได้ดีกว่าวิธีค่าคาดหมายสูงสุด

งานวิจัยนี้ศึกษาการเปรียบเทียบประสิทธิภาพในการทำนายผลการจำแนกข้อมูลสูญหายด้วยวิธีต่าง ๆ 3 วิธี คือ วิธีการประมาณค่าการถดถอย (regression imputation, RI) วิธีการประมาณค่าทดแทนพหุ (MI) และวิธีค่าคาดหมายสูงสุด (EM) เพื่อเป็นแนวทางสำหรับการบริหารจัดการข้อมูลในกรณีที่เกิดปัญหาข้อมูลสูญหาย ใช้วิธีในการจำแนกประกอบด้วย 4 วิธี คือ วิธีเพื่อนบ้านใกล้สุด k ตัว ใช้ขั้นตอนวิธีชนิด IBK วิธีต้นไม้ตัดสินใจใช้ขั้นตอนวิธีชนิด J48 (C4.5) วิธีโครงข่ายประสาทเทียมใช้ขั้นตอนวิธีชนิดเพอร์เซปตรอนแบบหลายชั้น และวิธีซัพพอร์ตเวกเตอร์แมชชีนใช้ขั้นตอนวิธีชนิด SMO เคอร์เนลแบบโพลิโนเมียลเคอร์เนล การเปรียบเทียบประสิทธิภาพของการจำแนกทั้ง 4 วิธี ใช้ค่าความแม่น (accuracy) ค่าคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) และค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย (mean absolute error, MAE)

2. วิธีการวิจัย

2.1 เครื่องมือที่ใช้ในการวิจัย

โปรแกรมที่ใช้ในการวิจัยครั้งนี้ คือ SPSS (Statistical Package for the Social Science for Windows) และ WEKA (Waikato Environment for Knowledge Analysis)

2.2 การเก็บรวบรวมข้อมูล การหาค่าสูญหาย การแบ่งข้อมูล การศึกษาขั้นตอนวิธี และการเปรียบเทียบประสิทธิภาพของวิธีการจำแนก

2.2.1 การเก็บรวบรวมข้อมูล โดยการศึกษาหาค่าข้อมูลสูญหายจากเว็บไซต์ UCI, Kaggle และ Vincentarelbundock จำนวน 6 ชุด ได้แก่

(1) โรคตับในรัฐอานธรประเทศ ประเทศอินเดีย (liver disease of Andhra Pradesh India) จำนวนข้อมูลทั้งหมด 583 ค่า พบค่าสูญหายจำนวน 11 ค่า คิดเป็นร้อยละ 1.89 เป็นชุดข้อมูลที่มีค่าสูญหายต่ำ (Ramana, 2012)

(2) การตรวจชิ้นเนื้อในผู้ป่วยมะเร็งเต้านม (biopsy data on breast cancer patient) จำนวนข้อมูลทั้งหมด 699 ค่า พบค่าสูญหายจำนวน 16 ค่า คิดเป็นร้อยละ 2.29 เป็นชุดข้อมูลที่มีค่าสูญหายต่ำ (Wolberg, 1992)

(3) การศึกษาระยะยาวของสารภูมิต้านทานโมโนโคลน (monoclonal gammopathy data) จำนวนข้อมูลทั้งหมด 892 ค่า พบค่าสูญหายจำนวน 38 ค่า คิดเป็นร้อยละ 4.26 เป็นชุดข้อมูลที่มีค่าสูญหายปานกลาง (Kyle *et al*., 2002)

(4) การตลาดของธนาคาร (issued and non-issued credit cards by a Bank) จำนวนข้อมูลทั้งหมด 598 ค่า พบค่าสูญหายจำนวน 26 ค่า คิดเป็นร้อยละ 4.35 เป็นชุดข้อมูลที่มีค่าสูญหายปานกลาง (Portuguese Banking Institution, 2012)

(5) ระดับสินเชื่อครอบครัวเดี่ยว (single family loan-level data) จำนวนข้อมูลทั้งหมด 783 ค่า พบค่าสูญหายจำนวน 71 ค่า คิดเป็นร้อยละ 9.07 เป็นชุดข้อมูลที่มีค่าสูญหายสูง (Saravananselvamohan, 1992)

(6) โรคหลอดเลือดหัวใจในเมือง Framingham รัฐ Massachusetts (cardiovascular disease in Framingham, Massachusetts) จำนวนข้อมูลทั้งหมด 1,000 ค่า พบค่าสูญหายจำนวน 97 ค่า คิดเป็นร้อยละ 9.70 เป็นชุดข้อมูลที่มีค่าสูญหายสูง (Ajmera, 2017)

2.2.2 การหาค่าสูญหาย

(1) วิธีการประมาณค่าการถดถอย (RI) คือ การแทนที่ข้อมูลสูญหายด้วยสมการประมาณค่าที่เป็นสมการเส้นตรงที่เกิดจากข้อมูลที่ไม่สูญหาย (ฐณัฐ, 2559) จากสมการประมาณค่า

$$\widehat{A}_{B^*} = a + bB^*$$

$$b = \frac{\sum_{i=1}^{k} A_i B_i - k\bar{B}\bar{A}}{\sum_{i=1}^{k} B_i^2 - k\bar{B}^2}$$

$$a = \bar{A} - b\bar{B}$$

โดยที่ B^* คือ ตำแหน่งของข้อมูลที่สูญหาย; $\widehat{A}_{B^*}$ คือ ค่าประมาณของข้อมูลที่สูญหาย ตำแหน่งที่ ${B_i}^*$; A_i คือ ข้อมูลที่ไม่สูญหายในแถวที่พบข้อมูลสูญหาย ตำแหน่งที่ I; B_i คือ ตำแหน่งของข้อมูลที่ไม่สูญหายในแถวที่พบข้อมูลสูญหาย; $i = 1, 2, \ldots, k$; k คือ จำนวนข้อมูลทั้งหมด; a คือ ค่ากำหนดตำแหน่งของเส้นตรง มีค่าเป็นค่าคงที่; b คือ ความชันของสมการเส้นตรง มีค่าเป็นค่าคงที่

(2) วิธีการประมาณค่าทดแทนพหุ (MI)

(2.1) ขั้นตอนที่ 1 คือการประมาณค่าข้อมูลสูญหาย ซึ่งจะประมาณค่าสูญหายให้กับข้อมูลที่มีการสูญหายหลายชุดข้อมูล ทำซ้ำเพื่อให้ได้ชุดข้อมูล m ชุด (m > 1)

(ก) ให้ตัวแปรที่มีค่าสูญหายเป็นตัวแปรตาม (Y) และตัวแปรที่เหลือเป็นตัวแปรอิสระ (X) แล้วสร้างตัวแบบการถดถอย ดังสมการ $\widehat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_P X_p$ เพื่อหาค่าประมาณพารามิเตอร์ $\hat{\beta}$ และเมทริกซ์ความแปรปรวนร่วม $\hat{\sigma}_j^2 V_j$ เมื่อ V_j เป็นเมทริกซ์ผกผันของ $X'X$

(ข) คำนวณหาค่าประมาณพารามิเตอร์ใหม่ $\hat{\beta}_* = (\hat{\beta}_{*0}, \hat{\beta}_{*1}, \ldots, \hat{\beta}_{*p})$ และคำนวณค่า $\hat{\sigma}_{*j}^2$ จากค่าประมาณพารามิเตอร์ที่ได้จากข้อ ก ดังสมการ $\hat{\beta}_* = \hat{\beta} + \sigma_{*j} V'_{hj} Z$

$$\hat{\sigma}_{*j}^2 = \frac{\hat{\sigma}_j^2 (n_j - p - 1)}{g}$$

โดยที่ $V_j = V'_{hj} V_{hj}$, $g = {\chi^2}_{n_j - p - 1}$; V_{hj} เป็นเมทริกซ์สามเหลี่ยมบน (upper triangular matrix); Z เป็นเวกเตอร์ของตัวแปรอิสระแบบสุ่ม ขนาด p+1; n_j เป็นจำนวนค่าสังเกตที่ไม่มีค่าสูญหาย; p เป็นจำนวนของตัวแปรอิสระหรือตัวแปรที่ไม่มีค่าสูญหาย

(ค) ประมาณค่าสูญหายจากสมการ $\widehat{Y} = \beta_{*0} + \beta_{*1} X_{*1} + \beta_{*2} X_{*2} + \ldots + \beta_{*p} X_{*P}$

$+Z_j\sigma_{ij}$ เมื่อสุ่มค่า Z_j ให้มีจำนวนเท่ากับสัมประสิทธิ์การถดถอย (ศศิธร, 2555)

(2.2) ขั้นตอนที่ 2 คือ วิเคราะห์ข้อมูลแต่ละชุดแยกกัน เพื่อประมาณค่าพารามิเตอร์ β_i เมื่อ i = 1, ..., m ด้วยวิธีกำลังสองน้อยสุดสามัญ (ordinary least squares method, OLS)

(2.3) ขั้นตอนที่ 3 คือ นำค่าที่ประมาณพารามิเตอร์ที่ได้จากแต่ละชุดข้อมูลที่มีการประมาณค่าสูญหาย มาสรุปให้กับชุดข้อมูลที่มีการสูญหาย

$$\beta = \frac{\beta_1+\beta_2+\cdots+\beta_m}{m}$$

(3) วิธีค่าคาดหมายสูงสุด (EM) กระบวนการหาค่าคาดหมายและการหาค่าสูงสุด หรือ EM-algorithm เป็นวิธีการเชิงตัวเลขที่อาศัยหลักการวนซ้ำ ๆ (iterative procedure) เพื่อหาค่าประมาณภาวะน่าจะเป็นสูงสุด (maximum likelihood) ของพารามิเตอร์ θ ของการแจกแจงในกรณีที่มีข้อมูลสูญหาย โดย EM-algorithm แบ่งเป็น 2 ขั้นตอน คือ กระบวนการหาค่าคาดหมาย (expectation, E-step) และกระบวนการหาค่าสูงสุด (maximization, M-step) โดยกำหนดให้ Y_{obs} แทนเซตของข้อมูลที่สังเกตได้ของการแจกแจงหนึ่ง และ Y_{mis} แทน ค่าที่สูญหาย ดังนั้น $Y = (Y_{obs}, Y_{mis})$ ซึ่งเป็นข้อมูลที่สมบูรณ์ (complete data) ให้ $P(Y|\theta)$ แทนฟังก์ชันความน่าจะเป็นของการแจกแจงที่มีพารามิเตอร์ θ จากฟังก์ชันภาวะน่าจะเป็นของการแจกแจง จะได้

$$L(\theta) = \prod_Y P(Y|\theta)$$

และล็อกฟังก์ชันภาวะน่าจะเป็น

$$\log L(\theta) = \sum_Y \log\left(P(Y|\theta)\right)$$

เมื่อมีข้อมูลสูญหายจะเรียกล็อกฟังก์ชันภาวะน่าจะเป็นที่ข้อมูลไม่สมบูรณ์ (observed incomplete log-likelihood) ซึ่งในกระบวนการหาค่าคาดหมายจะประมาณค่าคาดหมายของล็อกฟังก์ชันภาวะน่าจะเป็นภายใต้ข้อมูลที่สมบูรณ์ จะได้ว่า

$$Q(\theta) = E(\log L(\theta)|Y_{obs};\theta) = \hat{Y}_{mis}$$

โดยมีค่าคาดหมายของข้อมูลที่สูญหายและล็อกฟังก์ชันภาวะน่าจะเป็นดังนี้

$$Q(\theta) = E(Y_{mis}|Y_{obs};\theta)\log\left(P(Y_{mis};\theta)\right) + \sum_{Y_{obs}} \log\left(P(Y_{obs};\theta)\right)$$

สำหรับกระบวนการหาค่าสูงสุดจะแทนค่าคาดหมายของข้อมูลที่สูญหายจากขั้นตอน E-step และประมาณค่าคาดหมายจากล็อกฟังก์ชันภาวะน่าจะเป็นภายใต้ข้อมูลที่สมบูรณ์ (unobserved complete log-likelihood) ซึ่งมีขั้นตอนดังนี้

(3.1) ขั้นตอนที่ 0 กำหนดค่า $\hat{\theta}^{(r)}$ รอบที่ 0 และกำหนดค่าเริ่มต้น $r = 0$ โดยที่ r แทนจำนวนรอบ

(3.2) ขั้นตอนที่ 1 แทนค่า $\hat{\theta}^{(r)}$ ลงใน $\hat{Y}_{mis}^{(r+1)}$

(3.3) ขั้นตอนที่ 2 แทนค่า $\hat{Y}_{mis}^{(r+1)}$ ที่ได้จากขั้นตอนที่ 1 ลงใน $\hat{\theta}^{(r+1)}$ และให้ $r = r + 1$ จากขั้นตอนที่ 2 จะได้ค่า $\hat{\theta}$ ที่ปรับค่าแล้ว จากนั้นนำค่า $\hat{\theta}$ ดังกล่าวโดยในรอบที่ $r = r + 1$ กลับไปแทนค่าในขั้นตอนที่ 1 เพื่อหา $\hat{Y}_{mis}$ อีกครั้ง โดยวนซ้ำ ๆ จนกระทั่งค่าประมาณ $\hat{\theta}$ มีการลู่เข้า จึงจะได้ตัวประมาณพารามิเตอร์ของ θ

2.2.3 การแบ่งข้อมูล

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

2.2.4 การศึกษาขั้นตอนวิธี

(1) วิธีเพื่อนบ้านใกล้สุด k ตัว เป็นวิธีที่ใช้ในการจัดแบ่งคลาส (class) โดยเทคนิคนี้จะ

ตัดสินใจว่าคลาสใดที่จะแทนเงื่อนไขหรือกรณีใหม่ ๆ ได้บ้าง โดยการตรวจสอบจำนวนบางจำนวนในขั้นตอนวิธีเพื่อนบ้านใกล้สุด k ตัว ของกรณีหรือเงื่อนไขที่เหมือนกันหรือใกล้เคียงกันมากที่สุด โดยจะหาผลรวม (count up) ของจำนวนเงื่อนไข หรือกรณีต่าง ๆ สำหรับแต่ละคลาส และกำหนดเงื่อนไขใหม่ ๆ ให้คลาสที่เหมือนกันกับคลาสที่ใกล้เคียงกันมากที่สุด (นที และภาสกร, 2562)

(2) วิธีต้นไม้ตัดสินใจ เป็นวิธีการเรียนรู้ของเครื่องที่นิยมใช้มากที่สุดแบบหนึ่ง โดยการจำแนกข้อมูลออกเป็นคลาสต่าง ๆ โดยใช้คุณลักษณะ (attribute) ของข้อมูลในการจำแนกว่าคุณลักษณะใดของข้อมูลที่เป็นตัวกำหนดการจำแนกและลักษณะแต่ละตัวของข้อมูลมีการวัดความสำคัญอย่างไร ต้นไม้ตัดสินใจประกอบด้วยโหนดภายใน (internal node) คือ คุณลักษณะต่าง ๆ ของข้อมูล ใช้ในการตัดสินใจว่าข้อมูลจะไปอยู่ในกรณีไหน โดยโหนดภายในที่เป็นโหนดเริ่มต้นเรียกว่าโหนดกิ่ง (branch, link) เป็นค่าคุณลักษณะหรือเงื่อนไขของคุณลักษณะในโหนดที่ใช้ในการจำแนกข้อมูล ซึ่งโหนดภายในจะแตกกิ่งเป็นจำนวนเท่ากับจำนวนค่าคุณลักษณะของโหนดภายในนั้น โหนดใบ (leaf node) คือคลาสต่าง ๆ ซึ่งเป็นผลลัพธ์ในการจำแนกข้อมูล (อโณทัย, 2554)

(3) วิธีโครงข่ายประสาทเทียม โดยแนวคิดเริ่มต้นของเทคนิคนี้ได้มาจากการศึกษาโครงข่ายไฟฟ้าชีวภาพ (bioelectric network) ในสมอง ซึ่งประกอบด้วยเซลล์ประสาท (neuron) และจุดประสานประสาท (synapse) แต่ละเซลล์ประสาทประกอบด้วยปลายในการรับกระแสประสาท (dendrite) ซึ่งเป็นข้อมูลเข้า (input data) และปลายในการส่งกระแสประสาทเรียกว่าแกนประสาท (axon) ซึ่งเป็นเหมือนข้อมูลออก (output data) ของเซลล์ เซลล์เหล่านี้ทำงานด้วยปฏิกิริยาไฟฟ้าเคมี เมื่อมีการกระตุ้นด้วยสิ่งเร้าภายนอกหรือกระตุ้นด้วยเซลล์ด้วยกัน กระแสประสาทจะวิ่งผ่านปลายในการรับกระแสประสาทเข้าสู่นิวเคลียสซึ่งจะเป็นตัวตัดสินว่าต้องกระตุ้นเซลล์อื่น ๆ ต่อหรือไม่ ถ้ากระแสประสาทแรงพอ นิวเคลียสก็จะกระตุ้นเซลล์อื่น ๆ ต่อไปผ่านทางแกนประสาท (วิทยา, 2555)

(4) วิธีซัพพอร์ตเวกเตอร์แมชชีน เป็นสมการที่ใช้จำแนกค่าคุณลักษณะของ 2 กลุ่ม ที่วางตัวอยู่ในพื้นที่คุณลักษณะ (feature space) ออกจากกันโดยจะสร้างเส้นแบ่ง (plane) ที่เป็นเส้นตรงขึ้นมา และเพื่อให้ทราบว่าเส้นตรงที่แบ่ง 2 กลุ่มออกจากกันนั้น เส้นตรงใดที่เป็นเส้นที่ดีที่สุด โดยเส้นตรงนั้นจะเพิ่มเส้นขอบ (margin) ออกไปทั้งสองข้าง โดยเส้นขอบที่เพิ่มนั้นจะขนานกับเส้นเดิมเสมอ เส้นขอบที่เพิ่มขึ้นมานี้จะขยายออกไปจนกว่าจะสัมผัสกับค่าของกลุ่มตัวอย่างที่ใกล้ที่สุดเคอร์เนล (kernel) ที่สามารถเปลี่ยนข้อมูลที่มีมิติ (dimension) ที่ต่ำกว่าให้มีมีติสูงขึ้นเพื่อให้การแบ่งข้อมูลแบบ linear model ในโลกความเป็นจริงนั้นข้อมูล 2 กลุ่ม ไม่ได้วางตัวในพื้นที่คุณลักษณะ และไม่สามารถแบ่งโดยเส้นตรง แต่ข้อมูลอาจจับกลุ่มกันในตำแหน่งต่าง ๆ ดังนั้นจึงเป็นปัญหาทำให้ไม่สามารถที่จะใช้สมการซัพพอร์ตเวกเตอร์แมชชีนแบบเชิงเส้น ดังนั้นจะต้องมีเครื่องมือมาช่วยให้ข้อมูลเหล่านั้นเรียงตัวใหม่ในพื้นที่เรียกว่ พื้นที่หลายมิติ (higher dimensional space) (สุรวัชร และสายชล, 2560)

หลังจากนั้นนำข้อมูลที่แบ่งออกเป็น 3 ส่วน มาวิเคราะห์โดยใช้โปรแกรม WEKA ซึ่งวิเคราะห์จากวิธีการจำแนกทั้ง 4 วิธี ข้างต้น

2.2.5 การเปรียบเทียบประสิทธิภาพของวิธีการจำแนก

นำผลการวิเคราะห์ของแต่ละวิธีทั้ง 4 วิธี มาเปรียบเทียบประสิทธิภาพโดยพิจารณาจากเมทริกซ์ความสับสน (confusion matrix) ซึ่งเป็นรูปแบบตารางที่เฉพาะเจาะจงที่นำผลลัพธ์จากการ

ทำนายมาใส่ในตารางเมทริกซ์ความสับสนซึ่งจะช่วยให้ง่ายต่อการมองเห็นค่าทำนายของขั้นตอนวิธีดังตารางที่ 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) ค่าความแม่นในการทำนาย คือ การแสดงการวัดที่ได้มีความแม่น ในรูปอัตราส่วนโดยคิดเป็นร้อยละ (สุรวัชร และสายชล, 2560) โดยนำค่าจากตารางเมทริกซ์ความสับสนมาแทนในสมการ

Accuracy = จำนวนข้อมูลที่จำแนกถูกว่าคำตอบเป็นบวกและลบ x 100 %

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

(2) ค่าคลาดเคลื่อนกำลังสองเฉลี่ย (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

(3) ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) คือ ค่าวัดความแม่นของการทำนายที่วัดจากค่าความคลาดเคลื่อนโดยไม่คำนึงถึงทิศทางของความคลาดเคลื่อน 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 โดยนำชุดข้อมูลจำนวน 3 ชุด มาวิเคราะห์ข้อมูล ซึ่งจะสุ่มแบ่งข้อมูลออกเป็น 3 ส่วน คือ ส่วนที่ 1 ข้อมูลเรียนรู้ นำไปสร้างตัวแบบร้อยละ 70 ข้อมูลส่วนที่ 2 ข้อมูลตรวจสอบความถูกต้อง นำไปประเมินความผิดพลาดของตัวแบบร้อยละ 20 และข้อมูลส่วนที่ 3 ข้อมูลทดสอบ นำไปทดสอบตัวแบบร้อยละ 10 และผู้วิจัยได้นำมาเปรียบเทียบประสิทธิภาพในการทำนายผลของวิธีการจำแนกโดยพิจารณาจากค่าความแม่น ค่าคลาดเคลื่อนกำลังสองเฉลี่ย และค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย ซึ่งวิธีที่ใช้ในการทดสอบครั้งนี้มี 4 วิธี คือ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม วิธีซัพพอร์ตเวกเตอร์แมชชีน ผลการวิเคราะห์ข้อมูลได้ดังนี้

3.1.1 ข้อมูลชุดที่ 1 ซึ่งเป็นกรณีที่มีข้อมูลสูญหายต่ำ พบว่าวิธีโครงข่ายประสาทเทียมแทนค่าข้อมูลสูญหายด้วยวิธีการประมาณค่าการถดถอย ให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.18812 และวิธีซัพพอร์ตเวกเตอร์แมชชีนแทนค่าข้อมูลสูญหายด้วยวิธีการประมาณค่าการถดถอย วิธีการประมาณค่าทดแทนพหุและวิธีค่าคาดหมายสูงสุด ให้ค่าความแม่นสูงสุด คือ ร้อยละ 72.20338

ค่าคลาด เคลื่อนสัมบูรณ์เฉลี่ยต่ำสุด คือ ร้อยละ 0.27796

3.1.2 ข้อมูลชุดที่ 2 ซึ่งเป็นกรณีที่ข้อมูลสูญหายต่ำ พบว่าวิธีซัพพอร์ตเวกเตอร์แมชชีนแทนค่าข้อมูลสูญหายด้วยวิธีการประมาณค่าการถดถอย และวิธีค่าคาดหมายสูงสุด ให้ค่าความแม่นสูงสุด คือ ร้อยละ 98.28570 ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.01714 และค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยต่ำสุด คือ ร้อยละ 0.01716

3.1.3 ข้อมูลชุดที่ 3 ซึ่งเป็นกรณีที่ข้อมูลสูญหายปานกลาง พบว่าวิธีโครงข่ายประสาทเทียมแทนค่าข้อมูลสูญหายด้วยวิธีการประมาณค่าทดแทนพหุ ให้ค่าความแม่นสูงสุดคือร้อยละ 80 วิธีค่าคาดหมายสูงสุด ให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.13566 และวิธีซัพพอร์ตเวกเตอร์แมชชีนแทนค่าข้อมูลสูญหายด้วยวิธีการประมาณค่าการถดถอย วิธีการประมาณค่าทดแทนพหุและวิธีค่าคาดหมายสูงสุดให้ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ยต่ำสุด คือ 0.20444

3.1.4 ข้อมูลชุดที่ 4 ซึ่งเป็นกรณีที่ข้อมูลสูญหายปานกลาง พบว่าวิธีวิธีต้นไม้ตัดสินใจแทนค่าสูญหายด้วยวิธีค่าคาดหมายสูงสุดให้ค่าความแม่นสูงสุด คือ ร้อยละ 89.66666 และค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.08804 วิธีซัพพอร์ตเวกเตอร์แมชชีนแทนค่าข้อมูลสูญหายด้วยวิธีค่าคาดหมายสูงสุดให้ค่าความแม่นสูงสุดคือร้อยละ 89.66666 และค่าคลาดเคลื่อนกำลังสองสัมบูรณ์เฉลี่ยต่ำสุด คือ 0.10334

3.1.5 ข้อมูลชุดที่ 5 ซึ่งเป็นกรณีที่ข้อมูลสูญหายสูง พบว่าวิธีต้นไม้ตัดสินใจแทนค่าข้อมูลสูญหายด้วยวิธีประมาณค่าการถดถอย วิธีการประมาณค่าทดแทนพหุและวิธีค่าคาดหมายสูงสุดให้ค่าความแม่นสูงสุด คือ ร้อยละ 95.44306 และแทนค่าสูญหายด้วยวิธีการประมาณค่าทดแทนพหุให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.04296 วิธีซัพพอร์ตเวกเตอร์แมชชีนแทนค่าข้อมูลสูญหายด้วยวิธีประมาณค่าการถดถอย วิธีการประมาณค่าทดแทนพหุและวิธีค่าคาดหมายสูงสุดให้ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยต่ำสุด คือ 0.04810

3.1.6 ข้อมูลชุดที่ 6 ซึ่งเป็นกรณีที่ข้อมูลสูญหายสูง พบว่าวิธีซัพพอร์ตเวกเตอร์แมชชีนแทนค่าข้อมูลสูญหายด้วยวิธีประมาณค่าการถดถอย วิธีการประมาณค่าทดแทนพหุและวิธีค่าคาดหมายสูงสุดให้ค่าความแม่นสูงสุด คือ ร้อยละ 84.8 ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด คือ 0.15600 และค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยต่ำสุด คือ 0.15600

4. สรุปผลการวิจัย

งานวิจัยนี้ได้ศึกษาประสิทธิภาพในการหาค่าข้อมูลสูญหาย 3 วิธี คือ วิธีการประมาณค่าการถดถอย วิธีการประมาณค่าทดแทนพหุ และวิธีค่าคาดหมายสูงสุด ด้วยวิธีการจำแนก 4 วิธี คือ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม และวิธีซัพพอร์ตเวกเตอร์แมชชีน โดยพิจารณาจากค่าความแม่น ค่าคลาดเคลื่อนกำลังสองเฉลี่ย และค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย โดยใช้ข้อมูลที่มีค่าสูญหาย 6 ชุด

4.1 ชุดข้อมูลสูญหายต่ำ จำนวน 2 ชุด คือ ข้อมูลชุดที่ 1 โรคตับของรัฐอานธรประเทศประเทศอินเดีย วิธีที่มีประสิทธิภาพสูงสุด คือ การจำแนกด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน โดยวิธีการประมาณค่าการถดถอย วิธีการประมาณค่าทดแทนพหุ และวิธีค่าคาดหมายสูงสุด ข้อมูลชุดที่ 2 การตรวจชิ้นเนื้อในผู้ป่วยมะเร็งเต้านม วิธีที่มีประสิทธิภาพสูงสุด คือ การจำแนกด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน โดยวิธีการแทนค่าสูญหายด้วยวิธีการประมาณค่าการถดถอย และวิธีค่าคาดหมายสูงสุด เมื่อพิจารณาจากตารางที่ 2

4.2 ชุดข้อมูลสูญหายปานกลาง จำนวน 2 ชุด คือ ข้อมูลชุดที่ 3 การศึกษาระยะยาวของ

สารภูมิต้านทานโมโนโคลน วิธีที่มีประสิทธิภาพสูงสุดคือการจำแนกด้วยวิธีโครงข่ายประสาทเทียม โดยวิธีการประมาณค่าทดแทนพหุ ข้อมูลชุดที่ 4 การตลาดของธนาคาร วิธีที่มีประสิทธิภาพสูงสุด คือ การจำแนกด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน โดยวิธีการแทนค่าสูญหายด้วยวิธีค่าคาดหมายสูงสุด โดยพิจารณาจากตารางที่ 3

4.3 ชุดข้อมูลสูญหายสูง จำนวน 2 ชุด คือ ข้อมูลชุดที่ 5 ระดับสินเชื่อครอบครัวเดี่ยว วิธีที่มีประสิทธิภาพสูงสุด คือ การจำแนกด้วยวิธีต้นไม้ตัดสินใจโ ดยวิธีการแทนค่าสูญหายด้วยวิธีการประมาณค่าทดแทนพหุ และข้อมูลชุดที่ 6 ข้อมูลโรคหลอดเลือดหัวใจ วิธีที่มีประสิทธิภาพสูงสุด คือ การจำแนกด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน โดยวิธีการประมาณค่าการถดถอย วิธีการประมาณค่าทดแทนพหุ และวิธีค่าคาดหมาย โดยพิจารณาจากตารางที่ 4

วิธีแทนค่าสูญหายโดยวิธีประมาณค่าทดแทนพหุแทนค่าได้ดีในข้อมูลที่ 3 และข้อมูลที่ 5 เป็นข้อมูลที่มีตัวแปรส่วนใหญ่เป็นตัวแปรเชิงปริมาณ และวิธีแทนค่าสูญหายโดยวิธีประมาณค่าคาดหมายสูงสุดแทนค่าได้ดีในข้อมูลที่ 4 เป็นข้อมูลที่มีตัวแปรส่วนใหญ่เป็นตัวแปรเชิงคุณภาพ

ตารางที่ 2 ผลการเปรียบเทียบประสิทธิภาพของวิธีการจำแนกโรคตับของรัฐอานธรประเทศ ประเทศอินเดีย (ข้อมูลชุดที่ 1) และการตรวจชิ้นเนื้อในผู้ป่วยมะเร็งเต้านม (ข้อมูลชุดที่ 2) เป็นชุดข้อมูลที่มีค่าสูญหายต่ำ

วิธีการแทนค่าข้อมูลสูญหาย	ข้อมูลชุดที่ 1			ข้อมูลชุดที่ 2		
	ค่าความแม่น	ค่า MSE	ค่า MAE	ค่าความแม่น	ค่า MSE	ค่า MAE
การจำแนกด้วยวิธีเพื่อนบ้านใกล้สุด k ตัว						
วิธีการประมาณค่าการถดถอย	65.42372	0.34410	0.34652	96.85714	0.02776	0.03114
วิธีการประมาณค่าทดแทนพหุ	65.76272	0.34074	0.34314	96.85716	0.03131	0.03260
วิธีค่าคาดหมายสูงสุด	65.42372	0.34410	0.34652	96.85714	0.02776	0.03114
การจำแนกด้วยวิธีต้นไม้ตัดสินใจ						
วิธีการประมาณค่าการถดถอย	67.79658	0.24006	0.33948	95.42858	0.04441	0.06626
วิธีการประมาณค่าทดแทนพหุ	67.45760	0.23688	0.34470	95.14286	0.04637	0.06728
วิธีค่าคาดหมายสูงสุด	67.11862	0.24950	0.34674	95.42858	0.04441	0.06626
การจำแนกด้วยวิธีโครงข่ายประสาทเทียม						
วิธีการประมาณค่าการถดถอย	66.77966	**0.18812**	0.33984	96.85714	0.02348	0.03252
วิธีการประมาณค่าทดแทนพหุ	66.44068	0.18814	0.33972	97.14286	0.02095	0.03028
วิธีค่าคาดหมายสูงสุด	66.44068	0.18814	0.33974	97.14286	0.02345	0.03264
การจำแนกด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน						
วิธีการประมาณค่าการถดถอย	**72.20338**	0.27796	**0.27796**	**98.28570**	**0.01714**	**0.01716**
วิธีการประมาณค่าทดแทนพหุ	**72.20338**	0.27796	**0.27796**	97.99998	0.01999	0.02002

วิธีค่าคาดหมายสูงสุด	**72.20338**	0.27796	**0.27796**	**98.28570**	**0.01714**	**0.01716**

ตารางที่ 3 ผลการเปรียบเทียบประสิทธิภาพของวิธีการจำแนกการศึกษาระยะยาวของสารภูมิต้านทานโมโนโคลน (ข้อมูลชุดที่ 3) และการตลาดของธนาคาร (ข้อมูลชุดที่ 4) เป็นชุดข้อมูลที่มีค่าสูญหายปานกลาง

วิธีการแทนค่าข้อมูลสูญหาย	ข้อมูลชุดที่ 3			ข้อมูลชุดที่ 4		
	ค่าความแม่น	ค่า MSE	ค่า MAE	ค่าความแม่น	ค่า MSE	ค่า MAE
การจำแนกด้วยวิธีเพื่อนบ้านใกล้สุด k ตัว						
วิธีการประมาณค่าการถดถอย	76.44444	0.23498	0.23638	82.52800	0.16764	0.17006
วิธีการประมาณค่าทดแทนพหุ	76.66668	0.23276	0.23418	76.00000	0.23998	0.24610
วิธีค่าคาดหมายสูงสุด	76.44444	0.23498	0.23638	85.00002	0.14928	0.15166
การจำแนกด้วยวิธีต้นไม้ตัดสินใจ						
วิธีการประมาณค่าการถดถอย	77.77780	0.16442	0.26502	86.44808	0.11248	0.18342
วิธีการประมาณค่าทดแทนพหุ	78.22224	0.16320	0.26180	88.00000	0.10554	0.22154
วิธีค่าคาดหมายสูงสุด	77.55556	0.16666	0.26700	**89.66666**	**0.08804**	0.16966
การจำแนกด้วยวิธีโครงข่ายประสาทเทียม						
วิธีการประมาณค่าการถดถอย	79.55556	0.13580	0.24054	86.12568	0.12484	0.14816
วิธีการประมาณค่าทดแทนพหุ	**80.00000**	0.13580	0.24210	86.00002	0.12880	0.20558
วิธีค่าคาดหมายสูงสุด	79.55556	**0.1356**6	0.24030	86.00000	0.11962	0.14164
การจำแนกด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน						
วิธีการประมาณค่าการถดถอย	79.55556	0.20442	**0.20444**	87.43170	0.12570	0.12570
วิธีการประมาณค่าทดแทนพหุ	79.55556	0.20442	**0.20444**	87.33332	0.12666	0.12668
วิธีค่าคาดหมายสูงสุด	79.55556	0.20442	**0.20444**	**89.66666**	0.10334	**0.10334**

5. อภิปรายผล

ข้อมูลการวิจัยชุดที่ 1, 3, 5 และ 6 สอดคล้องกับงานวิจัยของ ปูเป้ และคณะ (2561) ที่ให้วิธีการประมาณค่าทดแทนพหุมีประสิทธิภาพที่ดีที่สุด ถ้าขนาดตัวอย่างมากกว่า 300 ข้อมูลการวิจัยชุดที่ 1, 2 และ 6 สอดคล้องกับงานวิจัยของ ทัตดา (2554) ที่ให้วิธีการประมาณค่าการถดถอยมีประสิทธิภาพที่ดีที่สุด ถ้ากำหนดเปอร์เซ็นต์ของข้อมูลสูญหายของตัวอย่างเท่ากับ 5, 10, 15 และ 20 % ส่วนข้อมูลการวิจัยชุดที่ 1, 2, 4 และ 6 สอดคล้องกับงานวิจัยของ Dong และ Peng (2013) ที่ให้วิธีค่าคาดหมายสูงสุดแทนค่าสูญหายได้ดีกว่าวิธีการประมาณค่าทดแทนพหุ นอกจากนี้ข้อมูลการวิจัยชุดที่ 1, 3, 5 และ 6 สอดคล้องกับงานวิจัยของ Peng และ Zhu (2008) ที่ให้วิธีการประมาณค่าทดแทนพหุแทนค่าสูญหายได้ดีกว่าวิธีค่าคาดหมายสูงสุด

ข้อมูลการวิจัยชุดที่ 2 และ 4 ขัดแย้งกับงานวิจัยของ ปูเป้ และคณะ (2561) ข้อมูลการวิจัยชุดที่ 3, 4 และ 5 ขัดแย้งกับงานวิจัยของ ทัตดา (2554) ข้อมูลการวิจัยชุดที่ 3 และ 5 ขัดแย้งกับงานวิจัยของ Dong และ Peng (2013) และข้อมูลการวิจัยชุดที่ 2 และ 4 ขัดแย้งกับงานวิจัยของ Peng และ Zhu (2008) เนื่องจากงานวิจัยข้างต้นไม่ได้

แทนค่าสูญหายด้วยวิธีวิธีการประมาณค่าการถดถอย วิธีการประมาณค่าทดแทนพหุ และวิธีค่าคาดหมายสูงสุดครบทั้ง 3 วิธี

ตารางที่ 4 ผลการเปรียบเทียบประสิทธิภาพของวิธีการจำแนกระดับสินเชื่อครอบครัวเดี่ยว (ข้อมูลชุดที่ 5) และชุดข้อมูลโรคหลอดเลือดหัวใจ (ข้อมูลชุดที่ 6) เป็นชุดข้อมูลที่มีค่าสูญหายสูง

วิธีการแทนค่าข้อมูลสูญหาย	ข้อมูลชุดที่ 5			ข้อมูลชุดที่ 6		
	ค่าความแม่น	ค่า MSE	ค่า MAE	ค่าความแม่น	ค่า MSE	ค่า MAE
การจำแนกด้วยวิธีเพื่อนบ้านใกล้สุด k ตัว						
วิธีการประมาณค่าการถดถอย	90.88608	0.09079	0.09262	75.2	0.24732	0.24874
วิธีการประมาณค่าทดแทนพหุ	90.88608	0.09082	0.09262	75.4	0.21106	0.24674
วิธีค่าคาดหมายสูงสุด	90.63292	0.09332	0.09514	75.2	0.24732	0.24874
การจำแนกด้วยวิธีต้นไม้ตัดสินใจ						
วิธีการประมาณค่าการถดถอย	**95.44306**	0.04299	0.0712	80.0	0.16514	0.25960
วิธีการประมาณค่าทดแทนพหุ	**95.44306**	**0.04296**	0.0712	80.4	0.16266	0.25898
วิธีค่าคาดหมายสูงสุด	**95.44306**	0.04299	0.0712	80.0	0.16502	0.25910
การจำแนกด้วยวิธีโครงข่ายประสาทเทียม						
วิธีการประมาณค่าการถดถอย	94.43036	0.04624	0.06656	79.2	0.17218	0.22864
วิธีการประมาณค่าทดแทนพหุ	94.93672	0.04332	0.06648	80.6	0.16552	0.22308
วิธีค่าคาดหมายสูงสุด	94.43036	0.04675	0.06766	78.2	0.1795	0.23382
การจำแนกด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน						
วิธีการประมาณค่าการถดถอย	95.18988	0.04810	**0.04810**	**84.4**	**0.15600**	**0.15600**
วิธีการประมาณค่าทดแทนพหุ	95.18988	0.04810	**0.04810**	**84.4**	**0.15600**	**0.15600**
วิธีค่าคาดหมายสูงสุด	95.18988	0.04810	**0.04810**	**84.4**	**0.15600**	**0.15600**

6. ข้อเสนอแนะ

6.1 เพื่อให้ได้ข้อสรุปของผลการวิเคราะห์ข้อมูลที่มีความสมบูรณ์มากขึ้น ดังนั้นผู้วิจัยอาจวิเคราะห์ข้อมูลด้วยวิธีอื่น ๆ ได้แก่ วิธีฐานกฎ (rule-based) วิธีลาดลงสโตแคสติก (Stochastic gradient descent) วิธีนาอีฟเบย์ (Naive Bayes) วิธีเบย์เซียนเน็ตเวิร์ค (Bayesian network) วิธีการถดถอยลอจิสติกทวิภาค (binary logistic regression) วิธีเพอร์เซปตรอนให้คะแนน (voted perceptron) เป็นต้น

6.2 การแทนค่าข้อมูลสูญหายยังมีวิธีอื่น ๆ ที่สามารถแทนค่าข้อมูลสูญหาย เพื่อให้ได้ผลการแทนค่าข้อมูลสูญหายที่มีความสมบูรณ์มากขึ้น อาจแทนค่าข้อมูลสูญหายด้วยวิธีอื่น ๆ เช่น การประมาณค่าภาวะน่าจะเป็นสูงสุด (maximum likelyhood estimation, MLE)

6.3 การศึกษาวิธีการแทนค่าข้อมูลสูญหายอาจมีโปรแกรมอื่น ๆ ที่มีประสิทธิภาพ เช่น SAS, R commander, Excel, Mplus

7. การนำไปใช้ประโยชน์

การหาค่าข้อมูลสูญหายสามารถนำไปใช้แก้ปัญหาที่ข้อมูลที่รวบรวมมานั้นมีคำตอบไม่ครบถ้วนสมบูรณ์ ข้อมูลได้รับคำตอบเพียงบางส่วน ทำให้เกิดข้อมูลสูญหายซึ่งส่งผลให้ไม่สามารถใช้ประโยชน์จากข้อมูลชุดนั้นอย่างเต็มที่ จึงนำข้อมูลสูญหายที่หาได้นั้นไปแทนค่าข้อมูลดังกล่าว

8. รายการอ้างอิง

ฐณัฐ วงศ์สายเชื้อ, Replace Missing Value: การแทนค่าสูญหายในโปรแกรม SPSS, แหล่งที่มา : https://www.youtube.com/watch?v=WzaeJ_HAqtk, 13 เมษายน 2559.

ทัตดา หิรัญพต, 2554, การเปรียบเทียบประสิทธิภาพระหว่างวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยและวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยแบบสโทแคสติก, ปัญหาพิเศษปริญญาตรี, มหาวิทยาลัยบูรพา, ชลบุรี.

นที ไทยธรรม และภาสกร สุวรรณโท, K-Nearest Neighbors (KNN), แหล่งที่มา : https://www.youtube.com/watch?v=WzaeJ_HAqtk, 17 พฤศจิกายน 2562.

ปิยะภรณ์ ประสิทธิ์วัฒนเสรี และสุคนธ์ ประสิทธิ์วัฒนเสรี, Missing Data and Management, แหล่งที่มา : http://dmbj.ejnal.com/e-journal/showdetail/?show_detail=T&art_id=1234, 5 พฤศจิกายน 2562.

ปูเป้ สุดศิลา, อำไพ ทองธีรภาพ และบุญอ้อม โฉมที, 2561, การเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรอิสระในการวิเคราะห์การถดถอยโลจิสติกแบบ 2 กลุ่ม, น. 1717-1718, การประชุมวิชาการและนำเสนอผลงานวิชาการระดับชาติ UTCC Academic Day ครั้งที่ 2, คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์, กรุงเทพฯ.

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

วิทยา พรพัชรพงศ์, โครงข่ายประสาทเทียม (Artificial Neural Networks - ANN), แหล่งที่มา : https://www.gotoknow.org/posts/163433, 4 กรกฎาคม 2555.

ศศิธร สมพงศ์นวกิจ, 2555, การเปรียบเทียบวิธีการประมาณค่าสูญหายแบบร่วม, วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยเกษตรศาสตร์, กรุงเทพฯ.

สายชล สินสมบูรณ์ทอง, 2560, การทำเหมืองข้อมูล, บริษัทจามจุรีโปรดักท์, กรุงเทพฯ.

สุรวัชร ศรีเปารยะ และสายชล สินสมบูรณ์ทอง, 2560, การเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มการเป็นโรคไตเรื้อรัง : กรณีศึกษาโรงพยาบาลแห่งหนึ่งในประเทศอินเดีย, ว. วิทยาศาสตร์และเทคโนโลยี 25(5): 844-846.

อโณทัย ศิลเทพาเวทย์, 2554, แบบจำลองเพื่อพัฒนาคุณภาพของผลิตภัณฑ์เอชจีเอในโรงงานอุตสาหกรรมฮาร์ดดิสก์ด้วยเทคนิคต้นไม้ตัดสินใจ, วิทยานิพนธ์ปริญญาโท, จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ.

Ajmera, A., 2017, Cardiovascular Disease of Framingham Massachusetts, Available Source: https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset, January 22, 2020.

Dong, Y. and Peng, C.Y.J., 2013, Principled missing data methods for researchers, SpringerPlus 2: 222.

Kyle, R., Therneau, T., Rajkumar, V., Offord, J., Larson, D., Plevak, M. and Melton, L.J., 2002, Monoclonal Gammopathy Data, Available Source: https://vincentarelbundock.github.io/Rdatasets/doc/survival/mgus2.html, February 14, 2020.

Peng, C.Y.J. and Zhu, J., 2008, Comparison of two approaches for handling missing covariates in logistic regression, Edu. Psychol. Measure. 68: 58-77.

Portuguese Banking Institution, 2012, Bank Marketing Data Set, Available Source: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing, January 21, 2020.

Ramana, B.V., 2012, Data Set of Liver Disease in Andhra Pradesh, India, Available Source: https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset), December 12, 2019.

Saravananselvamohan, 1992, Single Family Loan-Level Data, Available Source: https://www.kaggle.com/saravananselvamohan/freddie-mac-singlefamily-loanlevel-dataset/metada, January 25, 2020.

Wolberg, W.H., 1992, Biopsy Data on Breast Cancer Patients, Available Source: https://vincentarelbundock.github.io/Rdatasets/doc/MASS/ biopsy.html., January 3, 2020.