



ผลกระทบของขนาดหน้าต่างข้อมูลในการแทนค่าสูญหายข้อมูล อนุกรมเวลาแบบหลายตัวแปร

An Effect of Window Size in Missing Value Imputation for Multivariate Time Series Data

กรศิริณัฐ โรจนวรรณ (Kornsirinut Rothjanawan)¹* ดร.พยุ่ง มีสัจ (Dr.Phayung Meesad)**

(Received: April 4, 2018; Revised: June 4, 2018; Accepted: June 9, 2018)

บทคัดย่อ

การเก็บข้อมูลที่ไม่ครบถ้วน หรือข้อมูลที่สูญหาย ไม่สามารถนำไปใช้ในการวิเคราะห์ได้อย่างมีประสิทธิภาพ จึงจำเป็นต้องทำอย่างอื่นที่ทำการแก้ปัญหาค่าสูญหายก่อนนำไปใช้ ในกระบวนการเติมค่าสูญหายสามารถใช้ข้อมูลที่มีอยู่เพื่อพยากรณ์ค่าสูญหาย โดยเลือกข้อมูลเป็นลักษณะหน้าต่างข้อมูล บทความวิจัยนี้ ศึกษาผลกระทบของการเปลี่ยนขนาดหน้าต่างข้อมูลสำหรับวิธีการเติมค่าสูญหายในข้อมูลอนุกรมเวลาจำนวนมากแบบหลายตัวแปร การวิจัยเน้นที่การค้นหามาตรฐานหน้าต่างข้อมูลที่เหมาะสมสำหรับการสอนตัวแบบการเติมค่าสูญหาย นำข้อมูลเข้าสู่โมเดลการแทนที่ค่าสูญหายด้วยรูปแบบวิธีการทำเหมืองข้อมูล ในการทดสอบกับข้อมูลระดับน้ำเหนือเขื่อนและท้ายเขื่อนจำนวน 12 ตัวแปร วิธีการดำเนินการวิจัยเริ่มจากการสุ่มจำลองค่าสูญหาย สร้างโมเดลทำนายค่าสูญหาย และวัดประสิทธิภาพการจำลองค่าสูญหายโดยเปรียบเทียบขนาดหน้าต่างขนาดต่างๆ ได้แก่ 3 วัน 7 วัน และ 14 วัน ผลวิจัยพบว่า ขนาดหน้าต่างที่เหมาะสม คือ ขนาดหน้าต่างแบบ 14 วัน เมื่อใช้ร่วมกับโมเดลพยากรณ์ให้ประสิทธิภาพในการเติมค่าสูญหายในชุดทดสอบได้ดีที่สุด

ABSTRACT

Insufficient or incomplete data collection cannot be used to analyze effectively. It is mandatory that the missing values must be taken care of. In the imputation process, data windowing is used for moving existing data to prediction-missing ones. This research is to study the impact of changing the data window size for missing values imputation in multivariate time series data. The research focuses on finding suitable the data window size for using to train data to imputation the missing values and import data into missing imputation values using four techniques: Row Average, K-Nearest Neighbor (KNN), Fuzzy Logic System and Artificial Neural Network (ANN). The data used in the experiment were the water level above the dam and end the dam 12 variables. The research methodology started with missing data randomization, imputation models creation, and model performance evaluation. The window sizes in the experiment were 3 days, 7 days, and 14 days past period. The results found that appropriate window size was 14 days past period. The prediction model yielded best performance of missing value imputation on test data when 14 days past period data were applied for missing value imputation.

คำสำคัญ: การเติมค่าสูญหาย การเลื่อนหน้าต่างข้อมูล ข้อมูลหลายตัวแปร

Keywords: Imputation, Sliding window, Multivariate data

¹Correspondent author: phayung.m@it.kmutnb.ac.th

* นักศึกษา หลักสูตรปรัชญาดุษฎีบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้า พระนครเหนือ กรุงเทพฯ

** รองศาสตราจารย์ ภาควิชาการจัดการเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้า พระนครเหนือ กรุงเทพฯ

บทนำ

การทำเหมืองข้อมูลได้รับความนิยมอย่างแพร่หลายในงานประยุกต์ด้านต่าง ๆ ข้อมูลถูกนำไปสร้างโมเดลในการตัดสินใจ เช่น การจำแนกข้อมูล การจัดกลุ่มข้อมูล และการพยากรณ์ เป็นต้น ในการสร้างโมเดลตัดสินใจจำเป็นต้องมีข้อมูลที่สมบูรณ์ครบถ้วนมาใช้ [1] เพื่อให้ได้ผลการตัดสินใจที่ถูกต้องมากที่สุด แต่การเก็บข้อมูลอาจมีบางส่วนที่มีข้อมูลสูญหายหรือไม่สมบูรณ์เกิดขึ้น ซึ่งจำเป็นต้องมีการจัดการกับค่าที่สูญหาย ด้วยเทคนิคทางเหมืองข้อมูล (Data Mining) และเครื่องจักรเรียนรู้ (Machine Learning) ในการแทนค่าสูญหาย เพื่อให้ข้อมูลมีความสมบูรณ์ก่อนนำมาวิเคราะห์ เพื่อให้ได้ผลลัพธ์ที่มีประสิทธิภาพ

ข้อมูลอนุกรมเวลาเป็นข้อมูลที่มีอยู่ในงานประยุกต์ทั่วไปซึ่งเกิดขึ้นโดยธรรมชาติ ข้อมูลอนุกรมเวลาเป็นข้อมูลที่มีลักษณะเป็นลำดับข้อมูลตามเวลา หรือช่วงเวลาการจัดเก็บข้อมูลจะมีตัวชี้ที่สำคัญคือเวลา (time index) ซึ่งอาจจะเป็นวันที่ นาฬิกา วัน เดือน ปี หรืออื่นๆ โดยทั่วไปข้อมูล อนุกรมเวลาเป็นลำดับข้อมูลที่เรียงต่อเนื่อง ในระยะเวลาเท่าๆ กัน ในเวลา ตัวอย่างข้อมูล อนุกรมเวลาที่พบทั่วไป เช่น ข้อมูลราคากาแฟ ข้อมูลระดับน้ำใน มหาสมุทร ข้อมูลปริมาณน้ำฝนในแต่ละวัน ข้อมูลระดับน้ำเหนือ เขื่อนกั้นน้ำ เป็นต้น

การวิเคราะห์ข้อมูลอนุกรมเวลามีความสำคัญซึ่งสามารถนำไปวางแผนหรือการตัดสินใจในการปฏิบัติงานต่างๆ ในการวิเคราะห์ข้อมูลเบื้องต้นอาจใช้หลักสถิติเบื้องต้น เช่น การหาค่าสูงสุด การหาค่าต่ำสุด การหาค่าเฉลี่ย การหาค่าเฉลี่ยเคลื่อนที่ การหาเส้นสมการถดถอยเชิงเส้น เป็นต้น ปัจจุบันมีงานวิจัยที่มีการใช้เครื่องจักรเรียนรู้มาวิเคราะห์แนวโน้มหรือพยากรณ์ค่าในอนาคตของข้อมูลอนุกรมเวลา เช่น การพยากรณ์ปริมาณน้ำฝน การพยากรณ์การขึ้นลงของหุ้น เป็นต้น ปัญหาสำคัญอย่างหนึ่ง [3] ที่พบของข้อมูลอนุกรมเวลา ได้แก่ การสูญหายของข้อมูลจำนวนมากสาเหตุของข้อมูลสูญหายอาจเกิดขึ้นจากการที่ผู้จัดเก็บข้อมูลไม่ได้ทำการจัดเก็บตามกำหนดเวลา หรือการจัดเก็บข้อมูลด้วยเซนเซอร์ หรือเครื่องจักรแต่เกิดการชำรุดไม่สามารถอ่านและจัดเก็บข้อมูลตามโปรแกรมที่กำหนด

งานวิจัยด้านการเติมค่าสูญหายในอนุกรมมามากมาย [4] ส่วนใหญ่มีการใช้เทคนิคมากมายหลายแบบ แบบง่ายสุดทางสถิติ ได้แก่ การเติมค่าด้วยค่าเฉลี่ย ค่าฐานนิยม และค่ามัธยฐาน การเติมค่าด้วยรูปแบบเชิงคณิตศาสตร์ ได้แก่ เทคนิคสมการถดถอยเชิงเส้น (linear regression) เทคนิคสมการถดถอยแบบไม่เป็นเชิงเส้น (nonlinear regression) ค่าเฉลี่ยแถว (row average) หรือรูปแบบเครื่องจักรการเรียนรู้ (machine learning) ได้แก่ เทคนิควิธีเพื่อนบ้านใกล้เคียง (K-Nearest Neighbor: KNN) เทคนิควิธีโครงข่ายประสาทเทียม (Artificial Neural Network) และ เทคนิควิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) เป็นต้น ซึ่งข้อมูลสูญหายในข้อมูลอนุกรมเวลา มีประเด็นที่น่าสนใจในกรณีข้อมูลอนุกรมเวลาที่มีการสูญหายจำนวนมาก การจัดรูปแบบข้อมูลที่เหมาะสมจะเป็นอย่างไร และตัวแบบการเติมค่าที่เหมาะสมจะเป็นอย่างไร จากงานวิจัยของ พยุง และกรสิริณัฐ [5] ได้นำเสนอการค้นหารูปแบบ ข้อมูลที่เหมาะสมสำหรับใช้สอนตัวแบบการเติมค่าสูญหาย ผลวิจัยพบว่า เทคนิควิธีโครงข่ายประสาทเทียม (Artificial Neural Network) ให้ผลการทำนายในชุดทดสอบได้ดีที่สุด และเมื่อนำไปใช้ในการแทนที่ค่าสูญหายให้ผลลัพธ์คล้ายค่าจริง แต่ในประเด็นการเลื่อนหน้าต่างข้อมูลของงานวิจัยนี้ยังไม่ได้มีการศึกษา จึงเป็นประเด็นที่สนใจและนำมาศึกษาในงานวิจัยฉบับนี้

สำหรับงานวิจัยครั้งนี้ ผู้วิจัยศึกษาผลกระทบของการเปลี่ยนขนาดหน้าต่างข้อมูลสำหรับวิธีการเติมค่าสูญหายในข้อมูลอนุกรมเวลาจำนวนมากแบบหลายตัวแปร ซึ่งจะเน้นที่การค้นหามาตรการหน้าต่างข้อมูลที่เหมาะสมสำหรับใช้ในการสอนตัวแบบการเติมค่าสูญหาย นำข้อมูลเข้าสู่โมเดลการแทนที่ค่าสูญหายแบบต่างๆ [6] ทั้งนี้โมเดลที่จะใช้สำหรับ



การวัดเปรียบเทียบจะประกอบด้วย วิธีค่าเฉลี่ยแถว (Row Average) วิธีเพื่อนบ้านใกล้เคียง (K-Nearest Neighbor: KNN) ระบบฟัซซี่ลอจิก (Fuzzy Logic System) และ โครงข่ายประสาทเทียม (Artificial Neural Network)

วรรณกรรมที่เกี่ยวข้อง

1. ข้อมูลอนุกรมเวลา

ข้อมูลอนุกรมเวลา [7] เป็นข้อมูลลำดับของค่าที่กำหนดไว้ในช่วงเวลาที่เว้นระยะเท่ากัน เช่น ชั่วโมง วัน เดือน ปี เป็นต้น รูปแบบทางคณิตศาสตร์ของข้อมูลอนุกรมเวลาแสดงดังสมการ (1) และตัวอย่างภาพข้อมูลอนุกรมเวลาราคาหุ้น แสดงได้ดังภาพที่ 1

$$\sum_{i=-\infty}^{+\infty} d_i = \dots + d_{-2} + d_{-1} + d_0 + d_1 + d_2 + \dots \quad (1)$$

เมื่อ d_i คือ ข้อมูลที่เวลา i



ภาพที่ 1 ตัวอย่างข้อมูลอนุกรมเวลาราคาหุ้น

การวิเคราะห์ข้อมูลอนุกรมเวลาถูกใช้งานอย่างกว้างขวาง เช่น การพยากรณ์ทางเศรษฐกิจ การคาดการณ์ยอดขายการวิเคราะห์งบประมาณ การวิเคราะห์ตลาดหุ้น และกระบวนการและการควบคุมคุณภาพ การศึกษาสินค้าคงคลัง การคาดการณ์ปริมาณงาน การศึกษา สาธารณูปโภค และการสำรวจสำมะโนประชากร เป็นต้น

การวิเคราะห์ข้อมูลอนุกรมเวลามีหลายเทคนิควิธี อาจจะใช้หลักการรูปแบบทางคณิตศาสตร์หรือทางสถิติ รวมถึงโมเดลเครื่องจักรเรียนรู้ ซึ่งล้วนเป็นการสร้างโมเดลสำหรับพยากรณ์ค่าในอนาคตได้อย่างถูกต้องและแม่นยำ [8] ปัญหาสำคัญอย่างหนึ่งในการสร้างโมเดลพยากรณ์คือ ข้อมูลสูญหายซึ่งเป็นประเด็นสำคัญที่เป็นอุปสรรคในการสร้างโมเดลพยากรณ์ข้อมูลอนุกรมเวลา ข้อมูลที่สูญหายอาจจะส่งผลกระทบต่อกระบวนการสร้างโมเดลที่เหมาะสมสำหรับการพยากรณ์ค่าในอนาคต ความรุนแรงของผลกระทบขึ้นอยู่กับองค์ประกอบจากหลายส่วน หากจำนวนของข้อมูลที่สูญหายมีจำนวนมากก็อาจไม่สามารถสร้างโมเดลพยากรณ์ได้จึงต้อง มีการแก้ปัญหาข้อมูลสูญหาย ก่อนการนำข้อมูลไปสร้างโมเดลพยากรณ์

2. การแทนค่าข้อมูลสูญหาย

ปัจจุบันการแทนค่าข้อมูลสูญหายจะมีงานวิจัยหลายเทคนิควิธีที่ถูกนำไปใช้ เช่น เทคนิคทางสถิติ เทคนิคทางเหมืองข้อมูล (Data Mining) และเครื่องจักรเรียนรู้ (Machine Learning) เทคนิคเหล่านี้ถูกนำไปใช้สำหรับการแทนที่ค่าข้อมูลสูญหายอย่างกว้างขวางตามรายงานเอกสารการวิจัย จากการศึกษางานวิจัยที่เกี่ยวข้องพบว่า ได้มีงานวิจัยจำนวนมากได้นำเสนอวิธีการแทนค่าข้อมูลที่สูญหาย [9-10] โดยการนำเทคนิค เช่น ค่าเฉลี่ยแถว (Row Average) เพื่อนบ้านใกล้เคียง (K-Nearest Neighbor: KNN) ระบบคลุมเครือ (Fuzzy Logic Systems) โครงข่ายประสาทเทียม (Artificial Neural Network) เป็นต้น

2.1 ค่าเฉลี่ยแถว

ค่าเฉลี่ยแถวเป็นวิธีการทางสถิติแบบง่ายๆ [11] ที่นิยมใช้เนื่องจากไม่มีความซับซ้อนและใช้งานง่าย ซึ่งใช้สำหรับการแทนค่าข้อมูลที่สูญหาย เป็นการคำนวณหาค่าเฉลี่ยของตัวแปรเดียวกันเพื่อแทนที่ค่าสูญหาย ดังสมการ (2)

$$\bar{d} = \frac{1}{N} \sum_{t=0}^N d_t \quad (2)$$

เมื่อ $\bar{d}$ คือ ค่าเฉลี่ยแถวหรือตัวแปร
 d_t คือ ข้อมูลที่เวลา t

วิธีการนี้เป็นวิธีการที่ง่ายต่อการทำงาน แต่ผลลัพธ์จากวิธีการนี้ อาจถูกโน้มน้าวเนื่องจากค่าที่อยู่นอกกลุ่มได้จึงให้ประสิทธิภาพไม่ค่อยดีนัก

อัลกอริทึมที่ 1 : ค่าเฉลี่ยแถว (Row Average Impute)

ขั้นตอนวิธีการแทนค่าด้วยเฉลี่ยแถว (Row Average Impute)

ขั้นที่ 1 : จัดเตรียมข้อมูลในรูปแบบเมทริกซ์ของ Attribute Input คือ x สำหรับชุดข้อมูลสอน ตั้งค่า Vector ของ Attribute Target คือ t สำหรับการพยากรณ์ค่าที่สูญหาย

ขั้นที่ 2 : ค้นหาข้อมูลสูญหายของ Attribute Target คือ t_m

ขั้นที่ 3 : จากนั้น แต่ละเป้าหมายที่มีค่าข้อมูลสูญหาย ซึ่งจะอ่านจากค่าของข้อมูล ป้อนเข้าของ Attribute Target คือ x_m

ขั้นที่ 4 : จากนั้น ส่งค่าที่มีการเติมข้อมูลสูญหายคืน, t_{mi} โดยการคำนวณจากค่าเฉลี่ยของ t

2.2 เพื่อนบ้านใกล้เคียง

เทคนิคเพื่อนบ้านใกล้เคียง (K-Nearest Neighbor: KNN) [12] เป็นเทคนิควิธี ที่ได้รับความนิยมในการใช้งานอย่างมากเช่นกัน เนื่องจากเป็นวิธีการที่ง่ายและมีประสิทธิภาพซึ่งสามารถนำไปประยุกต์ใช้กับงานได้อย่างหลากหลาย เช่น งานทางด้านการจำแนกข้อมูล (Classification) รวมถึงงานทางด้านการแทนที่ข้อมูลที่สูญหาย (Missing Values Imputation) อัลกอริทึมสำหรับการใช้ KNN ในการแทนที่ค่าสูญหาย แสดงดังอัลกอริทึมที่ 2

อัลกอริทึมที่ 2 : เพื่อนบ้านใกล้เคียง (K-Nearest Neighbor Impute)

ขั้นตอนวิธีการแทนค่าด้วยเพื่อนบ้านใกล้เคียง (K-Nearest Neighbor Impute)

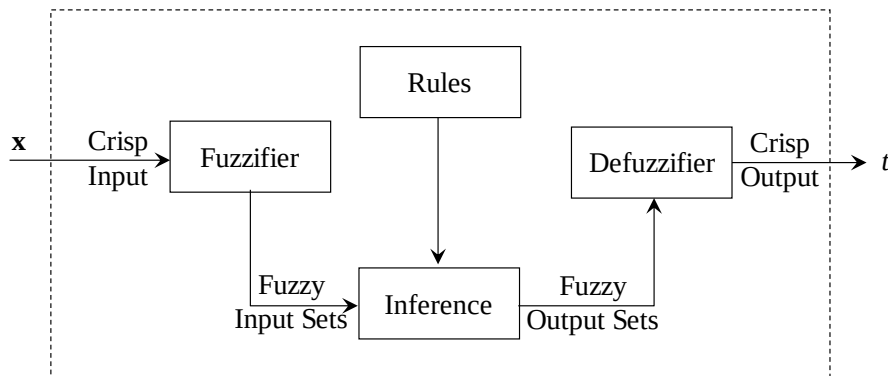
ขั้นที่ 1 : จัดเตรียมข้อมูลในรูปแบบเมทริกซ์ของ Attribute Input คือ x สำหรับชุดข้อมูลสอน กำหนดค่า Vector ของ Attribute Target คือ t สำหรับการพยากรณ์ค่าที่สูญหาย กำหนดค่า K



- ขั้นที่ 2 : ค้นหาข้อมูลสูญหายของ Attribute Target คือ t_m
- ขั้นที่ 3 : จากนั้น แต่ละเป้าหมายที่มีค่าข้อมูลสูญหาย ซึ่งจะอ่านจากค่าของข้อมูลป้อนเข้าของ Attribute Target คือ x_m
- ขั้นที่ 4 : คำนวณค่าระยะห่างระหว่างคุณลักษณะ ของแถว x_m และแต่ละหลักของ x
- ขั้นที่ 5 : เรียงลำดับจากน้อยไปมาก
- ขั้นที่ 6 : จากนั้นส่งค่าที่มีการเติมข้อมูลสูญหายคืน, t_{mi} โดยคำนวณจากค่าเฉลี่ยของข้อมูลใกล้เคียงสุดของ x_m
- จำนวน K ข้อมูล

2.3 ระบบฟัซซีลอจิก

ฟัซซีลอจิก [13] เป็นตรรกศาสตร์ภายใต้แนวคิดเหตุการณ์ต่างๆ จะมีความไม่แน่นอน (uncertain) แต่มีความคลุมเครือ (fuzzy) ซึ่งเป็นพื้นฐานของระบบฟัซซีลอจิก (fuzzy logic system) มีหลักการให้เหตุผลเลียนแบบการตัดสินใจของมนุษย์ และสามารถนำไปประยุกต์ใช้งานในการตัดสินใจต่างๆ อย่างมากมาย ตัวอย่างระบบฟัซซีลอจิกแสดงดังภาพที่ 2 และอัลกอริทึมสำหรับการแทนค่าตามอัลกอริทึมที่ 3



ภาพที่ 2 ระบบฟัซซีลอจิก

อัลกอริทึมที่ 3 : ระบบฟัซซีลอจิก (Fuzzy Logic System)

ขั้นตอนวิธีการแทนค่าด้วยวิธีระบบฟัซซีลอจิก (Fuzzy Logic System)

ขั้นที่ 1 : จัดเตรียมข้อมูลในรูปเมทริกซ์ของ Attribute Input คือ x สำหรับชุดข้อมูลสอน กำหนดค่า Vector ของ Attribute Target คือ t สำหรับการพยากรณ์ค่าที่สูญหาย กำหนดโมเดลสำหรับสอน

ขั้นที่ 2 : สอนโมเดลโดยกำหนดชุดสอน x_{tr} , t_{tr}

ขั้นที่ 3 : ค้นหาข้อมูลสูญหายของ Attribute Target คือ t_m

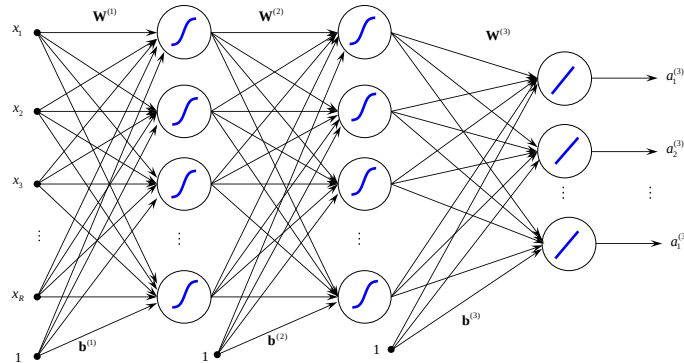
ขั้นที่ 4 : จากนั้น แต่ละเป้าหมายที่มีค่าข้อมูลสูญหาย ซึ่งจะอ่านจากค่าของข้อมูลป้อนเข้าของ Attribute Target คือ x_m

ขั้นที่ 5 : จากนั้นส่งค่าที่มีการเติมข้อมูลสูญหายคืน, t_{mi} โดยใช้ระบบฟัซซีลอจิก นำชุดสอนโมเดลระบบฟัซซีลอจิกไปใช้กับข้อมูลสูญหายอินพุต x_m

2.4 โครงข่ายประสาทเทียม

โครงข่ายประสาทเทียม [14] เป็นเครื่องจักรการเรียนรู้ซึ่งมีที่มาจากการจำลองสมองมนุษย์ ซึ่งเป็นการจำลองการทำงานของสมองมนุษย์ ด้วยสมการทางคณิตศาสตร์ โดยสามารถเรียนรู้ข้อมูลเกี่ยวกับสิ่งที่ต้องการสอนให้โครงข่ายเทียม เพื่อนำไปพยากรณ์ข้อมูลใหม่ในอนาคตได้ โครงข่ายประสาทเทียมมีความสามารถในการเรียนรู้เพื่อจำแนกกลุ่มข้อมูล (classification) และสามารถพยากรณ์ข้อมูลที่เป็นค่าทศนิยม (regression) จึงทำให้ได้รับความนิยม

อย่างแพร่หลายใน การประยุกต์ใช้งานด้านต่าง ๆ เช่น ระบบควบคุมทางวิศวกรรม ระบบตัดสินใจทางด้านการบริหาร ระบบพยากรณ์ทางด้านเศรษฐศาสตร์ เป็นต้น โครงข่ายประสาทเทียมที่นิยมใช้ ได้แก่ โครงข่ายประสาทเทียมเปอร์เซปตรอนแบบหลายชั้น ดังตัวอย่างในภาพที่ 3



ภาพที่ 3 โครงข่ายประสาทเทียมแบบหลายชั้น

อัลกอริทึมที่ 4 : โครงข่ายประสาทเทียม (Artificial Neural Network)

ขั้นตอนวิธีการแทนค่าด้วยโครงข่ายประสาทเทียม (Artificial Neural Network)

ขั้นที่ 1 : จัดเตรียมข้อมูลในรูปแบบเมทริกซ์ของ Attribute Input คือ x สำหรับชุดข้อมูลสอน กำหนดค่า Vector ของ Attribute Target คือ t สำหรับการพยากรณ์ค่าที่สูญหาย กำหนดโมเดลสำหรับสอน

ขั้นที่ 2 : สอนโมเดลโดยกำหนดชุดสอน x_{tr}, t_{tr}

ขั้นที่ 3 : ค้นหาข้อมูลสูญหายของ Attribute Target คือ t_m

ขั้นที่ 4 : จากนั้น แต่ละเป้าหมายที่มีค่าข้อมูลสูญหาย ซึ่งจะอ่านจากค่าของข้อมูลป้อนเข้าของ Attribute Target คือ x_m

ขั้นที่ 5 : จากนั้นส่งค่าที่มีการเติมข้อมูลสูญหายคืน, t_{mi} โดยโครงข่ายประสาทเทียมเปอร์เซปตรอนแบบหลายชั้น นำชุดสอนโมเดลโครงข่ายประสาทเทียมไปใช้กับข้อมูลสูญหายอินพุต x_m

2.5 วิธีการวิเคราะห์ความถูกต้องของตัวแบบ (K-fold Cross-Validation)

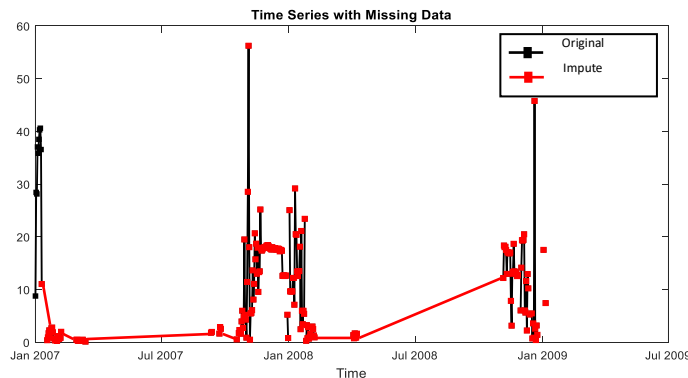
การตรวจสอบไขว้กัน (Cross-Validation) [15] เป็นวิธีการตรวจสอบค่าความผิดพลาดในการคาดการณ์ของตัวแบบ โดยพื้นฐานของวิธีการตรวจสอบไขว้กัน คือ การสุ่มตัวอย่าง โดยเริ่มจากการแบ่งชุดข้อมูลออกเป็น ส่วนๆ และนำบางส่วนจากชุดข้อมูลนั้นไปสอนตัวแบบพยากรณ์และอีกส่วนของข้อมูลถูกนำไปใช้ทดสอบตัวแบบที่ถูกสอน เป็นการตรวจสอบไขว้กัน วิธีการแบบนี้มีความนิยมซึ่งมักเรียกว่า K-fold Cross-Validation โดยจะแบ่งข้อมูลออกเป็น K ชุด เท่าๆ กัน เช่น $K = 5$ หมายถึง จะมีชุดข้อมูลจำนวน 5 ชุด ซึ่งจะทำให้การคำนวณค่าความผิดพลาด 5 รอบ โดยแต่ละรอบของการคำนวณ จะเลือกข้อมูลออกมา 1 ชุดเพื่อเป็นข้อมูลทดสอบ และข้อมูลอีก 4 ชุด จะถูกใช้เป็นข้อมูลสำหรับการเรียนรู้ ดังภาพที่ 4

รอบที่ 1 : ชุดสอน	2	3	4	5	ชุดทดสอบ	1
รอบที่ 2 : ชุดสอน	3	4	5	1	ชุดทดสอบ	2
รอบที่ 3 : ชุดสอน	4	5	1	2	ชุดทดสอบ	3
รอบที่ 4 : ชุดสอน	5	1	2	3	ชุดทดสอบ	4
รอบที่ 5 : ชุดสอน	1	2	3	4	ชุดทดสอบ	5

ภาพที่ 4 การตรวจสอบไขว้กัน กำหนด $K=5$

2.6 ปัญหาดังเดิมข้อมูลสูญหายจำนวนมาก

ข้อมูลที่มีการสูญหายจำนวนมาก [16] มีความยากในการทำนายค่าแทนที่สูญหาย เทคนิคดั้งเดิมส่วนมากสามารถประยุกต์ใช้ในการแทนที่ค่าสูญหายได้ แต่ข้อมูลอาจจะไม่ถูกต้องเท่าที่ควร ดังในภาพที่ 5 ซึ่งเป็นตัวอย่างการออกแบบโมเดลการแทนที่ค่าสูญหายที่ไม่เหมาะสม ซึ่งสังเกตได้จากการแทนค่าข้อมูลที่มีการสูญหายจำนวนมาก จะมีเป็นลักษณะเป็นเส้นตรง



ภาพที่ 5 ตัวอย่างปัญหาการแทนที่ค่าสูญหายจำนวนมากในข้อมูลอนุกรม

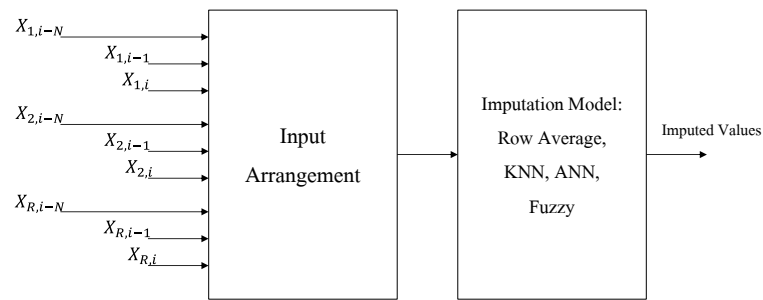
วิธีการดำเนินการวิจัย

การวิจัยครั้งนี้มีจุดมุ่งหมายเพื่อศึกษาผลกระทบของการเปลี่ยนขนาดหน้าต่างข้อมูลสำหรับวิธีการเติมค่าสูญหายในข้อมูลอนุกรมเวลาจำนวนมากแบบหลายตัวแปร การวิจัยเน้นที่การค้นหามาตรฐานหน้าต่างข้อมูล (Sliding window) ที่เหมาะสมสำหรับการสอนตัวแบบการเติมค่าสูญหาย นำข้อมูลเข้าสู่โมเดลการแทนที่ค่าสูญหายแบบต่างๆ โดยมีเงื่อนไขเบื้องต้น [17] คือ ตัวแปรต้นต้องมีความสัมพันธ์กับตัวแปรที่สูญหายหรือตัวแปรตาม วิธีการจัดรูปแบบข้อมูลสำหรับการใช้ในการแทนที่ค่าสูญหาย

1. รูปแบบการแทนที่ค่าสูญหาย

ในการวิจัยผู้วิจัยได้ออกแบบขั้นตอนการเปรียบเทียบวิธีการแทนที่ค่าสูญหายจำนวนมาก ในข้อมูลอนุกรมแบบหลายตัวแปร โดยเริ่มจากการออกแบบรูปแบบข้อมูล สำหรับประยุกต์ใช้กับเทคนิค การแทนที่ค่าสูญหาย ได้แก่ เทคนิคค่าเฉลี่ยแถว (Row Average) เทคนิคเพื่อนบ้านใกล้เคียง (K-Nearest Neighbor: KNN) เทคนิคระบบฟัซซีลอจิก (Fuzzy Logic System) และเทคนิคโครงข่ายประสาทเทียม (Artificial Neural Network) โดยมีคำถามวิจัยว่าเทคนิคใดที่มีความเหมาะสมในการแทนที่ค่าสูญหายจำนวนมากได้ดีที่สุด

สำหรับกระบวนการวิจัยเริ่มต้นจาก [18] 1) การวิเคราะห์และการสำรวจตรวจสอบข้อมูล 2) การจัดข้อมูลผิดปกติ 3) การค้นหาข้อมูลสูญหาย 4) การจัดข้อมูลอนุกรมให้อยู่ในรูปแบบสำหรับการสอนเครื่องจักรเรียนรู้ 5) การสอนเครื่องจักรเรียนรู้ด้วยข้อมูลชุดสอน 6) การทดสอบโมเดลการแทนที่ค่าสูญหายด้วยชุดข้อมูลทดสอบ และ 7) การประยุกต์ใช้โมเดลแทนที่ค่าข้อมูลสูญหาย



ภาพที่ 6 รูปแบบวิธีการแทนที่ค่าสูญหายในอนุกรมเวลาหลายตัวแปร

รูปแบบวิธีการแทนค่าสูญหายจำนวนมากหลายตัวแปรในข้อมูลอนุกรมเวลาที่ผู้วิจัยนำเสนอ ในการวิจัยครั้งนี้ แสดงดังภาพที่ 6 โดยมีขั้นตอนการทำงานดังนี้ 1) รับเข้าข้อมูลอนุกรมเวลาแบบหลายตัวแปร 2) เลือกตัวแปรเป้าหมาย สำหรับแทนที่ค่าสูญหาย 3) สำหรับตัวแปรอินพุตทำการจัด ข้อมูลอนุกรมโดยเลื่อนหน้าต่างข้อมูล แต่ละตัวแปร ย้อนหลังจำนวน N วัน เพื่อจัดเป็นเวกเตอร์อินพุต และตั้งค่าเป้าหมายที่เวลาล่วงหน้า 1 วัน ($i + 1$) จากตัวแปรที่เลือกไว้ เป็นค่าเป้าหมายสำหรับการแทนที่ค่าสูญหาย

ในการสอนเครื่องจักรการเรียนรู้ ผู้วิจัยนำเสนอวิธีการจัดรูปแบบข้อมูลโดยใช้ข้อมูลในอดีต เพื่อพยากรณ์ ข้อมูลในอนาคต ซึ่งต้องทำการเลื่อนหน้าต่างข้อมูล โดยจัดให้อยู่ในรูปเวกเตอร์ และเมทริกซ์ตามสมการ (3) – (7) จากนั้นนำข้อมูลที่จัดรูปแบบแล้ว เข้าสู่โมเดลทั้ง 4 แบบ เพื่อพยากรณ์ข้อมูลในอนาคต ดังภาพที่ 6

$$\mathbf{x}_{j,i} = [x_{j,i-N} \quad \cdots \quad x_{j,i-1} \quad x_{j,i}], j = 1, \dots, R \quad (3)$$

$$\mathbf{p}_i = [\mathbf{x}_{1,i} \quad \mathbf{x}_{2,i} \quad \cdots \quad \mathbf{x}_{R,i}], i = 1, \dots, M \quad (4)$$

$$t_i = x_{Q,i}, Q \notin \{j = 1, \dots, R\} \quad (5)$$

$$\mathbf{P} = [\mathbf{p}_1^T \quad \mathbf{p}_2^T \quad \cdots \quad \mathbf{p}_M^T] \quad (6)$$

$$\mathbf{t} = [t_1 \quad t_2 \quad \cdots \quad t_M] \quad (7)$$

$$\hat{y}_i = f(\mathbf{p}_i^T, parameters) \quad (8)$$

เมื่อ	$\mathbf{x}_{j,i}$	คือ เวกเตอร์หน้าต่างข้อมูลที่มีตัวประกอบเป็นข้อมูลจาก ตัวแปร j ที่เวลา $i-N, \dots, i-1$, และ i
	$\mathbf{p}_i$	คือ เวกเตอร์ที่ประกอบด้วย $\mathbf{x}_{j,i}$
	$\mathbf{P}$	คือ เมทริกซ์ที่มีคอลัมน์เป็นอินพุตเวกเตอร์
	$\mathbf{t}$	คือ เป็นเวกเตอร์ค่าเป้าหมายตามคอลัมน์ของเมทริกซ์ $\mathbf{P}$
	$\hat{y}_i$	คือ ค่าสูญหายที่ถูกแทนที่
	f	คือ โมเดลการแทนที่ค่าสูญหาย
	N	คือ จำนวนวันย้อนหลังในอดีต
	M	คือ จำนวนเรคคอร์ด
	R	คือ จำนวนตัวแปรอินพุต
	Q	คือ ตัวแปรเป้าหมายซึ่งต้องการแทนที่ค่าสูญหาย
	$parameters$	คือ ค่าพารามิเตอร์ของโมเดลแทนที่ค่าสูญหาย



2. ข้อมูลสำหรับการทดลอง

สำหรับข้อมูลที่ใช้ในการวิจัยครั้งนี้เป็นข้อมูลจากกรมชลประทานที่ 15 โครงการพระราชดำริลุ่มน้ำปากพนัง ในช่วงระยะเวลาปี 2550 - 2556 โดยเป็นข้อมูลอนุกรมเวลา มี 12 ตัวแปร ได้แก่ Var1: ข้อมูลระดับ เหนื่อน้ำเวลา 6.00 น. Var2: ข้อมูลระดับเหนื่อน้ำเวลา 12.00 น. Var3: ข้อมูลระดับเหนื่อน้ำเวลา 18.00 น. Var4: ข้อมูลระดับท้ายน้ำเวลา 6.00 น. Var5: ข้อมูลระดับท้ายน้ำเวลา 12.00 น. Var6: ข้อมูลระดับท้ายน้ำ เวลา 18.00 น. Var7: ระดับน้ำสูงสุดหน้าประตู Var8: ระดับน้ำสูงสุดท้ายประตู Var9: อัตราระบายน้ำ/วินาที Var10: อัตราระบายน้ำ/วัน Var11: ปริมาณน้ำฝน และ Var12: ปริมาณน้ำเก็บกัก

ในการวิจัยครั้งนี้ ได้ศึกษาผลกระทบของการเปลี่ยนขนาดหน้าต่างข้อมูลสำหรับวิธีการเดิค่าสูญหายในข้อมูลอนุกรมเวลาจำนวนมากแบบหลายตัวแปร โดยเน้นที่การค้นหามาตรฐานหน้าต่างข้อมูลที่เหมาะสมสำหรับการใช้ในการสอนตัวแบบการเดิค่าสูญหาย และนำข้อมูลเข้าสู่โมเดลการแทนที่ค่าสูญหายแบบต่างๆ โดยใช้ข้อมูลหลายตัวแปร เป็นข้อมูลอินพุตสำหรับการแทนที่ค่าสูญหาย และใช้หลักการข้อมูลในอดีตทำนายค่าในอนาคตด้วยหลักการเลื่อนหน้าต่างข้อมูล (Sliding window) ซึ่งจะทำให้การคัดแยกข้อมูลสูญหายออกไว้อีกก่อน และนำข้อมูลที่สมบูรณ์มาใช้ในการสร้างโมเดล ซึ่งจะกำหนดขนาดหน้าต่างข้อมูล 3 รูปแบบ คือ รูปแบบที่ 1 คือ 3 วัน รูปแบบที่ 2 คือ 7 วัน และรูปแบบที่ 3 คือ 14 วัน ตามลำดับนั้นคือ ใช้ข้อมูลวันปัจจุบันร่วมกับข้อมูลในอดีตอีก 2 วัน 6 วัน และ 13 วัน ย้อนหลังตามลำดับ เพื่อทำการเปรียบเทียบขนาดหน้าต่างข้อมูลรูปแบบที่เหมาะสม ก่อนนำไปเป็นอินพุตและค่าเป้าหมาย สำหรับสอน โมเดลการแทนค่าสูญหายของตัวแปรที่มีข้อมูลสูญหาย โดยจะแบ่งข้อมูลสมบูรณ์ออกเป็น ชุดสอน (Training Data) และชุดทดสอบ (Test Data) เพื่อให้การสอนโมเดลมีความน่าเชื่อถือ โมเดลจะใช้วิธีการสอนแบบตรวจสอบไขว้ (k-fold cross-validation) ซึ่งวิธีการที่ได้รับการยอมรับว่าเหมาะสำหรับการสอนและทดสอบเครื่องจักรการเรียนรู้ และในการวิจัยครั้งนี้ได้กำหนด $k = 5$ นั่นคือ แบ่งข้อมูลเป็น 5 ส่วน โดยใช้ข้อมูล 4 ส่วนทำการสอนโมเดลใช้ข้อมูล 1 ส่วนทดสอบโมเดล และเวียนสอนและทดสอบไขว้ 5 รอบ จนครบทุกส่วน สำหรับเทคนิคการแทนที่ค่าสูญหายใช้ 4 เทคนิค คือ เทคนิคค่าเฉลี่ยแถว (Row Average) เทคนิคเพื่อนบ้านใกล้เคียง (K-Nearest Neighbor: KNN) เทคนิคโครงข่ายประสาทเทียม (Artificial Neural Network) และระบบฟัซซีลอจิก (Fuzzy Logic System) สำหรับการวัดประสิทธิภาพของโมเดลแทนที่ค่า ผู้วิจัยใช้วิธีคำนวณรากที่สองของค่าเฉลี่ยผิดพลาดยกกำลังสอง (root mean squared error: *rmse*) ดังสมการ (9)

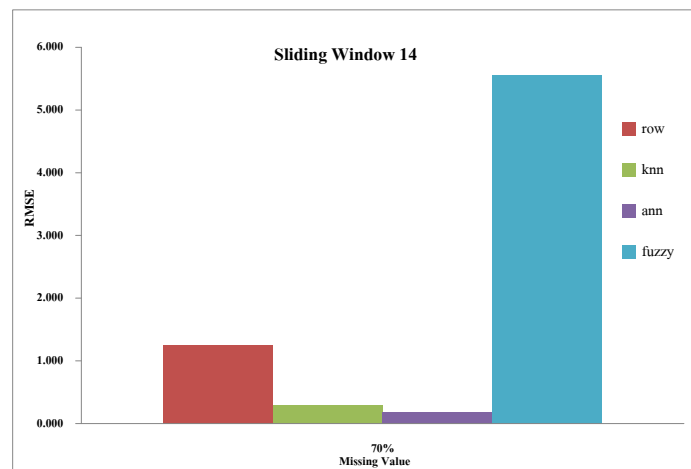
$$rmse = \sqrt{\frac{1}{L} \sum_{i=1}^L (t_i - y_i)^2} \quad (9)$$

เมื่อ	t_i	คือ ค่าเป้าหมายจริงของข้อมูลทดสอบ
	y_i	คือ ค่าพยากรณ์ จากระบบการแทนที่ค่าสูญหาย
	L	คือ จำนวนข้อมูลทดสอบ

ผลการดำเนินงานวิจัย

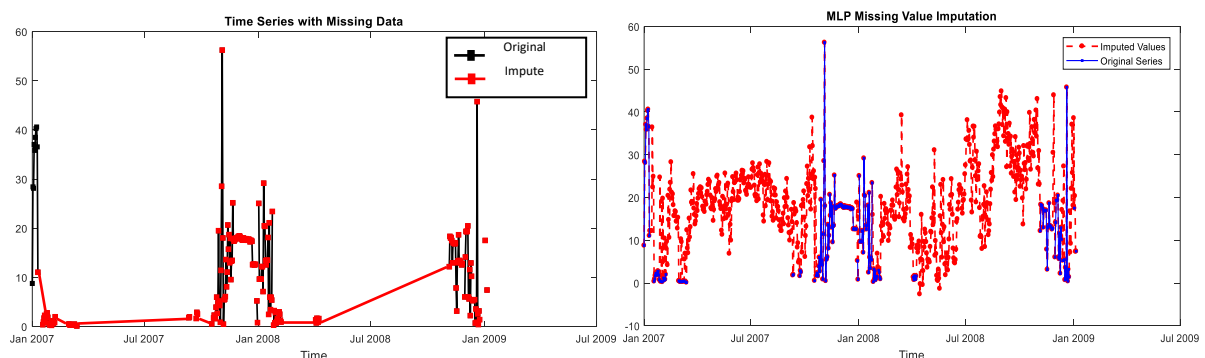
ในการดำเนินการวิจัยครั้งนี้ ผู้วิจัยศึกษาผลกระทบของการเปลี่ยนขนาดหน้าต่างข้อมูลสำหรับวิธีการเดิค่าสูญหายในข้อมูลอนุกรมเวลาจำนวนมากแบบหลายตัวแปร โดยเน้นที่การค้นหามาตรฐานหน้าต่างข้อมูลที่เหมาะสมสำหรับการใช้ในการสอนตัวแบบการเดิ ค่าสูญหาย โดยกำหนดขนาดหน้าต่างข้อมูล 3 รูปแบบ คือ รูปแบบที่ 1 คือ 3 วัน รูปแบบที่ 2 คือ 7 วัน และรูปแบบที่ 3 คือ 14 วัน ตามลำดับนั้นคือ ใช้ข้อมูลวันปัจจุบันร่วมกับข้อมูลในอดีตอีก 2 วัน 6 วัน และ 13 วัน ย้อนหลังตามลำดับ ตามภาพที่ 6 โดยจัดรูปแบบสมการ การเลื่อนข้อมูลดังสมการที่ (3) – (7) โดยเขียนสคริปต์

ด้วยภาษา Matlab ก่อนจะนำข้อมูลเข้าสู่โมเดลการแทนที่ค่าสูญหายทั้ง 4 เทคนิค คือ เทคนิคค่าเฉลี่ย (Row Average) ตามอัลกอริทึมที่ 1 และเทคนิคเพื่อนบ้านใกล้เคียง (KNN) ตามอัลกอริทึมที่ 2 สำหรับเทคนิควิธีโครงข่ายประสาทเทียม (Artificial Neural Network) ตามอัลกอริทึมที่ 4 ใช้ Neural Network Toolbox ชนิด Feed Forward Multilayer Perceptron Network (MLP) แบบ 2 ชั้นซ่อน โดยกำหนดจำนวนชั้นละ 15 นิวรอนแบบ Hyperbolic Tangent Sigmoidal (tansig) ทั้งสองชั้นซ่อน และ ชั้นเอาต์พุต เป็นแบบเชิงเส้น และสอนด้วยวิธี trainlm (Levenberg-Marquardt) ส่วนระบบฟัซซี่ (Fuzzy Logic System) ตามอัลกอริทึมที่ 3 ใช้เป็นแบบ ANFIS ใน Fuzzy System Toolbox กำหนดใช้ฟังก์ชันความเป็นสมาชิกของฟัซซี่เซตแบบ Gaussian Bell Membership Function (gbellmf) จำนวน 5 ฟัซซี่เซต และสอนด้วยวิธี Hybrid สำหรับการวัดประสิทธิภาพของโมเดลแทนที่ค่าผู้วิจัยใช้วิธีคำนวณรากที่สองของค่าเฉลี่ยผิดพลาดยกกำลังสอง (root mean squared error: *rmse*)



ภาพที่ 7 ผลการเติมค่าสูญหายในข้อมูลอนุกรมเวลาทั้ง 4 เทคนิค โดยการเลื่อนหน้าต่างข้อมูลแบบ 14 วัน ของตัวแปรที่ 10 (Var10)

จากภาพที่ 7 การเลื่อนหน้าต่างข้อมูล (Sliding window) แบบ 14 วัน ด้วยเทคนิคโครงข่ายประสาทเทียมของชุดข้อมูลตัวแปรที่ 10 (Var10) ด้วยค่ารากที่สองของค่าเฉลี่ยผิดพลาดยกกำลังสอง (*rmse*) ต่ำที่สุด เมื่อเปรียบเทียบกับเทคนิคค่าเฉลี่ยแถว (Row Average) เทคนิคเพื่อนบ้านใกล้เคียง (K-Nearest Neighbor: KNN) และระบบฟัซซี่ลอจิก (Fuzzy Logic System) ที่ค่าสูญหาย 70%



ภาพที่ 8 ผลการเติมค่าสูญหายในข้อมูลอนุกรมเวลาด้วยเทคนิคโครงข่ายประสาทเทียมตัวแปรที่ 10



ภาพที่ 8 ทางขวามือเป็นผลการเติมค่าสูญหายด้วยเทคนิคโครงข่ายประสาทเทียมด้วยการเลื่อนขนาดหน้าต่างข้อมูลของรูปแบบ 14 วัน กับชุดข้อมูลตัวแปรที่ 10 (Var10) ให้ผลลัพธ์การแทนที่ค่าสูญหายคล้ายค่าจริงมากที่สุด ค่า rmse มีค่าต่ำ และภาพที่ 8 ทางซ้ายมือจะเป็นการเติมค่าด้วยเทคนิคที่ไม่เหมาะสม กราฟจะเป็นลักษณะเส้นตรง และไม่มี ความสอดคล้องกับค่าจริง

สรุป และข้อเสนอแนะ

สำหรับงานวิจัยครั้งนี้ ผู้วิจัยศึกษาผลกระทบของการเปลี่ยนขนาดหน้าต่างข้อมูลสำหรับวิธีการเติมค่าสูญหาย ในข้อมูลอนุกรมเวลาจำนวนมากแบบหลายตัวแปร ซึ่งจะเน้นที่การค้นหามาตรฐานหน้าต่างข้อมูลที่เหมาะสมสำหรับใช้ในการสอนตัวแบบการเติมค่าสูญหาย นำข้อมูลเข้าสู่โมเดลการแทนที่ค่าสูญหายแบบต่างๆ ทั้งนี้โมเดลที่จะใช้สำหรับการวัดเปรียบเทียบจะประกอบด้วย วิธีค่าเฉลี่ยแถว (Row Average) วิธีเพื่อนบ้านใกล้เคียง (K-Nearest Neighbor: KNN) ระบบฟัซซี่ลอจิก (Fuzzy Logic System) และโครงข่ายประสาทเทียม (Artificial Neural Network) โดยกำหนดการเลื่อนหน้าต่างข้อมูล (Sliding window) 3 รูปแบบ คือ รูปแบบที่ 1 คือ 3 วัน รูปแบบที่ 2 คือ 7 วัน และ รูปแบบที่ 3 คือ 14 วัน ตามลำดับ นั่นคือ สำหรับรูปแบบที่ 1 คือ 3 วัน จะใช้ข้อมูลวันปัจจุบันร่วมกับข้อมูล ในอดีตอีก 2 วัน ย้อนหลัง สำหรับรูปแบบที่ 2 คือ 4 วัน จะใช้ข้อมูลวันปัจจุบันร่วมกับข้อมูล ในอดีตอีก 3 วัน ย้อนหลัง และ รูปแบบที่ 3 คือ 5 วัน จะใช้ข้อมูลวันปัจจุบันร่วมกับข้อมูล ในอดีตอีก 4 วัน ย้อนหลังตามลำดับ พบว่าเทคนิคโครงข่ายประสาทเทียมด้วยการเลื่อนหน้าต่างแบบ 14 วัน ให้ผลลัพธ์การแทนที่ค่าสูญหายคล้ายค่าจริงมากที่สุด และมีค่า rmse มีค่าต่ำ เมื่อเปรียบเทียบกับเทคนิคค่าเฉลี่ยแถว เทคนิคเพื่อนบ้านใกล้เคียง และระบบฟัซซี่ลอจิก

เอกสารอ้างอิง

1. Song H, Miao C, Roel W, Shen Z, Catthoor F. "Implementation of Fuzzy Cognitive Maps Based on Fuzzy Neural Network and Application in Prediction of Time Series," *IEEE Trans. Fuzzy Syst.*, vol. 18, no. 2, 233–250, Apr. 2010.
2. Bai X, Zhang F, Hou J, Xia F, Tolba A, Elashkar E. "Implicit Multi-Feature Learning for Dynamic Time Series Prediction of the Impact of Institutions," *IEEE Access*, vol. 5, 16372–16382, 2017.
3. Pratama I, Permanasari AE, Ardiyanto I, Indrayani R. "A review of missing values handling methods on time-series data," in *2016 International Conference on Information Technology Systems and Innovation (ICITSI)*, 2016, 1–6.
4. Afrianti YS. "Imputation Algorithm Based on Copula for Missing," vol. 2014, 252–257.
5. Meesud P, Rothjanawan K. "The Imputation Many Missing Value in Time Series Data Use Multivariate Relationships," *Natl. Conf. Inf. Technol.*, vol. 2017, no. 9, Nov. 2017.
6. Nogueira BM, Santos TRA, Zarate LE. "Comparison of Classifiers Efficiency on Missing Values Recovering: Application in a Marketing Database with Massive Missing Data," in *2007 IEEE Symposium on Computational Intelligence and Data Mining*, 2007, 66–72.
7. Zhou Y, Yu J, Wang X. "Time Series Prediction Methods for Depth-Averaged Current Velocities of Underwater Gliders," *IEEE Access*, vol. 5, 5773–5784, 2017.

8. Wang H, Yang J, Wang Z, Wang Q. "A binary granular algorithm for spatiotemporal meteorological data mining," in *2015 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICS DM)*, 2015, 5–11.
9. Dhargalkar SA, Bapat AU. "Determining missing values in dimension incomplete databases using spatial-temporal correlation techniques," in *2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies*, 2014, 601–606.
10. Shan Y, Deng G. "Kernel PCA regression for missing data estimation in DNA microarray analysis," in *2009 IEEE International Symposium on Circuits and Systems*, 2009, 1477–1480.
11. Li Y, Ngom A, Rueda L. "Missing value imputation methods for gene-sample-time microarray data analysis," in *2010 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, 2010, 1–7.
12. Keerin P, Kurutach W, Boongoen T. "Cluster-based KNN missing value imputation for DNA microarray data," in *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2012, 445–450.
13. Ichihashi H, Honda K, Notsu A, Yagi T. "Fuzzy c-Means Classifier with Deterministic Initialization and Missing Value Imputation," in *2007 IEEE Symposium on Foundations of Computational Intelligence*, 2007, 214–221.
14. Setiawan NA, Venkatachalam PA, Hani AFM. "Missing Attribute Value Prediction Based on Artificial Neural Network and Rough Set Theory," in *2008 International Conference on BioMedical Engineering and Informatics*, 2008, vol. 1, 306–310.
15. Ku-Mahamud KR, Zakaria N, Katuk N, Shbier M. "Flood Pattern Detection Using Sliding Window Technique," in *2009 Third Asia International Conference on Modelling Simulation*, 2009, 45–50.
16. Doreswamy, Gad I, Manjunatha BR. "Performance evaluation of predictive models for missing data imputation in weather data," in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2017, 1327–1334.
17. Santos CB. dos, Pedroso B, Guimaraes AM, Carvalho DR, Pilatti LA. "Forecasting of Human Development Index of Latin American Countries Through Data Mining Techniques," *IEEE Lat. Am. Trans.*, vol. 15, no. 9, 1747–1753, 2017.
18. Susanti SP, Azizah FN. "Imputation of missing value using dynamic Bayesian network for multivariate time series data," in *2017 International Conference on Data and Software Engineering (ICoDSE)*, 2017, 1–5.