การแทนค่าสูญหายจำนวนมากในข้อมูลอนุกรมเวลา ใช้ความสัมพันธ์หลายตัวแปร

The Imputation Many Missing Value in Time Series Data Use Multivariate Relationships

พยุง มีสัจ[1] และ กรสิริณัฐ โรจนวรรณ์[2]

[1]ภาควิชาการจัดการเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ

[2]ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ

มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ กรุงเทพฯ

ABSTRACT-- Time series information is important in many tasks. This is useful for predicting trends in business judgment. The problem of data collection is losing many potential data in the field. This make data analyzing cannot be performed efficiently. This paper proposes how to fill in the missing time value in the time series data. Many use multi-variable relationships by using existing data, create a fill missing value. The research focused on finding the right data model for teaching the missing value model by comparison of four alternative techniques: Row Average, K-Nearest Neighbor(KNN), Fuzzy Logic Systems and Artificial Neural Network. The research found Artificial Neural Network technique provides predictive effect and when used to replace lost values, the results are similar. That using some of the available data from multiple variables. It can be used to create a lossless representation. Variable data many variables will directly affect the formatting of data and models. Good data before creating a replacement for lost values.

KEY WORDS—Imputation value; Time series; Multivariate data; Artificial neural network; Data mining; Learn Machine

บทคัดย่อ—ข้อมูลอนุกรมเวลามีความสำคัญในงานต่างๆ มากมายหลายประเภท ซึ่งมีประโยชน์ในการพยากรณ์แนวโน้ม เพื่อประกอบการตัดสินในธุรกิจ ปัญหาของการเก็บข้อมูลที่ไม่ครบถ้วน ข้อมูลที่สูญหายจำนวนมาก จึงไม่สามารถนำไปใช้ในการวิเคราะห์แนวโน้วได้อย่างมีประสิทธิภาพ บทความวิจัยนี้นำเสนอวิธีการเติมค่าสูญหายในข้อมูลอนุกรมเวลาจำนวนมากใช้ความสัมพันธ์หลายตัวแปร โดยใช้ข้อมูลที่มีอยู่สร้างตัวแบบการเติมค่าสูญหาย การวิจัยเน้นที่การค้นหารูปแบบข้อมูลที่เหมาะสมสำหรับใช้สอนตัวแบบการเติมค่าสูญหาย ด้วยการเปรียบเทียบเทคนิคการแทนค่า จำนวน 4 เทคนิค ได้แก่ ค่าเฉลี่ยแถว (Row Average) เพื่อนบ้านใกล้เคียง (K-Nearest Neighbor: KNN) ระบบคลุมเครือ (Fuzzy Logic Systems) โครงข่ายประสาทเทียม (Artificial Neural Network) ผลวิจัยพบว่า โครงข่ายประสาทเทียมให้ผลการทำนายในชุดทดสอบได้ดีที่สุด และเมื่อทำไปใช้ในการแทนที่ค่าสูญหายให้ผลลัพธ์คล้ายค่าจริง ซึ่งการใช้ข้อมูลที่มีอยู่บางส่วนจากหลายตัวแปรสามารถนำไปใช้สำหรับการสร้างตัวแบบแทนค่าสูญหายได้อย่างมีประสิทธิภาพ ทั้งนี้ข้อมูลตัวแปรหลายตัวแปรจะมีผล โดยตรงต่อการจัดรูปแบบข้อมูลและตัวแบบ หรือไม่นั้นควรทำการวิเคราะห์ทำความเข้าใจในข้อมูลอย่างดีก่อนการนำไปสร้างตัวแบบการแทนค่าสูญหาย

คำสำคัญ-- การเติมค่าสูญหาย; ข้อมูลอนุกรมเวลา; ข้อมูลหลายตัวแปร; โครงข่ายประสาทเทียม; การทำเหมืองข้อมูล; เครื่องจักรเรียนรู้

1. บทนำ

ข้อมูลอนุกรมเวลา (Time Series) มีเกิดขึ้นโดยธรรมชาติ ซึ่งเป็นข้อมูล ที่มีลักษณะเป็นลำดับข้อมูลตามเวลา หรือช่วงเวลา การจัดเก็บข้อมูล จะมีตัวชี้ที่สำคัญคือเวลา (time index) ซึ่งอาจจะเป็นวินาที นาที วัน เดือน ปี หรืออื่นๆ โดยทั่วไป ข้อมูลอนุกรมเวลาเป็นลำดับข้อมูลที่เรียงต่อเนื่องในระยะเวลาเท่าๆกันในเวลา [1] ตัวอย่างข้อมูลอนุกรมเวลาที่พบทั่ว ทั่วไป เช่น ข้อมูลราคาหุ้น ข้อมูลระดับน้ำในมหาสมุทร ข้อมูลปริมาณน้ำฝนในแต่ละวัน ข้อมูลระดับน้ำเหนือเขื่อนกั้นน้ำ เป็นต้น การวิเคราะห์ข้อมูลอนุกรมเวลามีความสำคัญ ซึ่งสามารถนำผลการวิเคราะห์ไปวางแผน หรือ การตัดสินใจในการปฏิบัติงานต่างๆ ในการวิเคราะห์ข้อมูลเบื้องต้นอาจใช้ หลักสถิติเบื้องต้น [2] เช่น การหาค่าสูงสุด การหาค่าต่ำสุด การหาค่าเฉลี่ย การหาค่าเฉลี่ยเคลื่อนที่ การหาเส้นสมการ ถดถอยเชิงเส้น เป็นต้น ปัจจุบันมีงานวิจัยที่มีการใช้เครื่อง จักรเรียนรู้มาวิเคราะห์แนวโน้ว หรือพยากรณ์ค่าในอนาคตของข้อมูลอนุกรมเวลา เช่น การพยากรณ์ปริมาณน้ำฝน การพยากรณ์การขึ้นลงของหุ้น เป็นต้น

ปัญหาสำคัญอย่างหนึ่งที่พบของข้อมูลอนุกรมเวลาได้แก่ การสูญหายของข้อมูลจำนวนมาก [3] สาเหตุของข้อมูลสูญหายอาจเกิดขึ้น จากการที่ผู้จัดเก็บข้อมูลไม่ได้ทำการจัดเก็บตามกำหนดเวลาหรือ การจัดเก็บข้อมูลด้วยเซนเซอร์ หรือเครื่องจักรแต่เกิดการชำรุดไม่สามารถอ่านและจัดเก็บข้อมูลตามโปรแกรมกำหนด มีงานวิจัยด้านการเติมค่าสูญหายในอนุกรมมากมาย ซึ่งมีการใช้เทคนิคมากมายหลายแบบ แบบง่ายสุดทางสถิติ ได้แก่ การเติมค่า ด้วยค่าเฉลี่ย ค่าฐานนิยม และค่ามัฐยฐาน การเติมค่าด้วยรูปแบบเชิงคณิตศาสตร์ ได้แก่ เทคนิคสมการถดถอยเชิงเส้น (Linear Regression) เทคนิคสมการถดถอยแบบไม่เป็นเชิงเส้น (Nonlinear Regression) ค่าเฉลี่ยแถว (Row Average) หรือรูปแบบเครื่องจักรการเรียนรู้ [4] ได้แก่ เทคนิควิธีเพื่อนบ้านใกล้เคียง เทคนิควิธีโครงข่ายประสาทเทียม และเทคนิควิธีซัพพอร์ตเวกเตอร์แมชชีน เป็นต้น ข้อมูลสูญหายในข้อมูลอนุกรมเวลามีประเด็นที่น่าสนใจ ในกรณีข้อมูลอนุกรมเวลามีการสูญหายจำนวนมาก การจัดรูปแบบข้อมูลที่เหมาะสมจะเป็นอย่างไรและตัวแบบการเติมค่าที่เหมาะสมจะเป็นอย่างไร

สำหรับงานวิจัยครั้งนี้ ผู้วิจัยนำเสนอเทคนิคการจัดรูปแบบข้อมูลสำหรับอนุกรมเวลาหลายตัวแปรเพื่อประยุกต์ใช้ในการเติมค่าสูญหายจำนวนมากซึ่งมี องค์ประกอบหลายตัวแปร ซึ่งเป็นปัญหาที่ยากอย่างหนึ่ง ผู้วิจัยได้ทำการทดลอง ใช้ ในการสอนโมเดลการเติมค่าแบบต่างๆ ได้แก่ วิธีค่าเฉลี่ยแถว วิธีเพื่อนบ้านใกล้เคียง วิธีระบบฟัซซีและวิธีโครงข่าย ประสาทเทียมแบบหลายชั้น

2. วรรณกรรมที่เกี่ยวข้อง

2.1 การทำเหมืองข้อมูล

การทำเหมืองข้อมูล คือ การนำข้อมูลมาวิเคราะห์จากข้อมูลที่มีจำนวนมาก เพื่อหาความสัมพันธ์ของข้อมูลที่ซ่อนอยู่ โดยทำการจำแนกรูปแบบ จำแนกประเภท เชื่อมโยงข้อมูลที่มีความสัมพันธ์กัน และองค์ความรู้ใหม่จากข้อมูลเดิม โดยมีขั้นตอนดังนี้

1. ทำความเข้าใจปัญหา โดยการเลือกข้อมูลให้มีความเหมาะสมกับอัลกอริทึมที่ใช้งาน จำนวนที่ต้องการ และค่าเป้าหมายเพื่อให้ได้ผลลัพธ์ที่ต้องการ

2. ทำความเข้าใจข้อมูล โดยการรวบรวมตรวจสอบความถูกต้องและกำหนดคุณสมบัติที่ต้องการให้กับข้อมูล

3. เตรียมข้อมูล โดยการคัดเลือกข้อมูลเพื่อทำการแปลงให้อยู่ในรูปแบบที่เหมาะสมต่อการวิเคราะห์ข้อมูลด้วยเทคนิคต่างๆ

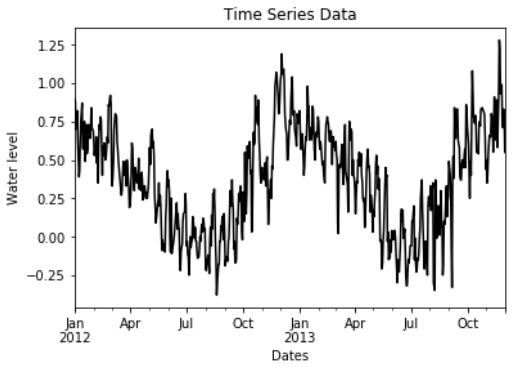
4. สร้างแบบจำลอง โดยแบ่งเป็น 2 ประเภท คือ 1) การสร้างแบบจำลองเพื่อการทำนาย (Predictive Data Mining) เป็นการคาดคะแนนลักษณะหรือประมาณค่าที่ชัดเจนของข้อมูลที่จะเกิดขึ้น โดยใช้ข้อมูลในอดีต 2) การสร้างแบบจำลองเพื่อใช้บรรยาย (Descriptive Data Mining) เพื่อหาแบบจำลองมาอธิบายลักษณะบางอย่างของข้อมูล

2.2 ข้อมูลอนุกรมเวลา

ข้อมูลอนุกรมเวลา หมายถึง เป็นข้อมูลลำดับของค่าที่กำหนดไว้ [5] ในช่วงเวลาที่เว้นระยะเท่ากัน เช่น ชั่วโมง วัน เดือน ปี เป็นต้น มีรูปแบบทางคณิตศาสตร์ดังสมการ (1) และตัวอย่างภาพ ข้อมูลอนุกรมเวลาแสดงได้ดังรูปที่ 1

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

เมื่อ d_i หมายถึงข้อมูลที่เวลา i



รูปที่ 1 ข้อมูลที่แสดงความสัมพันธ์กับเวลา

การวิเคราะห์ข้อมูลอนุกรมเวลาถูกใช้งานอย่างกว้างขวาง เช่น การพยากรณ์ทางเศรษฐกิจ การคาดการณ์ยอดขาย การวิเคราะห์ งบประมาณ การวิเคราะห์ตลาดหุ้น กระบวนการและการควบคุม คุณภาพ การศึกษาสินค้าคงคลัง การคาดการณ์ปริมาณงาน การศึกษา สาธารณูปโภค และการสำรวจสำมะโนประชากร เป็นต้น

การวิเคราะห์ข้อมูลอนุกรมเวลามีหลายเทคนิควิธี เช่น AR, MA, ARMA, ARIMA หรือการสร้างโมเดลด้วยวิธีเครื่องจักรเรียนรู้ เป็นต้น ซึ่งเป็นความพยายามในการสร้างโมเดลที่สามารถคาดการณ์ค่าในอนาคตได้อย่างถูกต้องและแม่นยำ ประเด็นสำคัญที่เป็นอุปสรรคต่อการสร้างโมเดลพยากรณ์ข้อมูลอนุกรมเวลา คือ ข้อมูลที่มีการสูญหาย

ในการสร้างโมเดลในการพยากรณ์ค่าในอนาคต ข้อมูลที่สูญหายจะส่งผลต่อกระทบการสร้างโมเดลที่เหมาะสม ความรุน แรงของผลกระทบขึ้นอยู่กับองค์ประกอบจากหลายส่วน โดยเฉพาะอย่างยิ่งจำนวนของข้อมูลที่สูญหาย หากมีจำนวนมากก็อาจส่งผลถึงการที่ไม่สามารถ สร้างโมเดลพยากรณ์ได้ ดังนั้นจึงต้องมีการแก้ปัญหาข้อมูลสูญหายก่อนการนำข้อมูลไปสร้างโมเดลพยากรณ์

2.3 การแทนค่าข้อมูลสูญหาย

การแทนค่าข้อมูลสูญหายมีหลายเทคนิควิธีสามารถนำเอาเทคนิคทางสถิติ หรือเทคนิคทางเหมืองข้อมูลอย่างง่าย ๆ มาใช้สำหรับ การแทนที่ค่าข้อมูลที่ขาดหาย จากการศึกษางานวิจัยที่เกี่ยวข้องพบว่า ได้มีงานวิจัยจำนวนมากได้นำเสนอวิธีการแทนข้อมูลที่ขาดหาย [6] [7] โดยการนำเทคนิค เช่น ค่าเฉลี่ยแถว (Row Average) เพื่อนบ้าน ใกล้เคียง (K-Nearest Neighbor: KNN) ระบบคลุมเครือ (Fuzzy Logic Systems) โครงข่ายประสาทเทียม (Artificial Neural Network) เป็นต้น มาใช้เพื่อแทนที่ค่าข้อมูลขาดหาย

2.3.1. ค่าเฉลี่ยแถว

เป็นวิธีการทางสถิติแบบง่ายๆ [8] ที่ใช้สำหรับการแทนค่าข้อมูลที่ขาดหาย เป็นการคำนวณหาค่าเฉลี่ยของตัวแปรเดียวกันเพื่อแทนที่ค่าสูญหายดังสมการ (2)

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

เมื่อ $\bar{d}$ คือ ค่าเฉลี่ยแถวหรือตัวแปร

d_t คือ ข้อมูลที่เวลา t

วิธีการนี้เป็นวิธีการที่ง่ายต่อการทำงาน แต่ผลลัพธ์จากวิธีการนี้ อาจถูกโน้มเอียงจากค่าที่อยู่นอกกลุ่มได้จึงให้ประสิทธิภาพไม่ค่อยดีนัก

Algorithm 1: Row_Average_Impute

Step 1: Prepare the matrix of input attributes, **X,** for training set; Set the vector of target attribute, ***t,*** for missing value prediction;

Step 2: Find missing data indexes of the target attribute, t_m;

Step 3: For each missing target, read in the input attribute values of the missing target, $\mathbf{x}_m$;

Step 4: Return the imputed missing value, t_{mi}, by calculating the average value of ***t***.

2.3.2. เพื่อนบ้านใกล้เคียง

เทคนิคเพื่อนบ้านใกล้เคียง (K-Nearest Neighbor: KNN) เป็นเทคนิควิธี ที่ได้รับความนิยมในการใช้งานอย่างมาก [9] เนื่องจากเป็นวิธีการที่ง่าย และมีประสิทธิภาพซึ่งสามารถนำไปประยุกต์ใช้กับงานได้อย่าง หลากหลาย เช่น งานทางด้านการจำแนกข้อมูล (Classification) รวมถึง งานทางด้านการแทนที่ข้อมูลที่สูญหาย (Missing Values Imputation) อัลกอริธึมสำหรับการใช้ KNN ในการแทนที่ค่าสูญหาย แสดงดัง Algorithm 2

Algorithm 2: KNN_Impute

Step 1: Prepare the matrix of input attributes, **X,** for training set; Set the vector of target attribute, **t**, for missing value prediction; Set *K* integer value;

Step 2: Find missing data indexes of the target attribute, t_m;

Step 3: For each missing target, read in the input attribute values of the missing target,$\mathbf{x}_m$;

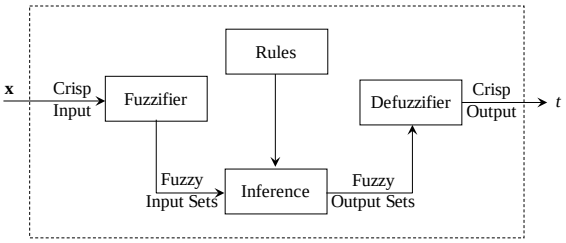
Step 4: Calculate Euclidean distance between record $\mathbf{x}_m$ and each records in **X**;

Step 5: Sort distance ascending;

Step 6: Return the imputed missing value, t_{mi}, by calculating the average value of the *K* nearest neighbor of $\mathbf{x}_m$.

2.3.3. ระบบฟัซซีลอจิก

ฟัซซีลอจิกเป็นตรรกศาสตร์ ภายใต้แนวคิดว่าเหตุการณ์ต่างๆ จะมีความไม่แน่นอน (uncertain) แต่มีความคลุมเครือ (fuzzy) ซึ่งเป็นพื้นฐานของระบบฟัซซีลอจิก (fuzzy logic system) มีหลักการให้เหตุผลเลียนแบบการตัดสินใจของมนุษย์และสามารถนำไปประยุกต์ใช้งาน [10] ในการตัดสินใจต่างๆ อย่างมากมาย ตัวอย่างระบบ ฟัซซีลอจิกแสดงดังรูปที่ 2 และอัลกอริธึมสำหรับการแทนค่า Algorithm 3



รูปที่ 2 ระบบฟัซซีลอจิก [11]

Algorithm 3: Fuzzy_Impute

Step 1: Prepare the matrix of input attributes, **X,** for training set; Set the vector of target attribute, **t**, for missing value prediction; Set fuzzy system training parameters;

Step 2: Train fuzzy system using training set, $\mathbf{X}_{tr}$, $\mathbf{t}_{tr}$;

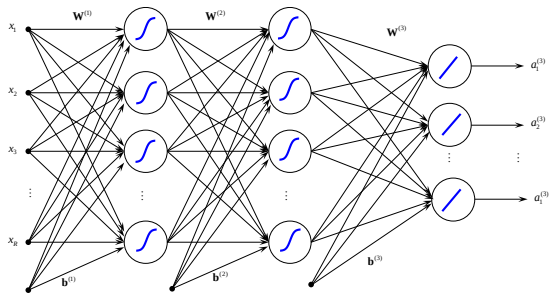
Step 3: Find missing data indexes of the target attribute, t_m;

Step 4: For each missing target, t_m, read in the input attribute values of the missing target, $\mathbf{x}_m$;

Step 5: Return the imputed missing value, t_{mi}, by applying trained fuzzy system with input $\mathbf{x}_m$.

2.3.4. โครงข่ายประสาทเทียม

โครงข่ายประสาทเทียมเป็นเครื่องจักรการเรียนรู้ซึ่งมีที่มาจากการจำลองสมองมนุษย์ ซึ่งเป็นการจำลองการทำงานของสมองมนุษย์ด้วยสมการทางคณิตศาสตร์ โดยสามารถเรียนรู้ข้อมูลเกี่ยวกับสิ่งที่ต้องการสอนให้โครงข่ายเทียมเพื่อนำไปพยากรณ์ข้อมูลใหม่ในอนาคตได้ [12] โครงข่ายประสาทเทียม มีความสามารถในการเรียนรู้เพื่อจำแนกกลุ่มข้อมูล (classification) และสามารถ พยากรณ์ข้อมูลที่เป็นค่าทศนิยม (regression) จึงทำให้ได้รับ ความนิยมอย่างแพร่หลายในการประยุกต์ใช้งานด้านต่างๆ เช่น ระบบควบคุมทางวิศวกรรม ระบบตัดสินใจทางด้านบริหาร ระบบพยากรณ์ทางด้านเศรษฐศาสตร์ เป็นต้น โครงข่ายประสาทเทียม ที่นิยมใช้ ได้แก่ โครงข่ายประสาทเทียมเปอร์เซ็ปตรอนแบบหลายชั้น ดังตัวอย่างในรูปที่ 3



รูปที่ 3 โครงข่ายประสาทเทียมแบบหลายชั้น [11]

Algorithm 4: ANN_Impute

Step 1: Prepare the matrix of input attributes, **X,** for training set; Set the vector of target attribute, **t**, for missing value prediction; Set ANN training parameters;

Step 2: Train ANN using training set, $\mathbf{X}_{tr}$, $\mathbf{t}_{tr}$;

Step 3: Find missing data indexes of the target attribute, t_m;

Step 4: For each missing target, t_m, read in the input attribute values of the missing target, $\mathbf{x}_m$;

Step 5: Return the imputed missing value, t_{mi}, by applying trained ANN with input $\mathbf{x}_m$.

2.4 วิธีการวิเคราะห์ความแม่นยำของตัวแบบ (K-fold Cross-Validation)

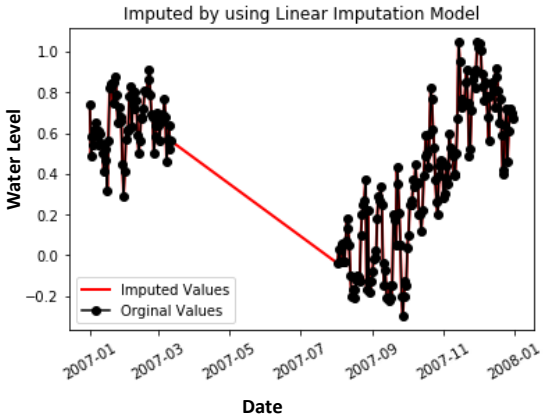
การตรวจสอบไขว้กัน (Cross-Validation) เป็นวิธีการตรวจสอบค่าความผิดพลาดในการคาดการณ์ของตัวแบบ โดยพื้นฐานของวิธีการตรวจสอบไขว้กัน คือ การสุ่มตัวอย่าง โดยเริ่มจากการแบ่งชุดข้อมูลออกเป็นส่วนๆ และนำบางส่วนจากชุดข้อมูลนั้นมาตรวจสอบ ผลลัพธ์จากการทำการตรวจสอบไขว้กัน มักถูกใช้เป็นตัวเลือกในการกำหนดตัวแบบ ในกรณีการทำ K-fold Cross-Validation จะแบ่งข้อมูลออกเป็น K ชุด เท่าๆ กัน เช่น K = 5 หมายถึง จะมีชุดข้อมูลจำนวน 5 ชุด ซึ่งจะทำการคำนวณค่าความผิดพลาด 5 รอบ โดยแต่ละรอบของการคำนวณ จะเลือกข้อมูลออกมา 1 ชุดเพื่อเป็นข้อมูลทดสอบ และข้อมูลอีก 4 ชุด จะถูกใช้เป็นข้อมูลสำหรับการเรียนรู้ ดังรูปที่ 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

รูปที่ 4 การตรวจสอบไขว้กัน กำหนด K=5 [13]

2.5 ปัญหาดั้งเดิมข้อมูลสูญหายจำนวนมาก

ข้อมูลที่มีการสูญหายจำนวนมากมีความยากในการทำนายค่าแทนที่สูญหาย เทคนิคดั้งเดิมส่วนมากสามารถประยุกต์ใช้ในการแทนที่ ค่าสูญหายได้แต่ข้อมูลอาจจะไม่ถูกต้องเท่าที่ควร ดังในรูปที่ 4 ซึ่งเป็นตัวอย่างออกแบบโมเดลการแทนที่ค่าสูญหายที่ไม่เหมาะ สังเกตจากข้อมูลที่มีอยู่จะมีการสวิงขึ้นลงแต่มีแนวโน้วลดลง ในขณะที่ข้อมูลแทนที่ค่าสูญหายจำนวนมากมีเป็นลักษณะลดลงแบบเป็นเส้นตรง



รูปที่ 5 ตัวอย่างปัญหาการแทนที่ค่าสูญหายจำนวนมากในข้อมูลอนุกรม

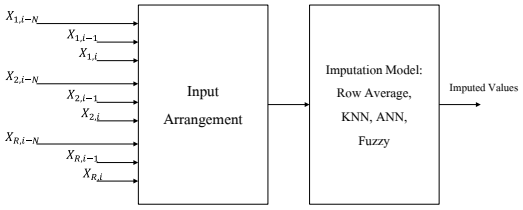
3. วิธีการดำเนินการวิจัย

การวิจัยครั้งนี้มีจุดมุ่งหมายในการค้นหาวิธีการที่เหมาะสมสำหรับการแทนที่ค่าสูญหายจำนวนมาก โดยใช้ข้อมูลหลายตัวแปรเป็นข้อมูลอินพุต สำหรับการแทนที่ค่าสูญหายและใช้หลักการข้อมูลในอดีตทำนายค่าในอนาคต โดยมีเงื่อนไขเบื้องต้นคือ ตัวแปรต้นต้องมีความสัมพันธ์กันกับตัวแปรที่สูญหาย หรือตัวแปรตามวิธีการจัดรูปแบบข้อมูลสำหรับใช้ในการแทนที่ค่าสูญหาย ตามหัวข้อ 3.1

3.1. รูปแบบการแทนที่ค่าสูญหาย

ในการวิจัยผู้วิจัยได้ออกแบบขั้นตอนวิธีการแทนที่ค่าสูญหายจำนวนมาก ในข้อมูลอนุกรมแบบหลายตัวแปร โดยเริ่มจากการออกแบบรูปแบบข้อมูลสำหรับประยุกต์ใช้กับเทคนิคการแทนที่ค่าสูญหาย ตามหัวข้อที่ 2 ได้แก่ ค่าเฉลี่ยแถว (Row Average) เพื่อนบ้านใกล้เคียง (K-Nearest Neighbor: KNN) ระบบคลุมเครือ (Fuzzy Logic Systems) โครงข่ายประสาท เทียม (Artificial Neural Network)

โดยมีคำถามวิจัยว่าเทคนิคใดที่มีความเหมาะสมในการแทน ค่าสูญหายจำนวนมากได้ดีที่สุดสำหรับกระบวนการวิจัย เริ่มต้นจาก 1) การวิเคราะห์ และการสำรวจตรวจสอบข้อมูล 2) การขจัดข้อมูลผิดปกติ 3) การค้นหาข้อมูลสูญหาย 4) การจัดข้อมูลอนุกรมให้อยู่ใน รูปแบบสำหรับการสอนเครื่องจักรเรียนรู้ 5) การสอนเครื่องจักรเรียนรู้ด้วยข้อมูลชุดสอน 6) การทดสอบโมเดลการแทน ที่ค่าสูญหายด้วยชุดข้อมูลทดสอบ และ7) การประยุกต์ใช้โมเดลแทนที่ค่าข้อมูลสูญหาย



รูปที่ 6 รูปแบบวิธีการแทนที่ค่าสูญหายในอนุกรมเวลาหลายตัวแปร

รูปแบบวิธีการแทนค่าสูญหายจำนวนมากหลายตัวแปรในข้อมูลอนุกรมเวลาที่ผู้วิจัยนำเสนอ ในการวิจัยครั้งนี้แสดงดังรูปที่ 6 โมเดลการแทนที่ค่าสูญหายในรูปที่ 5 มีขั้นตอนการทำงาน ดังนี้ 1) รับเข้าข้อมูลอนุกรมเวลาแบบหลายตัวแปร 2) เลือกตัวแปรเป้าหมายสำหรับแทนที่ค่าสูญหาย 3) สำหรับตัวแปรอินพุตทำการจัดข้อมูลอนุกรม โดยเลื่อนหน้าต่างข้อมูลแต่ละตัวแปรย้อนหลังจำนวน N วัน เพื่อจัดเป็น เวกเตอร์อินพุต และตั้งค่าเป้าหมายที่เวลาล่วงหน้า 1 วัน (i + 1) จากตัวแปรที่เลือกไว้เป็น ค่าเป้าหมายสำหรับการแทนที่ ค่าสูญหาย

ในการสอนเครื่องจักรการเรียนรู้ ผู้วิจัยนำเสนอวิธีการจัดรูปแบบ ข้อมูลโดยใช้ข้อมูลในอดีต เพื่อพยากรณ์ข้อมูลในอนาคต ซึ่งต้องทำการเลื่อนหน้าต่างข้อมูล โดยจัดให้อยู่ในรูปเวกเตอร์ และเมทริกซ์ตามสมการ (3) – (7)

$$\mathbf{x}_{j,i} = \left[x_{j,i-N} \quad \cdots \quad x_{j,i-1} \quad x_{j,i} \right], j = 1, \ldots, R \tag{3}$$

$$\mathbf{p}_i = \left[\mathbf{x}_{1,i} \quad \mathbf{x}_{2,i} \quad \cdots \quad \mathbf{x}_{R,i} \right], i = 1, \ldots, M \tag{4}$$

$$t_i = x_{Q,i}, \quad Q \notin \{ j = 1, \ldots, R \} \tag{5}$$

$$\mathbf{P} = \left[\mathbf{p}_1^T \quad \mathbf{p}_2^T \quad \cdots \quad \mathbf{p}_M^T \right] \tag{6}$$

$$\mathbf{t} = \left[t_1 \quad t_2 \quad \cdots \quad t_M \right] \tag{7}$$

$$\hat{y}_i = f(\mathbf{p}_1^T, parameters) \tag{8}$$

เมื่อ $\mathbf{x}_{j,i}$ คือ เวกเตอร์หน้าต่างข้อมูลที่มีตัวประกอบเป็นข้อมูลจากตัวแปร j ที่เวลา i-N, …, 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

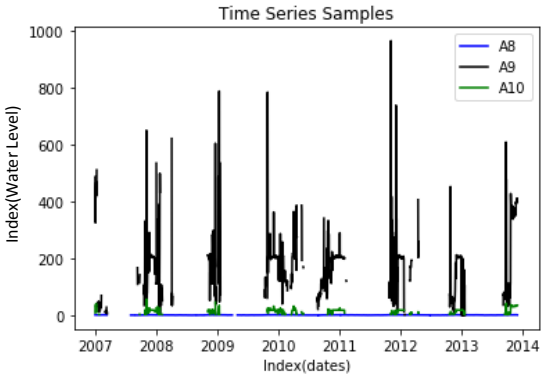
R คือ จำนวนตัวแปรอินพุต

Q คือ ตัวแปรเป้าหมายซึ่งต้องการแทนที่ค่าสูญหาย

parameters คือ ค่าพารามิเตอร์ของโมเดลแทนที่ค่าสูญหาย

3.2. ข้อมูลสำหรับการทดลอง

ข้อมูลที่ใช้ในการวิจัยครั้งนี้เป็นข้อมูลจากกรมชลประทานที่ 15 โครงการพระราชดำริลุ่มน้ำปากพนัง ในช่วงระยะเวลาปี 2550 - 2556 โดยเป็นข้อมูลอนุกรมเวลา มี 12 ตัวแปร ได้แก่ Var1: ข้อมูลระดับ เหนือน้ำเวลา 6.00 น. Var2: ข้อมูลระดับเหนือน้ำเวลา 12.00 น. Var3: ข้อมูลระดับเหนือน้ำเวลา 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 โดยมีตัวอย่างข้อมูล Var8, Var9 และ Var10 ดังรูปที่ 6



รูปที่ 7 ตัวอย่างข้อมูลอนุกรมเวลาในการทดสอบแนวคิดโมเดลที่นำเสนอ

รูปที่ 7 แสดงตัวอย่างบางส่วนของข้อมูลอนุกรมเวลาข้อมูลจากกรมชลประทานที่ 15 โครงการพระราชดำริลุ่มน้ำปากพนังซึ่งจะมีข้อมูลสูญหายจำนวนมาก และปัญหาที่ท้าทายสำหรับการสร้างโมเดลแทนที่ค่าสูญหาย

ในการวิจัยครั้งนี้เสนอหลักการสร้างโมเดลพยากรณ์ด้วยวิธีเครื่องจักรการเรียนรู้ โดยทำการคัดแยกข้อมูลสูญหายออกไว้ก่อน และนำข้อมูลที่สมบูรณ์มาใช้ในการสร้างโมเดล เพื่อนำไปเป็นอินพุตและค่าเป้าหมายสำหรับสอนโมเดลการแทนค่าสูญหายของตัวแปรที่มีข้อมูลสูญหาย โดยแบ่งข้อมูลสมบูรณ์ออกเป็นชุดสอนและชุดทดสอบ เพื่อให้การสอนโมเดลมีความน่าเชื่อถือโมเดลจะใช้วิธีการสอน แบบตรวจสอบไขว้ (k-fold cross-validation) ซึ่งเป็นหลักการที่ยอมรับ ว่าเหมาะสำหรับการสอนและทดสอบเครื่องจักรการเรียนรู้ การวิจัยครั้งนี้ กำหนด k = 5 นั่นคือแบ่งข้อมูลเป็น 5 ส่วน โดยใช้ข้อมูล 4 ส่วนทำการสอนโมเดล ใช้ข้อมูล 1 ส่วนทดสอบโมเดล และเวียนสอนและทดสอบไขว้ 5 รอบจนครบทุกส่วน สำหรับเทคนิคการแทนที่ค่าสูญหายคัดเลือกด้วยการเปรียบเทียบเทคนิควิธีการแทนค่า จำนวน 4 เทคนิค ได้แก่ ค่าเฉลี่ยแถว (Row Average) เพื่อนบ้านใกล้เคียง (K-Nearest Neighbor: KNN) ระบบคลุมเครือ (Fuzzy Logic Systems) โครงข่ายประสาทเทียม (Artificial Neural Network) สำหรับการวัดประสิทธิภาพของโมเดลแทนที่ค่า ผู้วิจัยใช้วิธีคำนวณรากที่สองของค่าเฉลี่ยผิดพลาดยกกำลังสอง (root mean squared error: *rmse*) ดังสมการ (9)

$$rmse = \sqrt{\frac{1}{L}\sum_{i=1}^{L}\left(t_i - \hat{y}_i\right)^2} \qquad (9)$$

เมื่อ t_i คือ ค่าเป้าหมายจริงของข้อมูลทดสอบ

$\hat{y}_i$ คือ ค่าพยากรณ์ จากระบบการแทนที่ค่าสูญหาย

L คือ จำนวนข้อมูลทดสอบ

4. ผลการดำเนินงานวิจัยและการอภิปราย

ในการดำเนินการวิจัยครั้งนี้ผู้วิจัยเสนอเทคนิควิธีการแทนที่ค่าข้อมูลที่มีการสูญหายจำนวนมากในข้อมูลอนุกรมเวลา ใช้หลักการเลื่อนหน้าต่างข้อมูล ตามรูปที่ 5 โดยจัดรูปแบบสมการการเลื่อนข้อมูล ดังสมการที่ (3) – (7) โดยเขียนสคริปต์ด้วยภาษา Matlab โมเดลการแทนที่ค่าสูญหายแบบ Row Average และ KNN เขียนเป็น Matlab สคริปต์ ตาม Algorithm 1 และ 2 ตามลำดับ ส่วนโครงข่ายประสาทเทียมใช้ Matlab Neural Network Toolbox ชนิด Feed Forward Multilatyer Perceptron Network (MLP) แบบ 2 ชั้นซ่อน โดยกำหนดจำนวนชั้นละ 15 นิวรอนแบบ Hyperbolic Tangent Sigmoidal (tansig) ทั้งสอง ชั้นซ่อน และชั้นเอาต์พุตเป็นแบบเชิงเส้น และสอนด้วยวิธี trainlm (Levenberg-Marquardt) ส่วนระบบฟัซซีใช้เป็นแบบ ANFIS ใน Matlab Fuzzy System Toolbox กำหนดใช้ฟังก์ชันความเป็นสมาชิกของฟัซซีเซตแบบ Gaussian Bell Membership Function (gbellmf) จำนวน 5 ฟัซซีเซต และสอนด้วยวิธี Hybrid ในการทดลองได้ กำหนดขนาดหน้าต่างข้อมูลเท่ากับ 3 นั่นคือใช้ข้อมูลวันปัจจุบันร่วมกับข้อมูลในอดีตอีกสองวันย้อนหลัง

เพื่อให้การทดสอบมีความคงทนต่อข้อมูลแปลกปลอม ผู้วิจัย ทำการสร้างโมเดลพยากรณ์แต่ละแบบด้วยการสุ่มเติมค่าสูญหายโดยค่อยๆ สุ่มจำนวนข้อมูลสูญหายเพิ่มเติมตั้งแต่ 5% ถึง 95% เพิ่มขั้นละ 5% จากตารางที่ 1 ถึง 4 เป็นผลการทดลองสุ่มค่าสูญหาย จากข้อมูลสมบูรณ์โดยแต่ละแถว ตัวเลข 0, 5 และ 10 หมายถึง ข้อมูลสูญหาย 0%, 5% และ 10% ตามลำดับ เป็นต้น ส่วนแต่ละคอลัมน์เป็นตัวแปรที่ 1 ถึง 12 ซึ่งต้องการหาค่าแทนที่ค่าสูญหายโดยค่าผิดพลาด (rmse) จะอยู่ในแต่ละเซลของตาราง

จากผลการแทนที่ค่าสูญหายเทียบกับค่าจริงของแต่ละตัวแปร จากผลการดำเนินงานวิจัยเปรียบเทียบ 4 เทคนิค ดังตาราง ที่ 1 ถึง 4 มีข้อสังเกตว่ากรณีไม่มีค่าสูญหายเลย (%Missing Values = 0) โมเดลแทนค่าสูญหายทุกโมเดลสามารถแทนที่ค่าได้อย่างถูกต้องนั่นคือ ค่า rmse = 0 ทุกตัวแปร และเมื่อทำการสุ่มให้มีค่าสูญหายเพิ่มมากขึ้นทีละ 5% ค่าพยากรณ์ที่เติมแทนที่จะมีค่าผิดพลาด หรือคลาดเคลื่อนมากขึ้นเรื่อยๆ เป็นสัดส่วนแปรผันตรงตามค่าสูญหายที่เพิ่มขึ้น พบว่าโครงข่ายประสาทเทียมให้ผลการเติมค่าใกล้เคียงค่าจริงมากที่สุด รองลงมาเป็นวิธี KNN, Row Average, และ Fuzzy System ตามลำดับ

ตารางที่ 1 ผลการเติมค่าสูญหายในข้อมูลอนุกรมเวลาด้วยวิธี Row Average

% Missing Values	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	Var11	Var12
0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.14	0.15	0.14	0.07	0.10	0.09	0.14	0.07	26.42	2.02	4.33	2.80
10	0.20	0.20	0.21	0.10	0.15	0.13	0.20	0.11	38.64	2.81	6.74	3.80
15	0.24	0.25	0.25	0.13	0.18	0.16	0.25	0.13	42.64	3.61	8.53	5.04
20	0.28	0.30	0.28	0.14	0.21	0.18	0.30	0.15	53.23	4.18	10.48	5.78
25	0.32	0.32	0.33	0.16	0.23	0.20	0.33	0.17	58.98	4.71	10.23	6.60
30	0.35	0.36	0.35	0.18	0.25	0.22	0.36	0.18	65.14	5.25	12.47	6.96
35	0.37	0.39	0.38	0.20	0.28	0.24	0.39	0.20	71.63	5.66	13.09	7.63
40	0.40	0.41	0.41	0.21	0.29	0.25	0.41	0.21	73.84	6.12	13.50	8.11
45	0.42	0.44	0.43	0.22	0.31	0.27	0.44	0.22	80.04	6.20	14.58	8.69
50	0.44	0.46	0.45	0.23	0.33	0.29	0.47	0.23	83.59	6.66	14.94	8.91
55	0.46	0.49	0.48	0.24	0.34	0.30	0.49	0.25	87.12	7.08	16.05	9.59
60	0.48	0.50	0.50	0.25	0.36	0.31	0.51	0.26	92.42	7.32	16.96	9.85
65	0.51	0.52	0.52	0.26	0.37	0.33	0.54	0.27	94.75	7.65	17.49	10.36
70	0.52	0.55	0.54	0.28	0.39	0.34	0.55	0.28	98.52	7.94	17.55	10.70
75	0.54	0.56	0.55	0.28	0.40	0.35	0.58	0.29	102.03	8.24	18.82	11.14
80	0.56	0.58	0.58	0.29	0.42	0.36	0.60	0.30	106.09	8.79	19.14	11.48
85	0.57	0.60	0.59	0.30	0.43	0.37	0.61	0.31	109.42	8.95	20.19	11.78
90	0.60	0.61	0.61	0.31	0.45	0.38	0.63	0.32	112.72	9.52	20.56	12.16
95	0.61	0.63	0.63	0.32	0.45	0.40	0.65	0.33	115.24	9.87	21.04	12.46

ตารางที่ 2 ผลการเติมค่าสูญหายในข้อมูลอนุกรมเวลาด้วยวิธี KNN

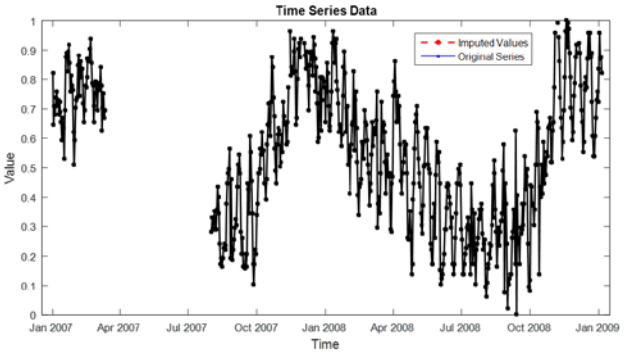
% Missing Values	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	Var11	Var12
0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	8.20	0.75	0.11	0.16	0.20	0.18	0.12	0.13	26.92	3.79	10.10	1.33
10	12.33	6.13	3.97	0.11	0.23	0.22	0.16	0.19	39.81	6.09	11.15	1.77
15	24.07	13.32	7.23	3.40	0.28	0.27	0.24	0.23	56.94	11.26	17.42	3.33
20	31.76	31.20	30.53	16.00	9.13	6.62	0.40	0.35	75.78	11.20	22.40	5.94
25	35.29	34.94	28.99	23.15	18.08	10.33	1.16	0.34	86.76	13.24	20.34	7.37
30	43.44	42.15	44.11	43.66	42.03	31.24	13.10	1.55	103.22	19.92	23.37	6.60
35	40.88	40.96	41.75	40.80	41.76	39.27	33.70	13.68	134.57	30.13	31.30	8.13
40	45.21	45.96	45.06	45.75	45.38	42.12	37.96	21.81	116.69	20.69	26.23	7.86
45	48.61	48.30	48.16	48.62	49.39	48.51	47.42	30.22	118.44	33.58	32.73	8.94
50	53.85	53.52	53.30	52.99	52.81	53.06	53.25	48.72	139.26	39.67	25.31	11.82
55	55.25	55.15	54.80	55.54	54.51	54.80	55.50	52.55	133.57	37.17	32.66	9.87
60	49.09	49.37	48.82	50.14	49.33	49.32	49.03	48.29	71.93	33.61	33.84	13.31
65	62.02	62.11	61.51	61.43	62.06	61.46	61.50	61.60	85.54	44.03	30.46	10.86
70	55.11	55.35	55.15	55.20	55.17	55.63	54.86	51.91	76.55	29.77	31.75	12.18
75	60.84	61.03	61.13	60.93	61.21	61.15	60.84	60.69	77.02	32.79	38.56	13.06
80	64.31	63.57	63.90	64.17	64.21	64.37	64.21	64.12	79.23	41.00	40.70	13.49
85	57.27	56.98	57.02	57.18	57.41	57.59	57.19	56.79	80.09	44.03	32.39	15.98
90	68.63	68.66	68.38	68.83	68.52	68.56	68.47	68.55	84.35	61.60	35.46	13.21
95	59.18	59.22	58.93	59.42	59.50	59.53	59.20	59.18	84.56	56.23	26.25	15.56

ตารางที่ 3 ผลการเติมค่าสูญหายในข้อมูลอนุกรมเวลาด้วยวิธี Fuzzy System

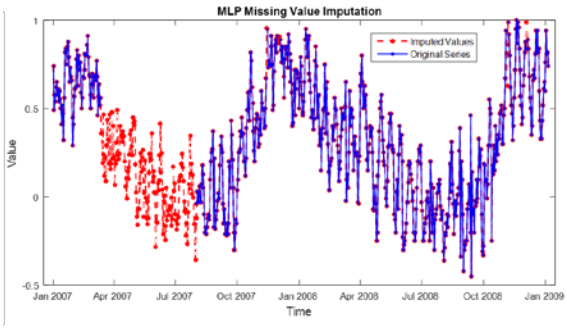
% Missing Values	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	Var11	Var12
0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.02	0.02	0.02	0.02	0.04	0.03	0.02	0.03	5.80	0.24	1.98	0.32
10	0.03	0.02	0.02	0.04	0.05	0.05	0.03	0.04	9.33	0.55	4.21	1.02
15	0.03	0.04	0.03	0.05	0.07	0.06	0.04	0.05	13.37	0.58	4.11	1.65
20	0.03	0.04	0.04	0.07	0.09	0.07	0.04	0.07	12.73	0.75	5.53	1.59
25	0.04	0.05	0.04	0.08	0.09	0.08	0.04	0.07	31.88	1.01	5.47	72.20
30	0.05	0.05	0.05	0.09	0.11	0.09	0.06	0.09	18.37	1.11	5.29	2.63
35	0.05	0.05	0.05	0.10	0.12	0.11	0.06	0.10	519.82	1.16	6.54	4.81
40	0.06	0.07	0.06	0.11	0.13	0.11	0.06	0.10	4255.50	1.64	7.73	99.24
45	0.07	0.07	0.07	0.11	0.14	0.12	0.07	0.10	29585.00	2.05	7.33	173.97
50	0.06	0.08	0.07	0.12	0.16	0.13	0.07	0.12	4.92E+07	48.89	6569.60	2346.20
55	0.07	0.08	0.08	0.15	0.18	0.16	0.08	0.13	4.77E+12	1.25E+06	8.45	2819.30
60	0.08	0.10	0.09	0.16	0.19	0.16	0.09	0.14	1.79E+17	6.78E+11	195.30	7.13E+13
65	0.09	0.20	0.10	0.17	0.21	0.18	0.09	0.15	4.75E+24	4.12E+13	9.36	1.41E+05
70	0.10	0.12	0.12	0.19	0.23	0.20	0.10	0.17	1.69E+55	7.88E+21	12.07	1.90E+16
75	0.11	0.13	0.15	0.25	0.25	0.23	0.11	0.18	4.88E+38	5.42E+27	10.37	8.07E+08
80	0.18	0.14	0.50	0.22	0.29	0.23	0.13	0.20	1.09E+81	1.89E+42	107.38	1.27E+12
85	0.17	1.29E+25	8.01E+25	0.25	0.32	0.27	9.73E+17	0.22	2.65E+82	1.59E+57	7.52E+79	6.47E+20
90	9.33E+24	4.78E+22	1.70E+41	1.09E+37	1.29E+14	4.38E+59	1.07E+22	0.24	2.50E+81	6.58E+76	8.65E+80	9.95E+75
95	1.12E+73	1.32E+74	2.70E+74	6.22E+75	8.97E+75	5.28E+75	1.03E+74	1.33E+75	1.74E+83	1.49E+79	5.00E+79	2.50E+77

ตารางที่ 4 ผลการเติมค่าสูญหายในข้อมูลอนุกรมเวลาด้วยวิธี ANN

% Missing Values	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	Var11	Var12
0	0	0	0	0	0	0	0	0	0	0	0	0
5	0.02	0.02	0.02	0.03	0.04	0.04	0.02	0.03	12.45	0.86	2.97	0.58
10	0.03	0.03	0.03	0.05	0.06	0.05	0.02	0.05	19.56	1.20	4.17	0.98
15	0.04	0.04	0.03	0.06	0.07	0.07	0.03	0.06	23.62	1.58	5.27	1.13
20	0.04	0.05	0.04	0.07	0.08	0.07	0.04	0.07	26.41	1.72	6.11	1.49
25	0.05	0.05	0.05	0.08	0.10	0.09	0.04	0.08	31.89	2.12	6.33	1.73
30	0.06	0.06	0.05	0.10	0.11	0.10	0.05	0.09	34.23	2.17	7.34	1.98
35	0.07	0.07	0.06	0.10	0.12	0.11	0.06	0.10	38.54	2.48	8.20	2.41
40	0.07	0.07	0.06	0.12	0.13	0.12	0.06	0.11	42.39	2.80	8.30	2.66
45	0.08	0.08	0.08	0.13	0.14	0.14	0.07	0.12	45.02	2.84	9.17	3.21
50	0.08	0.09	0.08	0.14	0.16	0.15	0.08	0.13	47.63	3.06	9.85	3.62
55	0.09	0.10	0.09	0.16	0.17	0.16	0.08	0.14	51.05	3.27	10.35	3.83
60	0.10	0.11	0.10	0.17	0.18	0.18	0.09	0.16	54.60	3.59	10.41	4.74
65	0.11	0.12	0.11	0.19	0.21	0.19	0.10	0.17	58.56	3.79	10.86	5.28
70	0.12	0.13	0.13	0.21	0.22	0.22	0.12	0.20	60.17	4.10	11.11	5.74
75	0.14	0.15	0.14	0.22	0.24	0.24	0.14	0.22	63.10	4.49	11.98	6.24
80	0.16	0.18	0.16	0.24	0.28	0.27	0.16	0.24	67.28	5.26	12.33	6.78
85	0.19	0.21	0.20	0.27	0.31	0.30	0.21	0.28	70.82	5.77	12.62	7.88
90	0.25	0.27	0.28	0.29	0.33	0.34	0.27	0.31	72.52	6.29	13.12	8.99
95	0.45	0.41	0.40	0.32	0.47	0.39	0.50	0.36	76.12	7.58	13.43	11.97



รูปที่ 7 ค่าสูญหายในข้อมูลอนุกรมเวลาตัวแปรที่ 8



รูปที่ 8 ผลการเติมค่าสูญหายในข้อมูลอนุกรมเวลาด้วยวิธีโครงข่ายประสาทเทียม

รูปที่ 7 แสดงข้อมูลอนุกรมเวลาของตัวแปรที่ 8 ซึ่งสังเกตเห็นได้ชัดว่ามีค่าสูญหายจำนวนมาก สำหรับรูปที่ 8 แสดงผลการแทนค่าสูญหายด้วยวิธีโครงข่ายประสาทเทียม ซึ่งข้อมูลที่ได้จากการแทนที่ค่าสูญหายมีอนุกรมคล้ายข้ออนุกรมที่มีอยู่เดิม

5. สรุป และข้อเสนอแนะ

ในการประยุกต์ใช้เครื่องจักรเรียนรู้สำหรับแทนที่ค่าสูญหายในข้อมูลอนุกรมเวลา จำเป็นต้องทำการจัดเตรียมข้อมูลให้เหมาะสมเพื่อการสร้างโมเดลการแทนที่ค่าสูญหาย ปัจจุบันเทคนิคการแทนที่ค่าสูญหายจำนวนมากยังไม่มีประสิทธิภาพเพียงพอ การวิจัยครั้งนี้ได้เน้นที่การแก้ปัญหาข้อมูลสูญหายจำนวนมากในข้อมูลอนุกรมเวลาโดยใช้ข้อมูลจากตัวแปรอื่นๆ ที่มีความสัมพันธ์กัน ผู้วิจัยนำเสนอวิธีการแทนที่ค่าสูญหายจำนวนมาก ในข้อมูลอนุกรมเวลาหลายตัวแปรมีแนวคิดคือ ข้อมูลหลายตัวแปรที่สัมพันธ์กันสามารถทำนายค่าไขว้กันได้ และข้อมูลในอดีตสามารถทำนายอนาคตได้เนื่องจากพฤติกรรมการเกิดเวียนซ้ำตามฤดูกาล ผลการดำเนินการวิจัย โดยนำข้อมูลอนุกรมหลายตัวแปรมาทำการจัดรูปแบบ และป้อนเข้าสู่โมเดลการแทนที่ค่าสูญหายแบบต่างๆ ได้แก่ ค่าเฉลี่ยแถว (Row Average) เพื่อนบ้านใกล้เคียง (K-Nearest Neighbor: KNN) ระบบคลุมเครือ (Fuzzy Logic Systems) โครงข่ายประสาทเทียม (Artificial Neural Network) ผลการวิจัย พบว่าโครงข่ายประสาทเทียมให้ผลการทำนาย ในชุดทดสอบได้ดีที่สุด และเมื่อทำไปใช้ในการแทนที่ค่าสูญหายให้ผลลัพธ์คล้ายค่าจริง

เอกสารอ้างอิง

[1] H. Song, C. Miao, W. Roel, Z. Shen, and F. Catthoor, “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, pp. 233–250, Apr. 2010.

[2] X. Bai, F. Zhang, J. Hou, F. Xia, A. Tolba, and E. Elashkar, “Implicit Multi-Feature Learning for Dynamic Time Series Prediction of the Impact of Institutions,” *IEEE Access*, vol. 5, pp. 16372–16382, 2017.

[3] I. Pratama, A. E. Permanasari, I. Ardiyanto, and R. Indrayani, “A review of missing values handling methods on time-series data,” in *2016 International Conference on Information Technology Systems and Innovation (ICITSI)*, pp. 1–6, 2016.

[4] Y. S. Afrianti, “Imputation Algorithm Based on Copula for Missing,”, pp. 252–257, 2014.

[5] J. D. Velasquez, “Adaptive Multidimensional Neuro-Fuzzy Inference System for Time Series Prediction,” *IEEE Lat. Am. Trans.*, vol. 13, no. 8, pp. 2694–2699, Aug. 2015.

[6] W. Insuwan, U. Suksawatchon, and J. Suksawatchon, “Improving missing values imputation in collaborative filtering with user-preference genre and singular value decomposition,” in *2014 6th International Conference on Knowledge and Smart Technology (KST)*, pp. 87–92, 2014.

[7] G. Chang, Y. Zhang, and D. Yao, “Missing data imputation for traffic flow based on improved local least squares,” *Tsinghua Sci. Technol.*, vol. 17, no. 3, pp. 304–309, Jun. 2012.

[8] Y. Li, A. Ngom, and L. Rueda, “Missing value imputation methods for gene-sample-time microarray data analysis,” in *2010 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pp. 1–7, 2010.

[9] P. Keerin, W. Kurutach, and T. Boongoen, “Cluster-based KNN missing value imputation for DNA microarray data,” in *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 445–450, 2012.

[10] H. Ichihashi, K. Honda, A. Notsu, and T. Yagi, “Fuzzy c-Means Classifier with Deterministic Initialization and Missing Value Imputation,” in *2007 IEEE Symposium on Foundations of Computational Intelligence*, pp. 214–221, 2007.

[11] พยุง มีสัจ, ระบบฟัซซีและโครงข่ายประสาทเทียม, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ. 2555.

[12] N. A. Setiawan, P. A. Venkatachalam, and A. F. M. Hani, “Missing Attribute Value Prediction Based on Artificial Neural Network and Rough Set Theory,” presented at the 2008 International Conference on

BioMedical Engineering and Informatics, vol. 1, pp. 306–310, 2008.

[13] R. and R. D. Little, “Statistical Analysis with Missing Data,” *Wiley, New York.*, p. 381, 1987.