การเลือกคุณลักษณะที่เหมาะสมสำหรับการแทนค่าข้อมูลสูญหาย ของข้อมูลอนุกรมเวลาด้วยเทคนิคเหมืองข้อมูล

Feature Selection Methods for Imputation Missing Values of Time Series Data using Data Mining

กรสิริณัฐ โรจนวรรณ์[1*], วิยุดา เพชรจิรโชติกุล[1]

Kronsirinut Rothjanawan[1*], Wiyuda Phetjirachotkul[1]

(Received: December 22, 2020; Revised: February 24, 2021; Accepted: March 26, 2021)

บทคัดย่อ

งานวิจัยนี้นำเสนอการเลือกคุณลักษณะที่เหมาะสม สำหรับการแทนค่าข้อมูลสูญหาย โดยใช้ข้อมูลอนุกรมเวลาของกรมชลประทานที่ 15 โครงการพระราชดำริลุ่มน้ำปากพนัง ในช่วงระยะเวลา 5 ปี จำนวน 12 ตัวแปร โดยใช้วิธีการเลือกคุณลักษณะโดยการโหวตจาก 5 วิธี ได้แก่ 1) Principal Components Analysis (PCA), 2) Correlation-based Feature Selection (CFS), 3) ReliefF algorithm (ReliefF), 4) Gain Ratio (GR) และ 5) Information Gain (IG) ผู้วิจัยประยุกต์ใช้โครงข่ายประสาทเทียมเพื่อแทนค่าข้อมูลสูญหาย จากการจำลองการสูญหายของข้อมูล 5, 10, 15, 20, 25 และ 30% ซึ่งผลการคัดเลือกตัวแปรที่เหมาะสมได้จำนวน 9 ตัวแปร คือตัวแปรที่ 1 ตัวแปรที่ 2 ตัวแปรที่ 3 ตัวแปรที่ 4 ตัวแปรที่ 5 ตัวแปรที่ 6 ตัวแปรที่ 7 ตัวแปรที่ 8 และตัวแปรที่ 10 ร่วมกับโมเดลโครงข่ายประสาทเทียม (Artificial Neural Network) เพื่อสร้างโมเดลการแทนที่ค่าสูญหาย 10 รูปแบบ คือ 9-3-1, 9-5-1, 9-10-1, 9-15-1, 9-20-1, 9-25-1, 9-30-1, 9-35-1, 9-40-1 และ 9-45-1 ซึ่งหมายถึง จำนวนตัวแปรอินพุต-จำนวนนิวรอนชั้นซ่อน-จำนวนเอาต์พุต การสอนโครงข่ายประสาทเทียมใช้วิธี 10-Fold-Cross-Validation โดยมีการทำซ้ำ 10 รอบ ต่อรูปแบบ พบว่า รูปแบบที่ให้ประสิทธิภาพดีที่สุด คือ รูปแบบ 9-30-1 ให้ค่าผล MSE ต่ำสุด เท่ากับ 0.669

คำสำคัญ: การเลือกคุณลักษณะ การแทนค่า ข้อมูลสูญหาย ข้อมูลอนุกรมเวลา เทคนิคเหมืองข้อมูล

---

[1]คณะวิศวกรรมศาสตร์ มหาวิทยาลัยนราธิวาสราชนครินทร์

[1]Faculty of Engineering, Princess of Naradhiwas University.

*Corresponding Author: kornsirinut.r@pnu.ac.th

Abstract

This research proposes a feature selection method for time series data of Royal Irrigation Department, Royal Thai Army, over a period of 5 year consisting of 12 variables. The proposed feature selection is a voting scheme based on 5 techniques: Principal Components Analysis (PCA), Correlation-based Feature Selection (CFS), ReliefF algorithm (ReliefF), Gain Ratio (GR), and Information Gain (IG) Multilayer Perceptron neural network was used as the missing values imputation model. To test the efficiency of the proposed method, the researchers used the complete data to randomly force the data to be missing for 5, 10, 15, 20, 25 and 30%, respectively. From the experiments, 9 out of 12 variables that are variables 1, 2, 3, 4, 5, 6, 7, 8 and 10, were selected. In addition, 10 Multilayer Perceptron neural network models that are 9-3-1, 9-5-1, 9-10-1, 9-15-1, 9-20-1, 9-25-1, 9-30-1, 9-35-1, 9-40-1 and 9-45-1 (inputs-hidden neurons-outputs) were used in the experiments. Using 10-fold-cross-validation, the best performance was the 9-30-1 model, yielding the lowest MSE equaled 0.669.

Keyword: Feature Selection, Data Imputation, Missing Values, Time Series Data, Data Mining

บทนำ

การพยากรณ์ข้อมูลอนุกรมเวลา (Albayrak, Turhan, & Kurt, 2017) เป็นปัญหาที่ทำได้ไม่ง่ายนัก เนื่องจากข้อมูลอนุกรมเวลามีความเป็นพลวัตและโกลาหล (Dynamics and chaos) มีงานวิจัยจำนวนมากได้นำเสนอการพยากรณ์ข้อมูลอนุกรมเวลา (Jangyodsuk, Seo, Elmasri, & Gao, 2015) เทคนิคที่นิยมได้แก่ เพื่อนบ้านใกล้เคียง (K-Nearest Neighbors: KNN) โครงข่ายประสาทเทียม (Artificial Neural Network: ANN) การถดถอยแบบไม่เป็นเชิงเส้น (Regression) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) และระบบฟัซซีลอจิก (Fuzzy Systems) เป็นต้น การแทนค่าสูญหายจะถูกใช้ก่อนนำข้อมูลไปสร้างโมเดลสำหรับการพยากรณ์ค่าอนุกรมเวลา การพยากรณ์ข้อมูลอนุกรมเวลาถูกนำไปประยุกต์ใช้กับงานหลายด้าน (Wu, Zhou, & Wang, 2020) เช่น การวิเคราะห์แนวโน้ม การวิเคราะห์การถดถอย ขั้นตอนวิธีทางพันธุกรรม การพยากรณ์การไหลของน้ำ เป็นต้น ความยุ่งยากในการพยากรณ์ข้อมูลอนุกรมเวลาจะมีมากยิ่งขึ้น หากข้อมูลอนุกรมเวลาเกิดการสูญหายจำนวนมาก โดยเฉพาะอย่างยิ่งกรณีหลายตัวแปร

ข้อมูลสูญหาย (Missing Data) (Li, Li, & Fishbune, 2016) เป็นปัญหาใหญ่ในการสร้างโมเดลเหมืองข้อมูล ผลของโมเดลที่ได้จากการสร้างด้วยข้อมูลที่ไม่ครบถ้วน อาจจะไม่สามารถนำไปใช้ในการ

ตัดสินใจได้อย่างถูกต้อง ปัญหาของข้อมูลที่ขาดหายเป็นไปได้หลายอย่าง เช่น อาจเกิดการผิดพลาดระหว่างการป้อนข้อมูลด้วยมือ หรืออาจเกิดจากการวัดที่ไม่ถูกต้องและความผิดพลาดของอุปกรณ์ เป็นต้น ซึ่งวิธีการแก้ไขปัญหาดังกล่าว สามารถทำได้หลายวิธี (Li et al., 2017) ซึ่งวิธีการที่ง่ายที่สุด คือ การตัดข้อมูลบางส่วนที่ไม่สมบูรณ์ทิ้งไป แต่วิธีการนี้จะทำให้สูญเสียสาระสำคัญบางอย่างที่อาจส่งผลต่อข้อสรุปก็เป็นได้ ดังนั้นการจัดการกับข้อมูลสูญหายมีหลายวิธีการให้เลือกซึ่งการพิจารณาเลือกใช้วิธีการใดขึ้นอยู่กับลักษณะของข้อมูลสูญหายที่เกิดขึ้น หากเลือกวิธีการที่ไม่เหมาะสมอาจเป็นการเพิ่มค่าความคลาดเคลื่อนของผลลัพธ์ ซึ่งงานวิจัยจำนวนมากได้ศึกษาและพัฒนาวิธีการแทนค่าข้อมูลที่ขาดหายสำหรับปัญหาต่าง ๆ โดยใช้เทคนิคทางสถิติการทำเหมืองข้อมูล (Widiasari, Nugroho, & Widyawan, 2017) ซึ่งเป็นกระบวนการการกลั่นกรองสารสนเทศที่ซ่อนอยู่ในฐานข้อมูลเพื่อใช้ในการทำนายแนวโน้มและพฤติกรรม โดยอาศัยข้อมูลในอดีตเพื่อค้นหารูปแบบความสัมพันธ์ และองค์ความรู้ใหม่จากข้อมูล สร้างแบบจำลองเพื่อการทำนาย (Predictive Data Mining) เป็นการคาดคะแนนลักษณะหรือประมาณค่าที่ชัดเจนของข้อมูลที่จะเกิดขึ้น โดยใช้ข้อมูลในอดีตทำนายข้อมูลในอนาคต หรือข้อมูลที่มีการสูญหายเพื่อเพิ่มประสิทธิภาพในการวิเคราะห์ข้อมูลให้มีความแม่นยำมากขึ้น

เทคนิคการลดขนาดข้อมูล (Data Reduction) (Duong & Tran, 2015) การลดขนาดข้อมูลเป็นกระบวนการหนึ่งในขั้นตอนการเตรียมข้อมูล นั่นคือการทำให้ข้อมูลตั้งต้นมีขนาดลดลงโดยสูญเสียลักษณะสำคัญของข้อมูลน้อยที่สุดและสูญเสียความถูกต้องของผลลัพธ์น้อยที่สุด เนื่องจากข้อมูลแต่ละตัวจะมีความสำคัญต่อการจัดกลุ่มข้อมูลไม่เท่ากัน ด้วยเทคนิคการเลือกข้อมูลที่ดีทำให้สามารถเลือกข้อมูลที่มีความสำคัญและสามารถใช้เป็นตัวแทนของข้อมูลส่วนใหญ่ได้ และในความเป็นจริงมักจะเกิดเหตุการณ์ที่ เรียกกันว่า Curse of dimensionality ขึ้นเสมอ นั้นหมายความว่า จำเป็นต้องลดขนาดมิติของข้อมูลลง (Dimensionality Reduction) เพื่อให้การจำแนก (Classifier) สามารถทำงานได้ถูกต้องมากขึ้น

ในงานวิจัยครั้งนี้นำเสนอ วิธีการลดมิติข้อมูลแบบโหวตจากตัวเลือกข้อมูลจาก 5 เทคนิค ได้แก่ Principal Components Analysis (PCA), Correlation-based Feature Selection (CFS), ReliefF algorithm (ReliefF), Gain Ratio (GR) และInformation Gain (IG) เพื่อลดขนาดมิติของข้อมูลลงให้เหมาะกับการแทนค่าข้อมูลสูญหายร่วมกับโครงข่ายประสาทเทียมแบบหลายชั้น (Multilayer Perceptron Artificial Neural Network) สำหรับการแทนค่าข้อมูลสูญหาย เพื่อนำไปประยุกต์ใช้ในการพยากรณ์ระดับน้ำ (Khairalla, Ning, & AL-Jallad, 2018) ซึ่งงานวิจัยนี้ใช้ข้อมูลระดับน้ำที่ประตูอุทกวิภาชประสิทธิของโครงการพระราชดำริลุ่มน้ำปากพนัง จังหวัดนครศรีธรรมราช

วัตถุประสงค์

เพื่อเลือกคุณลักษณะที่เหมาะสมสำหรับการแทนค่าข้อมูลสูญหายของข้อมูลอนุกรมเวลาด้วยเทคนิคเหมืองข้อมูล ประยุกต์ใช้กับข้อมูลระดับน้ำที่ประตูอุทกวิภาชประสิทธิของโครงการพระราชดำริลุ่มน้ำปากพนัง จังหวัดนครศรีธรรมราช

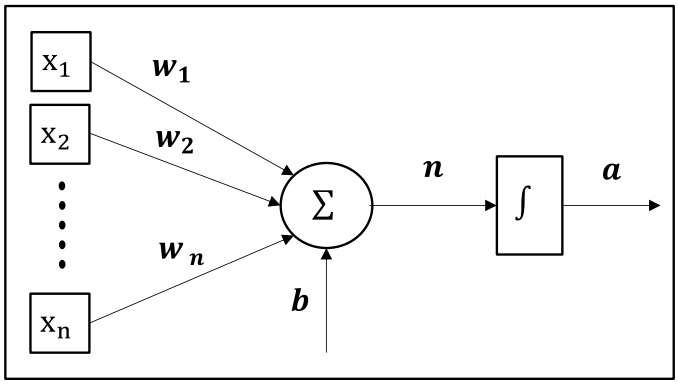
แนวคิดทฤษฎีที่เกี่ยวข้อง

1. การจำแนกข้อมูล (Classification)

การจำแนกข้อมูล (Classification) เป็นเทคนิคการทำเหมืองข้อมูลที่สำคัญเทคนิคหนึ่ง (Wan, Xiao, Zhang, & Leung, 2015) เป็นกระบวนการสร้างโมเดลจำแนกหรือแยกข้อมูลตามกลุ่มที่มีการกำหนดไว้ก่อนล่วงหน้าโดยการใช้ข้อมูลในอดีตเพื่อสร้างโมเดลจำแนกข้อมูล และนำไปใช้ทำนายกลุ่มของข้อมูลที่ไม่เคยพบมาก่อน เทคนิคที่นิยมนำมาใช้สร้างเป็นตัวจำแนกข้อมูล ดังต่อไปนี้

โครงข่ายประสาทเทียม (Artificial Neural Network: ANN) (Liu, Zhang, Quek, & Lin, 2017) มีพื้นฐานมาจากการจำลองการทำงานของสมองมนุษย์ ด้วยโปรแกรมคอมพิวเตอร์ เนื่องจากต้องการให้คอมพิวเตอร์มีความชาญฉลาดในการเรียนรู้เหมือนมนุษย์ที่สามารถเรียนรู้จากข้อมูลได้ สามารถฝึกฝนได้ และสามารถนำความรู้และทักษะ รวมทั้งสามารถนำไปประยุกต์ใช้ในงานต่าง ๆ ได้แก่ การจำแนกข้อมูล การพยากรณ์ค่าจริง และการจัดกลุ่ม

โครงข่ายประสาทเทียมมักถูกเรียกว่า “Black box” เนื่องโครงสร้างมีลักษณะซับซ้อนซึ่งทำให้ไม่ทราบองค์ความรู้หรือเหตุผลของการตัดสินได้ง่าย การทำงานของนิวรอลเน็ตเวิร์คเริ่มจากการส่งข้อมูล ($X_1, X_2,\ldots,X_n$) เข้ามายังอินพุต ส่งต่อผ่านไซแนป (Synapse) ซึ่งแทนด้วยค่าน้ำหนัก ($W_1, W_2,\ldots, W_n$) ส่งต่อผ่านฟังก์ชันกระตุ้น และส่งออกเป็นเอาต์พุต ดังแสดงในภาพที่ 1

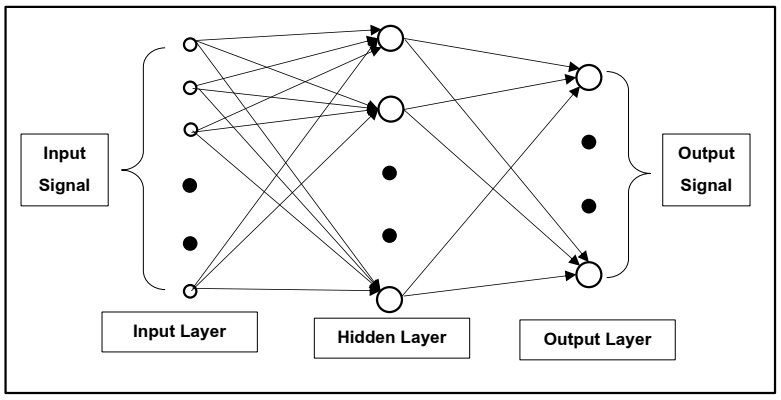


ภาพที่ 1 โครงข่ายประสาทเทียมเพอร์เซ็ปตรอน Function ผลรวม (Summation Function)

$$n = \sum_{i=1}^{z} x_i w_i + b \qquad (1)$$

โดยที่ ตัวแปร n คือ ผลรวมที่ได้จากฟังก์ชันผลรวม
ตัวแปร x_i คือ ค่าข้อมูลเข้าตัวที่ i
ตัวแปร w_i คือ ค่าน้ำหนักของนิวรอลตัวที่ i
ตัวแปร z คือ จำนวนนิวรอลชั้นข้อมูลเข้า
ตัวแปร b คือ ค่าความโน้มเอียง
ตัวแปร i มีค่าตั้งแต่ 1 ถึง z

โครงข่ายประสาทเทียมแบบหลายชั้น (Multilayer Perceptron: MLP) (Sim, Bae, & Choi, 2019) เป็นโครงข่ายประสาทเทียมที่มีการฝึกฝนเป็นแบบ Supervise หรือมีค่าเป้าหมายที่ถูกใช้สำหรับการปรับค่าน้ำหนักของโครงข่ายเพื่อลดค่าผิดพลาดในการทำนาย ในกระบวนการฝึกฝนเริ่มต้นจากการป้อนอินพุต (Input signal) เข้าโครงข่าย คำนวณค่าอินพุตคูณกับค่าน้ำหนักในโครงข่าย ส่งผ่านข้อมูลสู่ฟังก์ชันกระตุ้น และส่งออกเป็นเอาต์พุต (Output signal) ส่วนนี้เรียกว่า Forward pass เอาต์พุตของโครงข่ายจะถูกเปรียบเทียบกับค่าเป้าหมายซึ่งเป็นค่าผิดพลาด และค่าผิดพลาดจะถูกส่งค่าย้อนกลับไปปรับค่าน้ำหนักชั้นนอกสุด ค่าน้ำหนักชั้นซ่อน (Hidden Layer) และค่าน้ำหนักชั้นอินพุต (ส่วนหลังเรียกว่า Backward pass) อัลกอริทึมในการฝึกฝนนี้เป็นที่รู้จักกันดีในชื่อ Backpropagation Algorithm ภาพที่ 2 เป็นตัวอย่างของโครงข่ายประสาทเทียมแบบหลายชั้น



ภาพที่ 2 โครงข่ายประสาทเทียม Multilayer Perceptron แบบ 1 hidden layer

2. การคัดเลือกมิติข้อมูล (Feature selection)

Feature (Makaba & Dogo, 2019) การเลือกแอททริบิวต์ที่มีความสำคัญน้อยออก เพื่อดูประสิทธิภาพในการทำนายหลังจากที่ได้ตัดแอททริบิวต์บางตัวออก ซึ่งส่วนใหญ่จะให้ค่าความถูกต้องสูงขึ้นเพราะแอททริบิวต์ที่เหลือเป็นแอททริบิวต์โดยมีความสำคัญในงานวิจัยนี้ได้นำเสนอการลดมิติข้อมูล โดยเลือกวิธีดังต่อไปนี้

2.1 Correlation-based feature selection (Cfs)

การเลือกคุณสมบัติของแอททริบิวต์นั้น ต้องใช้การพิจารณาบนพื้นฐานความสัมพันธ์ (Correlation-based feature selection: Cfs) (Wang, Yang, Wang, & Wang, 2015) เป็นการหากลุ่มคุณลักษณะที่ได้รับการประเมินค่าจากความสามารถในการคาดการณ์ โดยคุณลักษณะที่ถูกคัดเลือกใช้สำหรับจำประเภทของข้อมูลสามารถจัดการกับคุณลักษณะที่ไม่เกี่ยวข้องกัน จะให้ประสิทธิภาพในการจำแนกประเภทข้อมูลที่ต่ำคำนวณ

2.2 Principal Component Analysis (PCA)

(Ma, Li, & Cottrell, 2020) เป็นเทคนิคที่ใช้ในการลดมิติของเวกเตอร์ลักษณะโดยการฉาย (Project) เวกเตอร์ไปบนแกนใหม่ที่เรียกว่าแกนองค์ประกอบหลัก (Principal component) ซึ่งแกนเหล่านี้มีความสำคัญแตกต่างกันลงไปตามค่าความแปรปรวน (Variance) บนแต่ละแกน

2.3 ReliefFAttributeEval

Relief (Xu, Chong, Li, Arabo, & Xiao, 2018) เป็นเทคนิคคัดเลือกคุณลักษณะที่มีพื้นฐานจาก Instance-based Learning เป็นการใช้การถ่วงน้ำหนักของคุณลักษณะ Relief โดยอาศัยหลักการทางสถิติและมีประสิทธิภาพของการทำงานที่ดี การทำงานของ Relief ใช้เวลาเป็นเชิงเส้น ซึ่งขึ้นกับจำนวนของคุณลักษณะ

2.4 Gain Ratio (GR)

Gain Ratio (Pasha & Mohamed, 2020) เป็นเทคนิคเพื่อประเมินความน่าเชื่อถือของมิติข้อมูล โดยการวัด Gain Ratio ในแต่ละคลาสการคำนวณ GR โดยใช้ค่า SplitINFO ในสมการที่ 2 และการคำนวณค่าวัด Gain Ratio ดังสมการที่ 3

$$SplitINFO = \sum_{i=1}^{k} \frac{n_i}{n} \log_2 \frac{ni}{n} \quad (2)$$

$$GainRatio = \frac{\Delta\, INFO}{SplitINFO} \quad (3)$$

2.5 Information Gain (IG)

Information Gain (Xu & Jiang, 2015)เป็นการประเมินค่าเพื่อใช้ในการแบ่งข้อมูลด้วยการคำนวณค่า Gain สำหรับแต่ละมิติข้อมูลถ้ามิติข้อมูลใดมีค่า Gain สูงสุด จะถูกเลือกให้เป็นกลุ่มย่อยที่มีอำนาจจำแนก ดังสมการที่ 4 แสดงการคำนวณค่า Entropy และคำนวณค่า Gain ดังสมการที่ 5

$$Entropy\,(p) = -\sum_{i=0}^{c-1} p(j|t)\log_2 p(j|t) \tag{4}$$

โดยที่ $\sum_i$คือ ผลรวมของความน่าจะเป็นของค่า j ทีเกิดในคลาส t

$p(j|t)$ คือค่าความถีทีมีความสัมพันธ์ของกลุ่ม j กับโหนด t

$$Gain = Entropy\,(p) - \left(\sum_{i=1}^{k} \frac{ni}{n}\, Entropy\,(i)\right) \tag{5}$$

โดยที่ $Entropy\,(p)$ คือ ค่า Entropy ของตัว Root

$\sum_{i=1}^{k} \frac{ni}{n}\, Entropy\,(i)$คือ ค่า Entropy ในแต่ละโหนดย่อย

2.6 Mean Squared Error (MSE)

Mean Squared Error (Caparino, Sison, & Medina, 2018) เป็นค่าวัดความถูกต้องของการพยากรณ์ที่วัดจากขนาดของค่าความคลาดเคลื่อนของการพยากรณ์ที่ได้ จากกำลังสองของค่าความคลาดเคลื่อน ค่านี้จะมีหน่วยวัดเป็นกำลังสองของหน่วยวัดของค่าสังเกต

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \tag{6}$$

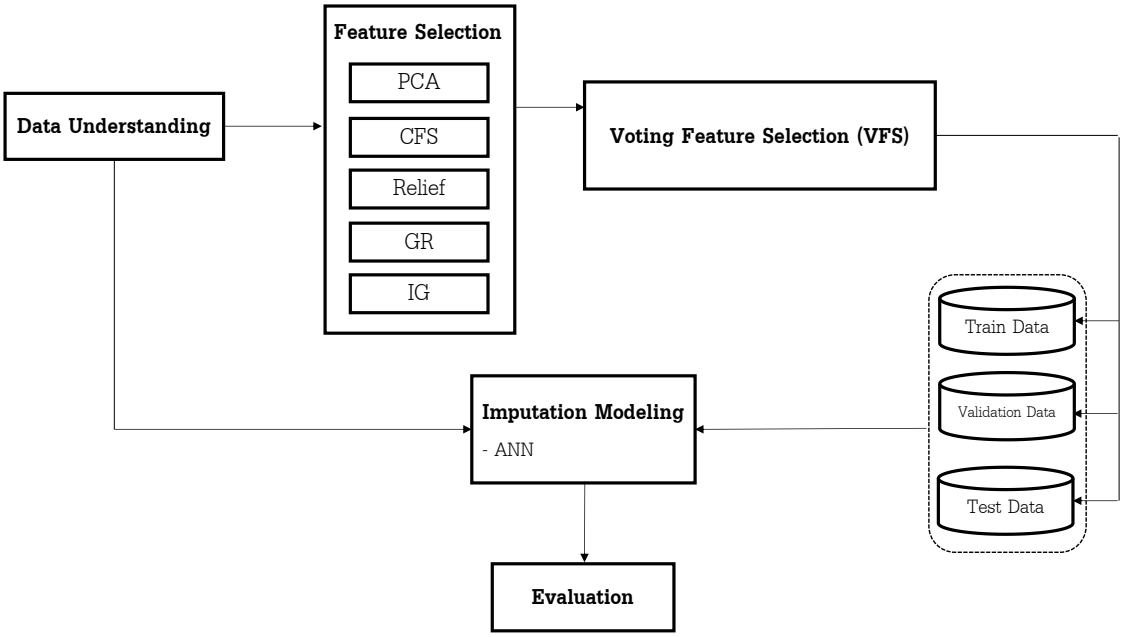
โดยที่ y_i คือ ค่าเป้าหมายจริงของข้อมูล

$\tilde{y}_i$ คือ ค่าพยากรณ์จากการแทนที่ค่าสูญหาย

n คือ จำนวนข้อมูล

วิธีการวิจัย

ในงานวิจัยนี้ได้นำเสนอการเลือกคุณลักษณะที่เหมาะสมสำหรับการแทนค่าข้อมูลสูญหายของข้อมูลอนุกรมเวลาซึ่งมีการเปรียบเทียบเทคนิคต่าง ๆ ที่ใช้ในการเลือกคุณลักษณะร่วมกับโครงข่ายประสาทเทียม เพื่อจะนำมาใช้ในการพยากรณ์ระดับน้ำ โดยมีกระบวนการดำเนินการวิจัย ดังภาพที่ 3



ภาพที่ 3 กระบวนการดำเนินการวิจัย

1. ศึกษาข้อมูลที่ใช้ในงานวิจัย (Data Understanding)

เป็นขั้นตอนของการจัดเตรียมข้อมูลเพื่อการวิเคราะห์โดยนำข้อมูลมาจากกรมชลประทานที่ 15 โครงการพระราชดำริลุ่มน้ำปากพนัง ในช่วงระยะเวลา 5 ปี ตั้งแต่ปี พ.ศ.2557 - พ.ศ. 2562 มีจำนวนทั้งหมด 12 ตัวแปร โดยมีการเก็บข้อมูลด้วยการจดบันทึกลงคอมพิวเตอร์ ซึ่งตัวแปรจะประกอบด้วย 1) ข้อมูลระดับเหนือน้ำเวลา 6.00 น. 2) ข้อมูลระดับเหนือน้ำเวลา 12.00 น. 3) ข้อมูลระดับเหนือน้ำเวลา 18.00 น. 4) ข้อมูลระดับท้ายน้ำเวลา 6.00 น. 5) ข้อมูลระดับท้ายน้ำเวลา 12.00 น. 6) ข้อมูลระดับท้ายน้ำเวลา 18.00 น. 7) ระดับน้ำสูงสุดหน้าประตู 8) ระดับน้ำสูงสุดท้ายประตู 9) อัตราระบายน้ำ/วินาที 10) อัตราระบายน้ำ/วัน 11) ปริมาณน้ำฝน และ12) ปริมาณน้ำเก็บกักตัวแปรทั้งหมดจะถูกนำมาใช้ในการวิเคราะห์เลือกคุณลักษณะที่มีความสำคัญส่งผลต่อความสัมพันธ์กับตัวแปรเป้าหมายเพื่อนำมาใช้ในการแทนค่าข้อมูลสูญหายของข้อมูลอนุกรมเวลา แสดงดังตารางที่ 1

ตารางที่ 1 แสดงข้อมูลที่ใช้ในการวิจัย

ลำดับ	ชื่อตัวแปร
1	A1: ระดับเหนือน้ำ เวลา 6.00 น.
2	A2: ระดับเหนือน้ำ เวลา 12.00 น.
3	A3: ระดับเหนือน้ำ เวลา 18.00 น.
4	A4: ระดับท้ายน้ำ เวลา 6.00 น.
5	A5: ระดับท้ายน้ำ เวลา 12.00 น.
6	A6: ระดับท้ายน้ำ เวลา 18.00 น.
7	A7: ระดับน้ำสูงสุดหน้าประตู
8	A8: ระดับน้ำสูงสุดท้ายประตู
9	A9: อัตราการระบายน้ำ / วินาที
10	A10: อัตราระบายน้ำ/วัน
11	A11: ปริมาณน้ำฝน
12	A12: ปริมาณน้ำเก็บกัก

2. การเลือกคุณลักษณะ (Feature Selection)

การเลือกคุณลักษณะ เพื่อจะทำการคุณลักษณะที่สำคัญของข้อมูล โดยการเลือกคุณลักษณะที่มีความสัมพันธ์กับค่าเป้าหมาย โดยนำข้อมูลมาจากกรมชลประทานที่ 15 โครงการพระราชดำริลุ่มน้ำปากพนังในช่วงระยะเวลา 5 ปี ตั้งแต่ปี พ.ศ. 2557 - พ.ศ. 2562 ซึ่งมีตัวแปรทั้งหมด 12 ตัวแปร ดังนี้ 1) ข้อมูลระดับเหนือน้ำเวลา 6.00 น. 2) ข้อมูลระดับเหนือน้ำเวลา 12.00 น. 3) ข้อมูลระดับเหนือน้ำเวลา 18.00 น. 4) ข้อมูลระดับท้ายน้ำเวลา 6.00 น. 5) ข้อมูลระดับท้ายน้ำเวลา 12.00 น. 6) ข้อมูลระดับท้ายน้ำเวลา 18.00 น. 7) ระดับน้ำสูงสุดหน้าประตู 8) ระดับน้ำสูงสุดท้ายประตู 9) อัตราระบายน้ำ/วินาที 10) อัตราระบายน้ำ/วัน 11) ปริมาณน้ำฝน และ 12) ปริมาณน้ำเก็บกัก โดยอัลกอริทึมในการลดขนาดมิติของข้อมูล แบบ Principal Components Analysis (PCA),Correlation-based Feature Selection (CFS), ReliefF algorithm (ReliefF), Gain Ratio (GR) ดังสมการที่ 2-3 และ Information Gain (IG) ดังสมการที่ 4-5 มาทำการเปรียบเทียบผลกับการเลือกใช้แอทริบิวต์ทั้งหมด โดยทำการโหวตคุณลักษณะที่สำคัญ

(Voting Feature Selection) มา 3ใน 5 จากเทคนิคในการคุณลักษณะที่สำคัญของข้อมูลสามารถสรุปผลจากการคุณลักษณะที่สำคัญของข้อมูลได้ดังตารางที่ 2

ตารางที่ 2 ผลของจำนวน Attribute จากการคุณลักษณะที่สำคัญของข้อมูล

Feature Selection	Attribute	Variable
CFS	3	2,3,10
PCA	9	1,2,3,4,5,6,7,8,10
Relife	10	1,2,3,4,5,6,7,8,9,11
GR	7	1,2,3,4,5,6,7
IR	8	1,4,5,6,7,8,9,10
Voting Feature Selection	9	1,2,3,4,5,6,7,8,10

จากตารางที่ 2 จะเห็นว่า ผลจากการเอาคุณลักษณะทั้งหมด 12 ตัวแปร โดยอัลกอริทึมในการลดขนาดมิติของข้อมูล เพราะการทำให้ข้อมูลมีขนาดลดลงโดยสูญเสียลักษณะสำคัญของข้อมูลน้อย และสูญเสียความถูกต้องของผลลัพธ์น้อย ซึ่งข้อมูลแต่ละตัวจะมีความสำคัญต่อการจัดกลุ่มข้อมูลไม่เท่ากัน ด้วยเทคนิคการเลือกข้อมูลที่ดีจะทำให้สามารถเลือกข้อมูลที่มีความสำคัญ และสามารถใช้เป็นตัวแทนของข้อมูลส่วนใหญ่ได้ผลจากการเปรียบเทียบเลือกตัวแปรกับการเลือกใช้แอทริบิวต์ทั้งหมด โดยทำการโหวตคุณลักษณะที่สำคัญ (Voting Feature Selection) มา 3ใน 5 จากเทคนิคในการคุณลักษณะที่สำคัญของข้อมูล จะได้ตัวแปร 9 ตัว คือ ตัวแปรที่ 1 ตัวแปรที่ 2 ตัวแปรที่ 3 ตัวแปรที่ 4 ตัวแปรที่ 5 ตัวแปรที่ 6 ตัวแปรที่ 7 ตัวแปรที่ 8 ตัวแปรที่ 10 นำเข้าสู่ขั้นตอนการออกแบบโมเดลการแทนค่าข้อมูลสูญหาย

การออกแบบโมเดลการแทนค่าข้อมูลสูญหาย

นำชุดข้อมูลของระดับน้ำที่ประตูอุทกวิภาชประสิทธิ ของโครงการพระราชดำริลุ่มน้ำปากพนัง จังหวัดนครศรีธรรมราชโดยใช้วิธีการเลือกคุณลักษณะ (Feature Selection) แบบ CFS (Correlation-based feature selection, PCA (Principal Component Analysis), Relief (Relief algorithm), Gain Ratio (GR) และ Information Gain (IG) ร่วมกับโมเดลโครงข่ายประสาทเทียม (Artificial Neural Network) 10 รูปแบบ คือ 9-3-1, 9-5-1, 9-10-1, 9-15-1, 9-20-1, 9-25-1, 9-30-1, 9-35-1, 9-40-1และ 9-45-1 และทำการแบ่งชุดข้อมูลที่ใช้สำหรับการเรียนรู้ หรือข้อมูลชุดสอน (Training Data Set) เป็นข้อมูลส่วนที่มีความสมบูรณ์ คิดเป็น 70% ของข้อมูลทั้งหมด เพื่อให้แบบจำลองได้เรียนรู้จากชุดข้อมูลจริง สำหรับ

ข้อมูลชุดทดสอบ (Test Data Set) คิดเป็น 15% ของข้อมูลทั้งหมด และข้อมูลชุดตรวจสอบ (Validation Data Set) คิดเป็น 15% ของข้อมูลทั้งหมด

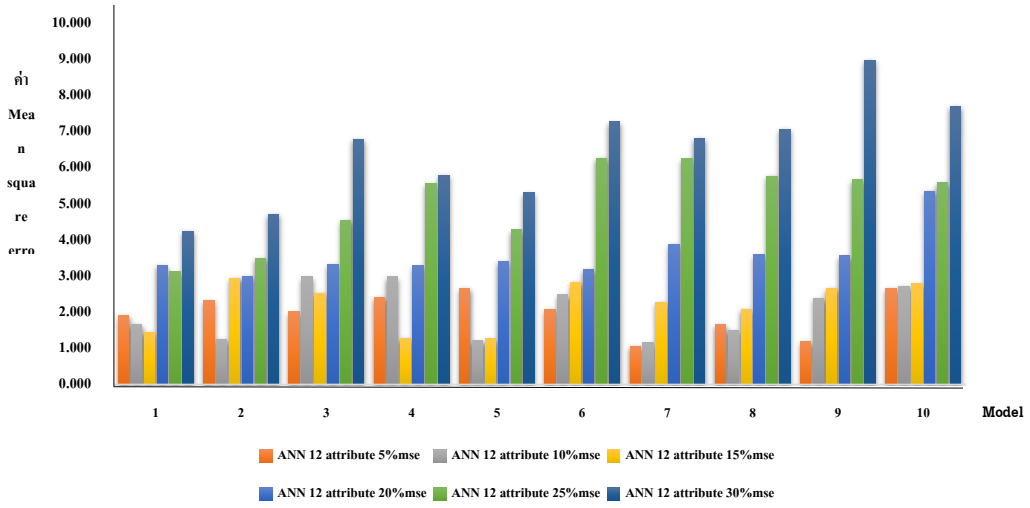
ค่าสูญหายที่ใช้เป็นค่าตั้งแต่ 5 - 30% โดยเพิ่มทีละ 5% ทำการแบ่งข้อมูลเป็น 3 ชุด คือ ชุดข้อมูลสอน (Training Data Set) ชุดข้อมูลทดสอบ (Testing Data Set) และชุดข้อมูลตรวจสอบ (Validation Data Set) โดยหาประสิทธิภาพของโมเดลด้วยค่าคลาดเคลื่อนเฉลี่ยกำลังสอง (Mean square error: MSE) โดยโปรแกรมที่ใช้ในงานวิจัยครั้งนี้ ผู้วิจัยเลือกใช้โปรแกรม WEKA และ MATLAB ซึ่งเป็นซอฟต์แวร์ด้านการทำเหมืองข้อมูล โดยคัดเลือกคุณลักษณะของข้อมูล (Feature selection) โดยอัลกอริธึมในการเลือกข้อมูลของโปรแกรม WEKA และใช้โปรแกรม MATLAB ในการเปรียบเทียบระสิทธิภาพการลดมิติข้อมูลเพื่อหาค่าความถูกต้องจากกระบวนเหล่านั้นด้วยเทคนิคการจำแนกข้อมูลแบบโครงข่ายประสาทเทียม (Artificial Neural Network)

การออกแบบโครงข่ายประสาทเทียม เพื่อแทนค่าข้อมูลสูญหายของข้อมูลระดับน้ำในงานวิจัยครั้งนี้ ผู้วิจัยได้ทำการสุ่มสร้างค่าสูญหายที่ 5 - 30% ตามลำดับ จากนั้นนำข้อมูลเข้าโมเดลโครงข่ายประสาทเทียมแบบสามชั้น ประกอบด้วย 1) ชั้นอินพุต 2) ชั้นซ่อน และ 3) ชั้นเอ้าต์พุต โดยชั้นอินพุตมีจำนวน 9 อินพุต เท่ากับจำนวน 10 ตัวแปรอินพุตลบด้วย 1 ตัวแปรเอ้าต์พุต ส่วนชั้นซ่อนจะกำหนดชนิดของฟังก์ชันเป็นแบบ Log Sigmoid และเอ้าต์พุตนิวรอนมีหนึ่งโหนด กำหนดฟังก์ชันเป็นแบบ Pure Linear โครงสร้างของโครงข่ายประสาทเทียมที่ใช้เป็น 10 รูปแบบเปลี่ยนแปลงตามจำนวนนิวรอนในชั้นซ่อน ได้แก่ 1) รูปแบบ 9-3-1 (9 อินพุต 3 นิวรอนชั้นซ่อน และ 1 เอ้าต์พุต) 2) 9-5-1 3) รูปแบบ 9-10-1 4) รูปแบบ 9-15-1 5) รูปแบบ 9-20-1 6) รูปแบบ 9-25-1 7) รูปแบบ 9-30-1 8) รูปแบบ 9-35-1 9) รูปแบบ 9-40-1 10) รูปแบบ 9-45-1เพื่อทดสอบความเหมาะสมของค่าสูญหายที่เปอร์เซ็นต์ต่าง ๆ สำหรับอัลกอริธึมในการฝึกฝนโครงข่ายประสาทเทียมใช้ LM (Levenberg–Marquardt) การจบการฝึกฝนโดยใช้ข้อมูลชุดตรวจสอบ (Validation set) หากค่าความคลาดเคลื่อนหรือค่าผิดพลาดใน ชุดตรวจสอบไม่มีการลดลงจบกระบวนการฝึกฝน ซึ่งในแต่ละรูปแบบจะมีการทดลองซ้ำ 10 ครั้ง โดยหาประสิทธิภาพของโมเดลด้วยค่าคลาดเคลื่อนเฉลี่ยกำลังสอง (Mean square error: MSE)

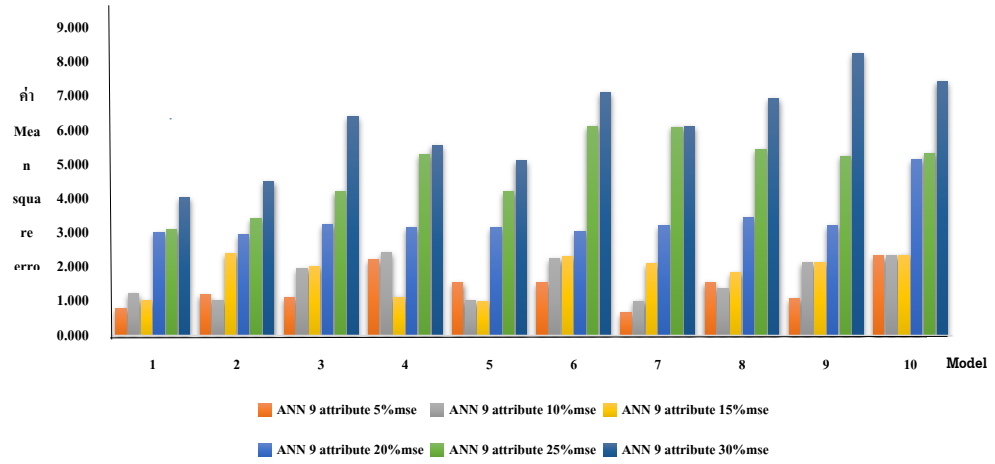
ผลการวิจัย

จากผลการทดลอง จะเห็นว่าภาพที่ 4 เป็นข้อมูลที่ไม่ได้มีการเลือกคุณลักษณะที่สำคัญของข้อมูลมีตัวแปร 12 ตัวแปร และภาพที่ 5 ข้อมูลที่มีการเลือกคุณลักษณะที่สำคัญของข้อมูลมีตัวแปร 9 ตัวแปร ที่ได้จากการโหวตคุณลักษณะที่สำคัญ (Voting Feature Selection) มา 3ใน 5 จากเทคนิคในการคุณลักษณะที่สำคัญของข้อมูล ด้วยเทคนิคการเลือกข้อมูลที่ดีจะทำให้สามารถเลือกข้อมูลที่มี

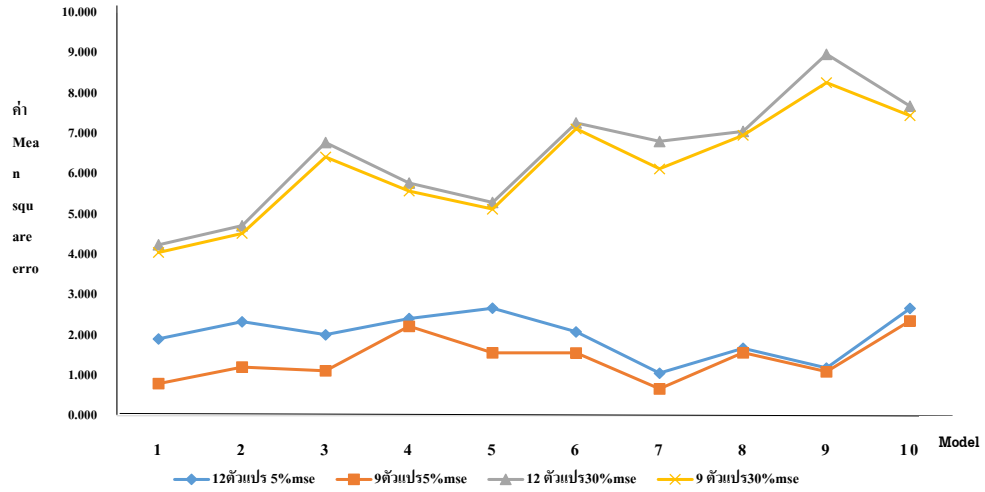
ความสำคัญ และสามารถใช้เป็นตัวแทนของข้อมูลส่วนใหญ่ได้ จากนั้นนำข้อมูลมาทำการสุ่มสร้างข้อมูลสูญหายที่ 5-30% ตามลำดับ และนำเข้าสู่โมเดล จากการหาประสิทธิภาพของโมเดลด้วยค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง (Mean square error: MSE) จะค่อย ๆ เพิ่มขึ้นเมื่อมีการสูญหายของข้อมูลที่เพิ่มขึ้น และทำให้ทราบว่าโมเดลสามารถทนต่อค่าสูญหายที่เปอร์เซ็นต์เพิ่มสูงขึ้นพบว่า จากภาพที่ 4 ค่าสูญหายที่ 5% รูปแบบที่มีประสิทธิภาพ คือ รูปแบบ 9-30-1ให้ผล MSE ต่ำสุด คือ 1.058 และค่าสูญหายที่ 30% รูปแบบที่ 9-40-1 ให้ผล MSE สูงสุด คือ 8.974 และภาพที่ 5 ค่าสูญหายที่ 5% รูปแบบที่มีประสิทธิภาพ คือ รูปแบบ 9-30-1 ให้ผล MSE ต่ำสุด คือ 0.669 และค่าสูญหายที่ 30% รูปแบบที่ 9-40-1 ให้ผล MSE สูงสุด คือ 8.269 ซึ่งการเลือกคุณลักษณะที่เหมาะสมจะส่งผลต่อแทนค่าข้อมูลสูญหายของข้อมูลอนุกรมเวลา ดังภาพที่ 6



ภาพที่ 4 กราฟแสดงผลการประสิทธิภาพของโมเดลแบบ 12 ตัวแปร



ภาพที่ 5 กราฟแสดงผลการประสิทธิภาพของโมเดลแบบ 9 ตัวแปร



ภาพที่ 6 กราฟแสดงผลการเปรียบเทียบประสิทธิภาพของโมเดลแบบ 9 ตัวแปร และแบบ 12 ตัวแปร ที่ค่าสูญหายที่ 5 และค่าสูญหายที่ 30%

อภิปรายผล

จากงานวิจัยนี้นำเสนอการเลือกคุณลักษณะที่เหมาะสมสำหรับการแทนค่าข้อมูลสูญหายของข้อมูลอนุกรมเวลา เพื่อทำการเลือกคุณลักษณะที่สำคัญของข้อมูล โดยเลือกตัวแปรที่มีความสัมพันธ์กับตัวแปรเป้าหมาย ซึ่งจากการศึกษางานวิจัยที่เกี่ยวกับการพยากรณ์ข้อมูล จะเอาข้อมูลที่มีอยู่มาใช้ในการพยากรณ์ข้อมูลล่วงหน้า แต่ในงานวิจัยนี้ได้มีการนำเทคนิคการเลือกคุณลักษณะที่เหมาะสมสำหรับการแทนค่าข้อมูลสูญหาย ก่อนการพยากรณ์ข้อมูล โดยได้เลือกใช้เทคนิคแบบ Principal Components Analysis (PCA), Correlation-based Feature Selection (CFS), ReliefF algorithm (ReliefF), Gain Ratio (GR) และ Information Gain (IG) มาทำการเปรียบเทียบผลกับการเลือกใช้แอทริบิวต์ทั้งหมด และโหวตคุณลักษณะที่สำคัญ (Voting Feature Selection) มา 3ใน 5 จากเทคนิคในการคุณลักษณะที่สำคัญของข้อมูล จาก 12 ตัวแปรเหลือ 9 ตัวแปร จากนั้นนำเข้าสู่โมเดลเข้าสู่การแทนค่าข้อมูลสูญหาย ร่วมกับโมเดลโครงข่ายประสาทเทียม (Artificial Neural Network) 10 รูปแบบ คือ 9-3-1, 9-5-1, 9-10-1, 9-15-1, 9-20-1, 9-25-1, 9-30-1, 9-35-1, 9-40-1 และ 9-45-1 ทำการสุ่มที่ค่าสูญหายตั้งแต่ 5 - 30% และทำการแบ่งชุดข้อมูลที่ใช้สำหรับการเรียนรู้ หรือข้อมูลชุดสอน (Training Data Set) เป็นข้อมูลส่วนที่มีความสมบูรณ์ คิดเป็น 70% ของข้อมูลทั้งหมด เพื่อให้แบบจำลองได้เรียนรู้จากชุดข้อมูลจริง สำหรับข้อมูลชุดทดสอบ (Test Data Set) คิดเป็น 15% ของข้อมูลทั้งหมด และข้อมูลชุดตรวจสอบ (Validation Data Set) คิดเป็น 15% ของข้อมูลทั้งหมดซึ่งในแต่ละเปอร์เซ็นต์ของค่าสูญหายจะทำ 10 รอบ แล้วหาค่าเฉลี่ย

ออกมา โดยหาประสิทธิภาพของโมเดลด้วยค่าคลาดเคลื่อนเฉลี่ยกำลังสอง (Mean square error: MSE) แล้วทำการเปรียบเทียบประสิทธิภาพของโมเดล ที่นำเข้าตัวแปร 12 ตัวแปร และตัวแปร 9 ตัวแปร พบว่า การเลือกคุณลักษณะที่เหมาะสม และเปอร์เซ็นต์การสูญหายของข้อมูล จะส่งผลต่อแทนค่าข้อมูลสูญหายของข้อมูลอนุกรมเวลา

สรุป

จากวัตถุประสงค์ของงานวิจัยเพื่อเลือกคุณลักษณะที่เหมาะสมสำหรับการแทนค่าข้อมูลสูญหายของข้อมูลอนุกรมเวลาด้วยเทคนิคเหมืองข้อมูล โดยใช้ข้อมูลจากกรมชลประทานที่ 15 โครงการพระราชดำริลุ่มน้ำปากพนัง ในช่วงระยะเวลา 5 ปี โดยเป็นข้อมูลอนุกรมเวลา ประกอบด้วย 1) ข้อมูลระดับเหนือน้ำเวลา 6.00 น. 2) ข้อมูลระดับเหนือน้ำเวลา 12.00 น.3) ข้อมูลระดับเหนือน้ำเวลา 18.00 น. 4) ข้อมูลระดับท้ายน้ำเวลา 6.00 น. 5) ข้อมูลระดับท้ายน้ำเวลา 12.00 น. 6) ข้อมูลระดับท้ายน้ำเวลา 18.00 น. 7) ระดับน้ำสูงสุดหน้าประตู 8) ระดับน้ำสูงสุดท้ายประตู 9) อัตราระบายน้ำ/วินาที 10) อัตราระบายน้ำ/วัน 11) ปริมาณน้ำฝน และ 12) ปริมาณน้ำเก็บกัก โดยใช้วิธีการเลือกคุณลักษณะ (Feature Selection) แบบ CFS (Correlation-based feature selection), PCA (Principal Component Analysis), Relief (Relief Attribute Eval), Gain Ratio (GR) และ Information Gain (IG) ร่วมกับโมเดลโครงข่ายประสาทเทียม (Artificial Neural Network) 10 รูปแบบ คือ 9-3-1, 9-5-1, 9-10-1, 9-15-1, 9-20-1, 9-25-1, 9-30-1, 9-35-1, 9-40-1 และ 9-45-1 ที่ค่าสูญหายตั้งแต่ 5 - 30% ซึ่งการทดสอบโมเดลแต่ละรูปแบบได้ทำซ้ำ 10 รอบ พบว่าการเลือกคุณลักษณะที่เหมาะสมจะส่งผลต่อแทนค่าข้อมูลสูญหายของข้อมูลอนุกรมเวลา โดยการทดลองจาก 12 ตัวแปร ค่าสูญหายที่ 5% รูปแบบที่มีประสิทธิภาพ คือ รูปแบบ 9-30-1ให้ผล MSE ต่ำสุด คือ 1.058 และค่าสูญหายที่ 30% รูปแบบที่ 9-40-1 ให้ผล MSE สูงสุด คือ 8.974 และการทดลองจาก 9 ตัวแปร ค่าสูญหายที่ 5% รูปแบบที่มีประสิทธิภาพ คือ รูปแบบ 9-30-1ให้ผล MSE ต่ำสุด คือ 0.669 และค่าสูญหายที่ 30% รูปแบบที่ 9-40-1 ให้ผล MSE สูงสุด คือ 8.269 ซึ่งเมื่อเปรียบเทียบการทดลองตัวแปร 12 ตัว และ ตัวแปร 9 ตัว พบว่า ผลการทดลองจากตัวแปร 9 ตัว ที่ผ่านวิธีการเลือกคุณลักษณะ (Feature Selection) ให้ประสิทธิภาพดีกว่า

ข้อเสนอแนะ

ในงานวิจัยครั้งต่อไป เพื่อให้ทราบถึงความเหมาะสมของโมเดลที่ใช้ในการวิเคราะห์ เนื่องจากปัจจัยของแต่ละพื้นที่มีความแตกต่างกัน ควรเพิ่มปัจจัยที่เกี่ยวข้องกับการพยากรณ์ระดับน้ำ และควรมีการเปรียบเทียบโมเดลเช่น ระบบผสมโครงข่ายประสาทเทียม ฟัซซี ซัพพอร์ตเวกเตอร์ และรีเกรซชัน

เป็นต้น ตลอดถึงการพยากรณ์ค่าอนาคต ของระดับน้ำด้วยเทคนิควิธีต่าง ๆ เพื่อเป็นแนวทางนำไปใช้ประโยชน์ในภาคเกษตรกรรม และการแจ้งเตือนภัยน้ำของหน่วยงานที่เกี่ยวข้องต่อไป

รายการอ้างอิง (References)

Albayrak, M., Turhan, K., & Kurt, B. (2017). *A missing data imputation approach using clustering and maximum likelihood estimation.* 2017 Medical Technologies National Congress (TIPTEKNO), 1–4.

Caparino, E. T., Sison, A. M., & Medina, R. P. (2018). *A Modified Imputation Method to Missing Data as a Preprocessing Technique.* 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), 1–6.

Duong, T. V., & Tran, D.Q. (2015). A fusion of data mining techniques for predicting movement of mobile users. *Journal of Communications and Networks*, *17*(6), 568–581.

Jangyodsuk, P., Seo, D. J., Elmasri, R., & Gao, J. (2015). *Flood Prediction and Mining Influential Spatial Features on Future Flood with Causal Discovery.* 2015 IEEE International Conference on Data Mining Workshop (ICDMW), 1462–1469.

Khairalla, M., Ning, X., & AL-Jallad, N. (2018). Modelling and optimisation of effective hybridisation model for time-series data forecasting. *The Journal of Engineering*, *2018*(2), 117–122.

Li, L., Zhang, J., Wang, Y., & Ran, B. (2017). *Multiple imputation for incomplete traffic accident data using chained equations.* 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), 1–5.

Li, X., Li, G., & Fishbune, R. (2016). *A Novel Missing-Rate-Oriented Selective Algorithm for Handling Missing Data by Minimizing Imputation.* 2016 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 234–237.

Liu, Z., Zhang, W., Quek, T. Q. S., & Lin, S. (2017). *Deep fusion of heterogeneous sensor data.* 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5965–5969.

Ma, Q., Li, S., & Cottrell, G. (2020). *Adversarial Joint-Learning Recurrent Neural Network for Incomplete Time Series Classification.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–1.

Makaba, T., & Dogo, E. (2019). *A Comparison of Strategies for Missing Values in Data on Machine Learning Classification Algorithms.* 2019 International Multidisciplinary Information Technology and Engineering Conference *(IMITEC)*, 1–7.

Pasha, S. J., & Mohamed, E. S. (2020). *Ensemble Gain Ratio Feature Selection (EGFS) Model with Machine Learning and Data Mining Algorithms for Disease Risk Prediction.* 2020 International Conference on Inventive Computation Technologies (ICICT), 590–596.

Sim, S., Bae, H., & Choi, Y. (2019). *Likelihood-based Multiple Imputation by Event Chain Methodology for Repair of Imperfect Event Logs with Missing Data.* 2019 International Conference on Process Mining (ICPM), 9–16.

Wan, D., Xiao, Y., Zhang, P., & Leung, H. (2015). *Hydrological Big Data Prediction Based on Similarity Search and Improved BP Neural Network.* 2015 IEEE International Congress on Big Data, 343–350.

Wang, H., Yang, J., Wang, Z., & Wang, Q. (2015). *A binary granular algorithm for spatiotemporal meteorological data mining.* 2015 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM), 5–11.

Widiasari, I. R., Nugroho, L. E., & Widyawan. (2017). *Deep learning multilayer perceptron (MLP) for flood prediction model using wireless sensor networkbased hydrology time series data mining.* 2017 International Conference on Innovative and Creative Information Technology (ICITech), 1–5.

Wu, Z., Zhou, Y., & Wang, H. (2020). Real-Time Prediction of the Water Accumulation Process of Urban Stormy Accumulation Points Based on Deep Learning. *IEEE Access*, *8*, 151938–151951.

Xu, J., & Jiang, H. (2015). *An Improved Information Gain Feature Selection Algorithm for SVM Text Classifier.* 2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, 273–276.

Xu, X., Chong, W., Li, S., Arabo, A., & Xiao, J. (2018). MIAEC: Missing Data Imputation Based on the Evidence Chain. *IEEE Access*, *6*, 12983–12992.