TSWE-03

การศึกษาการแทนที่ค่าสูญหายแบบหลายตัวแปรสำหรับการพยากรณ์ระดับน้ำลุ่มน้ำปากพนัง

*กรกฎ ถนิมกาญจน์[1] และ พยุง มีสัจ[2]

[1]*คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย*
109 หมู่ 2 ตำบลถ้ำใหญ่ อำเภอทุ่งสง นครศรีธรรมราช 80110
[2]*คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ*
1518 ถนนประชาราษฎร์ 1 แขวงวงศ์สว่าง เขตบางซื่อ กรุงเทพมหานคร 10800

ผู้เขียนติดต่อ: กรกฎ ถนิมกาญจน์ E-mail: goragot.1982@gmail.com

บทคัดย่อ

ประตูระบายน้ำอุทกวิภาชประสิทธิ เป็นประตูน้ำหลักที่ป้องกันไม่ให้น้ำเค็มทางด้านอ่าวปากพนังรุกล้ำเข้ามาในแม่น้ำปากพนังได้อย่างสิ้นเชิงโดยเฉพาะในฤดูแล้ง ซึ่งจะมีความสำคัญกับเกษตรกรบริเวณ อำเภอปากพนังแต่ปัญหาที่พบคือ ข้อมูลน้ำมีการสูญหาย (Missing Data) เป็นจำนวนมาก ทำให้ส่งผลต่อความแม่นยำในการพยากรณ์ระดับน้ำ ซึ่งการเตรียมข้อมูล (Data Preprocessing) จะเป็นขั้นตอนที่สำคัญและใช้เวลานานสำหรับการทำเหมืองข้อมูล (Data Mining) งานวิจัยนี้ได้ประยุกต์ใช้กับข้อมูลระดับน้ำ เพื่อนำเสนอรูปแบบการแทนที่ค่าสูญหายแบบหลายตัวแปรสำหรับการพยากรณ์ระดับน้ำ ด้วยวิธีสมการถดถอย (Regression Imputation) และวิธีการสมาชิกที่ใกล้ที่สุด (K-Nearest Neighbor: KNN) ซึ่งเป็นเทคนิคทางเหมืองข้อมูล เพื่อการพยากรณ์ระดับน้ำลุ่มน้ำปากพนัง การศึกษาครั้งนี้ได้ประเมินค่าความถูกต้องของเทคนิคโดย ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Root mean square error, RMSE) จากการศึกษาการแทนที่ค่าสูญหายแบบหลายตัวแปรสำหรับการพยากรณ์ระดับน้ำลุ่มน้ำปากพนัง พบว่า วิธีการสมาชิกที่ใกล้ที่สุด เป็นวิธีการที่มีประสิทธิภาพมากที่สุด เพราะมีค่า RMSE ต่ำกว่า วิธีสมการถดถอย

คำสำคัญ: การเปรียบเทียบประสิทธิภาพ; ค่าสูญหาย; การแทนค่าข้อมูล; วิธีการทางเหมืองข้อมูล

1. บทนำ

ประตูระบายน้ำอุทกวิภาชประสิทธิ เป็นประตูน้ำหลักที่ป้องกันไม่ให้น้ำเค็มทางด้านอ่าวปากพนังรุกล้ำเข้ามาในแม่น้ำปากพนังได้อย่างสิ้นเชิงโดยเฉพาะในฤดูแล้ง ซึ่งจะมีความสำคัญกับเกษตรกรบริเวณ อำเภอปากพนัง เนื่องจากการเก็บข้อมูลจะทำการเก็บด้วยตรวจวัด และบันทึกข้อมูลลงคอมพิวเตอร์ แต่ปัญหาที่พบคือ ข้อมูลน้ำมีการสูญหาย (Missing Data) เป็นจำนวนมากซึ่งเกิดจากหลายปัจจัย เช่น เกิดความผิดพลาดในการบันทึกข้อมูลโดยบุคคล และไม่สามารถกลับไปจดบันทึกได้ เพราะไม่สามารถย้อนเวลาได้ ทำให้ส่งผลต่อประสิทธิภาพการพยากรณ์ระดับน้ำ ซึ่งการเตรียมข้อมูล (Data Preprocessing) จะเป็นขั้นตอนที่สำคัญและใช้เวลานานสำหรับการทำเหมืองข้อมูล (Data Mining)

งานวิจัยจำนวนมากได้ศึกษาและพัฒนาวิธีการแทนค่าข้อมูลที่ขาดหายสำหรับปัญหาต่าง ๆ โดยใช้เทคนิคทางสถิติการทำเหมืองข้อมูล รวมไปถึงเทคนิคการเรียนรู้ของเครื่องจักรเพื่อเพิ่มประสิทธิภาพในการวิเคราะห์ข้อมูล เช่น Troyanskaya *et al.* [2] ได้ทำการศึกษาเปรียบเทียบวิธีการแทนที่ข้อมูลที่ขาดหายด้วยอัลกอริธึม แบบเพื่อนบ้านใกล้เคียง (K-Nearest Neighbor : KNN Imputation) การแตกค่าแบบเอกฐาน (Singular Value Decomposition: SVD) และแบบค่าเฉลี่ยแถวข้อมูล (Row Average) ผลที่ได้คือ KNN Impute ให้ประสิทธิภาพในการประมาณค่าข้อมูลที่ขาดหายดีกว่า Oba *et al.* ได้พัฒนาวิธีการสำหรับการประมาณค่าขาดหาย ด้วยวิธีเรียกว่า การวิเคราะห์องค์-ประกอบพื้นฐานแบบเบย์ (Bayesian Principal Component Analysis: BPCA impute) [3] ที่สามารถเอาชนะวิธีการ KNN และ SVD ได้

ในงานวิจัยนี้ต้องการจะศึกษาการแทนค่าข้อมูลที่สูญหาย ซึ่งเป็นวิธีการที่นิยมในการมาประยุกต์ใช้ในการแทนค่าสูญหายวิธีการหนึ่งก็คือ วิธีสมาชิกที่ใกล้ที่สุด ซึ่งพบว่าอยู่ในเกณฑ์ที่ดีและวิธีสมการถดถอย (Regression) เข้ามาเปรียบเทียบประสิทธิภาพ เพื่อใช้สำหรับการพยากรณ์ระดับน้ำ โดยนำข้อมูลจาก ที่ราบลุ่มน้ำปากพนัง มาทำการวิจัยในครั้งนี้ โดยเนื้อหาในบทความนี้ได้แบ่งเป็นส่วนดังนี้ ส่วนที่ 2

กล่าวถึงทฤษฎีที่เกี่ยวข้อง ส่วนที่ 3 วิธีการดำเนินการวิจัย ส่วนที่ 4 ผลการทดลอง ส่วนที่ 5 สรุปผลการทดลอง และส่วนที่ 6 ได้กล่าวถึงเอกสารอ้างอิงที่ได้ศึกษา

2. ทฤษฎีที่เกี่ยวข้อง

2.1 เหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูล ถือเป็นกระบวนการของการกลั่นกรองสารสนเทศ (Information) ที่ซ่อนอยู่ในฐานข้อมูลขนาดใหญ่ เพื่อใช้ในการทำนายแนวโน้ม และพฤติกรรม โดยอาศัยข้อมูลในอดีตเพื่อค้นหารูปแบบความสัมพันธ์ และองค์ความรู้ใหม่ จากข้อมูล [4] โดยมีขั้นตอนดังนี้

1) ทำความเข้าใจปัญหา โดยการเลือกข้อมูลให้มีความเหมาะสมกับอัลกอริทึมที่ใช้งานจำนวนที่ต้องการและค่าเป้าหมายเพื่อให้ได้ผลลัพธ์ที่ต้องการ

2) ทำความเข้าใจข้อมูล โดยการรวบรวมตรวจสอบความถูกต้องและกำหนดคุณสมบัติที่ต้องการให้กับข้อมูล

3) เตรียมข้อมูล โดยการคัดเลือกข้อมูลเพื่อทำการแปลให้อยู่ในรูปแบบที่เหมาะสมต่อการวิเคราะห์ข้อมูลด้วยเทคนิคต่างๆ

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

2.2 วิธีสมาชิกที่ใกล้ที่สุด

ปัจจุบันมีนักวิจัยหลายกลุ่มได้พยายามจัดการข้อมูลที่มีค่าสูญหายด้วยวิธีการต่างๆ ซึ่งวิธีสมาชิกที่ใกล้ที่สุด [1] เป็นวิธีการหนึ่งที่ถูกนำมาใช้หาความสัมพันธ์ระหว่างกลุ่มข้อมูลที่จะนำมาประมาณค่าที่สูญหาย [6] ใช้งานง่าย ไม่ซับซ้อน และนิยมใช้ในการแทนค่าข้อมูลที่ขาดหาย (Impute Missing Values) ขั้นตอนของ KNN Imputation มีดังนี้

ขั้นที่ 1 กำหนดค่า k เพื่อใช้ในการพิจารณาจำนวนสมาชิกที่ใกล้ที่สุด

ขั้นที่ 2 คำนวณหาระยะห่างระหว่างจุดด้วยวิธี Euclidian Distance ระหว่างข้อมูลกลุ่มทดสอบที่ต้องการพิจารณากับข้อมูลกลุ่มทดลอง ดังสมการที่ (1)

$$dist(X_i, X_j) = \sqrt{\sum_{k=1}^{n} (X_{i,j} - X_{j,k})^2} \quad ...(1)$$

โดยที่

$dist(X_i, X_j)$ คือ ระยะห่างระหว่างข้อมูลทดสอบแถวที่ i กับข้อมูลทดลองแถวที่ j

n คือ จำนวนข้อมูลทั้งหมดของตัวอย่าง

$X_{i,k}$ คือ ค่าแถวที่ i คอลัมน์ที่ k ของข้อมูลกลุ่มทดสอบ

$X_{j,k}$ คือ ค่าแถวที่ j คอลัมน์ที่ k ของข้อมูลกลุ่มทดลอง

ขั้นที่ 3 เรียงลำดับระยะห่างระหว่างจุดโดยพิจารณาจากข้อมูลที่ใกล้ที่สุดตามด้วยจำนวน k

ขั้นที่ 4 ประมาณค่าข้อมูลสูญหายจากค่าเฉลี่ยของข้อมูลที่อยู่ใกล้ที่สุด ดังสมการที่ (2)

$$X'_i = \frac{\sum_{i=1}^{k} X_i}{k} \quad ...(2)$$

โดยที่

X'_i คือ ค่าข้อมูลสูญหายที่ได้จากการประมาณ ค่าใหม่

k คือ จำนวนที่กำหนดไว้เพื่อพิจารณาสมาชิกที่อยู่ใกล้ที่สุด

X_i คือ ค่าข้อมูลกลุ่มทดลองที่ตรงกับข้อมูลสูญหายในข้อมูลกลุ่มทดสอบ

2.3 วิธีสมการถดถอย (Regression)

เป็นการศึกษาตัวแปรอิสระมีอิทธิพลอย่างไรต่อตัว แปรตามหรือตัวแปรอิสระที่มีผลทำให้ค่า Y ผันแปรไปในรูปแบบใดสามารถอธิบายลักษณะความสัมพันธ์ด้วยรูปแบบการถดถอย (Regression Model) [5] ดังสมการที่ (3)

$$Y = a + bX \quad ...(3)$$

โดยที่

Y คือ ตัวแปรตาม

X คือ ตัวแปรอิสระหรือตัวแปรต้น

a คือ ค่าคงที่

b คือ ความชัน

2.4 ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Root mean square error, RMSE)

ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) เป็นวิธีการวัดความคลาดเคลื่อนจากค่าที่พยากรณ์จากแบบจำลองกับค่าจริงที่เกิดขึ้นหากค่า RMSE มีค่าน้อย แสดงว่าแบบจำลองสามารถประมาณค่าได้ใกล้เคียงกับค่าจริง

ดังนั้นหากค่านี้มีค่าเท่ากับศูนย์แสดงว่าไม่เกิดความคลาดเคลื่อนในแบบจำลองนี้ สามารถคำนวณได้จากสมการที่ (4)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(predict - real)^2} \quad ...(4)$$

โดยที่

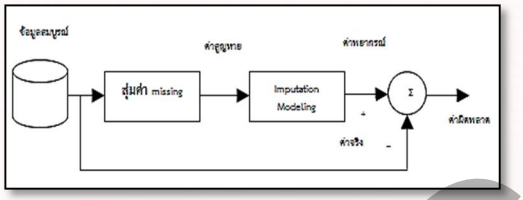
predi คือค่าประมาณจากแบบจำลองค่าข้อมูลที่ได้จากการพยากรณ์

real คือค่าจริงที่ได้จากข้อมูลจริง

n คือจำนวนของตัวอย่างที่ใช้ในการประมาณแบบจำลอง

3. วิธีการดำเนินการวิจัย

การทดลองครั้งนี้ นำข้อมูลระดับน้ำของพื้นที่ลุ่มน้ำปากพนัง ตั้งแต่ปี พ.ศ. 2554-2556 ซึ่งเป็นข้อมูลประเภท Numerical ซึ่งมีขั้นตอนดังรูปที่ 1



รูปที่ 1 ขั้นตอนการทดลองการประมาณค่าสูญหายที่ 1%, 2%, 3%, 4% และ5%ด้วยวิธี K-NN และวิธี Regression

ขั้นตอนการทดลองประมาณค่าสูญหาย ด้วยวิธี K-NN และวิธี Regression ดังนี้

ขั้นที่ 1 นำข้อมูลระดับน้ำที่มีค่าสมบูรณ์ ใช้เป็นข้อมูลนำเข้า

ขั้นที่ 2 สุ่มสร้าง Missing Value ที่มีความแตกต่างกัน คือ 1%, 2%, 3%, 4% และ5% เพื่อนำเข้าสู่โมเดลการแทนค่าข้อมูล

ขั้นที่ 3 แทนค่าข้อมูลที่สูญหายด้วยวิธี K-NN และวิธี Regression

จากนั้นนำคำตอบที่ได้ไปหาประสิทธิภาพ ด้วยค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Root mean square error, RMSE) ดังสมการที่ (4) ซึ่งเป็นวิธีการวัดความคลาดเคลื่อนจากค่าที่พยากรณ์กับค่าจริงที่เกิดขึ้น เพื่อใช้ในการพยากรณ์ระดับน้ำต่อไป

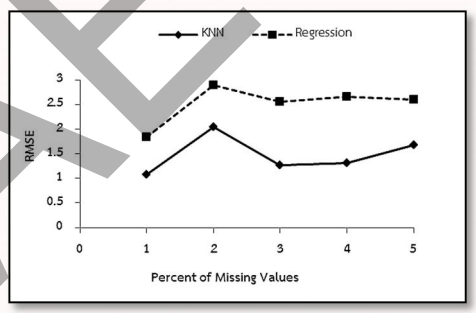
4. ผลการทดลอง

ในการทดลองครั้งนี้ผู้วิจัยได้ทำการศึกษาข้อมูลระดับน้ำลุ่มน้ำปากพนังซึ่งเป็นข้อมูลที่สมบูรณ์ไม่มีข้อมูลสูญหาย ของปี พ.ศ. 2554-พ.ศ. 2556 ซึ่งประกอบด้วยค่าระดับน้ำ จำนวนประตูระบายน้ำ อัตราการระบายน้ำ ปริมาณน้ำเก็บกักและปริมาณน้ำฝน เป็นต้น

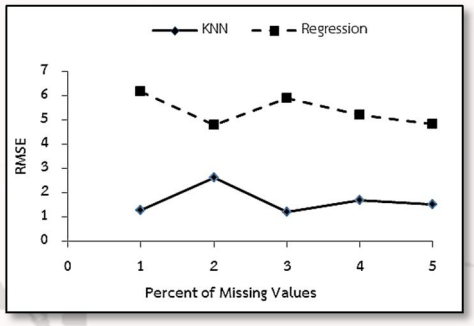
ผลการทดลองการแทนค่าข้อมูลสูญหายด้วยวิธี KNN และวิธี Regression โดยทำการสุ่มสร้าง Missing Value ตั้งแต่ 1%-5% ดังแสดงในตารางที่ 1 และรูปที่ 2-4

ตารางที่ 1 ค่า RMSE ของการพยากรณ์ค่าที่ขาดหายด้วยวิธี KNN และRegression สำหรับข้อมูลระดับน้ำลุ่มน้ำปากพนัง ปี พ.ศ. 2554-พ.ศ. 2556

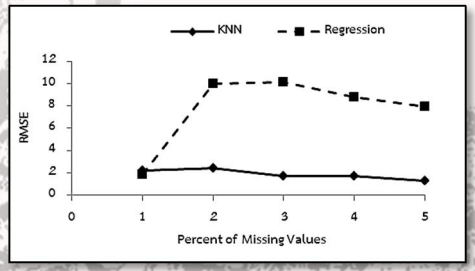
MISS %	ปี 2554		ปี 2555		ปี 2556	
	KNN	Regression	KNN	Regression	KNN	Regression
1%	1.08	1.84	1.27	6.19	2.21	1.86
2%	2.04	2.89	2.46	4.80	2.39	9.97
3%	1.27	2.56	1.20	5.89	1.73	10.11
4%	1.32	2.66	1.70	5.20	1.73	8.80
5%	1.68	2.61	1.51	4.82	1.27	7.90



รูปที่ 2 ค่า RMSE ของการพยากรณ์ค่าระดับน้ำที่สูญหายด้วยวิธี KNNและวิธี Regression ของพ.ศ. 2554



รูปที่ 3 ค่า RMSE ของการพยากรณ์ค่าระดับน้ำที่สูญหายด้วยวิธี KNN และวิธี Regression ของพ.ศ. 2555



รูปที่ 4 ค่า RMSE ของการพยากรณ์ค่าระดับน้ำที่สูญหายด้วยวิธี KNN และวิธี Regression ของพ.ศ. 2556

จากผลการทดลองในตารางที่ 1 เมื่อเปรียบเทียบวิธี KNN และวิธี Regression ดังรูปที่ 2-4 พบว่า มีแนวโน้มของค่าข้อมูลที่สูญหายไปในทิศทางเดียวกัน คือ ค่า RMSE ที่ 1%, 2%, 3%, 4% และ 5% วิธี KNN ต่ำกว่า วิธี Regression

5. สรุปผลการทดลอง

จากการทดลอง การแทนค่าข้อมูลสูญหายด้วยวิธีการทางเหมืองข้อมูล เพื่อพยากรณ์ระดับน้ำ โดยมีการสุ่มสร้าง Missing Value ตั้งแต่ 1% - 5% ด้วยวิธีสมาชิกที่ใกล้ที่สุด และวิธีสมการถดถอย ซึ่งนำมาประยุกต์ใช้กับข้อมูลระดับน้ำของพื้นที่ลุ่มน้ำปากพนัง จังหวัดนครศรีธรรมราช เพื่อนำเสนอรูปแบบการแทนที่ค่าสูญหายแบบหลายตัวแปรสำหรับการพยากรณ์ระดับน้ำ พบว่า วิธีสมาชิกที่ใกล้ที่สุด (K-Nearest Neighbor : KNN) มีประสิทธิภาพในการประมาณค่าสูญหายที่ดีกว่า วิธีสมการถดถอย (Regression) เมื่อประเมินค่าความถูกต้องกับทั้งสองวิธี โดยค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Square Error, RMSE) มีค่าต่ำกว่า

6. เอกสารอ้างอิง

[1] James E, S. Machleod, Andrew luk and D. Michael Titerington. (1987). A Re-Examination of the Distance Weighted k-Nearest Neighbor Classification Rule. IEEE Trans Man Cybernetics.

[2] Troyanskaya O., Cantor M., Sherlock G., Brown P., Hastie T., Tibshirani R., Botstein D., and Altman R.B. “Missing values estimation methods for DNA microarrays,” Bioinformatics, (2001). vol. 17, pp. 520-525.

[3] Oba S., Sato M.A., Takemasa I., Monden M., Matsubara K.I. and Ishii S.A. “A Bayesian missing value estimation method for gene expression profile data,” Bioinformatics, (2003). vol. 19, pp. 2088-2096.

[4] Witten, I.H., E. Frank, and M.A. Hall. (2011). Data Mining: Practical Machine Learning Tools and Techniques: Elsevier Science.

[5] สักการะ จู่ซู่ และธนพล เจนสุทธิเวชกุล (2010). “เปรียบเทียบการ พยากรณ์ปริมาณการยืมหนังสือห้องสมุดโรงเรียนนาบอน ด้วยการวิเคราะห์การถดถอย และโครงข่ายประสาทเทียม” in The 6th NCCIT. : 110-115

[6] อุมาภรณ์ สายแสงจันทร์ (2548). การประมาณค่าสูญหายของข้อมูลที่สูญหายโดยตัวแบบฟัซซี่เนียเรส. วิทยานิพนธ์สาขาวิทยาการคอมพิวเตอร์. มหาวิทยาลัยขอนแก่น.