การเพิ่มประสิทธิภาพซัพพอร์ตเวกเตอร์รีเกรสชันในการพยากรณ์อนุกรมเวลา

The Improvement of Support Vector Regression to Forecast Time Series

ธีร์ธวัช แก้ววิจิตร[1], นิตยา เกิดประสพ[2], กิตติศักดิ์ เกิดประสพ[2]

Teetawat Kaewwijit, Nittaya Kerdprasop, Kittisak Kerdprasop

Received: 7 November 2016 ; Accepted: 27 March 2017

บทคัดย่อ

ในปัจจุบันได้มีความพยายามในการหาเทคนิคใหม่ในการพยากรณ์ เพื่อให้การพยากรณ์มีความแม่นยำและความเร็วเพิ่มขึ้น โดยการคิดค้นเทคนิคใหม่ หรือการนำหลาย ๆ เทคนิคมาผสมกัน งานวิจัยนี้มีจุดประสงค์เพื่อเพิ่มประสิทธิภาพให้กับเทคนิคซัพพอร์ตเวกเตอร์รีเกรสชันในการพยากรณ์อนุกรมเวลา โดยใช้ค่าความคลาดเคลื่อนมาช่วยในการเพิ่มความแม่นยำให้กับตัวแบบ ซึ่งข้อมูลที่ใช้ในการวิเคราะห์เป็นข้อมูลอนุกรมเวลาทั้งหมด 5 ชุดข้อมูลประกอบไปด้วยชุดข้อมูลอุณหภูมิรายวันของแม่น้ำฟิชเชอร์ ข้อมูลปริมาณการผลิตน้ำนมของวัวในแต่ละเดือน ข้อมูลค่าความดันที่ระดับน้ำทะเลที่เมืองดาร์วิน ข้อมูลปริมาณคาร์บอนไดออกไซด์ที่ภูเขาไฟเมานาโลอา และข้อมูลค่าดัชนีที่คำนวณจากค่าความกดอากาศที่แตกต่างกันระหว่างจุด 2 จุดในตาฮีติและดาร์วิน โดยการเปรียบเทียบความแม่นยำของเทคนิคใหม่จะทำการเปรียบเทียบกับเทคนิค 2 แบบคือเทคนิคซัพพอร์ตเวกเตอร์รีเกรสชันดั้งเดิม และเทคนิค ARIMA และวัดค่าโดยใช้ค่ารากที่สองของค่าความคาดเคลื่อนกำลังสองเฉลี่ย และค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย ซึ่งจากผลการเปรียบเทียบพบว่าเทคนิคใหม่สามารถเพิ่มความแม่นยำให้กับเทคนิคซัพพอร์ตเวกเตอร์รีเกรสชันได้

คำสำคัญ : ซัพพอร์ตเวกเตอร์รีเกรสชัน ข้อมูลอนุกรมเวลา ค่าความคลาดเคลื่อน

Abstract

Currently, there are efforts to find new techniques in forecasting in order to improve precision and speed. The improvement is achieved by using new technique or a combination of techniques. This research aims to optimize support vector regression in forecasting time series by using the error to increase the accuracy of the model. The five datasets used in time series analysis are the daily temperature of the Fisher River, monthly milk production, the sea level pressure data at Darwin, carbon dioxide concentration at Mauna Loa mountain, and the atmospheric pressure difference between Tahiti and Darwin. The precision of the proposed model is compared against the traditional support vector regression and the ARIMA models using the root mean squared error and the mean absolute error metrics. From the experimental results, the proposed method can improve precision of the support vector regression technique.

Keywords : Support vector regression, Time Series, Error

บทนำ

การทำนาย ตามพจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. 2554 หมายถึง “การบอกเหตุการณ์หรือความเป็นไปที่จะเกิดในเบื้องหน้า” ซึ่งการทำนายนั้นได้มีมาตั้งแต่สมัยอดีต โดยการทำนายจะถูกใช้โดยผู้ทำพิธี ซึ่งอาจจะเป็นหมอดู พระ นักบวช คนทรง เป็นต้น โดยสิ่งที่ทำนายนั้นมักจะเป็นเรื่องเกี่ยวกับชีวิต ความรัก ความเป็นอยู่ ผลลัพธ์การพนันต่าง ๆ เป็นต้น โดยผลลัพธ์ในการทำนายนั้นอาจจะถูกต้องหรือไม่ถูกต้องก็ได้ เมื่อเวลาผ่านไปการทำนายได้มีการพัฒนาขึ้นโดยอาศัยหลักการต่าง ๆ เข้ามาช่วยเพื่อเพิ่มความแม่นยำในการทำนายเรียก

[1] นักศึกษาปริญญาโท, [2]รองศาสตราจารย์, สาขาวิชาวิศวกรรมคอมพิวเตอร์, มหาวิทยาลัยเทคโนโลยีสุรนารี, อ.เมือง, จ.นครราชสีมา

[1] Master degree student, [2]Associate Professor, Computer Engineering, Suranaree University of Technology, Muang District, Nakhon Ratchasima

** Corresponding author : starsskillers@gmail.com*

ว่าการพยากรณ์ โดยตามพจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. 2554 หมายถึง "ทำนายหรือคาดการณ์โดยอาศัยหลักวิชา" ซึ่งหลักวิชาที่ว่านี้คือเทคนิคต่าง ๆ ที่มาช่วยในการพยากรณ์ โดยทางสถิติและคอมพิวเตอร์คือการนำข้อมูลต่าง ๆ ที่มีความเกี่ยวข้องกับสิ่งที่ต้องการพยากรณ์มาช่วยในการหาค่าพยากรณ์ ซึ่งค่าที่นิยมใช้ในการพยากรณ์จะเป็นค่าข้อมูลที่ต้องการพยากรณ์ย้อนหลัง โดยลักษณะของข้อมูลที่ใช้จะอยู่ในรูปแบบของอนุกรมเวลา (Time series) และมีการใช้เทคนิคต่าง ๆ ในการพยากรณ์ โดยเทคนิคที่นิยมใช้คือ เทคนิคสมการเชิงเส้นทั่วไป (Generalized Linear) เทคนิค ARIMA เทคนิคการถดถอย (Regression) เทคนิคซัพพอร์ตเวกเตอร์รีเกรสชัน (Support Vector Regression) เทคนิคโครงข่ายประสาทเทียม (Artificial Neural Network) ซึ่งเทคนิคที่ได้กล่าวมานั้นเป็นเทคนิคที่ให้ผลลัพธ์ที่มีความแม่นยำสูง แต่ก็ยังมีความต้องการที่จะให้การพยากรณ์ที่ได้มีความแม่นยำที่สูงขึ้น ซึ่งในปัจจุบันได้มีนักวิจัยพัฒนาเทคนิคใหม่ ๆ เพื่อให้ได้ผลลัพธ์ที่ดีกว่าวิธีเดิม โดยมีการนำเทคนิคหลาย ๆ วิธีมาใช้ร่วมกัน เช่นการนำเทคนิควิธีเชิงพันธุกรรม (Genetic Algorithm) มาช่วยในการหาค่าพารามิเตอร์ในตัวแบบ (Model) เพื่อให้ได้ตัวแบบที่ดีที่สุด เป็นต้น เทคนิคหนึ่งที่น่าสนใจ และนำมาใช้ในการเพิ่มความแม่นยำคือการนำค่าความคลาดเคลื่อนที่ได้จากตัวแบบมาพิจารณาในการสร้างตัวแบบใหม่ โดยได้มีงานวิจัยที่ใช้เทคนิค ARIMA ในการสร้างตัวแบบ และนำเทคนิคเครือข่ายประสาทเทียมมาสร้างตัวแบบจากค่าความคลาดเคลื่อนที่ได้จากเทคนิค ARIMA จากนั้นนำตัวแบบทั้ง 2 มาสร้างเป็นตัวแบบใหม่ ซึ่งแนวคิดของงานวิจัยนี้คือตัวแบบนั้นประกอบด้วยส่วนที่เป็นเชิงเส้นและส่วนที่ไม่เป็นเชิงเส้น โดยใช้เทคนิค ARIMA ในการพยากรณ์ส่วนที่เป็นเชิงเส้น และเทคนิคโครงข่ายประสาทเทียมใช้ในการพยากรณ์ส่วนที่ไม่เป็นเชิงเส้น โดยตั้งชื่อเทคนิคว่าไฮบริด[1] (Hybrid) ซึ่งเทคนิคนี้ได้มีการพัฒนาต่อโดยการเปลี่ยนแปลงเทคนิคที่ใช้ เช่น Chen และ Wang[2] เสนอเทคนิคที่ดัดแปลงมาจากเทคนิคไฮบริด ซึ่งเป็นการรวมเทคนิคของ ARIMA และเทคนิคซัพพอร์ตเวกเตอร์รีเกรสชัน โดยกำหนดให้ตัวแบบแรกใช้เทคนิค ARIMA จากนั้นนำค่าความคลาดเคลื่อนของการทำนายมาใช้ในตัวแบบซัพพอร์ตเวกเตอร์รีเกรสชันเพื่อเพิ่มประสิทธิภาพในการพยากรณ์ขั้นสุดท้าย นอกจากนี้ยังได้มีการใช้เทคนิควิธีเชิงพันธุกรรมช่วยในการค้นหาค่าพารามิเตอร์ของซัพพอร์ตเวกเตอร์รีเกรสชัน Wang และคณะ[3] เสนอเทคนิคที่ดัดแปลงมาจากเทคนิคไฮบริดของ Zhang[1] โดยการใช้เทคนิค ARIMA ขั้นตอนเริ่มต้นจากการพยากรณ์โดยใช้เทคนิค ARIMA ก่อน แล้วจึงนำค่าความคลาดเคลื่อนที่ได้มาทำการสร้างตัวแบบ ARIMA อีกครั้ง จากนั้นนำตัวแบบทั้งสองตัวไปใช้ในการพยากรณ์ Khandelwal และคณะ[4] ได้มีการนำวิธีไฮบริดมาใช้ในการพยากรณ์ร่วมกับวิธีการแปลงข้อมูล Discrete Wavelet Transform (DWT) โดยการใช้ DWT มีวัตถุประสงค์เพื่อแยกข้อมูลออกเป็นข้อมูลความถี่สูงและความถี่ต่ำ จากนั้นนำวิธีการไฮบริดมาใช้ในการพยากรณ์ Oliveira และ Ludermir[5] ได้เสนอวิธีการเพิ่มเติมโดยการนำขั้นตอนวิธีหาค่าเหมาะสมที่สุดแบบกลุ่มอนุภาค (Particle Swarm Optimization : PSO) มาช่วยในการหาผลลัพธ์ที่ดีที่สุดและนำเทคนิคไฮบริดมาใช้ในการพยากรณ์ จากที่กล่าวมาข้างต้นพบว่าปัจจุบันได้มีความพยายามในการหาเทคนิคใหม่ ๆ เพื่อนำมาใช้ในการเพิ่มความแม่นยำในการพยากรณ์ โดยในงานวิจัยนี้ผู้วิจัยได้นำเสนอเทคนิคใหม่ โดยนำเทคนิคไฮบริดมาดัดแปลงใหม่ โดยมีสมมติฐานที่ว่าค่าความคลาดเคลื่อนที่ได้นั้นเป็นค่าความสัมพันธ์ที่ขาดหายไปในการพยากรณ์ครั้งแรก ดังนั้นเพื่อที่จะทำให้การพยากรณ์มีความแม่นยำมากขึ้นจึงได้ทำการนำค่าความคลาดเคลื่อนที่ได้มาทำการพยากรณ์อีกครั้งเพื่อเติมความสัมพันธ์ที่ขาดหายไป ซึ่งเทคนิคที่ใช้ในการทดลองคือเทคนิคซัพพอร์ตเวกเตอร์รีเกรสชัน และได้มีการเปรียบเทียบความแม่นยำกับเทคนิคซัพพอร์ตเวกเตอร์รีเกรสชันดั้งเดิม และเทคนิค ARIMA โดยเกณฑ์ที่ใช้ในการทดสอบความแม่นยำ คือ ค่ารากที่สองของค่าความคาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Squared Error : RMSE) และค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Error : MAE)

วัตถุประสงค์ของงานวิจัย

1. เพื่อศึกษาวิธีการเพิ่มความแม่นยำให้กับเทคนิคซัพพอร์ตเวกเตอร์รีเกรสชันดั้งเดิม โดยการนำค่าความคลาดเคลื่อนมาช่วยในการสร้างตัวแบบ
2. เพื่อเปรียบเทียบความแม่นยำในการพยากรณ์ของเทคนิคที่นำเสนอกับเทคนิคซัพพอร์ตเวกเตอร์รีเกรสชันดั้งเดิม และเทคนิค ARIMA ด้วยค่ารากที่สองของค่าความคาดเคลื่อนกำลังสองเฉลี่ย และค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย

วิธีดำเนินงานวิจัย

งานวิจัยนี้เป็นการทดลองเพื่อเพิ่มค่าความแม่นยำในการพยากรณ์ค่าให้กับตัวแบบโดยใช้ค่าความคาดเคลื่อนมาช่วยในการสร้างตัวแบบ ซึ่งตัวแบบที่ใช้ในการพยากรณ์สร้างโดยวิธีการซัพพอร์ตเวกเตอร์รีเกรสชัน ซึ่งมีวิธีขั้นตอนการวิจัยดังนี้

1. ซัพพอร์ตเวกเตอร์รีเกรสชัน

ซัพพอร์ตเวกเตอร์รีเกรสชัน[6] (Support Vector Regression : SVR) เป็นวิธีการหนึ่งที่ใช้พยากรณ์ค่าที่ได้รับ

ความนิยมอย่างมาก และมีการใช้อย่างแพร่หลาย เนื่องจากค่าที่พยากรณ์นั้นมีความแม่นยำสูง เป็นการดัดแปลงมาจากวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยมีสมการดังนี้

$$f(\boldsymbol{x}) =< \mathbf{w}, \boldsymbol{x} > + \boldsymbol{b} \quad (1)$$

โดยที่ w $\in \boldsymbol{X}$, b $\in \mathbb{R}$, < , > คือการกระทำดอทโปรดักระหว่าง w และ x โดย w คือค่าน้ำหนักของซัพพอร์ตเวกเตอร์ และ b คือค่าคงที่ ข้อมูลที่ใช้ในการพยากรณ์จะอยู่ในรูปแบบของ $\{(\boldsymbol{x_1}, \boldsymbol{y_1}), \ldots (\boldsymbol{x_l}, \boldsymbol{y_l})\} \subset \boldsymbol{X}$ x $\mathbb{R}$ โดยที่ $\boldsymbol{X}$ คือขนาดชนิดของข้อมูลนำเข้าและ $\mathbb{R}$ คือจำนวนจริง โดยเป้าหมายที่ต้องการคือหาค่าฟังก์ชัน $f(\boldsymbol{x})$ โดยสามารถหาค่าได้โดยการใช้สมการของลากรานจ์ (Lagrange function) โดยจะมีการเพิ่มตัวแปรในสมการซึ่งเรียกว่า ตัวคูณลากรานจ์ (Lagrange multipliers) ซึ่งผลลัพธ์ที่ได้มีสมการดังสมการที่ 2

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{l}} (\boldsymbol{\alpha_i} - \boldsymbol{\alpha_i^*}) < \boldsymbol{x_i}, \boldsymbol{x} > + \boldsymbol{b} \quad (2)$$

โดยที่ $\boldsymbol{\alpha_i}$ และ $\boldsymbol{\alpha_i^*}$ คือตัวคูณลากรานจ์แต่ในกรณีที่ไม่สามารถทำการพยากรณ์ข้อมูลได้ใน 2 มิติจะมีการนำเคอร์เนลเข้ามาช่วย โดยมีสมการดังสมการที่ 3

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{l}} (\boldsymbol{\alpha_i} - \boldsymbol{\alpha_i^*}) \mathbf{k}(\mathbf{x_i}, \mathbf{x}) + \boldsymbol{b} \quad (3)$$

โดยที่ $\mathbf{k}(\mathbf{x_i}, \mathbf{x})$ คือค่าเคอร์เนล

2. ARIMA

ARIMA (Autoregressive Integrated Moving Average) เป็นเทคนิคการพยากรณ์ที่เสนอโดยบ๊อกซ์และเจนกินส์ [7] ในปี 1976 โดยเทคนิคแรกที่ถูกเสนอเรียกว่า ARMA โดยจะแบ่งออกเป็น 2 ส่วนคือ Autoregressive (AR) และ Moving average (MA) โดยข้อมูลที่จะนำมาสร้างตัวแบบ ARMA จำเป็นต้องมีลักษณะนิ่ง (Stationary) ต่อมาจึงได้มีการพัฒนาเทคนิคARIMA ขึ้นมาโดยพัฒนาให้เทคนิค ARMA สามารถทำงานได้กับข้อมูลที่มีลักษณะไม่นิ่งได้ ซึ่งเทคนิค ARIMA เป็นเทคนิคที่ได้รับความนิยมสูงเนื่องจากให้ผลการพยากรณ์ที่มีความแม่นยำ

2.1 Autoregressive (AR) เป็นรูปแบบที่กำหนดว่าค่าพยากรณ์ที่เวลาใด ๆ ขึ้นอยู่กับค่าสังเกตก่อนหน้า โดยกำหนดว่าที่ค่าจริง $\boldsymbol{Y_t}$ ที่เวลา t ใด ๆ จะขึ้นกับค่าจริงที่เวลา t-1, t-2, t-3, ..., t-p โดยที่ค่า p คือค่าคาบเวลาที่ล้าหลัง ซึ่งสามารถแสดงเป็นสมการได้ดังสมการที่ 4

$$\boldsymbol{Y_t} = \boldsymbol{\theta_0} + \boldsymbol{\emptyset_1 Y_{t-1}} + \boldsymbol{\emptyset_2 Y_{t-2}} + \cdots + \boldsymbol{\emptyset_p Y_{t-p}} + \boldsymbol{\varepsilon_t} \quad (4)$$

เมื่อ $\boldsymbol{Y_t}$ *คือ ค่าจริงที่เวลา* t ใด ๆ
$\boldsymbol{\theta_0}$ คือ ค่าคงที่
$\boldsymbol{\emptyset_1, \emptyset_2, \ldots, \emptyset_p}$ คือ พารามิเตอร์ถ่วงน้ำหนัก
$\boldsymbol{\varepsilon_t}$ คือ ค่าคลาดเคลื่อนที่เวลา t

2.2 Moving average (MA)

เป็นรูปแบบที่กำหนดว่าค่าพยากรณ์ที่เวลาใด ๆ ขึ้นอยู่กับค่าคลาดเคลื่อนก่อนหน้า โดยกำหนดว่าที่ค่าสังเกต $\boldsymbol{Y_t}$ ที่เวลา t ใด ๆ จะขึ้นกับค่าคลาดเคลื่อนที่เวลา t-1, t-2, t-3, ..., t-q โดยที่ค่า q คือค่าคาบเวลาที่ล้าหลัง ซึ่งสามารถแสดงเป็นสมการได้ดังสมการที่ 5

$$\boldsymbol{Y_t} = \boldsymbol{\theta_0} + \boldsymbol{\varepsilon_t} - \boldsymbol{\theta_1 \varepsilon_{t-1}} - \boldsymbol{\theta_2 \varepsilon_{t-2}} - \cdots - \boldsymbol{\theta_q \varepsilon_{t-q}} \quad (5)$$

เมื่อ $\boldsymbol{Y_t}$ *คือ ค่าจริงที่เวลา* t ใด ๆ
$\boldsymbol{\theta_0}$ คือ ค่าคงที่
$\boldsymbol{\theta_1, \theta_2, \ldots, \theta_q}$ คือ พารามิเตอร์ถ่วงน้ำหนัก
$\boldsymbol{\varepsilon_t}$ คือ ค่าคลาดเคลื่อนที่เวลา t

โดยเมื่อทำการรวม AR(p) และ MA(q) จะได้เทคนิคที่เรียกว่า ARMA(p,q) โดยมีสมการดังสมการที่ 6

$$\boldsymbol{Y_t} = \boldsymbol{\theta_0} + \boldsymbol{\emptyset_1 Y_{t-1}} + \boldsymbol{\emptyset_2 Y_{t-2}} + \cdots + \boldsymbol{\emptyset_p Y_{t-p}} + \boldsymbol{\varepsilon_t} - \boldsymbol{\theta_1 \varepsilon_{t-1}} - \boldsymbol{\theta_2 \varepsilon_{t-2}} - \cdots - \boldsymbol{\theta_q \varepsilon_{t-q}} \quad (6)$$

2.3 ตัวแบบ ARIMA

ตัวแบบ ARIMA หรือ ARIMA(p,d,q) เป็นเทคนิคที่พัฒนามาจาก ARMA โดยทำให้สามารถใช้กับข้อมูลอนุกรมเวลาที่มีลักษณะนิ่งได้โดยทำการแปลงข้อมูลอนุกรมเวลาเดิม $\boldsymbol{Y_t}$ ซึ่งเป็นข้อมูลที่ไม่นิ่งให้เป็นอนุกรมเวลาค่าใหม่ $\boldsymbol{Z_t}$ ซึ่งเป็นข้อมูลที่นิ่ง โดยใช้การหาค่าผลต่างระหว่างค่าสังเกตในอนุกรมเวลาเดิม ดังสมการที่ 7

$$\boldsymbol{Z_t} = \boldsymbol{\nabla^d Y_t} \quad (7)$$

ถ้าให้ d = 1 จะได้ $\boldsymbol{Z_t = \nabla^1 Y_t = Y_t - Y_{t-1}}$
ถ้าให้ d = 2 จะได้
$\boldsymbol{Z_t = \nabla^2 Y_t = \nabla^1 Y_t - \nabla^1 Y_{t-1} = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2})}$
$\boldsymbol{= Y_t - 2Y_{t-1} + Y_{t-2}}$

3. เกณฑ์ที่ใช้ในการวัดความแม่นยำของตัวแบบ

3.1 ค่ารากที่สองของค่าความคาดเคลื่อนกำลังสองเฉลี่ย

ค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Squared Error : RMSE) เป็นวิธีที่การวัดค่าความคลาดเคลื่อนแบบมาตรฐาน ซึ่งนิยมใช้กันอย่างแพร่หลาย โดยมีสมการดังสมการที่ 8 ในการวัดค่าความแม่นยำจากวิธีการนี้ยิ่งค่าที่ได้มีค่าน้อยแสดงว่าตัวแบบที่ได้มีความแม่นยำในการทำนายสูง

$$\boldsymbol{RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2}} \quad (8)$$

โดยที่ RMSE คือค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย
n คือ จำนวนข้อมูลที่ใช้
$\mathbf{Y_t}$ คือ ค่าจริงที่เวลา t ใด ๆ
$\hat{\mathbf{Y}}_\mathbf{t}$ คือ ค่าที่ได้จากการพยากรณ์ที่เวลา t ใด ๆ

3.2 ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย

ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Deviation : MAD) หรือ (Mean Absolute Error : MAE) เป็นวิธีที่การวัดค่าความคลาดเคลื่อนที่นิยมอีกวิธีหนึ่ง ซึ่งวิธีนี้จะช่วยบอกถึงขนาดของความคลาดเคลื่อนรวมได้ โดยมีสมการดังสมการที่ 9 ในการวัดค่าความแม่นยำจากวิธีการนี้ยิ่งค่าที่ได้มีค่าน้อยแสดงว่าตัวแบบที่ได้จะมีความแม่นยำมาก

$$\boldsymbol{MAE = \frac{1}{n}\sum_{t=1}^{n}|Y_t - \hat{Y}_t|} \quad (9)$$

โดยที่ MAE คือค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย
n คือ จำนวนข้อมูลที่ใช้
$\mathbf{Y_t}$ คือ ค่าจริงที่เวลา t ใด ๆ
$\hat{\mathbf{Y}}_\mathbf{t}$ คือ ค่าที่ได้จากการพยากรณ์ที่เวลา t ใด ๆ

4. กรอบแนวคิดของงานวิจัย

แนวคิดของงานวิจัยนี้เกิดจากความต้องการที่จะเพิ่มความแม่นยำให้กับการพยากรณ์ โดยการใช้เทคนิคซัพพอร์ตเวกเตอร์รีเกรสชัน และเพิ่มความแม่นยำโดยการใช้ค่าความคลาดเคลื่อน ซึ่งกรอบแนวคิดในการวิจัยมีขั้นตอนดังนี้

1. นำชุดข้อมูลอนุกรมเวลามาใช้ในการสร้างตัวแบบซัพพอร์ตเวกเตอร์รีเกรสชัน
2. นำตัวแบบซัพพอร์ตเวกเตอร์รีเกรสชันที่ได้มาหาค่าความคลาดเคลื่อนโดยเทียบกับชุดข้อมูลอนุกรมเวลาที่ใช้ในการสร้างตัวแบบ
3. นำค่าความคลาดเคลื่อนที่ได้มาสร้างตัวแบบซัพพอร์ตเวกเตอร์รีเกรสชัน
4. สร้างตัวแบบผสมโดยใช้ตัวแบบซัพพอร์ตเวกเตอร์ รีเกรสชันจากชุดข้อมูล และตัวแบบซัพพอร์ตเวกเตอร์รีเกรสชันจากค่าความคลาดเคลื่อน

ซึ่งสามารถแสดงกรอบแนวคิดได้ดัง Figure 1

จาก Figure 1 ได้แสดงรายละเอียดของกรอบแนวคิดและขั้นตอนในการวิจัย ซึ่งสามารถอธิบายรายละเอียดต่าง ๆ ได้ดังต่อไปนี้

4.1 การแบ่งข้อมูลฝึกสอนและทดสอบ

ผู้วิจัยได้ทำการแบ่งชุดข้อมูลอนุกรมออกเป็น 2 ส่วน โดยในส่วนแรกเป็นข้อมูลฝึกสอน โดยใช้อัตราส่วน 70% ของข้อมูลทั้งหมด และส่วนที่สองเป็นข้อมูลทดสอบ โดยใช้อัตราส่วน 30% ของข้อมูลทั้งหมด เนื่องจากเป็นอัตราส่วนที่นิยมใช้ในงานสถิติ

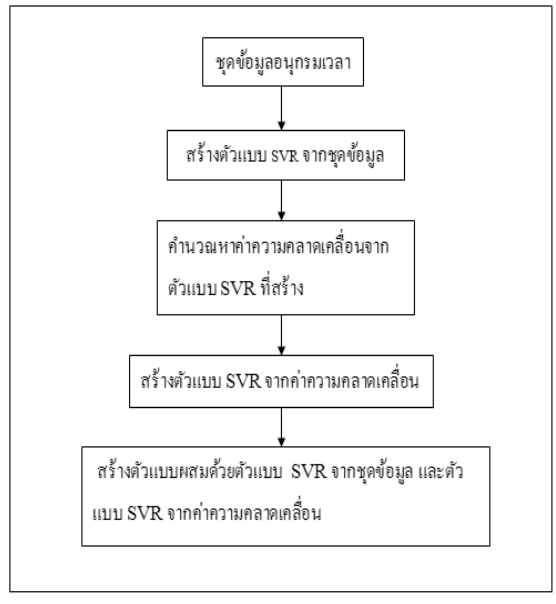

Figure 1 A framework of research methodology

4.2 การสร้างตัวแบบด้วยเทคนิคซัพพอร์ตเวกเตอร์รีเกรสชัน

ในการสร้างตัวแบบด้วยเทคนิคซัพพอร์ต

เวกเตอร์รีเกรสชันนั้นจำเป็นต้องกำหนดสมการที่ใช้ เนื่องจากข้อมูลมีลักษณะเป็นอนุกรมเวลาค่าที่ใช้ช่วยในการพยากรณ์ คือค่า วัน เดือน ปี โดยกำหนดให้มีสมการดังสมการที่ 10

$$\hat{\mathbf{Y}}_{\mathbf{t}} = f(x) =< w, x > +b \quad (10)$$

โดยที่ค่า $\hat{\mathbf{Y}}_{\mathbf{t}}$ คือ ค่าที่ได้จากการพยากรณ์

x คือ ข้อมูลนำเข้าประกอบด้วยข้อมูล วัน เดือน ปี และผลลัพธ์ ($\mathbf{Y}_{\mathbf{t}}$)

b คือ ค่าคงที่

โดยในงานวิจัยนี้ผู้วิจัยได้กำหนดให้ใช้ค่าเคอร์เนลเป็นเรเดียลเบซิสฟังก์ชัน ค่าแกมม่าเป็น 1/จำนวนมิติของข้อมูล ค่าคอสเป็น 1 ค่าเอปซิลอนเป็น 0.1 เพื่อใช้ในการสร้างตัวแบบพยากรณ์ โดยใช้ข้อมูลฝึกสอนในการสร้างตัวแบบ และเมื่อนำค่าที่ได้จากการพยากรณ์และค่าจริงจากข้อมูลฝึกสอนมาหาผลต่างจะได้ค่าความคลาดเคลื่อน ซึ่งหากค่าจริงมากกว่าค่าพยากรณ์ค่าความคลาดเคลื่อนจะมีค่าเป็นบวกและเป็นลบในกรณีตรงกันข้าม ดังสมการที่ 11

$$\mathbf{error} = |\mathbf{Y}_{\mathbf{t}} - \hat{\mathbf{Y}}_{\mathbf{t}}| \quad (11)$$

4.3 การนำค่าความคลาดเคลื่อนมาช่วยในการเพิ่มความแม่นยำ

การนำค่าความคลาดเคลื่อนมาช่วยในการเพิ่มความแม่นยำนั้นทำได้โดยการนำค่าความคลาดเคลื่อนมาสร้างตัวแบบพยากรณ์โดยค่าความคลาดเคลื่อนที่ได้ยังคงมีลักษณะเป็นอนุกรมเวลาอยู่ โดยมีสมการดังสมการที่ 12

$$\hat{\mathbf{Y}}_{\mathbf{error}} = f(x) =< w, x > +b \quad (12)$$

โดยที่ค่า $\hat{\mathbf{Y}}_{\mathbf{error}}$

คือ ค่าที่ได้จากการพยากรณ์

x คือ ข้อมูลนำเข้าประกอบด้วยข้อมูล วัน เดือน ปี และค่าความคลาดเคลื่อน

b คือ ค่าคงที่

โดยค่าอื่น ๆ ที่ใช้ในการสร้างตัวแบบเป็นค่าเดียวกับที่ใช้สร้างตัวแบบด้วยเทคนิคซัพพอร์ตเวกเตอร์รีเกรสชัน โดยผลลัพธ์ที่ได้นั้นจะเป็นตัวแบบที่ได้จากค่าความคลาดเคลื่อน ซึ่งจากสมมติฐานที่ว่าค่าความคลาดเคลื่อนคือค่าความสัมพันธ์ที่ขาดหายไป ดังนั้นเพื่อเพิ่มความแม่นยำให้ตัวแบบที่สร้างด้วยเทคนิคซัพพอร์ตเวกเตอร์ รีเกรสชันดั้งเดิมจึงได้นำตัวแบบที่ได้จากค่าความคลาดเคลื่อนมาช่วยในการพยากรณ์ ซึ่งตัวแบบใหม่ที่ได้แสดงดังสมการที่ 13 โดยเรียกตัวแบบนี้ว่าตัวแบบผสม (combine)

$$\hat{\mathbf{Y}}_{\mathbf{p}} = \hat{\mathbf{Y}}_{\mathbf{t}} + \hat{\mathbf{Y}}_{\mathbf{error}} \quad (13)$$

โดยที่ค่า $\hat{\mathbf{Y}}_{\mathbf{p}}$ คือ ค่าที่ได้จากการพยากรณ์ด้วยตัวแบบผสม

4.4 การทดสอบความแม่นยำของตัวแบบ

ในการทดสอบความแม่นยำของตัวแบบนั้นข้อมูลที่ใช้คือข้อมูลทดสอบ โดยใช้ตัวแบบผสมในการพยากรณ์ค่าซึ่งวัดค่าโดยใช้ค่ารากที่สองของค่าความคาดเคลื่อนกำลังสองเฉลี่ย และค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย โดยทำการเปรียบเทียบเทคนิคทั้งหมด 3 ชนิดคือ เทคนิค ซัพพอร์ตเวกเตอร์รีเกรสชันดั้งเดิม เทคนิคซัพพอร์ตเวกเตอร์รีเกรสชันที่เพิ่มความแม่นยำด้วยค่าความคลาดเคลื่อน และเทคนิค ARIMA

5. ข้อมูลที่ใช้

ข้อมูลที่ใช้ในการวิจัยเป็นข้อมูลที่อยู่ในรูปแบบของอนุกรมเวลา โดยเป็นข้อมูลจาก Data Market[8] (https://datamarket.com/data/) และ Duke University[9] (http://www2.stat.duke.edu/~mw/ts_data_sets.html) ซึ่งข้อมูลที่นำมาใช้มีทั้งหมด 5 ชุด ประกอบไปด้วยชุดข้อมูลอุณหภูมิรายวันของแม่น้ำฟิชเชอร์ (TEMP) ข้อมูลปริมาณการผลิตน้ำนมของวัวในแต่ละเดือน (MILK) ข้อมูลค่าความดันที่ระดับน้ำทะเลที่เมืองดาร์วิน (SLP) ข้อมูลปริมาณคาร์บอนไดออกไซด์ที่ภูเขาไฟเมานาโลอา (CO2) และข้อมูลค่าดัชนีที่คำนวณจากค่าความกดอากาศที่แตกต่างกันระหว่างจุด 2 จุดในตาฮีติและดาร์วิน (SOI)

ผลการทดลอง

จาก Table 1 และ Table 2 พบว่าเมื่อนำค่าความคลาดเคลื่อนมาช่วยในการทำนาย ทำให้โมเดลเข้าใกล้ค่าของข้อมูลฝึกสอนมากขึ้น โดยทุกข้อมูลให้ผลลัพธ์ที่ดีขึ้นทั้งหมด จากตารางที่ 3 เป็นการเปรียบเทียบโดยใช้ค่ารากที่สองของค่าความคาดเคลื่อนกำลังสองเฉลี่ย ซึ่งผลลัพธ์ที่ได้ คือ เทคนิคใหม่ให้ผลลัพธ์ดีที่สุด 3 ชุดข้อมูล เทคนิค ซัพพอร์ตเวกเตอร์รีเกรสชันให้ผลลัพธ์ดีที่สุด 1 ชุดข้อมูล และเทคนิค ARIMA ให้ผลลัพธ์ดีที่สุด 1 ชุด ซึ่งเมื่อทำการเทียบเฉพาะเทคนิคใหม่และเทคนิคซัพพอร์ตเวกเตอร์รีเกรสชัน

Table 1 A comparison of RMSE from the training dataset

data	SVR	Proposed method	% Improvement
temp	5.15720	5.11310	0.855115179
milk	22.4174	21.4962	4.109307948
SLP	1.0115	1.0099	0.158180919
co2	0.4751	0.434	8.650810356
SOI	1.5845	1.5772	0.460713159

Table 2 A comparison of MAE from the training dataset

data	SVR	Proposed method	% Improvement
temp	3.96800	3.90040	1.703629032
milk	16.9748	15.1543	10.72472135
SLP	0.7916	0.7864	0.656897423
co2	0.3925	0.3262	16.89171975
SOI	1.2274	1.2123	1.23024279

พบว่าเทคนิคใหม่ให้ผลลัพธ์ดีที่สุด 4 ชุดข้อมูล และเทคนิคซัพพอร์ตเวกเตอร์รีเกรสชันให้ผลลัพธ์ดีที่สุด 1 ชุดข้อมูล เมื่อทำการเทียบผลลัพธ์เทคนิคใหม่และเทคนิคซัพพอร์ตเวกเตอร์รีเกรสชันพบว่าชุดข้อมูลอุณหภูมิรายวันของแม่น้ำฟิชเชอร์ให้ผลดีขึ้น 1.34% ข้อมูลปริมาณการผลิตน้ำนมของวัวในแต่ละเดือนให้ผลดีขึ้น 17.15% ข้อมูล ค่าความดันที่ระดับน้ำทะเลที่เมืองดาร์วินให้ผลดีขึ้น 3.86% ข้อมูลปริมาณคาร์บอนไดออกไซค์ที่ภูเขาไฟเมานาโลอาให้ผลดีขึ้น 4.31% และ ข้อมูลค่าดัชนีที่คำนวณจากค่าความกดอากาศที่แตกต่างกันระหว่างจุด 2 จุดในตาฮีติและดาร์วินให้ผลแย่ลง 1.99% จากตารางที่ 4 เป็นการเปรียบเทียบโดยใช้ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย ซึ่งผลลัพธ์ที่ได้ คือ เทคนิคใหม่ให้ผลลัพธ์ดีที่สุด 3 ชุดข้อมูล เทคนิคซัพพอร์ตเวกเตอร์รีเกรสชันให้ผลลัพธ์ดีที่สุด 1 ชุดข้อมูล และเทคนิค ARIMA ให้ผลลัพธ์ดีที่สุด 1 ชุด ซึ่งเมื่อทำการเทียบเฉพาะเทคนิคใหม่และเทคนิคซัพพอร์ตเวกเตอร์รีเกรสชันพบว่าเทคนิคใหม่ให้ผลลัพธ์ดีที่สุด 4 ชุดข้อมูล และ

Table 3 A comparison of RMSE from the test dataset

data	SVR	Proposed method	ARIMA	% Improvement
temp	10.56466	10.42317	15.31096	1.33928
milk	60.28028	49.94115	85.42048	17.15176
SLP	1.921235	1.846996	2.548502	3.86413
co2	6.510226	6.229361	2.930608	4.31421
SOI	1.918531	1.956628	1.984341	-1.98574

Table 4 A comparison of MAE from the test dataset

data	SVR	Proposed method	ARIMA	% Improvement
temp	9.07392	8.96033	12.73791	1.25183
milk	48.35438	39.37634	66.36006	18.56717
SLP	1.57495	1.48809	2.155819	5.51510
co2	5.58294	5.27019	2.42601	5.60189
SOI	1.45858	1.49647	1.530834	-2.59773

เทคนิคซัพพอร์ตเวกเตอร์รีเกรสชันให้ผลลัพธ์ดีที่สุด 1 ชุดข้อมูล เมื่อทำการเทียบผลลัพธ์เทคนิคใหม่และเทคนิคซัพพอร์ตเวกเตอร์รีเกรสชันพบว่าชุดข้อมูลอุณหภูมิรายวันของแม่น้ำฟิชเชอร์ให้ผลดีขึ้น 1.25% ข้อมูลปริมาณการผลิตน้ำนมของวัวในแต่ละเดือนให้ผล ดีขึ้น 18.57% ข้อมูลค่าความดันที่ระดับน้ำทะเลที่เมืองดาร์วินให้ผลดีขึ้น 5.52% ข้อมูลปริมาณคาร์บอนได ออกไซค์ที่ภูเขาไฟเมานาโลอาให้ผลดีขึ้น 5.60% และ ข้อมูลค่าดัชนีที่คำนวณจากค่าความกดอากาศที่แตกต่างกันระหว่างจุด 2 จุดในตาฮีติและดาร์วินให้ผลแย่ลง 2.60%

สรุปผลการทดลอง

จากการทดลองพบว่าเทคนิคการเพิ่มประสิทธิภาพให้กับเทคนิคซัพพอร์ตเวกเตอร์รีเกรสชันดั้งเดิม โดยการนำค่าความคลาดเคลื่อนมาช่วยในการสร้างตัวแบบสามารถทำได้ เนื่องจากการนำค่าความคลาดเคลื่อนมาใช้ในการสร้างโมเดลจะทำให้โมเดลที่ได้นั้นเข้าใกล้ตัวของข้อมูลฝึกสอนมากขึ้น ซึ่งทำให้ผลลัพธ์ที่ได้เมื่อนำไปใช้กับข้อมูลทดสอบได้ผลลัพธ์ที่ดีขึ้น แต่ในกรณีของข้อมูลค่าดัชนีที่คำนวณจากค่าความกดอากาศที่แตกต่างกันระหว่างจุด 2 จุดในตาฮีติและดาร์วินที่ให้ผลลัพธ์แย่ลง อาจเนื่องมาจากเกิดการโอเวอร์ฟิตติ้ง (Overfitting) ซึ่งเป็นเหตุการณ์ที่โมเดลสามารถให้ผลลัพธ์ได้ดีมากในข้อมูลฝึกสอนแต่ให้ผลลัพธ์ที่แย่ลงกับข้อมูลทดสอบ

เอกสารอ้างอิง

1. Zhang, G. Peter. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing. 2003;50:159-175.
2. Chen, Kuan-Yu, and Cheng-Hua Wang. A hybrid SARIMA and support vector machines in forecasting the production values of the machinery industry in Taiwan. Expert Systems with Applications. 2007;32.1 :254-264.

3. Wang, Yuanyuan, et al. Application of residual modification approach in seasonal ARIMA for electricity demand forecasting: a case study of China. Energy Policy. 2012;48:284-294.
4. Khandelwal, Ina, Ratnadip Adhikari, and Ghanshyam Verma. Time Series Forecasting Using Hybrid ARIMA and ANN Models Based on DWT Decomposition. Procedia Computer Science. 2015;48:173-179.
5. de Oliveira, João FL, and Teresa B. Ludermir. A hybrid evolutionary decomposition system for time series forecasting. Neurocomputing. 2016;180:27-34.
6. Smola, Alex J., and Bernhard Schölkopf. A tutorial on support vector regression. Statistic and computing. 2004;14.3:199-222
7. BOX, George EP; JENKINS, Gwilym M. Time series analysis, control, and forecasting. San Francisco, CA: Holden Day. 1976;3226.3228:10.
8. DataMarket [Internet]. Iceland:DataMarket; 2008 [cited 2016 April 17]. Available from: https:// datamarket. com/data/
9. SOME TIME SERIES DATASETS [Internet]. America: Duke University [cited 2016 April 17]. Available from: http://www2.stat.duke.edu/~mw /ts_data_sets.html