การสร้างแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่อง เพื่อคาดการณ์ปริมาณน้ำท่า

นางสาวรติพร จันทร์กลั่น

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์
มหาวิทยาลัยเทคโนโลยีสุรนารี
ปีการศึกษา 2560

MODELING WITH MACHINE LEARNING TECHNIQUES TO PREDICT RUNOFF

Ratiporn Chanklan

A Thesis Submitted in Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy in Computer Engineering

Suranaree University of Technology

Academic Year 2017

การสร้างแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่องเพื่อคาดการณ์ปริมาณน้ำท่า

มหาวิทยาลัยเทคโนโลยีสุรนารี อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาดุษฎีบัณฑิต

คณะกรรมการสอบวิทยานิพนธ์

(รศ. ดร.กิตติศักดิ์ เกิดประสพ)
ประธานกรรมการ

(รศ. ดร.นิตยา เกิดประสพ)
กรรมการ (อาจารย์ที่ปรึกษาวิทยานิพนธ์)

(ผศ. ดร.ปรเมศวร์ ห่อแก้ว)
กรรมการ

(ผศ. ดร.ศุภกฤษฎิ์ นิวัฒนากูล)
กรรมการ

(ผศ. ดร.สายสุนีย์ จับโจร)
กรรมการ

(ศ. ดร.สันติ แม้นศิริ)
รองอธิการบดีฝ่ายวิชาการและพัฒนาความเป็นสากล

(รศ. ร.อ. ดร.กนต์ธร ชำนิประศาสน์)
คณบดีสำนักวิชาวิศวกรรมศาสตร์

รติพร จันทร์กลั่น : การสร้างแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่องเพื่อคาดการณ์ปริมาณน้ำท่า (MODELING WITH MACHINE LEARNING TECHNIQUES TO PREDICT RUNOFF) อาจารย์ที่ปรึกษา : รองศาสตราจารย์ ดร.นิตยา เกิดประสพ, 135 หน้า.

น้ำท่าคือปริมาณน้ำที่เกิดจากน้ำฝนตกลงมาในพื้นที่รับน้ำแล้วไหลลงสู่แม่น้ำ การคาดการณ์ปริมาณน้ำท่าในอนาคตเป็นสิ่งที่มีประโยชน์สามารถใช้ในการเฝ้าระวังหรือวางแผนรับมือการจัดการกับการขาดแคลนน้ำหรือการเกิดอุทกภัยได้ กระบวนการเกิดน้ำท่าเป็นกระบวนการที่ซับซ้อนดังนั้นการคาดการณ์ปริมาณน้ำท่าที่มีประสิทธิภาพจึงจำเป็นต้องใช้วิธีการที่เหมาะสม งานวิจัยนี้เสนอวิธีการ ANN-GS เพื่อใช้ในการคาดการณ์ปริมาณน้ำท่ารายเดือน วิธีการ ANN-GS เป็นการผสมผสานการคาดการณ์ปริมาณน้ำท่าโดยใช้วิธีการเรียนรู้ของเครื่อง 2 อัลกอริทึมได้แก่ โมเดลเชิงเส้นโดยนัยทั่วไป และซัพพอร์ตเวกเตอร์รีเกรสชัน โดยใช้ข้อมูลในการเรียนรู้ของเครื่องได้แก่ ปริมาณน้ำท่า ปริมาณน้ำฝน จำนวนวันที่ฝนตก ตัวเลขของเดือน และดัชนีผลต่างพืชพรรณซึ่งเป็นข้อมูลจากดาวเทียม NOAA-STAR งานวิจัยนี้ทดสอบประสิทธิภาพของอัลกอริทึมที่พัฒนาขึ้นด้วยการเปรียบเทียบกับอัลกอริทึมแบบอื่น ได้แก่ วิธีการการวิเคราะห์การถดถอยเชิงเส้น โครงข่ายประสาทเทียม อาริม่า โมเดลเชิงเส้นโดยนัยทั่วไป และซัพพอร์ตเวกเตอร์รีเกรสชัน ซึ่งเกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพของงานวิจัยนี้จะใช้ค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ยเพื่อแสดงความผิดพลาดของการคาดการณ์น้ำท่าในแต่ละอัลกอริทึม และใช้ค่าสหสัมพันธ์เป็นค่าที่ใช้แสดงความสัมพันธ์ระหว่างค่าที่คาดการณ์จากอัลกอริทึมกับค่าจริง ซึ่งผลการวิจัยพบว่าวิธีการที่นำเสนอมีประสิทธิภาพในการคาดการณ์น้ำท่าได้มีประสิทธิภาพดีกว่าวิธีการอื่นที่นำมาเปรียบเทียบ

สาขาวิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนักศึกษา รติพร จันทร์กลั่น

ปีการศึกษา 2560 ลายมือชื่ออาจารย์ที่ปรึกษา

RATIPORN CHANKLAN : MODELING WITH MACHINE LEARNING TECHNIQUES TO PREDICT RUNOFF. THESIS ADVISOR : ASSOC. PROF. NITTAYA KERDPRASOP, Ph.D., 135 PP.

RUNOFF PREDICTION/ARTIFICIAL NEURAL NETWORK/GENERALIZED LINEAR MODEL/SUPPORT VECTOR REGRESSION

Runoff is the flow of water, from rain, flow the Earth's surface into the river. The runoff prediction can help to estimate ahead of time water volume, which is useful to plan and manage plan for dealing with water shortages or floods. The runoff process is complex and needs an adequate modeling technique for efficient prediction. This research proposes an Artificial Neural Network with a Combined Generalized Linear Model and Support Vector Regression (ANN-GS) method to predict monthly runoff. The inputs for our neural network model are mixed prediction results made by Generalized Linear Model and Support Vector Regression. The inputs are runoff, rainfall, the number of rainy days, the number of month, and Normalized Difference Vegetation Index obtained from the NOAA STAR. Our work use Linear Regression Analysis, Artificial Neural Network, Autoregressive Integrated Moving Average model and Support Vector Regression to compare performance with the proposed ANN-GS model. We use two criteria to evaluate the runoff prediction performance root mean squared error (RMSE) to show error of prediction value, and correlation coefficient (R) to explore the relationship between actual runoff and prediction. The results show that our proposed method performs better than other methods.

School of Computer Engineering

Academic Year 2017

Student's Signature Ratiporn Chanklan

Advisor's Signature Nittaya Kerd

กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จได้ด้วยดี ผู้วิจัยขอขอบพระคุณ บุคคล และกลุ่มบุคคลต่างๆที่ได้กรุณาให้คำปรึกษา แนะนำ และช่วยเหลือ ทั้งในด้านวิชาการ และด้านการดำเนินงานวิจัยดังต่อไปนี้

รองศาสตราจารย์ ดร.กิตติศักดิ์ เกิดประสพ อาจารย์หัวหน้าสาขาวิชาวิศวกรรมคอมพิวเตอร์ และรองศาสตราจารย์ ดร.นิตยา เกิดประสพ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่คอยให้คำปรึกษาในการทำงานวิจัย ช่วยให้คำแนะนำ ตรวจทานความถูกต้อง และการจัดการรูปแบบของวิทยานิพนธ์ฉบับนี้

คุณปราณี กฐินใหม่ เลขานุการสาขาวิชาวิศวกรรมคอมพิวเตอร์ ที่ให้ความช่วยเหลือในการประสานงานด้านเอกสารต่าง ๆ ระหว่างศึกษา

ดร. นันทวุฒิ คะอังกุ และนักศึกษาบัณฑิตสาขาวิชาวิศวกรรมคอมพิวเตอร์ ทุกท่านที่ให้คำแนะนำ ตรวจความถูกต้องให้คำปรึกษาและให้ความช่วยเหลือมาโดยตลอด

นอกจากนี้ขอขอบคุณครู อาจารย์ทั้งในอดีตและปัจจุบันที่ให้ความรู้แก่ผู้วิจัยจนประสบความสำเร็จในชีวิต

และสุดท้าย ขอกราบขอบพระคุณ บิดา มารดา ที่ให้กำเนิด อบรม เลี้ยงดู ให้ความรัก และส่งเสริมการศึกษา ทำให้ผู้วิจัยมีความรู้ ความสามารถ มีจิตใจที่เข้มแข็ง และเป็นกำลังใจที่ให้แก่ผู้วิจัย จนทำให้ผู้วิจัยประสบความสำเร็จในชีวิตตลอดมา

รติพร จันทร์กลั่น

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญรูป

รูปที่ **หน้า**

สารบัญรูป (ต่อ)

รูปที่ **หน้า**

สารบัญรูป (ต่อ)

รูปที่ **หน้า**

บทที่ 1

บทนำ

1.1 ความสำคัญและที่มาของปัญหาการวิจัย

น้ำเป็นสิ่งที่มีความสำคัญมากในการดำรงชีวิตของสิ่งมีชีวิต มนุษย์ต้องการใช้น้ำในการดำรงชีวิตประจำวัน แต่หากน้ำมีปริมาณมากเกินไปหรือน้อยเกินไปอาจจะทำให้เกิดปัญหาที่เกี่ยวกับน้ำ ได้แก่ สภาวะน้ำท่วม และการขาดแคลนน้ำ

ในปัจจุบันสภาวะน้ำท่วม และการขาดแคลนน้ำเป็นปัญหาที่หลายประเทศ ให้ความสำคัญ เพราะในแต่ละปีปัญหาที่เกิดจากน้ำยิ่งทวีความรุนแรงและสร้างความเสียหายรุนแรงมากขึ้น ดังนั้น วิธีการป้องกันหรือวิธีการแก้ไขปัญหาที่เหมาะสมจึงเป็นสิ่งจำเป็น ปริมาณน้ำบนพื้นโลกคิดเป็น 2 ใน 3 ส่วนของพื้นผิวโลก แต่ประชากรบนโลกจำนวน 1 ใน 5 ของประชากรทั้งหมด กำลังประสบปัญหาการขาดแคลนน้ำ (ศูนย์วิจัยกสิกรไทย, 2547) ความต้องการในการใช้น้ำมีปริมาณเพิ่มมากขึ้นทุกวัน สถานการณ์ขาดแคลนน้ำทวีความรุนแรงมากขึ้น แหล่งน้ำผิวดินที่เคยมี ได้แก่ น้ำในแม่น้ำ สระ คลอง มีปริมาณลดลงและไม่เพียงพอสำหรับความต้องการของประชากรโลก นอกจากนี้การประสบปัญหาสภาวะน้ำท่วม ซึ่งเป็นปัญหาที่สำคัญซึ่งเกิดขึ้นตามธรรมชาติและตามฤดูกาล มีแนวโน้มรุนแรงเพิ่มมากขึ้นทุกปี ปัญหาที่เกิดจากน้ำเหล่านี้ก่อให้เกิดความเสียหายและอันตราย ต่อสิ่งมีชีวิต ทรัพย์สิน และมีผลกระทบต่อเศรษฐกิจทางอุตสาหกรรมและการเกษตรเป็นอย่างมาก ซึ่งการขาดแคลนน้ำทางการเกษตรหรือน้ำท่วมผลผลิตทางการเกษตรจะส่งผลให้ผลผลิตทางการเกษตรลดลงและส่งผลกระทบทางเศรษฐกิจ โดยทำให้เกิดการขาดแคลนอาหารของมนุษย์ และทำให้เกิดการเจ็บป่วยด้วยโรคต่าง ๆ ซึ่งอาจนำไปสู่การเสียชีวิตได้ ดังนั้นจึงจำเป็นต้องแก้ไขปัญหาทรัพยากรน้ำอย่างเป็นระบบ โดยจะต้องมีการบริหารจัดการทรัพยากรน้ำ และวางแผนการใช้น้ำเพื่อแก้ไขปัญหาที่เกิดจากน้ำ หรือเตรียมความพร้อมเพื่อป้องกันการเกิดปัญหาซึ่งสามารถลดความรุนแรงและผลกระทบให้เบาบางลงได้

ปริมาณน้ำในแม่น้ำที่เกิดจากน้ำฝนที่ตกลงมาในพื้นที่รับน้ำแล้วไหลลงสู่แม่น้ำ เรียกว่า น้ำท่า (Runoff) การคาดการณ์น้ำท่าในอนาคตที่มีประสิทธิภาพจะทำให้การบริหารจัดการน้ำมีประสิทธิภาพและทำให้ทราบว่ามีปริมาณน้ำเพียงพอกับความต้องการใช้น้ำหรือไม่ หรือมีปริมาณน้ำมาก ซึ่งทำให้สามารถวางแผนแก้ไขปัญหาที่เกิดจากการขาดแคลนน้ำและน้ำท่วมได้ล่วงหน้า หรือสามารถวางแผนรับมือกับปัญหาน้ำท่วมได้อย่างทันที จะเห็นได้ว่า น้ำท่าถือเป็นข้อมูลที่มีสำคัญในทางอุทกวิทยาอย่างมาก

การคาดการณ์ปริมาณน้ำท่าเป็นการวิเคราะห์ที่ค่อนข้างยากเพราะมีปัจจัยและกระบวนการเกิดที่ซับซ้อนและมีความสัมพันธ์ที่ไม่เป็นเชิงเส้น การคาดการณ์ปริมาณน้ำท่าที่แม่นยำจะช่วยในการเฝ้าระวังการขาดแคลนน้ำ วางแผนการแก้ไขปัญหาหรือบรรเทาความเสียหายจากการขาดแคลนน้ำหรือการเกิดน้ำท่วมได้ล่วงหน้า ปริมาณน้ำฝนย้อนหลังมักเป็นปัจจัยหลักที่นิยมใช้พิจารณาเพื่อการคาดการณ์ปริมาณน้ำท่า (Sajikumar et al., 1999; Agarwal et al., 2004; Maria et al., 2004) และในบางครั้งปริมาณน้ำท่าย้อนหลังถูกนำมาใช้ร่วมด้วยเพื่อคาดการณ์ปริมาณน้ำท่า (Dorum et al., 2010) นอกจากนี้ยังใช้ค่าจากปัจจัยอื่น ๆ เพื่อพิจารณาคาดการณ์ปริมาณน้ำท่า เช่น อุณหภูมิ, ค่าความชื้นในอากาศ เป็นต้น วิธีการวิเคราะห์การถดถอย (Regression Analysis) เป็นวิธีการทางสถิติที่มักจะถูกนำมาใช้ในการคาดการณ์น้ำท่า (Pilgrim et al., 1998; McIntyre et al., 2007; Patel et al., 2016) และในปัจจุบันการประยุกต์อัลกอริทึมที่เกี่ยวกับการเรียนรู้ของเครื่อง (Machine Learning) เพื่อนำมาใช้ในการคาดการณ์ปริมาณน้ำท่าเริ่มเป็นที่นิยมมากขึ้นในทางด้านอุทกวิทยา เพราะไม่จำเป็นต้องทราบข้อมูลที่เกี่ยวข้องกับลักษณะภูมิประเทศ เช่น ลักษณะดิน ความดูดซับน้ำของดิน ขนาดของพื้นที่รับน้ำ เป็นต้น ทำให้การคาดการณ์น้ำท่าไม่จำเป็นต้องเสียเวลารวบรวมข้อมูลเชิงภูมิศาสตร์และคำนวณโมเดลได้รวดเร็ว ทำให้มีเวลาในการวางแผนเตรียมรับมือกับปัญหาการขาดแคลนน้ำหรือน้ำท่วม การใช้โครงข่ายประสาทเทียม (Artificial Neural Network: ANN) ได้รับความนิยมและประสบความสำเร็จในการคาดการณ์น้ำท่า เพราะโครงข่ายประสาทเทียมสามารถใช้ได้กับข้อมูลที่ซับซ้อนเป็นเชิงเส้นและไม่เป็นเชิงเส้น (Sezin et al., 2000; Raid et al., 2004; Chen et al., 2013; Aichouri et al., 2015) นอกเหนือจากโครงข่ายประสาทเทียมเทคนิคการเรียนรู้ของเครื่องที่เรียกชื่อว่า ซัพพอร์ตเวกเตอร์รีเกรสชัน (Support Vector Regression: SVR) ได้รับความสนใจและถูกนำมาประยุกต์ใช้งานกับการคาดการณ์น้ำท่าเพราะเป็นวิธีการที่สามารถเลือกใช้เคอร์เนลให้เหมาะสมกับข้อมูลทำให้การคาดการณ์น้ำนั้นมีประสิทธิภาพ (Wu et al., 2008; Choubey et al., 2014) โดยวิธีการตรวจสอบประสิทธิภาพการคาดการณ์ปริมาณน้ำท่านิยมใช้มาตรวัดทางสถิติ เช่น ค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient: R) ค่าสัมประสิทธิ์การตัดสินใจ (Coefficient of Determination: R^2) ค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Squared Error: RMSE) เป็นต้น

ดังนั้นงานวิจัยนี้จึงได้เสนอการพัฒนาวิธีการคาดการณ์ปริมาณน้ำท่าด้วยวิธีการ ANN-GS ซึ่งเป็นเทคนิคการผสมผสานของวิธีการเรียนรู้ 2 อัลกอริทึมได้แก่ โมเดลที่ถูกพัฒนาจากวิธีการวิเคราะห์การถดถอย คือโมเดลเชิงเส้นโดยนัยทั่วไป (Generalized Linear Model: GLM) และซัพพอร์ตเวกเตอร์รีเกรสชัน จากนั้นนำผลการคาดการณ์ของทั้งสองโมเดลมาใช้ในการเรียนรู้ของโครงข่ายประสาทเทียม เพื่อใช้สร้างโมเดลในการคาดการณ์ปริมาณน้ำท่าที่มีประสิทธิภาพ เกณฑ์ที่

ใช้ในการประเมินประสิทธิภาพเป็นมาตรวัดทางสถิติ 2 มาตรวัด ได้แก่ ค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย และค่าสัมประสิทธิ์สหสัมพันธ์

สิ่งที่เป็นความก้าวหน้าใหม่ (Contribution) ของงานวิจัยนี้คือการพัฒนาอัลกอริทึมชื่อ ANN-GS ซึ่งเป็นอัลกอริทึมแบบไฮบริดที่ผสมผสานข้อเด่นของวิธีการทางสถิติของเทคนิคโมเดลเชิงเส้นโดยนัยทั่วไป และวิธีการด้านการเรียนรู้ของเครื่องของเทคนิคซัพพอร์ตเวกเตอร์รีเกรสชัน โดยใช้โครงข่ายประสาทเทียมทำหน้าที่เรียนรู้และผสมผสานผลการทำงานของ โมเดลเชิงเส้นโดยนัยทั่วไปและซัพพอร์ตเวกเตอร์รีเกรสชัน

งานวิจัยนี้ได้มีการทดสอบประสิทธิภาพของอัลกอริทึมที่พัฒนาขึ้นด้วยการเปรียบเทียบกับอัลกอริทึมอื่น ได้แก่ การวิเคราะห์การถดถอยเชิงเส้น โครงข่ายประสาทเทียม อาริม่า โมเดลเชิงเส้นโดยนัยทั่วไป และและซัพพอร์ตเวกเตอร์รีเกรสชัน เพื่อใช้ในการสร้างโมเดลในการคาดการณ์ปริมาณน้ำท่า

1.2 วัตถุประสงค์ของการวิจัย

จากแนวคิดในการทำงานวิจัย ผู้วิจัยมีวัตถุประสงค์ในการวิจัยดังนี้

1) เพื่อศึกษาและพัฒนาอัลกอริทึมสำหรับสร้างโมเดลการคาดการณ์น้ำท่าให้มีความแม่นยำสูงขึ้นกว่าวิธีการทางสถิติหรือวิธีการการเรียนรู้ของเครื่องที่ใช้อยู่ทั่วไป

2) เพื่อศึกษาแนวทางที่เหมาะสมในการให้ค่าในการเรียนรู้แก่โครงข่ายประสาทเทียมโดยใช้ค่าการคาดการณ์จาก 2 อัลกอริทึมประกอบด้วย โมเดลเชิงเส้นโดยนัยทั่วไป (GLM) และซัพพอร์ตเวกเตอร์รีเกรสชัน (SVR) เพื่อใช้ในการคาดการณ์ปริมาณน้ำท่า

1.3 ขอบเขตของการวิจัย

1) ในงานวิจัยนี้ใช้ค่าน้ำท่า น้ำฝน จำนวนวันที่ฝนตกในแต่ละเดือน จากเว็บไซต์ของศูนย์อุทกวิทยาชลประทานเป็นปัจจัยหลักในการสร้างโมเดล สามารถดาวน์โหลดข้อมูลได้จากเว็บไซต์ http://www.hydro-1.net และ http://www.hydro-4.net

2) ในงานวิจัยนี้ใช้ข้อมูลดัชนีผลต่างพืชพรรณซึ่งเป็นค่าจากการรับรู้ระยะไกลของดาวเทียม NOAA เป็นปัจจัยร่วมในการสร้างโมเดล สามารถดาวน์โหลดข้อมูลได้จากเว็บไซต์ http://www.star.nesdis.noaa.gov

3) งานวิจัยนี้พัฒนาอัลกอริทึม ANN-GS ที่ใช้สำหรับการคาดการณ์น้ำท่าโดยใช้โครงข่ายประสาทเทียม โดยใช้ค่าในการเรียนรู้จากโมเดลเชิงเส้นโดยนัยทั่วไป และซัพพอร์ตเวกเตอร์รีเกรสชัน และทดสอบโมเดล ANN-GS ด้วยการคาดการณ์น้ำท่าในลุ่มน้ำปิงและลุ่มน้ำมูล

4) การเปรียบเทียบประสิทธิภาพการคาดการณ์น้ำท่าระหว่างอัลกอริทึมที่นำเสนอกับอัลกอริทึมอื่น ๆ ใช้ค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย และค่าสัมประสิทธิ์สหสัมพันธ์

5) ในงานวิจัยนี้ใช้โปรแกรม IBM SPSS Modeler เป็นเครื่องมือในการพัฒนาอัลกอริทึม ANN-GS เพื่อการคาดการณ์ปริมาณน้ำท่า

1.4 ประโยชน์ที่คาดว่าจะได้รับ

จากการศึกษาและพัฒนางานวิจัยนี้ ผู้วิจัยคาดว่าวิธีการคาดการณ์น้ำท่าที่นำเสนอจะเกิดประโยชน์ต่อผู้ใช้ สามารถนำไปใช้ในการคาดการณ์น้ำท่าในลุ่มน้ำอื่น ๆ ได้และมีประสิทธิภาพที่ดีในการคาดการณ์ นอกจากนี้วิธีการผสมผสานเทคนิค GLM เข้ากับ SVR เพื่อใช้เพิ่มประสิทธิภาพการคาดการณ์ของ ANN จะเป็นแนวทางให้กับงานวิจัยด้านอื่น ๆ นอกเหนือจากด้านอุทกวิทยาปรับปรุงไปใช้ประโยชน์ได้

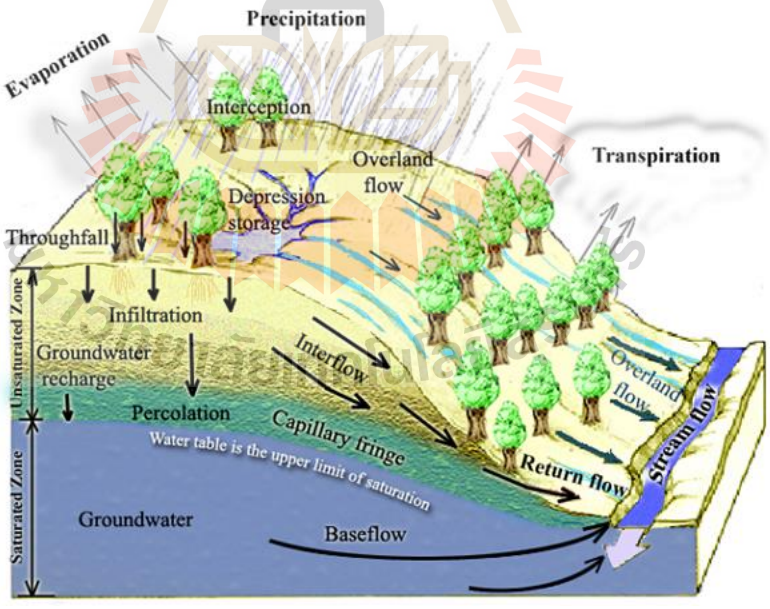
บทที่ 2

ปริทัศน์วรรณกรรมและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงปริทัศน์วรรณกรรมและงานวิจัยที่เกี่ยวข้อง โดยมีรายละเอียดของกระบวนการเกิดน้ำท่า การวิเคราะห์การถดถอยเชิงเส้น โมเดลเชิงเส้นโดยนัยทั่วไป โครงข่ายประสาทเทียม โมเดลอาริม่า ซัพพอร์ตเวกเตอร์รีเกรสชัน การประเมินความสามารถของโมเดลด้วยค่าสัมประสิทธิ์สหสัมพันธ์และค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย และงานวิจัยที่เกี่ยวข้องกับการคาดการณ์ปริมาณน้ำท่า

2.1 ความหมายของน้ำท่า (Runoff)

น้ำท่าหรือน้ำในแม่น้ำ เป็นข้อมูลที่สำคัญทางด้านอุทกวิทยา เพื่อใช้ในการวิเคราะห์ปริมาณน้ำสำหรับการวางแผนการจัดการน้ำท่าเพื่อให้เพียงพอต่อความต้องการของมนุษย์ ซึ่งกระบวนการเกิดน้ำท่าสามารถแสดงได้ดังรูปที่ 2.1 ดังนี้



รูปที่ 2.1 กระบวนการเกิดน้ำท่า (ที่มา : David et al., 2003)

จากรูปที่ 2.1 กระบวนการเกิดน้ำท่าหรือน้ำในแม่น้ำเรียกว่า Stream Flow เกิดจากน้ำในลักษณะของเหลวหรือของแข็งซึ่งเกิดจากก้อนเมฆบนท้องฟ้าตกมาบนพื้นโลก (Precipitation) ซึ่งจะหมายรวมถึง ฝน หิมะ และลูกเห็บ น้ำที่ตกจากท้องฟ้าถูกพืชสกัดกั้นไว้บางส่วน เรียกว่า น้ำ

พืชยึด (Interception) โดยน้ำที่เกาะบนใบไม้ กิ่งไม้ และลำต้นของต้นไม้แล้วไหลลงสู่ดิน เรียกว่า น้ำพืชหยด (Throughfall) น้ำส่วนใหญ่ที่ถูกพืชสกัดกั้นไว้จะระเหยกลับสู่ชั้นบรรยากาศ โดยน้ำที่กลับสู่ชั้นบรรยากาศอาจเกิดจากการคายน้ำของพืช (Trainspiration) และเกิดจากการระเหย (Evaporation) จากดินและน้ำพืชหยด (Throughfall) แหล่งน้ำต่าง ๆ เช่น แม่น้ำ ลำคลอง หนอง บึง ทะเลสาบ อ่างเก็บน้ำ และแหล่งน้ำสาธารณะอื่น ๆ เรียกว่า แหล่งน้ำผิวดิน (Surface Water) ซึ่งแหล่งน้ำผิวดินที่ทำให้เกิดน้ำท่าจะเกิดจากน้ำพืชหยดและการละลายของหิมะ (Snowmelt) หรือเกิดจากฝนที่สะสมบนแอ่งพักน้ำ (Depression Storage) หรือน้ำที่ไหลอยู่บนผิวดิน (Overland Flow) น้ำที่ซึมเข้าไปในดินอาจซึมลึกเข้าไปภายในชั้นดินและชั้นหินซึมลงไปถูกกักเก็บหรือสะสมตัวอยู่ใต้ดินเป็นน้ำบาดาล (Groundwater) น้ำในลำธารที่เกิดจากการไหลซึมออกมาของน้ำบาดาลเรียกว่า Baseflow ส่วนมากเกิดในช่วงฤดูแล้งหรือช่วงฝนตกในปริมาณน้อย น้ำใต้ผิวดินที่ไหลกลับไปบนผิวดินเรียกว่า Return Flow จะเห็นว่าเมื่อฝนตกลงสู่ดินจะเกิดการไหลซึ่งการไหลของน้ำสามารถแบ่งเป็น 3 แบบ ได้แก่ ไหลหน้าผิวดิน (Overland Flow) ไหลภายในดิน (Inter Flow) และไหลใต้ดิน (Groundwater Flow)

เนื่องจากเส้นทางการไหลของน้ำฝนที่ตกลงมามีความซับซ้อนและมีการเปลี่ยนแปลง การตรวจวัดปริมาณน้ำในแต่ละส่วนนั้นทำได้ยาก ในการวิเคราะห์ข้อมูลน้ำท่านิยมพิจารณาน้ำในแม่น้ำเป็น 2 ส่วนประกอบด้วย น้ำที่ไหลลงสู่แม่น้ำทันทีหลังฝนตกหรือไหลลงแม่น้ำอย่างรวดเร็ว (Direct Runoff) และน้ำที่ใช้เวลานานในการไหลมาในลำน้ำ (Base Flow)

Direct runoff ส่วนใหญ่เป็นน้ำฝนที่ตกลงมาแล้วไหลไปตามผิวดิน (Surface Runoff) นอกจากนี้ยังรวมถึงน้ำฝนที่ตกลงในแม่น้ำโดยตรง และน้ำที่ไหลภายในดินบางส่วนที่ไหลพ้นผิวดินขึ้นมา

Base Flow เป็นน้ำที่ไหลช้าในการเดินทางจากจุดที่ฝนตกลงมาจนกระทั่งถึงแม่น้ำอาจใช้เวลาเป็นหลาย ๆ วัน จนกระทั่งเป็นปี ปริมาณของน้ำส่วนนี้ในแม่น้ำค่อนข้างจะคงที่โดยมีการเปลี่ยนแปลงตามฤดูกาล แหล่งกำเนิดของน้ำในส่วนนี้เกิดจากน้ำที่ไหลทางใต้ดิน

การคำนวณเพื่อการคาดการณ์ปริมาณน้ำท่านิยมใช้วิธีการทางสถิติ เช่น การวิเคราะห์ถดถอยเชิงเส้น และการวิเคราะห์อนุกรมเวลาด้วยโมเดลอาริม่า ในระยะหลังเริ่มปรากฏงานวิจัยที่นำวิธีการเรียนรู้ของเครื่อง เช่น โครงข่ายประสาทเทียมซัพพอร์ตเวกเตอร์รีเกรสชัน มาใช้เพื่อเพิ่มความแม่นยำในการคาดการณ์ปริมาณน้ำท่า โดยรายละเอียดของวิธีการเหล่านี้อธิบายในหัวข้อ 2.2 ถึง 2.6

2.2 การวิเคราะห์การถดถอยเชิงเส้น (Linear Regression Analysis: LR)

การวิเคราะห์การถดถอย (Regression Analysis) เป็นวิธีทางสถิติที่ใช้หาความสัมพันธ์ระหว่างตัวแปรเชิงปริมาณตั้งแต่สองตัวขึ้นไป เพื่อใช้ในการทำนายค่าของตัวแปรหนึ่งจากตัวแปรอื่น ตัวแปรที่ใช้ในการวิเคราะห์การถดถอยมีสองแบบคือ ตัวแปรอิสระ (Independent Variable) ซึ่งเป็นตัวแปรที่ทราบค่า และตัวแปรตาม (Dependent Variable) เป็นตัวแปรที่ไม่ทราบค่า โดยสมมุติฐานจะคาดว่าตัวแปรอิสระจะมีผลต่อตัวแปรตามเพื่อที่จะไปคาดการณ์ค่าของตัวแปรตาม การวิเคราะห์การถดถอยระหว่างตัวแปรอิสระหนึ่งตัวกับตัวแปรตามหนึ่งตัวที่มีความสัมพันธ์เชิงเส้นตรงเรียกว่า การวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression Analysis) ถ้าตัวแปรตามหนึ่งตัวกับตัวแปรอิสระหลายตัวมีความสัมพันธ์กันแบบเชิงเส้นตรงเรียกว่า การวิเคราะห์การถดถอยเชิงเส้นพหุคูณ (Multiple Linear Regression Analysis)

2.2.1 การวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression Analysis)

การวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression Analysis) เป็นการศึกษาความสัมพันธ์ระหว่างสองตัวแปรในที่นี้ให้แทนด้วยตัวแปร X เป็นตัวแปรอิสระและ Y เป็นตัวแปรตามซึ่งสองตัวแปรนี้มีความสัมพันธ์กันในลักษณะเส้นตรง (Linear)

ในการวิเคราะห์สมการถดถอยอย่างง่ายค่าของตัวแปร X จะมีการกำหนดค่าไว้ และค่าของตัวแปร Y จะเปลี่ยนแปลงไปตามตัวแปร X ค่าของ X หนึ่งค่าจะมีค่า Y ที่เป็นคู่ของค่า X เมื่อนำค่า X และ Y หลาย ๆ ค่ามาพล็อตบนแกน X และ Y (ตัวอย่างดังรูป 2.2) แล้วลากเส้นตรงผ่านจุดที่พล็อตลงในกราฟ เส้นตรงที่ลากนั้นจะแสดงความสัมพันธ์ระหว่างค่าเฉลี่ยของตัวแปร X และตัวแปร Y ซึ่งเส้นตรงนี้เรียกว่า เส้นกราฟถดถอย (Regression Line) การสร้างเส้นกราฟถดถอยทำได้โดยนำข้อมูลจากตัวแปรที่ทำการศึกษามาวิเคราะห์หาความสัมพันธ์ที่สามารถบอกแนวโน้มของความสัมพันธ์โดยใช้แผนภาพเส้นตรง และจะทำการหาเส้นตรงที่ดีที่สุดเพื่อเป็นตัวแทนของรูปแบบความสัมพันธ์ของตัวแปรที่ศึกษา เส้นตรงที่ดีที่สุดจะมีเพียงเส้นเดียว การวิเคราะห์การถดถอยเชิงเส้นอย่างง่ายมีสมการดังนี้ (มนต์ชัย เทียนทอง, 2548)

สมการในกรณีใช้ข้อมูลประชากรทั้งหมด: $Y = \propto + \beta X + \varepsilon$ (2-1)

สมการในกรณีใช้ข้อมูลที่เป็นตัวแทนประชากร: $Y = a + bX + e$ (2-2)

สมการพยากรณ์: $\hat{Y} = a + bX$ (2-3)

เมื่อ $Y, \hat{Y}$ คือ ตัวแปรตาม

X คือ ตัวแปรอิสระหรือตัวแปรต้น

$\propto, a$ คือ ค่าคงที่ของสมการถดถอยหรือจุดตัดบนแกน Y

β, b คือ ความชันของเส้นถดถอย หรือสัมประสิทธิ์การถดถอยของตัวแปรต้น

ε, e คือ ค่าความคลาดเคลื่อน

โดย a และ b เป็นค่าประมาณแบบกำลังสองต่ำสุด (Least Square Method) ของ $\propto$ และ β ตามลำดับ ซึ่งหมายถึงจะหาค่า a และ b ที่ทำให้ผลรวมกำลังสองของค่าความคลาดเคลื่อน ($\sum e_i^2$) มีค่าน้อยที่สุด สามารถหาค่าของ a ได้จากสมการ (มนต์ชัย เทียนทอง, 2548)

$$a = \bar{Y} - b\bar{X} \tag{2-4}$$

และหาค่า b ได้จากสมการ

$$b = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{(\sum \mathrm{X}-\bar{X})^2} = \frac{n\sum \mathrm{XY}-(\sum \mathrm{X})(\sum \mathrm{Y})}{n\sum \mathrm{X}^2-(\sum \mathrm{X})^2} \tag{2-5}$$

เมื่อ $\bar{Y}$ คือ ค่าเฉลี่ยของตัวแปรตาม

$\bar{X}$ คือ ค่าเฉลี่ยของตัวแปรอิสระหรือตัวแปรต้น

n คือ จำนวนของกลุ่มตัวอย่าง

ค่าของ b ที่เป็นความชันของกราฟเส้นตรง ที่เกิดจากสมการเชิงเส้นถ้าทราบค่าของ b และ a จะสามารถพยากรณ์ค่าของตัวแปร Y ได้ เรียก b ว่า สัมประสิทธิ์การถดถอย (Regression Coefficient) หรือเรียกว่า สัมประสิทธิ์การพยากรณ์ และเรียก a ว่าจุดตัดบนแกน Y สรุปความสัมพันธ์ของ X และ Y จากค่าของ b ได้ดังนี้ (มนต์ชัย เทียนทอง, 2548)

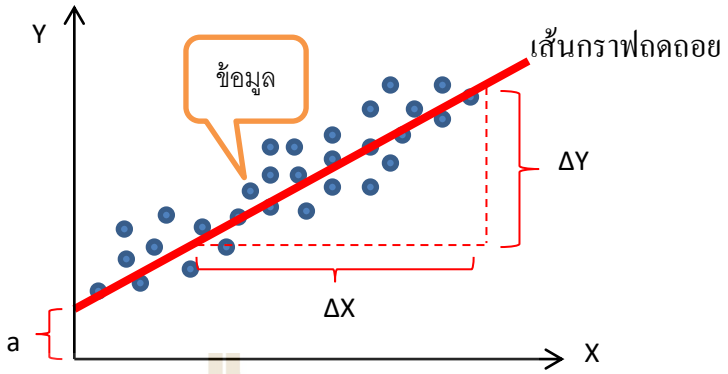
1. ถ้า b > 0 แสดงว่า X และ Y มี ความสัมพันธ์กันในทิศทางเดียวกัน ถ้า X มีค่าสูงขึ้น ค่า Y ก็มีค่าสูงขึ้น และถ้า X มีค่าต่ำลง ค่า Y ก็มีค่าต่ำลง

2. ถ้า b < 0 แสดงว่า X และ Y มีความสัมพันธ์กันในทิศทางตรงกันข้าม ถ้า X มีค่าสูงขึ้นค่าของ Y จะต่ำลง และถ้า X มีค่าต่ำลงค่าของ Y จะสูงขึ้น

3. ถ้า b มีค่าเข้าใกล้ 0 แสดงว่า X และ Y มีความสัมพันธ์กันน้อย

4. ถ้า b = 0 แสดงว่า X และ Y ไม่มีความสัมพันธ์กัน เส้นกราฟจะมีลักษณะเป็นเส้นตรงขนานกับแกน X นั้นคือค่าของ Y จะมีค่าเท่ากับ ค่าคงที่ a โดยไม่เปลี่ยนค่าไปตามค่าของ X

5. ถ้า b = 1 แสดงว่าความชันของเส้นกราฟมีค่าเท่ากับ 45 องศา ซึ่งหมายถึงการเปลี่ยนแปลงของค่า X เท่ากับการเปลี่ยนแปลงของค่า Y ทำให้แปลความหมายได้ว่า X และ Y มีความสัมพันธ์กันมากที่สุด



รูปที่ 2.2 จุดตัดแกน Y และความชันของเส้นกราฟถดถอย

ตารางที่ 2.1 ตัวอย่างข้อมูลแสดงการคำนวณเพื่อวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย

ตัวที่	1	2	3	4	5	6	7	8	9	10
X	90	112	132	148	150	162	171	180	211	222
Y	82	90	111	121	128	133	141	155	162	170

ตารางที่ 2.2 ผลรวม ค่าเฉลี่ย ผลคูณของตัวแปรต้นกับตัวแปรตาม และค่ายกกำลังสองของตัวแปรต้น

ตัวที่	X	Y	XY	X^2
1	90	82	7,380	8,100
2	112	90	10,080	12,544
3	132	111	14,652	17,424
4	148	121	17,908	21,904
5	150	128	19,200	22,500
6	162	133	21,546	26,244
7	171	141	24,111	29,241
8	180	155	27,900	32,400
9	211	162	34,182	44,521
10	222	170	37,740	49,284
ผลรวม	1578	1293	214,699	264,162
ค่าเฉลี่ย	157.8	129.3	21,469.9	26,416.2

จากข้อมูลตัวอย่างในตารางที่ 2.1 การคำนวณเพื่อหาความสัมพันธ์ของ X และ Y ในลักษณะการวิเคราะห์การถดถอยเชิงเส้นอย่างง่ายทำได้โดยเตรียมข้อมูลผลคูณตามตาราง 2.2 ค่าที่เป็นตัวแปรต้นคือ X ค่าที่เป็นตัวแปรตามคือ Y วิธีการเลือกตัวแปรสังเกตได้โดยตัวแปรที่ต้องการทราบค่าหรือเป็นคำตอบในอนาคต คือตัวแปรตาม (Y) ส่วนค่าที่ต้องใส่ในสมการเพื่อให้ได้คำตอบคือตัวแปรต้น (X) ทำการหาผลรวม ค่าเฉลี่ย ผลคูณ ผลรวมของผลคูณของตัวแปรต้นกับตัวแปรตามและค่ายกกำลังสองของตัวแปรต้น

คำนวณหาสัมประสิทธิ์ของตัวแปรต้นและค่าคงที่ โดยนำค่าที่คำนวณได้จากตารางที่ 2.2 นำมาคำนวณหาดังนี้

$$b = \frac{n\sum XY-(\sum X)(\sum Y)}{n\sum X^2-(\sum X)^2}$$

$$= \frac{10(214,699)-(1,578)(1,293)}{10(264,162)-(1,578)^2}$$

$$= \frac{2,146,990-2,040,354}{2,641,620-2,490,084}$$

$$= \frac{106,636}{151,536}$$

$$= 0.7037$$

$$a = \overline{Y} - b\overline{X}$$

$$= 129.3 - (0.7037)\ 157.8$$

$$= 129.3 - 111.044$$

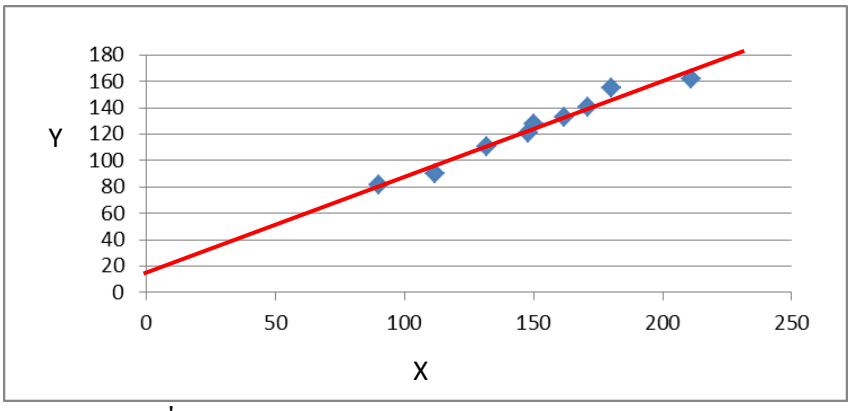
$$= 18.26$$

สมการถดถอยอย่างง่ายของข้อมูลสามารถแสดงเป็นภาพกราฟได้ดังรูปที่ 2.3 และเขียนเป็นสมการได้ดังนี้

$$Y = a + bX$$

$$= 18.26 + 0.704(X)$$

จากสมการที่ได้สามารถแปลความได้ว่า X และ Y มีความสัมพันธ์กันมากโดยสังเกตจากค่าความชัน 0.704 ที่มีค่าเข้าใกล้ 1 และความสัมพันธ์ของ X และ Y มีลักษณะแปรผันตามกัน เนื่องจากค่าความชันมีค่าเป็นบวก และเมื่อ X มีค่าเป็นศูนย์ Y จะมีค่าเท่ากับ 18.26 ซึ่งเป็นค่าของ a ที่หมายถึงจุดตัดบนแกน Y



รูปที่ 2.3 การกระจายของข้อมูลและเส้นกราฟถดถอย

2.2.2 การวิเคราะห์การถดถอยเชิงเส้นพหุคูณ (Multiple Linear Regression Analysis)

ในการหาความสัมพันธ์ของตัวแปรอิสระกับตัวแปรตามนั้นบางครั้งจำนวนตัวแปรอิสระที่สนใจในการศึกษามีมากกว่าหนึ่งตัวแปร ความสัมพันธ์นี้ไม่สามารถใช้การถดถอยเชิงเส้นอย่างง่ายในการวิเคราะห์ได้ สำหรับกรณีที่มีตัวแปรอิสระมากกว่า 1 ตัวที่มีความสัมพันธ์เชิงเส้นกับตัวแปรตาม สมการถดถอยสามารถเขียนให้อยู่ในรูปของสมการดังนี้ (พรสิน สุภวาลย์, 2556)

สมการในกรณีใช้ข้อมูลประชากรทั้งหมด: $$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon \quad (2\text{-}6)$$

สมการในกรณีใช้ข้อมูลที่เป็นตัวแทนประชากร: $$Y = b_0 + b_1 X_1 + \cdots + b_k X_k + e \quad (2\text{-}7)$$

สมการพยากรณ์: $$\hat{Y} = b_0 + b_1 X_1 + \cdots + b_k X_k \quad (2\text{-}8)$$

เมื่อ $Y, \hat{Y}$ คือ ตัวแปรตาม

X_i คือ ตัวแปรอิสระหรือตัวแปรต้น โดย i มี ค่าตั้งแต่ 1, 2, …,k

β_0, b_0 คือ ค่าคงที่ของสมการถดถอยหรือจุดตัดบนแกน Y

β_i, b_i คือ ความชันของเส้นถดถอย หรือสัมประสิทธิ์การถดถอยของตัวแปรต้นโดย i มีค่าตั้งแต่ 1, 2, …,k

ε, e คือ ค่าความคลาดเคลื่อน

การคำนวณสมการถดถอยพหุคูณนั้นทำได้โดยการใช้เมตริกซ์ซึ่งจะช่วยให้การคำนวณง่ายและรวดเร็ว ขนาดของเมตริกซ์ในการคำนวณจะใหญ่ขึ้นตามจำนวนตัวแปรอิสระดังนี้

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

การประมาณค่าพารามิเตอร์ของสมการถดถอยพหุคูณใช้วิธีการเช่นเดียวกับสมการถดถอยเชิงเส้นอย่างง่าย คือการใช้วิธีกำลังสองน้อยที่สุด แต่จะมีการกำหนดให้ Q เป็นค่าผลรวมกำลังสองของความคลาดเคลื่อนที่น้อยที่สุด (พรสิน สุภวาลย์, 2556)

$$Q = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_k X_{ik})^2$$

$$= \sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{k}\beta_j X_{ij})^2 \quad (2\text{-}9)$$

ในการคำนวณหาค่า Q น้อยสุดจะต้องทำการหาอนุพันธ์ย่อยเทียบกับค่า β_j แต่ละค่าโดยสามารถเขียนสมการทั้งสองได้ดังนี้ (พรสิน สุภวาลย์, 2556)

$$\left.\frac{\partial Q}{\partial \beta_0}\right|_{b_0,b_1,\ldots,b_k} = -2\sum_{i=1}^{n}\left(Y_i - b_0 - \sum_{j=1}^{k} b_j X_{ij}\right) = 0 \quad (2\text{-}10)$$

และ

$$\left.\frac{\partial Q}{\partial \beta_0}\right|_{b_0,b_1,\ldots,b_k} = -2\sum_{i=1}^{n}\left(Y_i - b_0 - \sum_{j=1}^{k} b_j X_{ij}\right)X_{ij} = 0 \quad (2\text{-}11)$$

ดังนั้นสมการปกติสามารถเขียนในรูปของเมตริกซ์ดังนี้

$$(X'X)b = X'Y \quad (2\text{-}12)$$

สามารถหาเวกเตอร์ b ได้ ดังนี้

$$b = (X'X)^{-1}X'Y \quad (2\text{-}13)$$

โดย X'X ต้องสามารถหาเมตริกซ์ผกผันได้และเป็นเมตริกซ์สมมาตรที่มีขนาด (k+1) × (k+1) และค่าของสมาชิกในแนวเฉียงเป็นผลรวมกำลังสองของค่าในแต่ละหลัก

ตารางที่ 2.3 ตัวอย่างข้อมูลเพื่อใช้แสดงการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ

ตัวที่	1	2	3	4	5	6
Y	1	2	3	4	5	6
X_1	2	4	6	8	10	12
X_2	6	5	4	3	2	1

วิธีการคำนวณการวิเคราะห์การถดถอยเชิงเส้นพหุคูณทำได้โดยการนำข้อมูลจากตารางที่ 2.3 เขียนให้อยู่ในรูปเมตริกซ์ได้ดังนี้

$$X = \begin{bmatrix} 1 & 2 & 6 \\ 1 & 4 & 5 \\ 1 & 6 & 4 \\ 1 & 8 & 3 \\ 1 & 10 & 2 \\ 1 & 12 & 1 \end{bmatrix} \qquad Y = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 4 & 6 & 8 & 10 & 12 \\ 6 & 5 & 4 & 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 6 \\ 1 & 4 & 5 \\ 1 & 6 & 4 \\ 1 & 8 & 3 \\ 1 & 10 & 2 \\ 1 & 12 & 1 \end{bmatrix} = \begin{bmatrix} 6 & 96 & 75 \\ 96 & 364 & 112 \\ 75 & 112 & 91 \end{bmatrix}$$

$$\det(X'X) = 198{,}744 + 806{,}400 + 806{,}400 - 2{,}047{,}500 - 89{,}304 - 838{,}656$$

$$= -1{,}163{,}916$$

$$Adj\,(X'X) = \begin{bmatrix} 6 & 96 & 75 \\ 96 & 364 & 112 \\ 75 & 112 & 91 \end{bmatrix}^t$$

$$= \begin{bmatrix} 6 & 96 & 75 \\ 96 & 364 & 112 \\ 75 & 112 & 91 \end{bmatrix}$$

$$= \begin{bmatrix} (-1)^{1+1}\begin{vmatrix} 364 & 122 \\ 122 & 91 \end{vmatrix} & (-1)^{2+1}\begin{vmatrix} 96 & 112 \\ 75 & 91 \end{vmatrix} & (-1)^{3+1}\begin{vmatrix} 96 & 364 \\ 75 & 112 \end{vmatrix} \\ (-1)^{1+2}\begin{vmatrix} 96 & 75 \\ 112 & 91 \end{vmatrix} & (-1)^{2+2}\begin{vmatrix} 6 & 75 \\ 75 & 91 \end{vmatrix} & (-1)^{3+2}\begin{vmatrix} 6 & 96 \\ 75 & 112 \end{vmatrix} \\ (-1)^{1+3}\begin{vmatrix} 96 & 75 \\ 364 & 112 \end{vmatrix} & (-1)^{2+3}\begin{vmatrix} 6 & 75 \\ 96 & 112 \end{vmatrix} & (-1)^{3+3}\begin{vmatrix} 6 & 96 \\ 96 & 364 \end{vmatrix} \end{bmatrix}$$

$$= \begin{bmatrix} 18{,}240 & -336 & -16{,}548 \\ -336 & -5{,}079 & 6{,}528 \\ -16{,}548 & 6{,}528 & -7{,}032 \end{bmatrix}$$

$$(X'X)^{-1} = \frac{1}{-1{,}163{,}916} \begin{bmatrix} 18{,}240 & -336 & -16{,}548 \\ -336 & -5{,}079 & 6{,}528 \\ -16{,}548 & 6{,}528 & -7{,}032 \end{bmatrix}$$

$$= \begin{bmatrix} -0.016 & 0.0003 & 0.014 \\ 0.0003 & 0.004 & -0.0056 \\ 0.014 & -0.0056 & 0.006 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 4 & 6 & 8 & 10 & 12 \\ 6 & 5 & 4 & 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{bmatrix} = \begin{bmatrix} 21 \\ 182 \\ 56 \end{bmatrix}$$

$$b = (X'X)^{-1}X'Y$$

$$= \begin{bmatrix} -0.016 & 0.0003 & 0.014 \\ 0.0003 & 0.004 & -0.0056 \\ 0.014 & -0.0056 & 0.006 \end{bmatrix} \begin{bmatrix} 21 \\ 182 \\ 56 \end{bmatrix}$$

$$= \begin{bmatrix} 0.5026 \\ 0.4207 \\ -0.3892 \end{bmatrix}$$

นั่นคือ $b_0 = 0.5026$, $b_1 = 0.4207$ และ $b_2 = -0.3892$
จากข้อมูลตัวอย่างในตารางที่ 2.3 สามารถวิเคราะห์ความสัมพันธ์ของ X และ Y และเขียนเป็นสมการการถดถอยเชิงเส้นพหุคูณได้ดังนี้

$$Y = 0.5026 + 0.4207 X_1 - 0.3892X_2$$

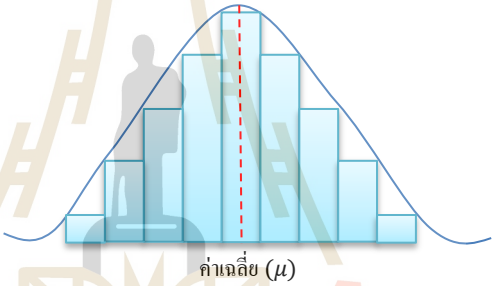
2.3 โมเดลเชิงเส้นโดยนัยทั่วไป (Generalized Linear Models: GLM)

โมเดลเชิงเส้นโดยนัยทั่วไป ถูกเสนอโดย Nelder และ Wedderburn ในปี ค.ศ.1972 เป็นวิธีการทางสถิติที่มีแนวคิดเพื่อหาความสัมพันธ์ระหว่างตัวแปรต้นกับตัวแปรตามที่ไม่มีข้อจำกัดว่าการกระจายตัวของค่าความคลาดเคลื่อนในตัวแปรตามจะต้องเป็นแบบปกติเท่านั้น และข้อมูลที่นำมาวิเคราะห์ความสัมพันธ์เป็นได้ทั้งข้อมูลตัวเลขและไม่เป็นตัวเลข โครงสร้างของโมเดลเชิงเส้นโดยนัยทั่วไปประกอบด้วย 3 องค์ประกอบคือ องค์ประกอบแบบสุ่ม องค์ประกอบเชิงระบบ และฟังก์ชันเชื่อมโยง (Fox, 2008)

องค์ประกอบแบบสุ่ม (Random Component) เป็นองค์ประกอบที่เกี่ยวข้องกับคุณลักษณะของการแจกแจงของตัวแปรตาม (Y_i) โดยตัวแปรตามจะมีค่าเป็นอิสระต่อกันและมีการแจกแจงแบบปกติ หรือไม่ปกติก็ได้แต่ต้องเป็นการแจกแจงที่อยู่ในชนิดของตระกูลเอกโพเนนเชียล (Exponential Family) ได้แก่ Gaussian (Normal), Binomial, Poisson, Gamma และ inverse-Gaussian ทั้งนี้การ

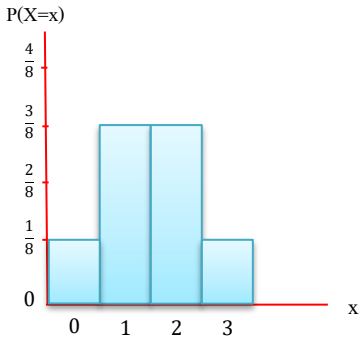
แจกแจงของตระกูลเอกโพเนนเชียลต้องมีคุณสมบัติคือ การแจกแจงสามารถเขียนให้อยู่ในรูปของค่าเฉลี่ยและความแปรปรวนได้ และความแปรปรวนเป็นฟังก์ชันของค่าเฉลี่ย

- การแจกแจงแบบปกติ (Normal Distribution) เป็นการแจกแจงของตัวแปรสุ่มแบบต่อเนื่อง สามารถปรากฏในข้อมูลหลากหลายประเภท เช่น คะแนนสอบของนักเรียนในชั้นเรียนขนาดใหญ่ ส่วนสูงของประชากร ค่าใช้จ่ายของครัวเรือน เป็นต้น กราฟของการแจกแจงปกติจะเป็นกราฟรูประฆังคว่ำ ส่วนปลายโค้งทั้ง 2 ด้านจะมีค่าเข้าใกล้ 0 และมีค่าเป็นอนันต์ มีจุดศูนย์กลางที่ค่าเฉลี่ย เส้นโค้งมีลักษณะสมมาตรรอบค่าเฉลี่ย มีค่าฐานนิยม ค่ามัธยฐาน และค่าเฉลี่ยเท่ากัน ลักษณะข้อมูลที่มีการแจกแจงปกติ แสดงตัวอย่างดังรูปที่ 2.4



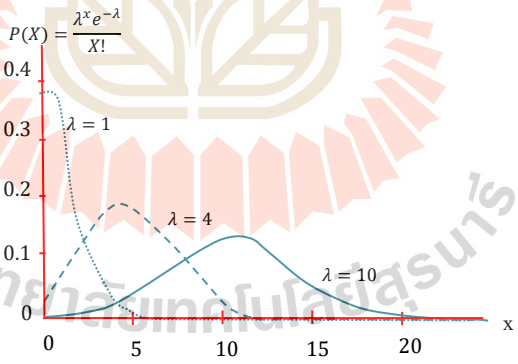
รูปที่ 2.4 การแจกแจงแบบปกติ

- การแจกแจงแบบทวินาม (Binomial Distribution) เป็นการแจกแจงของตัวแปรสุ่มแบบไม่ต่อเนื่องที่ใช้ได้กับข้อมูลที่เป็นการทดลองซ้ำ ๆ หลายครั้ง โดยมีผลลัพธ์เพียง 2 ค่า คือ สำเร็จและไม่สำเร็จ การทดลองแต่ละครั้งเป็นอิสระต่อกัน โดยกำหนดให้ p เป็นความน่าจะเป็นของความสำเร็จในการทดลองแต่ละครั้ง และ 1-p เป็นความน่าจะเป็นของการทดลองไม่สำเร็จ ตัวอย่างข้อมูล เช่น การโยนเหรียญ การฉีดวัคซีน การทดสอบแบตเตอรี่ เป็นต้น ตัวอย่างการแจกแจงข้อมูลแบบทวินามที่เกิดจากการโยนเหรียญ 3 เหรียญ แสดงได้ดังรูป 2.5



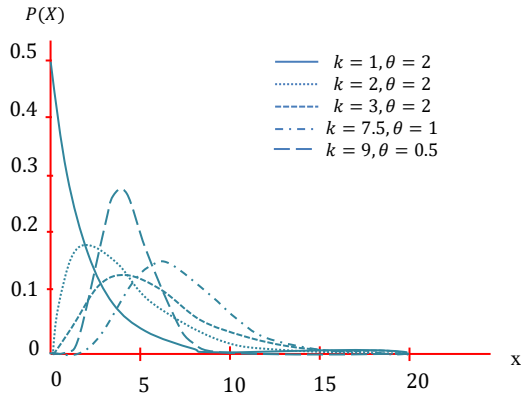
รูปที่ 2.5 การแจกแจงแบบทวินาม

- การแจกแจงแบบปัวส์ซอง (Poisson Distribution) เป็นการแจกแจงของตัวแปรสุ่มแบบไม่ต่อเนื่องที่ใช้ได้กับข้อมูลที่เป็นการทดลองการนับจำนวนครั้งของเหตุการณ์ที่สนใจในระยะเวลาหรือเกิดขึ้นในขอบเขตที่กำหนด ซึ่งขอบเขตที่กำหนดอาจจะเป็นช่วงเวลา พื้นที่ ปริมาตร เป็นต้น โดยกำหนดให้ λ คือค่าเฉลี่ยของจำนวนเหตุการณ์ที่เกิดขึ้นในขอบเขตที่กำหนด ตัวอย่างข้อมูล เช่น จำนวนครั้งของการใช้ตู้ ATM ช่วง 09.00-10.00 น. จำนวนคำผิดในหนึ่งหน้ากระดาษเอสี่ ของเสียที่เกิดจากเครื่องจักรในหนึ่งชั่วโมง ลักษณะการแจกแจงแบบปัวส์ซองแสดงตัวอย่างได้ดังรูปที่ 2.6



รูปที่ 2.6 การแจกแจงแบบปัวส์ซอง

- การแจกแจงแกมม่า (Gamma Didtribution) เป็นการแจกแจงของตัวแปรสุ่มแบบต่อเนื่อง การแจกแจงแกมม่าใช้ในการจำลองแบบเวลาที่ใช้ก่อนจะเกิดเหตุการณ์พัวส์ซอง k ครั้ง กรณี k=1 จะเท่ากับเวลาที่ใช้ก่อนที่เกิดเหตุการณ์พัวส์ซองถัดไป เช่น การโทรศัพท์เข้าในโรงพยาบาลซึ่งเป็นไปตามกระบวนการพัวส์ซองด้วยอัตรา 5 ครั้งต่อนาที ความน่าจะเป็นที่หนึ่งนาที่ผ่านไปก่อนที่จะมีการโทรเข้าครั้งที่สองมีค่าเป็น 0.9596 ตัวอย่างการแจกแจงแกมม่าแสดงได้ดังรูปที่ 2.7



รูปที่ 2.7 การแจกแจงแบบแกมม่า

องค์ประกอบเชิงระบบ (Systematic Component) เป็นการนำตัวแปรต้นมากำหนดเป็นฟังก์ชันเชิงเส้นเพื่อใช้ในการพยากรณ์ตัวแปรตาม

$$E(Y_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} \tag{2-14}$$

ซึ่งการรวมตัวในลักษณะเชิงเส้นของตัวแปรต้น $(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi})$ เรียกว่า ตัวพยากรณ์เชิงเส้น (Linear Predictor) ซึ่งเป็นการเขียนตัวแปรต้นให้อยู่ในรูปของตัวประมาณเชิงเส้น (η_i) แสดงในรูปสมการได้ดังนี้

$$\eta_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} \tag{2-15}$$

กำหนดให้ n คือขนาดของข้อมูลตัวอย่าง โดยประกอบด้วยตัวแปรตาม $Y_1, Y_2, \ldots, Y_i$ เมื่อ $i = 1, 2, \ldots, n$ และตัวแปรต้น X_j เมื่อ $j = 1, 2, \ldots, p$

ฟังก์ชันเชื่อมโยง (Link Function) เป็นฟังก์ชันอธิบายความสัมพันธ์ระหว่างองค์ประกอบแบบสุ่มและองค์ประกอบเชิงระบบ โดยฟังก์ชันเชื่อมโยง $(g(\cdot))$ มีคุณสมบัติสามารถหาอนุพันธ์ได้ (Differentiable) และเป็นฟังก์ชันทางเดียว (Monotonic) โดยฟังก์ชันเชื่อมโยงสามารถอธิบายด้วยค่าเฉลี่ย $E(Y_i) \equiv \mu_i$ ซึ่งขึ้นอยู่กับตัวทำนายเชิงเส้น

$$g(\mu_i) = \eta_i \tag{2-16}$$

ฟังก์ชันเชื่อมโยงสามารถเขียนในรูปแบบผกผันได้ดังนี้

$$\mu_i = g^{-1}(\eta_i) \tag{2-17}$$

ตารางที่ 2.4 รูปแบบของการกระจายตัวของตัวแปรตามและฟังก์ชันเชื่อมโยง

การกระจาย	ลักษณะข้อมูล	ชื่อฟังก์ชันเชื่อมโยง	ฟังก์ชันเชื่อมโยง $g(\mu_i) = \eta_i$	การผกผันของฟังก์ชันเชื่อมโยง $\mu_i = g^{-1}(\eta_i)$
Normal	real: $(-\infty, \infty)$	Identity	μ_i	n_i
Gamma	real: $(0, +\infty)$	Inverse	μ_i^{-1}	n_i^{-1}
Inverse Gaussian	real: $(0, +\infty)$	Inverse squared	μ_i^{-2}	$n_i^{-\frac{1}{2}}$
Poisson	integer: 0, 1, 2, …	Log	$\ln(\mu)$	$\exp(n_i)$

โมเดลเชิงเส้นโดยนัยทั่วไปจะทำการตรวจสอบการกระจายตัวของตัวแปรตาม (Y_i) จากนั้นเขียนสมการเชิงเส้นที่นำตัวแปรต้นมาใช้ในการพยากรณ์ตัวแปรตามแล้วเลือกใช้ฟังก์ชันเชื่อมโยงให้เหมาะสมกับการกระจายตัวของตัวแปรตาม รูปแบบการกระจายและฟังก์ชันเชื่อมโยงสรุปได้ดังตารางที่ 2.4

ตัวอย่างเช่น สมการเชิงเส้นคือ Y=2+6a+3b ถ้าการกระจายตัวของตัวแปรตามเป็นแบบ Normal จะได้โมเดลเชิงเส้นโดยนัยทั่วไปคือ Y = 2+6a+3b และถ้าการกระจายตัวของตัวแปรตามเป็นแบบ Gamma จะได้โมเดลเชิงเส้นโดยนัยทั่วไปคือ Y = -2-6a-3b เป็นต้น

2.4 โครงข่ายประสาทเทียม (Artificial Neural Network: ANN)

โครงข่ายประสาทเทียม (Artificial Neural Network) เป็นการจำลองการทำงานของเครือข่ายประสาทในสมองมนุษย์ด้วยโปรแกรมคอมพิวเตอร์ซึ่งสามารถปรับเปลี่ยนตัวเองต่อการตอบสนองของค่าอินพุตตามกฎการเรียนรู้ (Learning Rule) หลังจากที่เครือข่ายได้เรียนรู้สิ่งที่ต้องการ เครือข่ายนั้นสามารถทำงานที่กำหนดไว้ได้ โครงข่ายประสาทเทียมเป็นแนวความคิดที่ต้องการให้คอมพิวเตอร์มีความสามารถในการเรียนรู้ คิดวิเคราะห์ และตัดสินใจได้เหมือนมนุษย์ สมองมนุษย์มีนิวรอน (Neuron) หรือเซลล์ประสาท ประมาณ 10^{11} นิวรอนเชื่อมต่อกันเป็นโครงข่ายขนาดใหญ่ จึงสามารถกล่าวได้ว่าสมองมนุษย์เป็นคอมพิวเตอร์ที่มีการปรับตัวเอง (Adaptive) ไม่เป็นเชิงเส้น (Nonlinear) และมีการทำงานแบบขนาน (Parallel) ในการจัดการการทำงานร่วมกันของนิวรอน

โครงข่ายประสาทเทียมมีคุณลักษณะเด่นคือ ประกอบด้วยหน่วยประมวลผลย่อย ๆ ซึ่งเชื่อมต่อแบบขนานเป็นจำนวนมาก มีหน่วยประมวลผลย่อยแต่ละหน่วยที่มีโครงสร้างง่าย ๆ ที่ดูเสมือนมีความสามารถต่ำ แต่เมื่อหน่วยประมวลผลย่อยทำงานร่วมกันแบบกระจายทำให้โครงข่ายประสาทเทียมมีการทำงานที่มีประสิทธิภาพ ข้อดีของการเชื่อมต่อด้วยหน่วยประมวลผลย่อย ๆ จำนวนมาก ถ้าเครือข่ายบางส่วนเกิดความเสียหาย โครงข่ายประสาทเทียมนั้นจะยังคงสามารถทำงานได้ คุณสมบัติที่เด่นสุดคือ สามารถเรียนรู้และแก้ไขปัญหาได้อย่างมีประสิทธิภาพ ผลจากการเรียนรู้ด้วยตัวอย่างข้อมูลบางส่วนนำไปสู่การตอบสนองต่อข้อมูลอินพุตที่เข้ามาใหม่

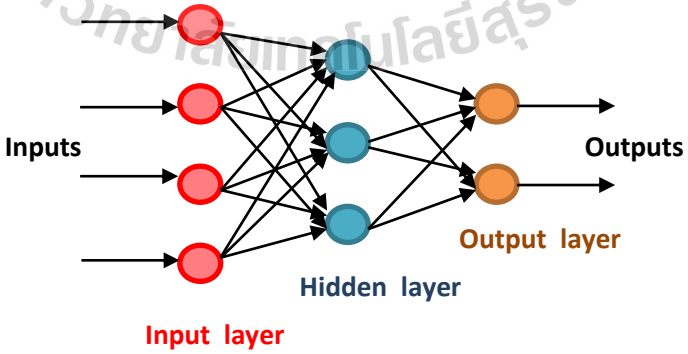
โครงข่ายประสาทเทียมประกอบไปด้วยเซตของโหนดและเส้นเชื่อมระหว่างโหนด แสดงดังรูปที่ 2.8 โดยโหนดแบ่งเป็น 3 ระดับ ได้แก่ ชั้นอินพุต (Input Layer) ชั้นซ่อน(Hidden Layer) และชั้นเอาท์พุต(Output Layer) ในแต่ละชั้นซ่อนอาจจะมีได้มากกว่า 1 ชั้น

โหนดที่อยู่ในชั้นอินพุตเรียกว่า Input node จำนวนโหนดในชั้นอินพุตจะเท่ากับจำนวนคุณสมบัติ (Attribute) ที่ใช้อธิบายข้อมูลแต่ละตัว โดยโหนดในชั้นนี้จะไม่มีการคำนวณดังนั้นข้อมูลที่ออกจากชั้นนี้จะไม่มีการแปลง

โหนดที่อยู่ในชั้นซ่อนเรียกว่า Hidden Node จำนวนชั้นและจำนวนโหนดจะขึ้นอยู่กับผู้ออกแบบโครงข่ายโดยอาจจะต้องทดลองหลาย ๆ แบบแล้วดูว่าแบบใดให้ประสิทธิภาพที่ดีที่สุด

โหนดที่อยู่ในชั้นเอาท์พุตเรียกว่า Output Node จำนวนโหนดจะเท่ากับจำนวนกลุ่มหรือจำนวนประเภทของข้อมูลที่ต้องการจำแนก

ในโครงข่ายจะมีเส้นเชื่อมจากทุกโหนดในชั้นอินพุตไปยังทุกโหนดในชั้นซ่อนและมีเส้นเชื่อมจากทุกโหนดในชั้นซ่อนไปยังทุกโหนดในชั้นเอาท์พุต เส้นเชื่อมแต่ละเส้นจะมีค่าน้ำหนัก (Weight)

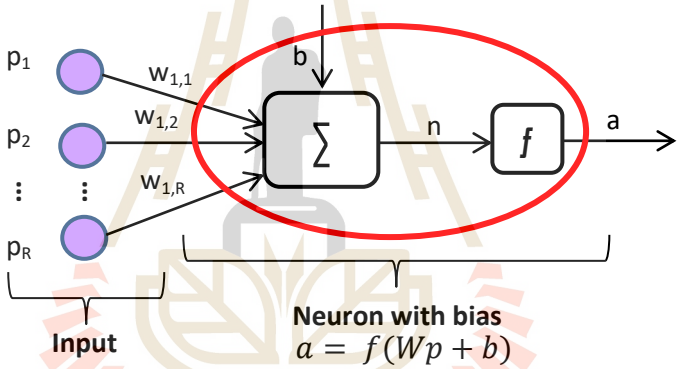


รูปที่ 2.8 สถาปัตยกรรมของโครงข่ายประสาทเทียม

การทำงานของแต่ละโหนดเทียบได้กับเซลล์ประสาทในสมองมนุษย์ 1 เซลล์ อินพุตที่เข้าสู่โหนดจะเป็นเวกเตอร์ของคุณสมบัติของข้อมูลตัวอย่างมีค่า $p = [p_1,p_2,...,p_R]$ ซึ่งเป็นค่าอินพุตที่ถูกป้อนมีจำนวน R องค์ประกอบ และเวกเตอร์น้ำหนัก $W=[w_1,w_2,...,w_R]$ มีค่าเอนเอียงหรือไบแอส b นำอินพุตมาคูณกับน้ำหนักของแต่ละเส้นเชื่อม ผลที่ได้จากอินพุตทุก ๆ เส้นเชื่อมของโหนดจะเอามารวมกันและรวมกับค่าไบแอสแล้วส่งต่อไปยังฟังก์ชันถ่ายโอน (Transfer Function) ทำให้เกิดเป็นค่าเอาท์พุต a ในที่นี้ f เป็นฟังก์ชันถ่ายโอนทำหน้าที่รับค่าอินพุต n เพื่อเปลี่ยนเป็นค่าเอาท์พุต a และค่าเอาท์พุต a สามารถคำนวณได้จากสมการ (Howard et al., 2000)

$$a = f(n) = f(Wp + b) \tag{2-18}$$

เมื่อ $n = w_{1,1}p_1 + w_{1,2}p_2 + \cdots + w_{1,R}p_R + b = Wp + b$

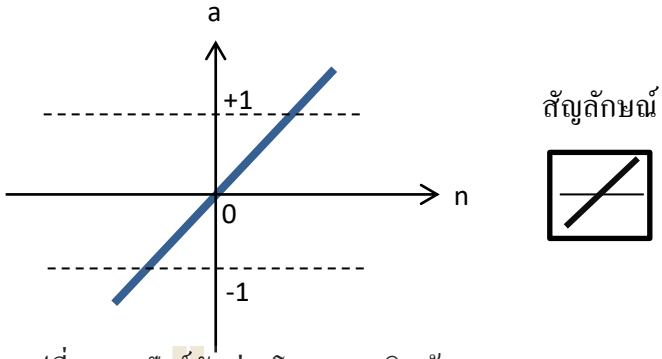


รูปที่ 2.9 โครงขายประสาทเทียมหนึ่งหน่วยแบบหลายอินพุต

2.4.1 ฟังก์ชันถ่ายโอน (Transfer Function)

ฟังก์ชันถ่ายโอนหรือฟังก์ชันกระตุ้น (Activation Function) เป็นฟังก์ชันที่ทำหน้าที่กำหนดค่าผลลัพธ์ของโครงข่ายประสาทเทียม ฟังก์ชันถ่ายโอนมีหลายแบบ การเลือกใช้ฟังก์ชันถ่ายโอนจะพิจารณาจากลักษณะผลลัพธ์ว่าเป็นค่าต่อเนื่องหรือไม่ และขอบเขตของผลลัพธ์เป็นอย่างไร โดยฟังก์ชั่นถ่ายโอนแบบต่าง ๆ มีดังนี้

1. ฟังก์ชันถ่ายโอนแบบเชิงเส้น (Linear Transfer Function)

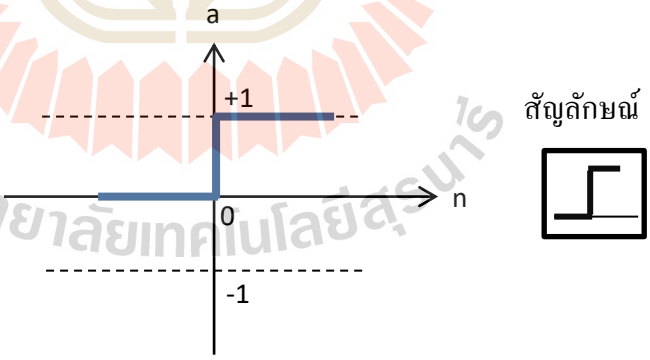


รูปที่ 2.10 ฟังก์ชันถ่ายโอนแบบเชิงเส้น

ค่าผลลัพธ์ของนิวรอนของฟังก์ชันถ่ายโอนแบบเชิงเส้นคือ a = n โดยที่ขอบเขตของ a คือ [-∞,∞] ตัวอย่างฟังก์ชันถ่ายโอนแบบเชิงเส้นแสดงดังรูปที่ 2.10 และสามารถเขียนในรูปสมการได้ดังนี้

$$a = \begin{cases} n, \text{ถ้า } n > 0 \\ 0, \text{ถ้า } n = 0 \\ -n, \text{ถ้า } n < 0 \end{cases} \qquad (2\text{-}19)$$

2. ฟังก์ชันถ่ายโอนแบบฮาร์ดลิมิต (Hard Limit Transfer Function)

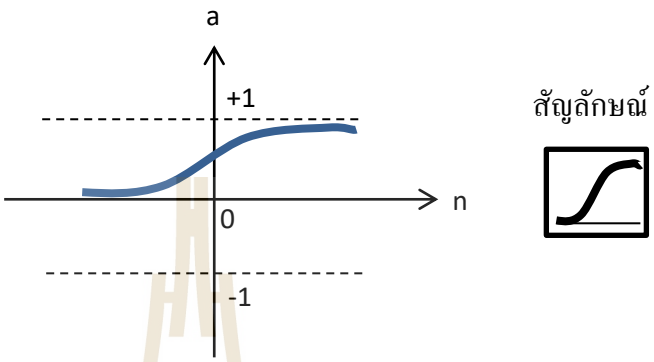


รูปที่ 2.11 ฟังก์ชันถ่ายโอนแบบฮาร์ดลิมิต

ค่าผลลัพธ์ของนิวรอนของฟังก์ชันถ่ายโอนแบบฮาร์ดลิมิตจะเป็นได้สองค่าคือ ถ้าอินพุต n ของนิวรอนมีค่าน้อยกว่า 0 ค่าผลลัพธ์หรือเอาท์พุต a มีค่าเป็น 0 แต่ถ้า n มีค่าเท่ากับ 0 หรือมากกว่า 0 ค่าผลลัพธ์คือ 1 โดยที่ขอบเขตของ a คือ [0,1] ลักษณะฮาร์ดลิมิตแสดงดังรูปที่ 2.11 และสามารถเขียนในรูปสมการได้ดังนี้

$$a = \begin{cases} 0, \text{ถ้า } n < 0 \\ 1, \text{ถ้า } n \geq 0 \end{cases} \quad (2\text{-}20)$$

3. ฟังก์ชันถ่ายโอนแบบล็อก-ซิกมอยด์ (Log Sigmoid Transfer Function)

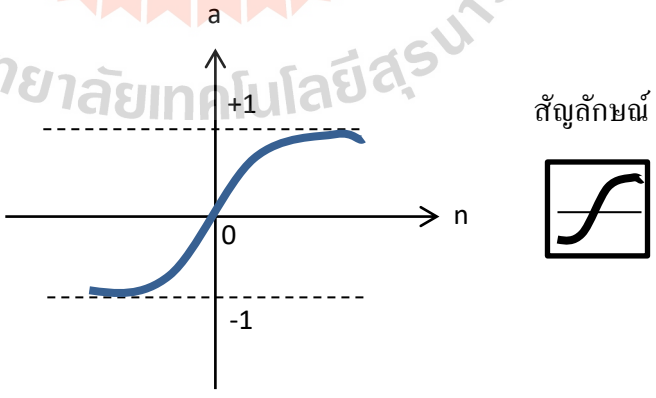


รูปที่ 2.12 ฟังก์ชันถ่ายโอนแบบล็อก-ซิกมอยด์

ค่าผลลัพธ์ของนิวรอนของฟังก์ชันถ่ายโอนแบบล็อก-ซิกมอยด์เป็นค่าต่อเนื่องแบบไม่เป็นเชิงเส้นที่เป็นเฉพาะค่าบวกดังรูปที่ 2.12 และสามารถคำนวณได้ดังสมการที่ 2-21 โดยที่ขอบเขตของ a คือ [0,1]

$$a = \frac{1}{1+e^{-n}} \quad (2\text{-}21)$$

4. ฟังก์ชันถ่ายโอนแบบแทนเจนท์-ซิกมอยด์ (Tan Sigmoid Transfer Function)

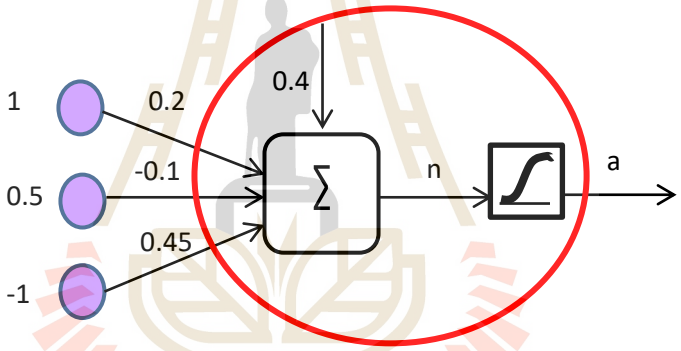


รูปที่ 2.10 ฟังก์ชันถ่ายโอนแบบแทนเจนท์-ซิกมอยด์

ค่าผลลัพธ์ของนิวรอนของฟังก์ชันถ่ายโอนแบบแทนเจนท์-ซิกมอยด์เป็นค่าต่อเนื่องแบบไม่เป็นเชิงเส้นที่เป็นได้ทั้งค่าบวกและค่าลบดังรูปที่ 2.13 และสามารถคำนวณได้ดังสมการที่ 2-22 โดยที่ขอบเขตของ a คือ [-1,1]

$$a = \frac{e^{n}-e^{-n}}{e^{n}+e^{-n}} \qquad (2\text{-}22)$$

ตัวอย่างการคำนวณภายในโหนดของโครงข่ายประสาทเทียมจำนวน 1 โหนด โดยกำหนดให้ข้อมูลอินพุต 1 แถวหรือเรคคอร์ด มี 3 ค่าแอตทริบิวต์ได้แก่ $p_1 = 1$, $p_2 = 0.5$ และ $p_3 = -1$ และเวกเตอร์น้ำหนักแต่ละเส้นเชื่อม $w_{1,1} = 0.2$, $w_{1,2} = -0.1$ และ $w_{1,3} = 0.45$ และค่าไบแอสมีค่า b = 0.4 โดยโหนดนี้ใช้ฟังก์ชันถ่ายโอนแบบล็อก-ซิกมอยด์ โครงสร้างของโหนดและค่าพารามิเตอร์ต่าง ๆ แสดงได้ดังรูปที่ 2.14



รูปที่ 2.14 ค่าข้อมูลอินพุต ค่าไบแอส และเวกเตอร์น้ำหนักบนแต่ละเส้นเชื่อมของโหนด

แสดงการคำนวณผลลัพธ์ที่ได้จากการทำงานของโหนดในโครงข่ายประสาทเทียมได้ดังนี้

หาค่า n จากสูตร $n = w_{1,1}p_1 + w_{1,1}p_2 + \cdots + w_{1,R}p_R + b$

แทนค่า n = (0.2 * 1) + (-0.1 * 0.5) + (0.45 * -1) + 0.4

= 0.2 + (-0.05) + (-0.45) + 0.4

= 0.1

หาค่า a จากสูตร $a = \frac{1}{1+e^{-n}}$

แทนค่า $a = \frac{1}{1+e^{-0.1}}$

$= \frac{1}{1+0.905}$

$$= \frac{1}{1.905}$$

$$= 0.525$$

ดังนั้นค่าผลลัพธ์ของโหนด คือ 0.525

2.4.2 การปรับพารามิเตอร์เพื่อให้โครงข่ายประสาทเทียมจดจำสิ่งที่เรียนรู้

สำหรับค่าน้ำหนัก w และค่าไบแอส b เป็นค่าพารามิเตอร์ที่สามารถปรับค่าได้ เพื่อให้โครงข่ายประสาทเทียมจดจำสิ่งที่เรียนรู้ การสอนโครงข่ายประสาทเทียมให้เรียนรู้คือการค้นหาค่าน้ำหนักของเส้นเชื่อมแต่ละเส้นที่เหมาะสมที่ทำให้สามารถจำแนกประเภทของข้อมูลตัวอย่างที่ใช้สอน (Training Data) ได้ถูกต้องมากที่สุด สิ่งสำคัญคือต้องทราบค่า Weight สำหรับสิ่งที่เราต้องการเพื่อให้คอมพิวเตอร์รู้จำซึ่งเป็นค่าที่ไม่แน่นอน แต่สามารถกำหนดให้คอมพิวเตอร์ปรับค่าเหล่านั้นได้โดยการสอนให้โครงข่ายประสาทเทียมรู้จักรูปแบบ (Pattern) ของสิ่งที่ต้องการให้รู้จำ ถ้าโครงข่ายประสาทเทียมให้ค่าเอาท์พุตผิด จะทำการปรับค่าน้ำหนักจนกว่าค่าความผิดพลาดจะน้อยลงหรืออยู่ในเกณฑ์ที่ยอมรับ สามารถคำนวณค่าความผิดพลาดดังสมการต่อไปนี้

$$e = t - y \tag{2-24}$$

เมื่อ t = ค่าเป้าหมาย (Target) หรือ คลาส (Class) ซึ่งเป็นค่าที่ถูกต้อง

y = ค่าเอาท์พุตจากโครงข่ายประสาทเทียม

ค่าน้ำหนักและ bias ที่ถูกปรับปรุงสามารถคำนวณโดยใช้สมการที่ (2-25) และ (2-26) แสดงดังต่อไปนี้

$$w^{new} = w^{old} + ep \tag{2-25}$$

$$b^{new} = b^{old} + e \tag{2-26}$$

เมื่อ w^{new} = ค่าน้ำหนักใหม่

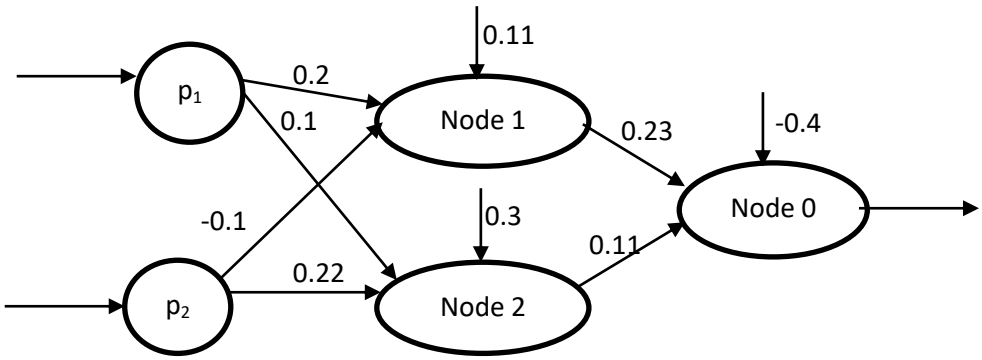
w^{old} = ค่าน้ำหนักเก่า

b^{new} = ค่าไบแอสใหม่

b^{old} = ค่าไบแอสเก่า

p = ข้อมูล

e = ค่าความผิดพลาด



รูปที่ 2.15 ค่าน้ำหนักและค่าไบแอสเริ่มต้นของโครงข่ายประสาทเทียมแบบ 2-2-1

ตารางที่ 2.5 ตัวอย่างข้อมูลที่ใช้ในการเรียนรู้ของโครงข่ายประสาทเทียมแบบ 2 อินพุต 2 โหนดในชั้นซ่อนและ 1 เอาท์พุท (2-2-1)

ข้อมูลที่	p_1	p_2	เป้าหมาย
1	0.5	-1	0.1
2	0.5	-0.8	0.08

การคำนวณโครงข่ายประสาทเทียมโดยค่าข้อมูลที่ใช้ในการเรียนรู้แสดงในตารางที่ 2.5 ภายในโหนดของโครงข่ายประสาทเทียมใช้ฟังก์ชันถ่ายโอนแบบล็อก-ซิกมอยด์ กำหนดค่าน้ำหนักเริ่มต้นดังแสดงในรูป 2.15 คำนวณได้ดังนี้

ข้อมูลอินพุตที่ 1 มีค่า $p_1 = 0.5, p_2 = -1$

Node 1

n = (0.2 * 0.5) + (-0.1 * -1) + 0.11

= 0.1 + 0.1 + 0.11

= 0.31

a = $\frac{1}{1+e^{-0.31}}$

= 0.58

Node 2

n = (0.1 * 0.5) + (0.22 * -1) + 0.3

= 0.05 - 0.22 + 0.3

= 0.13

a $= \frac{1}{1+e^{-0.13}}$

$= 0.53$

Node 0

n $= (0.23 * 0.58) + (0.11 * 0.53) - 0.4$

$= 0.13 + 0.06 – 0.4$

$= -0.21$

a $= \frac{1}{1+e^{-(-0.21)}}$

$= 0.45$

คำนวณค่าความผิดพลาด $e = t - y$

จะได้ $e = 0.1 - 0.45$

$= -0.35$

ปรับค่าน้ำหนักจากสมการ $w^{new} = w^{old} + ep$

$w_{1,1} = 0.2 + (-0.35 * 0.5) = 0.025$

$w_{1,2} = -0.1 + (-0.35 * -1) = 0.25$

$w_{2,1} = 0.1 + (-0.35 * 0.5) = -0.075$

$w_{2,2} = 0.22 + (-0.35 * -1) = 0.57$

$w_{0,1} = 0.23 + (-0.35 * 0.58) = 0.027$

$w_{0,2} = 0.11 + (-0.35 * 0.53) = -0.076$

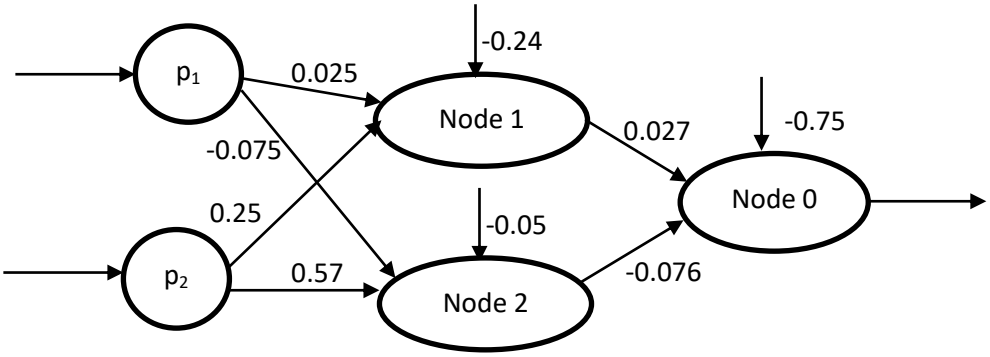
ปรับค่าไบแอสจากสมการ $b^{new} = b^{old} + e$

$b_1 = 0.11 + (-0.35) = -0.24$

$b_2 = 0.3 + (-0.35) = -0.05$

$b_0 = -0.4 + (-0.35) = -0.75$

แสดงค่าน้ำหนักและค่าไบแอสที่ถูกปรับจากการเรียนรู้ข้อมูลอินพุตที่ 1 ของโครงข่ายประสาทเทียมแบบ 2-2-1 ได้ดังรูป 2.16



รูปที่ 2.16 ค่าน้ำหนักและค่าไบแอสที่ปรับค่าครั้งที่ 1 ของโครงข่ายประสาทเทียมแบบ 2-2-1

ข้อมูลอินพุตที่ 2 มีค่า $p_1 = 0.5$, $p_2 = -0.8$

Node 1

n = (0.025 * 0.5) + (0.25 * -0.8) – 0.24

= 0.0125 - 0.2 – 0.24

= -0.43

a = $\frac{1}{1+e^{-(-0.43)}}$

= 0.39

Node 2

n = (-0.075 * 0.5) + (0.57 * -0.8) – 0.05

= -0.0375 - 0.456 - 0.05

= -0.54

a = $\frac{1}{1+e^{-(-0.54)}}$

= 0.37

Node 0

n = (0.027 * 0.39) + (-0.076 * 0.37) - 0.75

= 0.0105 - 0.0281 – 0.75

= -0.77

a = $\frac{1}{1+e^{-(-0.77)}}$

= 0.32

คำนวณค่าความผิดพลาด $e = t - y$

จะได้ $e =$ 0.08 - 0 = -0.24

ปรับค่าน้ำหนักจากสมการ $w^{new} = w^{old} + ep$

$w_{1,1}$ = 0.025 + (-0.24* 0.5) = -0.095

$w_{1,2}$ = 0.25 + (-0.24* -0.8) = 0.442

$w_{2,1}$ = -0.075 + (-0.24* 0.5) = -0.195

$w_{2,2}$ = 0.57 + (-0.24* -0.8) = 0.762

$w_{0,1}$ = 0.027+ (-0.24* 0.39) = -0.067

$w_{0,2}$ = -0.076+ (-0.24* 0.37) = -0.165

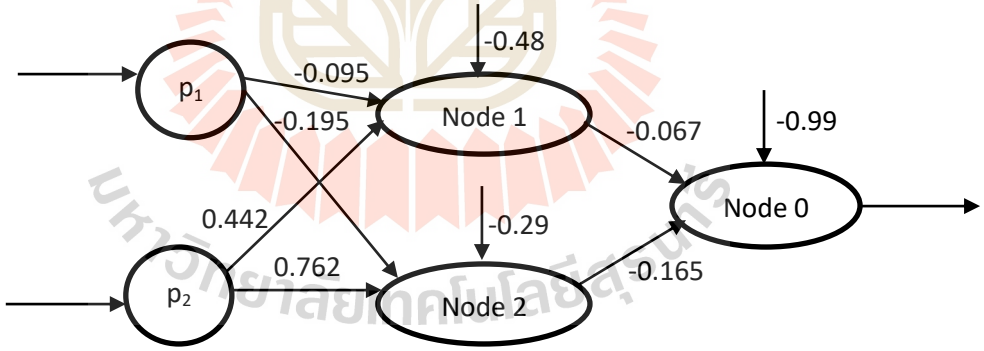
ปรับค่าไบแอสจากสมการ $b^{new} = b^{old} + e$

b_1 = -0.24 + (-0.24) = -0.48

b_2 = -0.05 + (-0.24) = -0.29

b_0 = -0.75 + (-0.24) = -0.99

แสดงค่าน้ำหนักและค่าไบแอสที่ถูกปรับจากการเรียนรู้ข้อมูลอินพุตที่ 2 ของโครงข่ายประสาทเทียมแบบ 2-2-1 ได้ดังรูปที่ 2.17



รูปที่ 2.17 ค่าน้ำหนักและค่าไบแอสที่ปรับค่าครั้งที่ 2 ของโครงข่ายประสาทเทียมแบบ 2-2-1

2.4.3 โครงข่ายประสาทเทียมแบบการแพร่ย้อนกลับ (Back-propagation ANN)

โครงข่ายประสาทเทียมแบบการแพร่ย้อนกลับ เป็นโครงข่ายประสาทเทียมที่ใช้กระบวนการย้อนกลับของการรู้จำในการฝึก Feed-Forward Neural Networks โดยจะมีการใช้อัลกอริทึมการแพร่ย้อนกลับ เพื่อช่วยในการปรับปรุงน้ำหนักคะแนนของเครือข่าย (Network Weight) หลังจากใส่รูปแบบข้อมูลสำหรับฝึกให้แก่เครือข่ายในแต่ละครั้งแล้ว ค่าที่ได้รับ (Output)

จากเครือข่ายจะถูกนำไปเปรียบเทียบกับผลที่คาดหวัง (Target) แล้วคำนวณหาค่าความผิดพลาด ซึ่งค่าความผิดพลาดนี้จะถูกส่งกลับเข้าสู่เครือข่ายเพื่อใช้แก้ไขค่าน้ำหนักคะแนนต่อไป

กระบวนการเรียนรู้และปรับปรุงแก้ไขนั้นเปลี่ยนไปแบบอัตโนมัติ ถ้าโครงข่ายประสาทเทียมให้คำตอบผิด ค่าน้ำหนักและค่าไบแอสจะถูกปรับจนกว่าจะมีค่าน้อยลงหรืออยู่ในเกณฑ์ที่ยอมรับได้ โครงข่ายประสาทเทียมมีลักษณะเป็นชั้น แต่ละชั้นเชื่อมต่อกันโดยโหนดในชั้นอินพุตมีเส้นเชื่อมไปยังทุกโหนดในชั้นซ่อนและโหนดในชั้นซ่อนจะมีเส้นเชื่อมไปยังทุกโหนดในชั้นเอาท์พุต เมื่อโครงข่ายประสาทเทียมได้รับข้อมูลป้อนเข้าในชั้นอินพุตแล้วจะส่งข้อมูลป้อนเข้าไปยังชั้นซ่อน แต่ละโหนดในชั้นซ่อนจะคำนวณค่าผลรวมน้ำหนักคูณด้วยข้อมูลเข้าและบวกด้วยค่าไบแอสแล้วส่งต่อไปยังฟังก์ชันถ่ายโอนเพื่อให้ได้เอาท์พุตในแต่ละโหนดในชั้นซ่อน แล้วส่งข้อมูลเอาท์พุตจากชั้นซ่อนไปยังชั้นเอาท์พุต แต่ละโหนดชั้นเอาท์พุตจะทำการคำนวณค่าเอาท์พุตเหมือนในโหนดชั้นซ่อน แล้วคำนวณค่าผลต่างระหว่างค่าเป้าหมายกับค่าผลลัพธ์ที่ได้จากโครงข่าย โครงข่ายจะมีการปรับค่าความผิดพลาดจากชั้นเอาท์พุต แล้วแพร่ย้อนกลับไปยังชั้นซ้อนจากนั้นแพร่ย้อนไปยังชั้นอินพุต

นอกจากนี้ยังมีอีกสองพารามิเตอร์ที่เป็นค่าคงที่ที่ผู้ใช้จำเป็นต้องกำหนดค่าได้แก่ อัตราการเรียนรู้ (Learning Rate: η) ซึ่งเป็นค่าที่จะบอกว่าจะปรับค่าน้ำหนักไปมากน้อยเพียงใด ถ้ากำหนดค่าอัตราการเรียนรู้มีค่ามากแสดงว่าแต่ละรอบของการเรียนรู้ค่าน้ำหนักจะถูกปรับไปมาก ซึ่งก็อาจจะใช้เวลาฝึกหรือเรียนรู้จากข้อมูลน้อยลง แต่บางครั้งถ้าค่าน้ำหนักถูกปรับมากเกินไปก็อาจจะทำให้ข้ามค่าน้ำหนักที่ดีได้ ในบางครั้งยิ่งเรียนรู้จากข้อมูลค่าความผิดพลาดยิ่งเพิ่มจึงจำเป็นจะต้องมีค่าโมเมนตัม (Momentum: $\propto$) ซึ่งเป็นพารามิเตอร์ที่ใช้ในการปรับค่าน้ำหนักเช่นเดียวกัน จะทำหน้าที่เพื่อเหนี่ยวนำค่าเอาท์พุตที่ได้ในแต่ละรอบ ปกติถ้าไม่มีค่านี้บางครั้งขณะเรียนรู้จากข้อมูล ค่าผิดพลาดเริ่มลดลงตามลำดับไปเรื่อย ๆ อาจมีค่ากลับเพิ่มขึ้นมาได้ ถ้าใช้ค่าโมเมนตัมจะสามารถช่วยควบคุมให้การเรียนรู้ไปในแนวทางเดียวกัน โดยเมื่อค่าผิดพลาดยิ่งมากขึ้นค่าน้ำหนักของอัตราการเรียนรู้ก็ควรจะกระโดดเร็วขึ้น หรือในทางกลับกันถ้าค่าผิดพลาดลดลงเรื่อย ๆ ก็ควรจะปรับค่าน้ำหนักให้น้อยลง

ค่าเอาท์พุตของแต่ละชั้นจะเป็นอินพุตให้กับชั้นถัดไป โครงข่ายในชั้นแรกคือชั้นอินพุตซึ่งจะรับค่าอินพุตโดยตรงจากภายนอกโครงข่าย ในชั้นนี้จะไม่มีการคำนวณค่าหรือเปลี่ยนแปลงค่าอินพุต นั่นคือ $S(x) = x$

อัลกอริทึมการแพร่ย้อนกลับจะใช้ข้อมูลคู่ระหว่างค่าของข้อมูลอินพุตกับค่าเป้าหมายป้อนให้โครงข่ายเรียนรู้ โดยกำหนดให้ Q เป็นจำนวนข้อมูลทั้งหมดของชุดข้อมูลฝึกสอน T ที่ประกอบด้วยคู่ของข้อมูล (X_k, D_k) โดยที่ k = 1, 2, …, Q เมื่อ X_k คือข้อมูลเข้าซึ่ง $X_k \in R^n$ ค่า n คือ

มิติของข้อมูลเข้าและ D_k คือค่าเป้าหมาย $D_k \in R^p$ ค่า p คือจำนวนเอาท์พุตของโครงข่าย ค่า q คือจำนวนโหนดในชั้นซ่อน สามารถเขียนรูปแบบความสัมพันธ์ได้ดังสมการต่อไปนี้ (Kumar, 2004)

$$T = \{(X_k, D_k)\}_{k=1}^{Q} \tag{2-27}$$

ขั้นตอนหลักในการทำงานของอัลกอริทึมการแพร่ย้อนกลับ อธิบายได้ดังต่อไปนี้

(1) ขั้นตอนการคำนวณค่าเอาท์พุต $\mathcal{S}(x)$ ที่ออกจากนิวรอน

1. คำนวณค่าเอาท์พุตของโหนดในชั้นอินพุท

$$\mathcal{S}(x_i^k) = x_i^k,\ i = 1, \dots, \mathrm{n} \tag{2-28}$$

$$\mathcal{S}(x_0^k) = x_0^k = 1 \tag{2-29}$$

เมื่อ i คือโหนดในชั้นอินพุต

n คือจำนวนโหนดทั้งหมดในชั้นอินพุต

x_i^k คือค่าแอตทริบิวต์ที่เข้าโหนด i ของข้อมูลอินพุต X_k ที่ป้อนเข้าในโครงข่าย

$\mathcal{S}(x_0^k)$ คือค่าเอาท์พุตของไบแอสในชั้นอินพุต

2. คำนวณค่าเอาท์พุตของโหนดในชั้นซ่อน

$$z_h^k = \sum_{i=0}^{n} w_{ih}^k \mathcal{S}(x_i^k) = \sum_{i=0}^{n} w_{ih}^k x_i^k,\ h = 1, \dots, \mathrm{q} \tag{2-30}$$

$$\mathcal{S}(z_h^k) = \frac{1}{1+\exp^{(-z_h^k)}},\ h = 1, \dots, \mathrm{q} \tag{2-31}$$

$$\mathcal{S}(z_0^k) = 1 \tag{2-32}$$

เมื่อ h คือโหนดในชั้นซ่อน

w_{0h}^k คือค่าน้ำหนักไบแอสของเส้นเชื่อมจากชั้นอินพุตไปยังชั้นซ่อนของโหนดในชั้นซ่อนที่ h

$\mathcal{S}(z_0^k)$ คือค่าเอาท์พุตของไบแอสในชั้นซ่อน

3. คำนวณค่าเอาท์พุตของโหนดในชั้นเอาท์พุต

$$y_j^k = \sum_{h=0}^{q} w_{hj}^k \mathcal{S}(z_h^k),\ j = 1, \dots, \mathrm{p} \tag{2-33}$$

$$\mathcal{S}(y_j^k) = \frac{1}{1+exp^{(-y_j^k)}},\ j = 1, \dots, \mathrm{p} \tag{2-34}$$

เมื่อ j คือโหนดในชั้นซ่อน

w_{0j}^{k} เป็นน้ำหนักของไบแอสของเส้นเชื่อมจากชั้นซ่อนไปยังชั้นเอาท์พุตของโหนดในชั้นเอาท์พุตที่ j

(2) ขั้นตอนการคำนวณค่าผิดพลาด δ และค่าน้ำหนัก w

1. คำนวณค่าผิดพลาดในชั้นเอาท์พุตและค่าน้ำหนักของเส้นเชื่อมจากชั้นซ่อนไปยังชั้นเอาท์พุต

$$\delta_j^k = \left(d_j^k - \mathcal{S}\left(y_j^k\right)\right)\mathcal{S}'\left(y_j^k\right), j = 1, \dots, \mathrm{p} \tag{2-35}$$

$$\Delta w_{hj}^k = \eta\delta_j^k\left(z_h^k\right), h = 0, \dots, \mathrm{q};\ j = 1, \dots, \mathrm{p} \tag{2-36}$$

เมื่อ $\mathcal{S}'\left(y_j^k\right) = \mathcal{S}\left(y_j^k\right)(1 - \mathcal{S}\left(y_j^k\right))$

2. คำนวณค่าผิดพลาดในชั้นซ่อนและค่าน้ำหนักของเส้นเชื่อมจากชั้นอินพุตไปยังชั้นซ่อน

$$\delta_h^k = \left(\sum_{j=1}^{p}\delta_j^k\, w_{hj}^k\right)\mathcal{S}'\left(z_h^k\right), \quad h = 0, \dots, \mathrm{q} \tag{2-37}$$

$$\Delta w_{ih}^k = \eta\delta_h^k(x_i^k) \quad i = 0, \dots, \mathrm{n};\ \mathrm{h} = 1, \dots, \mathrm{q} \tag{2-38}$$

เมื่อ $\mathcal{S}'\left(z_h^k\right) = \mathcal{S}\left(z_h^k\right)(1 - \mathcal{S}\left(z_h^k\right))$

(3) ขั้นตอนการคำนวณปรับค่าน้ำหนักของเส้นเชื่อม

1. คำนวณปรับค่าน้ำหนักของเส้นเชื่อมจากชั้นซ่อนไปยังชั้นเอาท์พุต

$$w_{hj}^{k+1} = w_{hj}^k + \Delta w_{hj}^k + \propto \Delta w_{hj}^{k-1} \tag{2-39}$$

เมื่อ $\Delta w_{hj}^0 = 0; h = 0, \dots, \mathrm{q};\ \mathrm{j} = 1, \dots, \mathrm{p}$

2. คำนวณปรับค่าน้ำหนักของเส้นเชื่อมจากชั้นอินพุตไปยังชั้นซ่อน

$$w_{ih}^{k+1} = w_{ih}^k + \Delta w_{ih}^k + \propto \Delta w_{ih}^{k-1} \tag{2-40}$$

เมื่อ $\Delta w_{ih}^0 = 0; i = 0, \dots, \mathrm{n};\ \mathrm{h} = 1, \dots, \mathrm{q}$

(4) ขั้นตอนการคำนวณความคลาดเคลื่อนรวมทั้งหมด

สามารถคำนวณค่าผิดพลาดรวมของแต่ละเอาท์พุตของนิวรอน ได้ดังนี้

$$E_k = D_k - \mathcal{S}(y_k) \tag{2-41}$$

เมื่อ $E_k = \left(e_1^k, e_2^k, \ldots, e_p^k\right)$

$= ((d_1^k - \mathcal{S}(y_1^k), \ldots, (d_p^k - \mathcal{S}(y_p^k))$

d = ค่าเป้าหมายของข้อมูล (target)

y = ค่าเอาท์พุตจากโครงข่าย

k = ข้อมูลเข้าตัวที่ k

p = จำนวนเอาท์พุตของโครงข่าย

รวมค่าความผิดพลาดทั้งหมดโดยใช้ผลรวมกำลังสองของความคลาดเคลื่อน (ε_k) คำนวณได้จากสมการ

$$\varepsilon_k = \frac{1}{2}\sum_{j=1}^{p}(d_j^k - \mathcal{S}(y_j^k))^2 \tag{2-42}$$

เมื่อ j = คือโหนดในชั้นเอาท์พุต

คำนวณค่าความผิดพลาดรวมทั้งหมดโดยใช้ผลรวมกำลังสองของความคลาดเคลื่อนแบบเฉลี่ย (ε_{av}) ซึ่งเป็นค่าความผิดพลาดเฉลี่ยของโครงข่ายจากการเรียนรู้จากข้อมูล k รอบ คำนวณได้จากสมการ

$$\varepsilon_{av} = \frac{1}{Q}\sum_{k=1}^{Q}\varepsilon_k \tag{2-43}$$

เมื่อ Q เป็นจำนวนข้อมูลทั้งหมด

การทำงานตามขั้นตอนอัลกอริทึมการแพร่ย้อนกลับที่อธิบายไว้ข้างต้น สามารถแสดงขั้นตอนแบบสรุปในรูปที่ 2.18 (ปรับปรุงจาก (Kumar, 2004))

ขั้นตอนวิธีการ : อัลกอริทึมการแพร่ย้อนกลับ (Back-propagation Algorithm)

กำหนดให้ : ข้อมูลฝึกสอน T ประกอบด้วยเวกเตอร์ $X_k \in R$ และค่าเอาท์พุตเวกเตอร์ $D_k \in R$ และ n-q-p เป็นค่าสถาปัตยกรรมของโครงข่ายประสาทเทียม

เริ่มต้น : สุ่มค่าน้ำหนักของชั้นอินพุตไปชั้นซ่อน w_{ih}^1 เป็นค่าน้อย ๆ ,

$$\Delta w_{ih}^0 = 0,\ i = 0, \dots, n;\ \ h = 1, \dots, q$$

สุ่มค่าน้ำหนักของชั้นซ่อนไปชั้นเอาท์พุต w_{hj}^1 เป็นค่าน้อย ๆ ,

$$\Delta w_{hj}^0 = 0,\ h = 0, \dots, q;\ \ h = 1, \dots, p$$

กำหนดค่า $k = 1, \eta, \propto$ และค่าความผิดพลาดที่รับได้คือ τ

วนรอบทำซ้ำ {

เลือกข้อมูลฝึกสอน $(X_k, D_k) \in T$

คำนวณสัญญาณไปข้างหน้า:

$$\mathcal{S}(x_i^k) = x_i^k, \qquad i = 1, \dots, \text{n}$$
$$\mathcal{S}(x_0^k) = 1$$
$$z_h^k = \sum_{i=0}^{n} w_{ih}^k x_i^k, \qquad h = 1, \dots, \text{q}$$
$$\mathcal{S}(z_h^k) = \frac{1}{1+\exp^{(-z_h^k)}}, \qquad h = 1, \dots, \text{q}$$
$$\mathcal{S}(z_0^k) = 1$$
$$y_j^k = \sum_{h=0}^{q} w_{hj}^k \mathcal{S}(z_h^k), \qquad j = 1, \dots, \text{p}$$
$$\mathcal{S}(y_j^k) = \frac{1}{1+\exp^{(-y_j^k)}}, \qquad j = 1, \dots, \text{p}$$

คำนวณเดลต้าหรือค่าความผิดพลาดที่นิวรอนชั้นเอาท์พุต

$$\delta_j^k = (d_j^k - \mathcal{S}(y_j^k))\mathcal{S}'(y_j^k) \qquad j = 1, \dots, \text{p}$$
$$\Delta w_{hj}^k = \eta \delta_j^k (z_h^k) \qquad h = 0, \dots, \text{q};\ \text{j} = 1, \dots, \text{p}$$

คำนวณเดลต้าหรือค่าความผิดพลาดที่นิวรอนชั้นซ่อน

$$\delta_h^k = (\sum_{j=1}^{p} \delta_j^k w_{hj}^k)\mathcal{S}'(z_h^k) \qquad h = 0, \dots, \text{q}$$
$$\Delta w_{ih}^k = \eta \delta_h^k (x_i^k) \qquad i = 0, \dots, \text{n};\ \text{h} = 1, \dots, \text{q}$$

ปรับปรุงค่าน้ำหนัก:

$$w_{hj}^{k+1} = w_{hj}^k + \Delta w_{hj}^k + \propto \Delta w_{hj}^{k-1} \qquad h = 0, \dots, \text{q};\ \text{j} = 1, \dots, \text{p}$$
$$w_{ih}^{k+1} = w_{ih}^k + \Delta w_{ih}^k + \propto \Delta w_{ih}^{k-1} \qquad i = 0, \dots, \text{n};\ \text{h} = 1, \dots, \text{q}$$

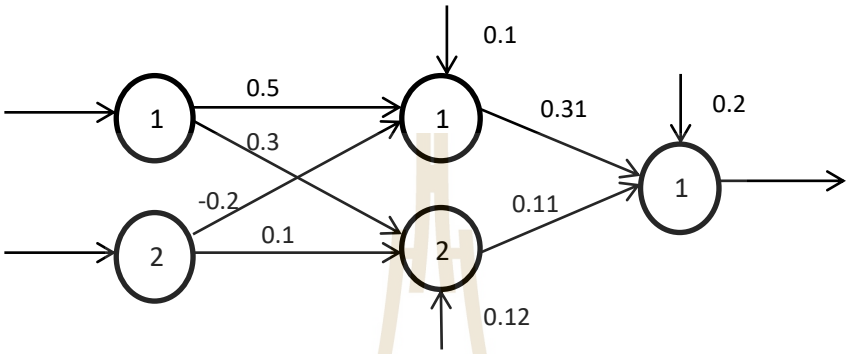
รวมค่าความผิดพลาด ε_k

} หยุดทำซ้ำจนกว่า $(\varepsilon_{av} = \frac{1}{Q}\sum_{k=1}^{Q} \varepsilon_k < \tau)$

รูปที่ 2.18 การทำงานของอัลกอริทึมการแพร่ย้อนกลับ

ตารางที่ 2.6 ข้อมูลฝึกสอนใช้ในการคำนวณตามตัวอย่างอัลกอริทึมการแพร่ย้อนกลับ

Pattern Index	x_1^k	x_2^k	d_1^k
1	0.5	-0.5	0.9



รูปที่ 2.19 โครงข่ายที่ใช้คำนวณอัลกอริทึมการแพร่ย้อนกลับจากการเรียนรู้ 1 รอบ

วิธีการคำนวณต่อไปนี้เป็นการแสดงวิธีคำนวณตามอัลกอริทึมการแพร่ย้อนกลับจากการเรียนรู้ 1 รอบ กำหนดค่าอัตราการเรียนรู้ $\eta = 1.2$ และค่าโมเมนตัม $\propto$=0.8 โดยกำหนดค่าน้ำหนักและค่าไบแอสของโครงข่ายตามรูปที่ 2.19 และใช้ข้อมูลฝึกสอนดังแสดงในตารางที่ 2.6

คำนวณโครงข่ายไปข้างหน้ารอบที่ 1 (k=1)

- **คำนวณค่าเอาท์พุตที่ออกจากโหนด**

ขั้นซ่อนใช้สูตร $z_h^k = \sum_{i=0}^{n} w_{ih}^k x_i^k$, $\mathcal{S}(z_h^k) = \frac{1}{1+\exp^{(-z_h^k)}}$

เมื่อ i = 0, 1, 2; h = 1, 2

โหนด 1 แทนค่า $z_1^1 = w_{01}^1 x_0^1 + w_{11}^1 x_1^1 + w_{21}^1 x_2^1$

$$= (0.1 * 1) + (0.5 * 0.5) + (-0.2 * -0.5) = 0.45$$

$$\mathcal{S}(z_1^1) = \frac{1}{1+\exp^{(-0.45)}} = 0.61$$

โหนด 2 แทนค่า $z_2^1 = w_{02}^1 x_0^1 + w_{12}^1 x_1^1 + w_{22}^1 x_2^1$

$$= (0.12 * 1) + (0.13 * 0.5) + (0.1 * -0.5) = 0.225$$

$$\mathcal{S}(z_2^1) = \frac{1}{1+\exp^{(-0.225)}} = 0.56$$

ชั้นเอาท์พุตใช้สูตร $y_j^k = \sum_{h=0}^{q} w_{hj}^k \mathcal{S}(z_h^k), \quad \mathcal{S}(y_j^k) = \frac{1}{1+exp^{(-y_j^k)}}$

เมื่อ j = 1; h = 0, 1, 2

โหนด 1 แทนค่า $y_1^1 = w_{01}^1 \mathcal{S}(z_0^1) + w_{11}^1 \mathcal{S}(z_1^1) + w_{21}^1 \mathcal{S}(z_2^1)$

$= (0.2 * 1) + (0.31 * 0.61) + (0.11 * 0.56)$

$= 0.27$

$$\mathcal{S}(y_1^1) = \frac{1}{1+\exp^{(-0.27)}} = 0.57$$

- **คำนวณค่าผิดพลาดในชั้นเอาท์พุต**

ใช้สูตร $\delta_j^k = (d_j^k - \mathcal{S}(y_j^k))\mathcal{S}(y_j^k)(1 - \mathcal{S}(y_j^k))$ เมื่อ j = 1

$\delta_1^1 = (d_1^1 - \mathcal{S}(y_1^1))\mathcal{S}(y_1^1)(1 - \mathcal{S}(y_1^1))$

$= (0.9 - 0.57) * 0.57 * (1 - 0.57)$

$= 0.08$

- **คำนวณค่าผิดพลาดในชั้นซ่อน**

ใช้สูตร $\delta_h^k = \left(\sum_{j=1}^{p} \delta_j^k w_{hj}^k\right)\mathcal{S}(z_h^k)\left(1 - \mathcal{S}(z_h^k)\right)$ เนื่องจากแทนค่าเข้าไปในสูตรโดยตรงอาจจะสับสน ใช้วิธีคำนวณทีละโหนดในชั้นซ่อนหรือดังนั้น $\delta_h^k = \delta H_h^k$ เมื่อ h = 1, 2

$\delta H_1^1 = (\delta_1^1 w_{11}^1)\mathcal{S}(z_1^1)\left(1 - \mathcal{S}(z_1^1)\right)$

$= (0.08 * 0.31) * 0.61 * (1 - 0.61)$

$= 0.006$

$\delta H_2^1 = (\delta_1^1 w_{21}^1)\mathcal{S}(z_2^1)(1 - \mathcal{S}(z_2^1))$

$= (0.08 * 0.11) * 0.56 * (1 - 0.56)$

$= 0.002$

- **คำนวณค่าน้ำหนักเส้นเชื่อมจากชั้นซ่อนไปชั้นเอาท์พุต**

ใช้สูตร $\Delta w_{hj}^k = \eta \delta_j^k (z_h^k)$

เมื่อ $\Delta w_{hj}^k = \Delta n_{hj}^k$; h = 0, 1, 2; j = 1

$\Delta n_{01}^1 = \eta \delta_1^1 (z_0^1) = 1.2 * 0.08 * 1 = 0.096$

$\Delta n_{11}^1 = \eta \delta_1^1 (z_1^1) = 1.2 * 0.08 * 0.61 = 0.059$

$\Delta n_{21}^1 = \eta \delta_1^1 (z_2^1) = 1.2 * 0.08 * 0.56 = 0.054$

- **คำนวณค่าน้ำหนักเส้นเชื่อมจากชั้นอินพุตไปชั้นซ่อน**

ใช้สูตร $\Delta w_{ih}^{k} = \eta \delta_{h}^{k}(x_{i}^{k})$

เมื่อ $\delta_{h}^{k} = \delta H_{h}^{k}$ และ $\Delta w_{ih}^{k} = \Delta m_{hj}^{k}$; i = 0, 1, 2; h = 1, 2

$$\Delta m_{01}^{1} = \eta \delta_{1}^{1}(x_{0}^{1}) = 1.2 * 0.06 * 1 = 0.072$$
$$\Delta m_{11}^{1} = \eta \delta_{1}^{1}(x_{1}^{1}) = 1.2 * 0.06 * 0.5 = 0.036$$
$$\Delta m_{21}^{1} = \eta \delta_{1}^{1}(x_{2}^{1}) = 1.2 * 0.06 * -0.5 = -0.036$$
$$\Delta m_{02}^{1} = \eta \delta_{2}^{1}(x_{0}^{1}) = 1.2 * 0.02 * 1 = 0.024$$
$$\Delta m_{12}^{1} = \eta \delta_{2}^{1}(x_{1}^{1}) = 1.2 * 0.02 * 0.5 = 0.012$$
$$\Delta m_{22}^{1} = \eta \delta_{2}^{1}(x_{2}^{1}) = 1.2 * 0.02 * -0.5 = -0.012$$

- **คำนวณปรับค่าน้ำหนักเส้นเชื่อมจากชั้นซ่อนไปชั้นเอาท์พุต**

ใช้สูตร $w_{hj}^{k+1} = w_{hj}^{k} + \Delta w_{hj}^{k} + \propto \Delta w_{hj}^{k-1}$

เมื่อ $w_{hj}^{k} = n_{hj}^{k}$; $\Delta w_{hj}^{0} = 0$; h = 0, 1, 2; j = 1

$$n_{01}^{2} = n_{01}^{1} + \Delta n_{01}^{1} + \propto \Delta n_{01}^{0} = 0.2 + 0.096 + (0.8 * 0) = 0.296$$
$$n_{11}^{2} = n_{11}^{1} + \Delta n_{11}^{1} + \propto \Delta n_{11}^{0} = 0.31 + 0.059 + (0.8 * 0) = 0.369$$
$$n_{21}^{2} = n_{21}^{1} + \Delta n_{21}^{1} + \propto \Delta n_{21}^{0} = 0.11 + 0.054 + (0.8 * 0) = 0.164$$

- **คำนวณปรับค่าน้ำหนักเส้นเชื่อมจากชั้นอินพุตไปชั้นซ่อน**

ใช้สูตร $w_{ih}^{k+1} = w_{ih}^{k} + \Delta w_{ih}^{k} + \propto \Delta w_{ih}^{k-1}$

เมื่อ $\Delta w_{ih}^{k} = \Delta m_{hj}^{k}$; $\Delta w_{ih}^{0} = 0$; i = 0, 1, 2; h =1, 2

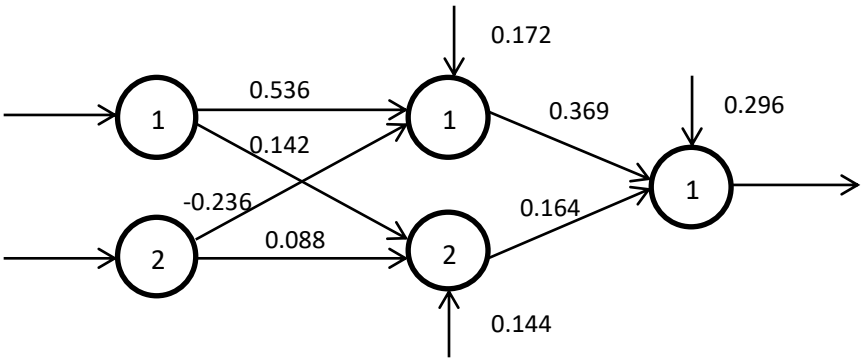
$$m_{01}^{2} = m_{01}^{1} + \Delta m_{01}^{1} + \propto \Delta m_{01}^{0} = 0.1 + 0.072 + (0.8 * 0) = 0.172$$
$$m_{11}^{2} = m_{11}^{1} + \Delta m_{11}^{1} + \propto \Delta m_{11}^{0} = 0.5 + 0.036 + (0.8 * 0) = 0.536$$
$$m_{21}^{2} = m_{21}^{1} + \Delta m_{21}^{1} + \propto \Delta m_{21}^{0} = (-0.2) + (-0.036) + (0.8 * 0) = -0.236$$
$$m_{02}^{2} = m_{02}^{1} + \Delta m_{02}^{1} + \propto \Delta m_{02}^{0} = 0.12 + 0.024 + (0.8 * 0) = 0.144$$
$$m_{12}^{2} = m_{12}^{1} + \Delta m_{12}^{1} + \propto \Delta m_{12}^{0} = 0.13 + 0.012 + (0.8 * 0) = 0.142$$
$$m_{22}^{2} = m_{22}^{1} + \Delta m_{22}^{1} + \propto \Delta m_{22}^{0} = 0.1 + (-0.012) + (0.8 * 0) = 0.088$$

- **คำนวณค่าความผิดพลาดกำลังสองของความคลาดเคลื่อน**

ใช้สูตร $\varepsilon_{k} = \frac{1}{2}\sum_{j=1}^{p}(d_{j}^{k} - \mathcal{S}(y_{j}^{k}))^{2}$ เมื่อ j = 1

$$\varepsilon_{k} = \frac{1}{2}\left(d_{1}^{1} - \mathcal{S}(y_{1}^{1})\right)^{2} = \frac{1}{2}(0.9 - 0.57)^{2} = 0.054$$

จากการเรียนรู้ข้อมูลที่ 1 โครงข่ายมีค่าผิดพลาดเท่ากับ 0.054 และโครงข่ายถูกปรับค่าน้ำหนักเส้นเชื่อมแสดงดังรูปที่ 2.20



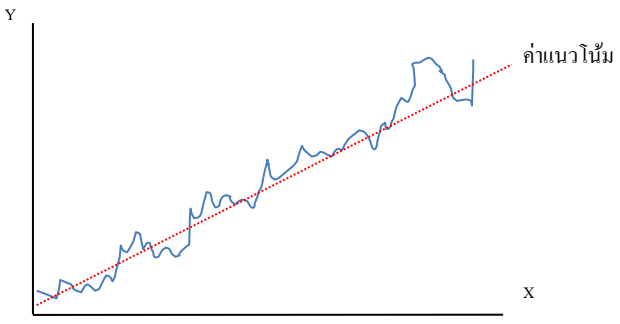
รูปที่ 2.20 โครงข่ายที่ใช้คำนวณอัลกอริทึมการแพร่ย้อนกลับที่ปรับใหม่จากวิธีการเรียนรู้ 1 รอบ

2.5 โมเดลอาริม่า (Autoregressive Integrated Moving Average Model: ARIMA)

ARIMA เป็นวิธีการสร้างแบบจำลองด้วยวิธีการ Box-Jenkins ที่เสนอโดย George E.P. Box และ Gwilym M. Jenkins ในปี ค.ศ. 1970 และได้รับการปรับปรุงในปี ค.ศ. 1994 ซึ่งในปัจจุบันเป็นวิธีการที่นิยมใช้อย่างมากในการวิเคราะห์ข้อมูลอนุกรมเวลา (จินดามาส สุทธิชัยเมธี, 2554) การพยากรณ์แบบ Box-Jenkins เป็นวิธีการวิเคราะห์อนุกรมเวลา (Time Series) ซึ่งหมายถึงการวิเคราะห์ข้อมูลที่มีการเปลี่ยนแปลงไปตามเวลา มีการเก็บต่อเนื่องในช่วงเวลาหนึ่ง ที่มีความถี่เท่า ๆ กัน เช่น รายวัน รายสัปดาห์ รายเดือน รายปี และข้อมูลที่นำมาวิเคราะห์ต้องเป็นลักษณะ คงที่ (Stationary) เพื่อให้แบบจำลองมีความแม่นยำในการทำนาย

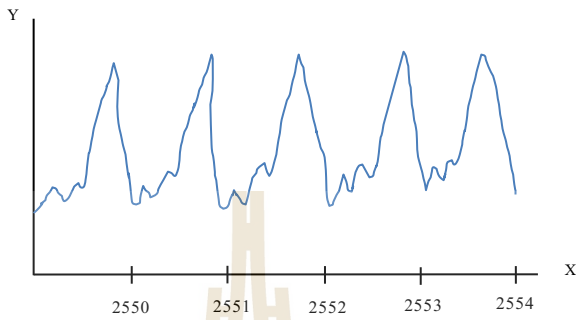
อนุกรมเวลามีส่วนประกอบหลัก 4 ส่วน ซึ่งเป็นสาเหตุของการแปรผันแบบต่าง ๆ ของข้อมูลอนุกรมเวลาซึ่งมีส่วนประกอบ ดังนี้

1. แนวโน้ม (Trend: T) เป็นค่าที่แสดงถึงการเคลื่อนไหวของข้อมูลในระยะยาว หรือการเปลี่ยนแปลงของข้อมูลในระยะยาวที่แสดงการเปลี่ยนแปลงแบบเพิ่มขึ้นหรือลดลง ลักษณะแนวโน้มอาจจะเป็นเส้นตรงหรือเส้นโค้งก็ได้ตัวอย่างแนวโน้มแสดงได้ดังรูปที่ 2.21



รูปที่ 2.21 ค่าแนวโน้มของอนุกรมเวลา

2. การแปรผันตามฤดูกาล (Seasonal: S) เป็นการเปลี่ยนแปลงของข้อมูลที่เกิดจากฤดูกาล ซึ่งในแต่ละปีจะเกิดขึ้นซ้ำ ๆ ในช่วงเวลาเดียวกัน โดยแต่ละรอบของฤดูกาลจะต้องจบภายใน 1 ปี เช่น ยอดขายของร้านสะดวกซื้อจะสูงใจช่วงต้นเดือน เป็นต้น



รูปที่ 2.22 การแปรผันตามฤดูกาลของอนุกรมเวลา

3. การแปรผันตามวัฏจักร (Cyclical: C) คือการเคลื่อนไหวของข้อมูลที่มีลักษณะเดิมซ้ำ ๆ กัน มีลักษณะคล้ายกับการแปรผันตามฤดูกาล แตกต่างกันที่ระยะเวลาของการเคลื่อนไหวของข้อมูลจะมีระยะเวลานานกว่า 1 ปี เช่น อาจจะใช้เวลา 3 ปีใน 1 รอบวัฏจักร เป็นต้น
4. การแปรผันเนื่องจากเหตุการณ์ไม่ปกติ (Irregular: I) เป็นการแปรผันที่มีการเคลื่อนไหวของข้อมูลที่มีรูปแบบเกิดขึ้นไม่แน่นอน ไม่สามารถคาดการณ์ได้ล่วงหน้า เช่น การชุมนุมประท้วง การประกาศนัดหยุดงาน เป็นต้น

ในการสร้างแบบจำลอง ARIMA ข้อมูลที่นำมาใช้ในการวิเคราะห์จะต้องเป็นข้อมูลที่มีลักษณะคงที่ (Stationary) คือข้อมูลที่มีค่าเฉลี่ย และความแปรปรวน มีค่าคงที่เท่ากันตลอดระยะเวลาที่ศึกษา ซึ่งค่าเหล่านี้จะขึ้นอยู่กับระยะหรือช่วงเวลา แบบจำลอง ARIMA(p, q, d) ประกอบด้วย 3 ส่วน ได้แก่ แบบจำลอง Auto Regressive (AR(p)) กระบวนการ Integrated (I(d)) และแบบจำลอง Moving Average (MA(q)) (กมลวรรณ สารพานิช, 2555) โดยรายละเอียดแต่ละส่วนมีลักษณะดังต่อไปนี้

1. แบบจำลอง Auto Regressive (AR(p)) เป็นรูปแบบที่แสดงว่าค่าสังเกต y_t ถูกกำหนดจากค่าสังเกตที่เกิดขึ้นก่อนหน้า p ค่า หรือ $y_{t-1}, y_{t-2}, \ldots, y_{t-p}$ โดยกระบวนการ Auto Regressive ที่มีค่าสังเกตที่อยู่ก่อนหน้า p ค่า หรือ AR(p) สามารถเขียนในรูปแบบสมการได้ดังนี้

$$\text{AR(p) คือ } y_t = \propto + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_p y_{t-p} + \varepsilon_t \qquad (2\text{-}36)$$

โดยที่ α คือ ค่าคงที่

β_j คือ พารามิเตอร์ตัวที่ j เมื่อ j=1, 2, .., p

ε_t คือค่าความคลาดเคลื่อน ณ เวลา t

ในกรณี AR(1) สามารถเขียนในรูปแบบสมการดังนี้

$$y_t = \propto + \beta_1 y_{t-1} + \varepsilon_t \quad (2\text{-}37)$$

โดย กำหนดให้ $|\beta_1| < 1$ แสดงให้เห็นว่าข้อมูลอนุกรมเวลามีคุณสมบัติคงที่

ในกรณี AR(2) สามารถเขียนในรูปแบบสมการดังนี้

$$y_t = \propto + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \varepsilon_t \quad (2\text{-}38)$$

โดย กำหนดให้ $|\beta_1|, |\beta_2|, |\beta_2 - \beta_1| < 1$ แสดงให้เห็นว่าข้อมูลอนุกรมเวลามีคุณสมบัติคงที่

2. แบบจำลอง Moving Average (MA(q)) เป็นการนำค่าความคลาดเคลื่อนจากการพยากรณ์ที่อยู่ก่อนหน้าหรือ $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \varepsilon_{t-q}$ มากำหนดแทนค่าสังเกต y_t โดยกระบวนการ Moving Average ที่มีค่าคลาดเคลื่อนก่อนหน้าที่ q หรือเรียกว่า MA(q) สามารถเขียนในรูปแบบสมการได้ดังนี้

MA(q) คือ $y_t = \delta + \varepsilon_t - \gamma_1 \varepsilon_{t-1} - \gamma_2 \varepsilon_{t-2} - \cdots - \gamma_q \varepsilon_{t-q}$ (2-39)

โดยที่ δ คือ ค่าคงที่

γ_j คือ พารามิเตอร์ตัวที่ j เมื่อ j=1, 2, .., q

ε_t คือ ค่าความคลาดเคลื่อน ณ เวลา t

ในกรณี MA(1) สามารถเขียนในรูปแบบสมการดังนี้

$$y_t = \delta + \varepsilon_t - \gamma_1 \varepsilon_{t-1} \quad (2\text{-}40)$$

โดยกำหนดให้ $|\gamma_1| < 1$ แสดงให้เห็นว่าข้อมูลอนุกรมเวลามีคุณสมบัติคงที่

ในกรณี MA(2) สามารถเขียนในรูปแบบสมการดังนี้

$$x_t = \delta + \varepsilon_t - \gamma_1 \varepsilon_{t-1} - \gamma_2 \varepsilon_{t-2} \quad (2\text{-}41)$$

โดย กำหนดให้ $|\gamma_1 - \gamma_2|, |\gamma_1| < 1$ แสดงให้เห็นว่าข้อมูลอนุกรมเวลามีคุณสมบัติคงที่

3. กระบวนการ Integrated (I(d)) เป็นการหาผลต่างของอนุกรมเวลาระหว่างข้อมูลที่เวลา t กับข้อมูลย้อนหลังไป d คาบเวลา แบบจำลอง ARIMA ต้องใช้ข้อมูลที่มีคุณสมบัติคงที่เท่านั้น ถ้าข้อมูลมีคุณสมบัติไม่คงที่จะต้องทำการเปลี่ยนให้มีคุณสมบัติคงที่ก่อน โดยการหาผลต่างของข้อมูลอนุกรมเวลา โดยทั่วไปแล้วจะต้องหาผลต่างอันดับที่ d สามารถเขียนสมการดังนี้

$$\text{I(d) คือ } \Delta_d y_t = \Delta_{d-1}(y_t - y_{t-1}) \tag{2-42}$$

ในกรณี I(1) สามารถเขียนในรูปแบบสมการดังนี้

$$\text{I(1) คือ } \Delta y_t = y_t - y_{t-1} \tag{2-43}$$

ในกรณี I(2) สามารถเขียนในรูปแบบสมการดังนี้

$$\text{I(2) คือ } \Delta_2 y_t = \Delta(y_t - y_{t-1}) \tag{2-44}$$

จากรายละเอียดต่าง ๆ ที่กล่าวข้างต้นถ้านำส่วนต่าง ๆ มาพิจารณาร่วมกันสามารถกำหนดรูปแบบของแบบจำลอง ARIMA(p, q, d) คือ

$$\Delta_d y_t = \delta + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_p y_{t-p} + \varepsilon_t - \gamma_1 \varepsilon_{t-1} - \gamma_2 \varepsilon_{t-2} - \cdots - \gamma_q \varepsilon_{t-q} \tag{2-45}$$

โดยที่ y_t คือค่าสังเกตในอนุกรมเวลาที่ t

p คือ อันดับของ Auto Regressive

q คือ อันดับของ Moving Average

δ คือ ค่าคงที่

t คือ เวลาที่สนใจ

Δ_d คือ ผลต่างอันดับที่ d

β คือ พารามิเตอร์ของ Auto Regressive

γ คือ พารามิเตอร์ของ Moving Average

ε_t คือ ค่าความคลาดเคลื่อน ณ เวลา t

ตัวอย่างเช่น ในกรณี ARIMA(0, 1, 1) หรือ IMA(1, 1) สามารถเขียนในรูปแบบสมการดังนี้

$$y_t - y_{t-1} = \delta + \varepsilon_t - \gamma_1 \varepsilon_{t-1} \tag{2-46}$$

ในกรณี ARIMA(1, 1, 0) หรือ ARI(1, 1) สามารถเขียนในรูปแบบสมการดังนี้

$$\begin{aligned} \Delta_d y_t &= \alpha + \beta_1 \Delta y_{t-1} + \varepsilon_t \\ y_t - y_{t-1} &= \alpha + \beta_1 (y_{t-1} - y_{t-2}) + \varepsilon_t \end{aligned} \tag{2-47}$$

ในกรณี ARIMA(1, 1, 1) สามารถเขียนในรูปแบบสมการดังนี้

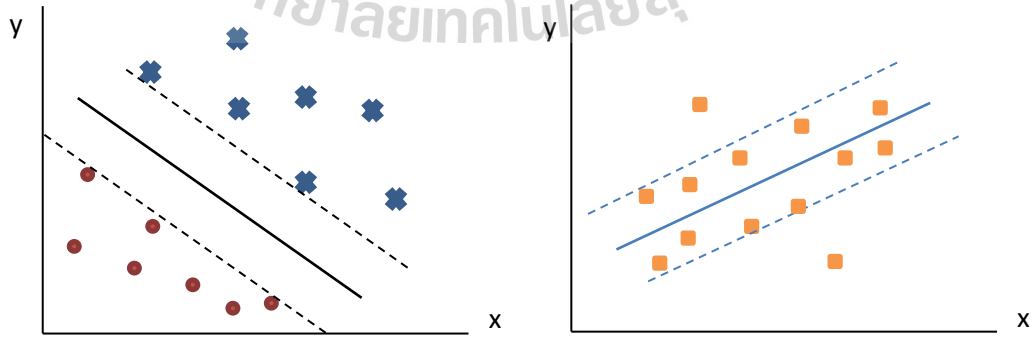
$$y_t - y_{t-1} - \beta_1 (y_{t-1} + y_{t-2}) = \alpha + \varepsilon_t - \gamma_q \varepsilon_{t-1} \tag{2-48}$$

ในกรณี ARIMA(0, 1, 0) สามารถเขียนในรูปแบบสมการดังนี้

$$y_t - y_{t-1} = \varepsilon_t \tag{2-49}$$

2.6 ซัพพอร์ตเวกเตอร์รีเกรสชัน (Support Vector Regression: SVR)

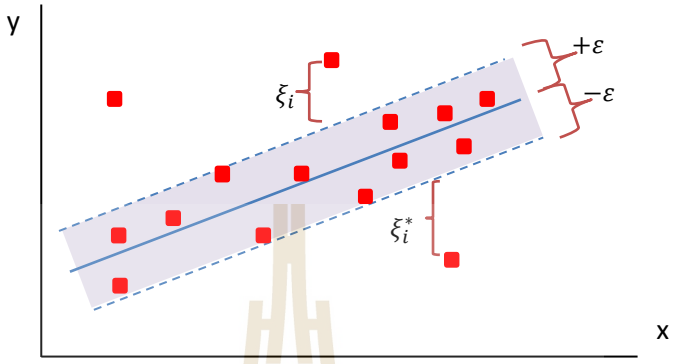
ซัพพอร์ตเวกเตอร์รีเกรสชัน มีแนวคิดพื้นฐานมาจากซัพพอร์เวกเตอร์แมชชีน (Support Vector Machine: SVM) ซึ่งเป็นวิธีการที่ใช้วิธีการถดถอยเพื่อรักษาคุณลักษณะหลักของข้อมูล SVR จะใช้หลักการเดียวกันกับ SVM แต่แตกต่างกันที่ประเภทการนำไปใช้งาน ซึ่ง SVM จะใช้ในการจำแนกข้อมูลออกเป็นกลุ่ม ส่วน SVR จะใช้ในการพยากรณ์ที่ให้ผลลัพธ์เป็นเลขจำนวนจริงซึ่งเป็นสิ่งที่ยากมากในการพยากรณ์ค่าจากข้อมูลที่มีอยู่โดยค่าของการพยากรณ์ที่เป็นเลขจำนวนจริงมีจำนวนที่เป็นไปได้มากมายไม่มีที่สิ้นสุด วิธีการ SVR จึงต้องค้นหาและสร้างขอบเขตที่ดีที่สุดให้กับการถดถอยโดยใช้ฟังก์ชันการสูญเสียเป็นค่าความผิดพลาดที่ยอมรับได้ซึ่งขอบเขตนี้จะอยู่ภายในบริเวณค่าที่แท้จริงอยู่ แผนภาพเปรียบเทียบแนวคิดของ SVM และ SVR แสดงดังรูปที่ 2.23



ก. ซัพพอร์เวกเตอร์แมชชีน ข. ซัพพอร์ตเวกเตอร์รีเกรสชัน

รูปที่ 2.23 ตัวอย่างซัพพอร์เวกเตอร์แมชชีนและซัพพอร์ตเวกเตอร์รีเกรสชัน

การสร้างขอบเขตที่ดีที่สุดของ SVR จะใช้ฟังก์ชันการสูญเสียเป็นค่าความผิดพลาดที่ยอมรับได้ซึ่งขอบเขตนี้จะอยู่ภายในบริเวณค่าที่แท้จริง ซึ่งฟังก์ชันนี้จะเรียกว่า “Epsilon Intensive-Loss Function”



รูปที่ 2.24 ขอบเขต ε-Band ของซัพพอร์ตเวกเตอร์รีเกรสชัน

จากรูป 2.24 พื้นที่ระหว่างเส้นประทั้ง 2 เส้นเรียกว่า ε-Band (Epsilon Band) ตัวแปรที่เป็นมาตรวัดสำหรับระบุความผิดพลาดในข้อมูลฝึกสอนคือ จุดข้อมูลที่อยู่นอก ε-Band ส่วนจุดที่อยู่ภายใน ε-Band จะให้ค่าความผิดพลาดเป็นศูนย์

ถ้ากำหนดให้ชุดข้อมูลฝึกสอน $\{(x_1, y_1), \ldots, (x_i, y_i)\} \subset X \times R$ เมื่อ x_i คือเวกเตอร์ข้อมูล y_i คือผลลัพธ์และ N คือจำนวนข้อมูลทั้งหมด สัญลักษณ์ $\langle \cdot, \cdot \rangle$ คือ dot product ใน X เป้าหมายของ SVR คือค้นหา $f(x)$ ที่ให้ค่า ε มากที่สุดจากค่า y_i สำหรับข้อมูลฝึกสอนและค่าความผิดพลาดต้องมีค่าไม่เกิน ε เมื่อ SVR เป็นเส้นตรงใช้สมการดังนี้ (Granata et al., 2016)

$$f(x) = \langle w, x \rangle + b \tag{2-50}$$

โดยต้องการค่า $||w||^2$ ที่น้อยที่สุด สามารถหาได้จากสมการ

$$minimize\ \frac{1}{2}\|w\|^2\ subject\ to \begin{cases} y_i - \langle w, x_i \rangle + b \leq \varepsilon \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon \end{cases} \tag{2-51}$$

สมการที่ 2-51 เป็นสมการที่สมมติว่า $f(x)$ มีค่าใกล้เคียงกับ (x_i, y_i) ทั้งหมด นั่นหมายถึงข้อมูลทั้งหมดอยู่ใน ε-Band ในกรณีที่มีข้อมูลอยู่นอก ε-Band สามารถเขียนสมการให้อยู่ในรูปของ Slack Variable (ξ_i, ξ_i^*) จะใช้สมการดังนี้

$$minimize\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)\ subject\ to \begin{cases} y_i - \langle w, x_i\rangle + b \le \varepsilon + \xi_i \\ \langle w, x_i\rangle + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0, i = 1,2,\dots,N \end{cases} \quad (2\text{-}52)$$

เมื่อ w คือเวกเตอร์ถ่วงน้ำหนัก และ C คือเป็นพารามิเตอร์ที่กำหนดค่าเอาไว้ เพื่อทำหน้าที่ควบคุมค่าใช้จ่ายในเมื่อมีการพยากรณ์ผิดพลาดซึ่งเป็นตัวควบคุมให้ความผิดพลาดอยู่ในช่วงที่ยอมรับได้ ถ้า C > 0 จะเป็นตัวความคุมความผิดพลาดที่ยอมรับได้ของ SVR ซึ่งการจัดการนี้เรียกว่า “ε- Intensive Loss Function” ดังนั้นค่า $|\xi|_\varepsilon$ อธิบายได้ดังสมการที่ 2-53

$$|\xi|_\varepsilon \begin{cases} 0 & if\ |\xi| < 0 \\ |\xi| - \varepsilon & \text{otherwise} \end{cases} \quad (2\text{-}53)$$

นำสมการที่ 2-52 มาใช้วิธีการคูณแบบลากรองจ์ (Lagrange multiplier method) เพื่อหาฟังก์ชันการถดถอยของ SVR ได้ดังสมการ

$$L = \frac{1}{2}||w||^2 + \mathrm{C}\sum_{i=1}^{N}(\xi_i + \xi_i^*) - \sum_{i=1}^{N}(\eta_i\xi_i + \eta_i^*\xi_i^*) - \sum_{i=1}^{N}\alpha_i(\varepsilon + \xi_i - y_i + \langle w, x_i\rangle + b) - \sum_{i=1}^{N}\alpha_i^*(\varepsilon + \xi_i^* + y_i - \langle w, x_i\rangle - b) \quad (2\text{-}54)$$

สมการที่ 2-52 ต้องมีเงื่อนไข $\alpha_i, \alpha_i^*, \eta_i, \eta_i^* \ge 0$ และหาอนุพันธ์ของ L เทียบด้วย w, b, ξ_i และ ξ_i^* แสดงได้ดังสมการ

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{N}(\alpha_i^* - \alpha_i)x_i = 0 \quad (2\text{-}55)$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{N}(\alpha_i^* - \alpha_i) = 0 \quad (2\text{-}56)$$

$$\frac{\partial L}{\partial \xi_i^{(*)}} = \mathrm{C} - \alpha_i^* - \eta_i^* = 0 \quad (2\text{-}57)$$

แทนค่าสมการ 2-55-2-57 ในสมการ 2-54 จะได้

$$maximize \begin{cases} -\frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i + \alpha_i^*)(\alpha_j + \alpha_j^*)\langle x_i, x_j\rangle \\ -\varepsilon\sum_{i=1}^{N}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i\,(\alpha_j + \alpha_j^*) \end{cases} subject\ to \begin{cases} \sum_{i=1}^{N}(\alpha_i + \alpha_i^*) = 0 \\ \alpha_i, \alpha_i^* \in [0, C] \end{cases} \quad (2\text{-}58)$$

เขียนสมการที่ 2-55 ได้ใหม่ดังสมการต่อไปนี้

$$w = \sum_{i=1}^{N}(\alpha_i^* - \alpha_i)x_i \quad (2\text{-}59)$$

จะได้สมการ SVR ในการพยากรณ์ดังนี้

$$f(x) = \sum_{i=1}^{N}(\alpha_i^* - \alpha_i)\langle x_i, x\rangle + b \quad (2\text{-}60)$$

เมื่อ α_i^* และ α_i คือตัวคูณแบบลากรองจ์ C เป็นจำนวนเต็มค่าคงที่สำหรับเป็นค่าใช้จ่ายเมื่อมีข้อผิดพลาดเกิดขึ้น ถ้าอินพุตเวกเตอร์เป็นซัพพอร์ตเวกเตอร์จะมีค่า $\alpha_i, \alpha_i^* > 0$ ส่วนที่ไม่ใช่ซัพพอร์ตเวกเตอร์คือ $\alpha_i, \alpha_i^* = 0$

นอกจากนี้ SVR ยังมีเคอร์เนลฟังก์ชันต่าง ๆ ให้เลือกใช้งานเพื่อสร้าง SVR ที่มีประสิทธิภาพในการพยากรณ์โดยสามารถใช้สมการ 2-60 และรูปแบบเคอร์เนลฟังก์ชันแบบต่าง ๆ ดังแสดงในตารางที่ 2.7

$$f(x) = \sum_{i=1}^{N}(\alpha_i^* - \alpha_i)K(x_i x) + b \quad (2\text{-}61)$$

เมื่อ

$$b = y_i - \langle w, x_i\rangle - \varepsilon \quad (2\text{-}62)$$

$$b = y_i - \langle w, x_i\rangle + \varepsilon \quad (2\text{-}63)$$

ตารางที่ 2.7 เคอร์เนลฟังก์ชันที่ใช้ร่วมกับซัพพอร์ตเวกเตอร์รีเกรชชัน

Kernel	$\boldsymbol{K(x_i x)}$
Linear	$x_i^T x$
Radial Basis Function	$\exp\left(-\gamma\left\|\|x_i - x\right\|\|^2\right), \gamma > 0$
Polynomial	$((x_i \cdot x) + \eta)^d$
Sigmoid	$tanh(\gamma(x_i \cdot x) + \eta), \gamma > 0$

2.7 การประเมินประสิทธิภาพโมเดล

ในส่วนนี้จะกล่าวถึงการประเมินประสิทธิภาพโมเดลของงานวิจัยนี้ โดยใช้มาตรวัด 2 ชนิด ได้แก่ ค่าสัมประสิทธิ์สหสัมพันธ์ และค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย ซึ่งค่าทั้งสองจะใช้ประเมินประสิทธิภาพโมเดลในมุมมองที่แตกต่างกัน ซึ่งรายละเอียดของมาตรวัดทั้งสองจะอธิบายในหัวข้อ 2.7.1 และ 2.7.2

2.7.1 ค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient: R)

ค่าสหสัมพันธ์ (Correlation) เป็นสถิติที่ใช้แสดงระดับความสัมพันธ์ระหว่างตัวแปร ค่าสหสัมพันธ์ที่คำนวณได้ เรียกว่า ค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) นิยมแทนด้วยสัญลักษณ์ R การหาความสัมพันธ์ระหว่างตัวแปรนั้นมักจะใช้สัญลักษณ์ของตัวแปรเป็น x และ y โดยค่าสหสัมพันธ์เป็นการวัดความสัมพันธ์เชิงเส้นจะมีค่าอยู่ระหว่าง -1 ถึง 1 และค่า R จะบอกถึงทิศทางและขนาดของความสัมพันธ์

ทิศทางของความสัมพันธ์ (Direction of the Relationship) ในการหาลักษณะความสัมพันธ์ระหว่างตัวแปรนั้นสามารถสร้างแผนภาพการกระจัดกระจาย (Scatter plot) เพื่อดูทิศทางของความสัมพันธ์ได้ โดยมีลักษณะความสัมพันธ์ 3 แบบ คือ

- สหสัมพันธ์เชิงบวก (Positive Correlation) ซึ่งหมายความว่าเมื่อตัวแปรตัวหนึ่งมีค่าเพิ่มหรือลดลงอีกตัวแปรหนึ่งก็จะมีค่าเพิ่มขึ้นหรือลดลงไปด้วย

- สหสัมพันธ์เชิงลบ (Negative Correlation) หมายถึงเมื่อตัวแปรตัวหนึ่งมีค่าเพิ่มขึ้นหรือลดลงอีกตัวหนึ่งจะมีค่าเปลี่ยนแปลงในทิศทางตรงข้ามเสมอ

- สหสัมพันธ์เป็นศูนย์ (Zero Correlation) หมายถึงตัวแปรสองตัวไม่มีความสัมพันธ์ซึ่งกันและกันเลย

ขนาดของความสัมพันธ์ จะใช้ตัวเลขของค่าสัมประสิทธิ์สหสัมพันธ์ ถ้าค่าสัมประสิทธิ์สหสัมพันธ์มีค่าที่เข้าใกล้ -1 หรือ 1 แสดงถึงการมีความสัมพันธ์กันในระดับสูง แต่หากมีค่าเข้าใกล้ 0 แสดงถึงการมีความสัมพันธ์กันในระดับน้อย หรือไม่มีเลย สำหรับการพิจารณาค่าสัมประสิทธิ์สหสัมพันธ์ โดยทั่วไปอาจใช้เกณฑ์ดังนี้ (Hinkle, 1998)

ตารางที่ 2.8 การแปลผลค่าสัมประสิทธิ์สหสัมพันธ์

ค่า R (ทั้งทิศทางบวกและลบ)	ระดับของความสัมพันธ์
0.90 - 1.00	มีความสัมพันธ์กันสูงมาก
0.70 - 0.90	มีความสัมพันธ์กันในระดับสูง
0.50 - 0.70	มีความสัมพันธ์กันในระดับปานกลาง
0.30 - .50	มีความสัมพันธ์กันในระดับต่ำ
0 - 0.30	มีความสัมพันธ์กันในระดับต่ำมาก

ค่าสหสัมพันธ์ (R) สามารถคำนวณได้จากสมการที่ (2-64)

$$R = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}} \quad (2\text{-}64)$$

โดย R คือ ค่าสัมประสิทธิ์สหสัมพันธ์

n คือ จำนวนข้อมูล

y_i คือ ค่าของข้อมูลจริง

x_i คือ ค่าของข้อมูลที่พยากรณ์

2.7.2 ค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Squared Error : RMSE)

ค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ยใช้ในการประเมินความผิดพลาดในการพยากรณ์ของโมเดล ถ้าค่า RMSE ยิ่งน้อยหมายถึงการพยากรณ์ยิ่งแม่นยำ ในทางกลับกันถ้าค่า RMSE ยิ่งมากหมายถึงการพยากรณ์มีประสิทธิภาพในการพยากรณ์ผิดพลาดสูง สามารถคำนวณได้ดังสมการต่อไปนี้

$$\text{RMSE} = \sqrt{\frac{\sum(T_i - O_i)^2}{N}} \quad (2\text{-}65)$$

เมื่อ T_i คือ ค่าของข้อมูลจริง

O_i คือ ค่าของข้อมูลที่พยากรณ์

N คือ จำนวนข้อมูลทั้งหมด

ตารางที่ 2.9 ตัวอย่างข้อมูลฝึกสอนที่ใช้ในการเรียนรู้

ข้อมูลที่	X_1	X_2	Y
1	0	0.270	1.81
2	0	0.246	4.34
3	14.4	0.190	11.11
4	44.7	0.156	28.13
5	138.6	0.171	43.42
6	96.2	0.213	27.83
7	91.5	0.214	27.83

ตารางที่ 2.10 ตัวอย่างข้อมูลทดสอบที่ใช้ในการเรียนรู้

ข้อมูลที่	X_1	X_2	Y
1	0	0.280	18.33
2	83.9	0.236	14.81
3	0	0.219	35.21

โดยทั่วไปการสร้างและทดสอบโมเดลจะแบ่งข้อมูลออกเป็นสองส่วนได้แก่ ข้อมูลฝึกสอนและข้อมูลทดสอบ ซึ่งข้อมูลทั้งสองชุดจะต้องมีแอตทริบิวต์เหมือนกันแต่ข้อมูลและจำนวนข้อมูลไม่จำเป็นต้องเหมือนกัน การสร้างโมเดลนั้นจะใช้ข้อมูลฝึกสอนเพื่อให้อัลกอริทึมเรียนรู้ข้อมูลเพื่อนำไปสร้างโมเดล จากนั้นทดสอบประสิทธิภาพโมเดลโดยการใช้ข้อมูลทดสอบซึ่งเป็นข้อมูลที่โมเดลไม่เคยเรียนรู้นำมาให้โมเดลพยากรณ์ค่า ตัวอย่างข้อมูลฝึกสอนและข้อมูลทดสอบแสดงในตารางที่ 2.8 และ 2.9 ตามลำดับ และเมื่อนำข้อมูลชุดฝึกสอนไปใช้ในการเรียนรู้ของอัลกอริทึม ANN GLM และ SVR เพื่อสร้างโมเดลในการพยากรณ์ ผลการพยากรณ์ข้อมูลทดสอบของทั้งสามโมเดลแสดงในตารางที่ 2.10

ตารางที่ 2.11 ผลการพยากรณ์ข้อมูลทดสอบของ ANN GLM และ SVR

ข้อมูลที่	ANN	GLM	SVR
1	1.830	8.004	7.872
2	23.159	28.309	11.099
3	2.380	2.094	8.757

ตารางที่ 2.12 ผลรวม ค่าเฉลี่ย ผลคูณของค่าของข้อมูลจริงกับค่าของโมเดล ANN

ข้อมูลที่	Y	Y^2	ANN	ANN^2	Y*ANN	$(Y-ANN)^2$
1	18.33	335.99	1.830	3.35	33.54	272.25
2	14.81	219.34	23.159	536.34	342.98	69.71
3	35.21	1239.74	2.380	5.66	83.80	1077.81
ผลรวม	68.35	1795.07	27.369	545.35	460.33	1419.76

จากผลการพยากรณ์ข้อมูลทดสอบจากตารางที่ 2.10 นำมาประเมินประสิทธิภาพโมเดล ANN โดยการคำนวณค่าสัมประสิทธิ์สหสัมพันธ์ และค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ยทำได้โดยเตรียมข้อมูลตามตาราง 2.2 ประกอบด้วย ผลรวม ค่ายกกำลังสอง ผลรวมยกกำลังสองของข้อมูลจริง (Y) กับข้อมูลที่พยากรณ์ด้วยโมเดล ANN ค่าผลคูณ และค่าผลต่างยกกำลังสองของข้อมูลจริงกับข้อมูลที่พยากรณ์ด้วยโมเดล ANN แสดงการคำนวณได้ดังนี้

$$R_{ANN} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

$$= \frac{3(460.33) - (27.369 * 68.35)}{\sqrt{[3(545.35) - (27.369)^2] * [3(1{,}795) - (68.35)^2]}}$$

$$= \frac{1{,}380.99 - 1{,}870.67}{\sqrt{886.99 * 713.48}}$$

$$= \frac{-489.68}{795.52}$$

$$= -0.616$$

$$\text{RMSE}_{ANN} = \sqrt{\frac{\sum(T_i - O_i)^2}{N}}$$

$$= \sqrt{\frac{1{,}419.76}{3}}$$

$$= 21.75$$

ตารางที่ 2.13 ผลรวม ค่าเฉลี่ย ผลคูณของค่าของข้อมูลจริงกับค่าของโมเดล GLM

ข้อมูลที่	Y	Y^2	GLM	GLM^2	Y*GLM	$(Y\text{-}GLM)^2$
1	18.33	335.99	8.004	64.06	146.71	106.63
2	14.81	219.34	28.309	801.40	419.26	182.22
3	35.21	1239.74	2.094	4.38	73.73	1096.67
ผลรวม	68.35	1795.07	38.407	869.85	639.70	1385.52

จากผลการพยากรณ์ข้อมูลทดสอบจากตารางที่ 2.10 นำมาประเมินประสิทธิภาพโมเดล GLM โดยเตรียมข้อมูลตามตาราง 2.12 เพื่อคำนวณค่าสัมประสิทธิ์สหสัมพันธ์ และค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ยแสดงการคำนวณได้ดังนี้

$$R_{GLM} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

$$= \frac{3(639.70) - (38.407 * 68.35)}{\sqrt{[3(869.85) - (38.407)^2] * [3(1{,}795.07) - (68.35)^2]}}$$

$$= \frac{1{,}919.1 - 2{,}625.12}{\sqrt{1{,}134.45 * 713.48}}$$

$$= \frac{-706.02}{899.67}$$

$$= -0.785$$

$$\text{RMSE}_{GLM} = \sqrt{\frac{\sum(T_i - O_i)^2}{N}}$$

$$= \sqrt{\frac{1{,}385.52}{3}}$$

$$= 21.49$$

ตารางที่ 2.14 ผลรวม ค่าเฉลี่ย ผลคูณของค่าของข้อมูลจริงกับค่าของโมเดล SVR

ข้อมูลที่	Y	Y^2	SVR	SVR^2	Y*SVR	$(Y-SVR)^2$
1	18.33	335.99	7.872	61.97	144.29	109.37
2	14.81	219.34	11.099	123.19	164.38	13.77
3	35.21	1239.74	8.757	76.69	308.33	699.76
ผลรวม	68.35	1795.07	27.728	261.84	617.0	822.90

จากผลการพยากรณ์ข้อมูลทดสอบจากตารางที่ 2.10 นำมาประเมินประสิทธิภาพโมเดล SVR โดยเตรียมข้อมูลตามตาราง 2.13 เพื่อคำนวณค่าสัมประสิทธิ์สหสัมพันธ์ และค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ยแสดงการคำนวณได้ดังนี้

$$R_{SVR} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

$$= \frac{3(617) - (27.728 * 68.35)}{\sqrt{[3(261.84) - (27.728)^2] * [3(1{,}795.07) - (68.35)^2]}}$$

$$= \frac{1{,}851 - 1{,}895.21}{\sqrt{16.68 * 713.48}}$$

$$= \frac{-44.21}{109.09}$$

$$= -0.405$$

$$\text{RMSE}_{SVR} = \sqrt{\frac{\sum(T_i - O_i)^2}{N}}$$

$$= \sqrt{\frac{822.90}{3}}$$

$$= 16.56$$

ค่าสัมประสิทธิ์สหสัมพันธ์ของ ANN GLM และ SVR มีค่า -0.616 -0.785 -0.405 ตามลำดับ แสดงให้เห็นว่าโมเดลทั้งสามให้การพยากรณ์ผกผันกับค่าจริง ส่วนค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ยของ ANN GLM และ SVR มีค่า 21.75 21.49 16.56 ตามลำดับ แสดงให้เห็นว่า SVR โมเดลมีความผิดพลาดในการพยากรณ์น้อยที่สุด จะเห็นได้ว่าถ้าใช้การประเมินประสิทธิภาพด้วยค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ยอย่างเดียวจะแสดงให้เห็นว่า

SVR มีประสิทธิภาพในการพยากรณ์ดีกว่า ANN และ GLM แต่จะไม่ทราบว่าโมเดล SVR ANN และ GLM ให้ค่าการพยากรณ์ผกผันจากค่าจริง ดังนั้นการประเมินประสิทธิภาพโมเดลควรใช้มาตรวัดทั้งสองพิจารณาควบคู่กันเพื่อแสดงประสิทธิภาพการพยากรณ์ของแต่ละโมเดลได้อย่างชัดเจน

2.8 งานวิจัยที่เกี่ยวข้อง

การคาดการณ์ปริมาณน้ำท่าเป็นสิ่งสำคัญที่สามารถช่วยให้ทราบค่าปริมาณน้ำในแหล่งกักเก็บว่ามีความเพียงพอต่อความต้องการใช้ในอนาคตหรือไม่ การคาดการณ์น้ำท่าล่วงหน้ามีประโยชน์ในการเฝ้าระวังปัญหาที่เกิดจากน้ำได้ แต่การวิเคราะห์ปริมาณน้ำท่าแบบดั้งเดิมนั้นจำเป็นต้องทราบข้อมูลหลายอย่างเพื่อนำมาใช้ในการคำนวณ เช่น สภาพภูมิประเทศ ขนาดพื้นที่ สภาพดิน การระเหย หรือข้อมูลภูมิอากาศ เป็นต้น การรวบรวมข้อมูลเพื่อจะนำมาใช้ในการคาดการณ์น้ำท่าจะต้องใช้เวลานานกว่าจะทราบปริมาณน้ำท่าล่วงหน้าจึงทำให้มีเวลาน้อยในการเตรียมการรับมือกับปัญหาที่เกิดขึ้น ดังนั้นจึงมีงานวิจัยจำนวนมากพยายามที่จะคาดการณ์ปริมาณน้ำท่าด้วยเทคนิคการเรียนรู้ของเครื่องเพื่อลดเวลาการรวบรวมข้อมูลที่ใช้ในการคาดการณ์น้ำท่าและเพิ่มประสิทธิภาพการคาดการณ์น้ำท่าให้ดีขึ้น

Heesung Yoon และคณะ (2011) เสนอโมเดลคาดการณ์น้ำท่าที่ใช้ ANN แบบเพอร์เซ็ปตรอนหลายชั้นและ SVM โดยใช้ข้อมูลอินพุตที่แตกต่างกัน 5 รูปแบบได้แก่ รูปแบบที่ 1 ใช้น้ำฝนกับระดับน้ำ รูปแบบที่ 2 ใช้น้ำท่า รูปแบบที่ 3 ใช้น้ำฝนกับน้ำท่า รูปแบบที่ 4 ใช้ระดับน้ำกับน้ำท่า และน้ำฝน และรูปแบบที่ 5 ใช้ระดับน้ำ กับน้ำท่า โดยใช้ข้อมูลที่เวลาย้อนหลัง 1, 2, 4, 6 และ 8 ชั่วโมง แล้วเลือกเอาชุดรูปแบบข้อมูลอินพุตที่ดีที่สุดแต่ละเวลาย้อนหลังมาใช้เป็นอินพุตให้อัลกอริทึม การวัดประสิทธิภาพใช้ค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Square error: RMSE), ค่าความผิดพลาดเฉลี่ย (Mean Error: ME), ค่าเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Percentage Error: MAPE), ค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient: R), ค่าสัมประสิทธิ์ประสิทธิผล (Nash–Sutcliffe Efficiency: NS) และเกณฑ์การคัดเลือกตัวแบบของอาไคเคะ (Akaike Information Criterion: AIC) ผลการทดลองพบว่า SVM มีประสิทธิภาพในการคาดการณ์น้ำท่าดีกว่าโครงข่ายประสาทเทียม

Imen Aichouri และคณะ (2015) ศึกษาการคาดการณ์ของน้ำท่าสำหรับลุ่มน้ำในพื้นที่กึ่งแห้งแล้งในแถบชายฝั่งทะเลเมดิเตอร์เรเนียนโดยใช้ ANN แบบเพอร์เซ็ปตรอนหลายชั้น (Multilayer Perceptron Network: MLPN) ซึ่งเป็นอัลกอริทึมที่ใช้ได้กับเหตุการณ์ที่เป็นเชิงเส้นและไม่เป็นเชิงเส้นโดยไม่จำเป็นต้องใช้สมมติฐานใด ๆ เปรียบเทียบผลกับการถดถอยเชิงเส้น (Multiple Linear Regression) โดยใช้ข้อมูลอินพุตเป็นปริมาณน้ำท่าและน้ำฝนรายวันย้อนหลัง 1- 7 วัน ทำการวัด

ประสิทธิภาพด้วยค่าเฉลี่ยความผิดพลาดกำลังสอง (Average Squared of Error: ASE), ค่าสัมประสิทธิ์การตัดสินใจ (Coefficient of Determination: R^2), ค่าความคลาดเคลื่อนสัมพัทธ์เฉลี่ย (Mean Absolute Relative Error: MARE) ผลการศึกษาแสดงให้เห็นว่า ANN มีประสิทธิภาพในการคาดการณ์น้ำท่าดีกว่าวิธีการถดถอยเชิงเส้น

Gokcen Uysal และคณะ (2016) ศึกษาการคาดการณ์ข้อมูลน้ำท่าในพื้นที่ที่มีหิมะปกคลุม โดยใช้ข้อมูลในช่วงเดือนที่เป็นต้นฤดูใบไม้ผลิต่อเนื่องไปถึงปลายฤดูร้อน (1 มีนาคม - 30 มิถุนายน) เพื่อใช้ในการพิจารณาบริหารจัดการทรัพยากรน้ำโดยเฉพาะอย่างยิ่งการเฝ้าระวังน้ำท่วม การคาดการณ์น้ำท่าใช้ข้อมูลน้ำฝน อุณหภูมิเฉลี่ยรายวัน และข้อมูลจากดาวเทียม MODIS บริเวณพื้นที่หิมะปกคลุม (Snow Cover Area: SCA) ซึ่งเป็นข้อมูลจากกราฟ (Snow Depletion Curves, SCD) ที่เวลาย้อนหลัง 1 วัน เปรียบเทียบโมเดล ANN แบบเพอร์เซ็ปตรอนหลายชั้น (Multilayer Perceptron Network: MLPN) และแบบเรเดียลเบสิสฟังก์ชัน (Radial Basis Function: RBF) วัดประสิทธิภาพโดยใช้ค่าสัมประสิทธิ์การตัดสินใจ (Coefficient of determination: R^2) ค่าความผิดพลาดเฉลี่ย (Mean Error: ME) ค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Squared error: RMSE) และค่าคลาดเคลื่อนสัมบูรณ์ (Mean Absolute Error: MAE) ผลการทดลองสรุปว่า ANN แบบเพอร์เซ็ปตรอนหลายชั้นนั้นมีประสิทธิภาพดีกว่าแบบเรเดียลเบสิสฟังก์ชัน

Ayob Katimon และคณะ (2017) เสนอการใช้โมเดล ARIMA เพื่อพยากรณ์ค่าต่าง ๆ ที่เกี่ยวกับคุณภาพของน้ำ ตัวอย่างเช่น น้ำฝน ปริมาณการไหลของน้ำ ค่า pH เป็นต้น โดยทำการกำหนดค่าพารามิเตอร์ให้ ARIMA โดยใช้ข้อมูลย้อนหลังของค่าคุณภาพของน้ำที่ต้องการพยากรณ์ วิธีการทดสอบประสิทธิภาพการพยากรณ์ใช้มาตรวัดอาไคเคะ ผลกาทดลองแสดงให้เห็นว่า ARIMA เหมาะกับการพยากรณ์ค่าอะลูมิเนียมไอออน (AL^{3+}) เฟอร์รัสไอออน (FE^{2+}) แอมโมเนียไอออน (NH^{4+}) และน้ำฝน ส่วนวิธีการ AR เหมาะที่จะใช้การพยากรณ์ แมงกานีสไอออน (MN^{2+}) ค่าพีเอช (pH) สีของน้ำ การไหลของน้ำ ความขุ่นของน้ำ และปริมาณสารแขวนลอย

Alireza Sharif และคณะ (2017) ศึกษาการทำนายน้ำท่ารายวันโดยใช้อัลกอริทึมเชิงเส้นและไม่เป็นเชิงเส้น ใช้ข้อมูลน้ำฝนและน้ำท่าย้อนหลัง ทำการเลือกข้อมูลย้อนหลังตั้งแต่ 1 ถึง 4 วัน การสร้างโมเดล ใช้ 3 วิธีการในการคัดเลือกข้อมูลนำเข้าได้แก่ Gamma test, Forward selection และ Factor analysis ทีมวิจัยนี้เลือกใช้อัลกอริทึม ANN SVM LR และ Adaptive Neuro-Fuzzy Inference System (ANFIS) จากนั้น ทำการเปรียบเทียบประสิทธิภาพอัลกอริทึมด้วย R^2 NS และ RMSE ผลการทดลองแสดงให้เห็นว่าการใช้โมเดล SVM กับข้อมูลที่คัดเลือกด้วย Gamma Test ให้ประสิทธิภาพดีที่สุดในการคาดการณ์น้ำท่ารายวัน

Zaini N. และคณะ (2018) เสนอการใช้โมเดล SVM ร่วมกับเทคนิคพาติเคิลสวอมออปติไมเซชัน (Particle Swarm Optimization: PSO) ในการพยากรณ์น้ำท่ารายวัน ตัวแปรที่ใช้เป็นข้อมูลอินพุต ได้แก่ ปริมาณน้ำฝน น้ำท่า และข้อมูลทางอุตุนิยมวิทยาที่เวลาย้อนหลัง 1วันจากเวลาที่สนใจ ได้แก่ อุณหภูมิ อุณหภูมิสูงสุด ความชื้นสัมพัทธ์เฉลี่ย การระเหย ความเร็วลมเฉลี่ย ทดลองเปรียบเทียบกับ SVM และ SVM-PSO แบบที่ใช้ข้อมูลเฉพาะ น้ำฝน น้ำท่า และ SVM แบบที่ใช้ข้อมูลทั้งหมด วัดประสิทธิภาพโดยใช้ R^2 และ RMSE ผลการทดลองสรุปได้ว่าวิธีการ SVM-PSO ร่วมกับตัวแปรปริมาณน้ำฝน น้ำท่า และข้อมูลทางอุตุนิยมวิทยาย้อนหลัง มีประสิทธิภาพดีที่สุดและสามารถทำนายล่วงหน้าได้ 1-7 วัน

จากการศึกษางานวิจัยที่เกี่ยวข้องพบว่าการใช้อัลกอริทึมเดียวสำหรับการสร้างโมเดลในการพยากรณ์น้ำท่าอาจจะยังไม่เพียงพอในการคาดการณ์ปริมาณน้ำท่าให้มีประสิทธิภาพสูงสุด ในงานวิจัยนี้ได้เสนอวิธีการชื่อ ANN-GS ซึ่งเป็นผสมผสานการพยากรณ์ของสองอัลกอริทึมได้แก่ อัลกอริทึมที่เป็นวิธีการทางสถิติคือ GLM และอัลกอริทึมที่เป็นการเรียนรู้ของเครื่องคือ SVR เพื่อการกำหนดข้อมูลที่เหมาะสมให้กับการเรียนรู้และเพิ่มประสิทธิภาพการคาดการณ์น้ำท่าด้วย ANN โดยอัลกอริทึมทั้งสองจะเป็นวิธีการที่หาสมการความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามแต่จะมีการเลือกสมการความสัมพันธ์ที่แตกต่างกัน อัลกอริทึม GLM จะพิจารณาค่าความผิดพลาดของสมการจากข้อมูลทั้งหมด แล้วเลือกสมการที่มีค่าความผิดพลาดน้อยที่สุด ดังนั้นสมการที่ได้จะเป็นการคาดการณ์น้ำท่าโดยดูจากพฤติกรรมโดยรวมของข้อมูลทั้งหมด แต่อัลกอริทึม SVR จะเลือกสมการที่ดีที่สุดด้วยการพยายามเก็บข้อมูลไว้ในระยะขอบของเส้นที่ใช้เป็นสมการความสัมพันธ์โดยจะเลือกสมการที่มีระยะที่มีข้อมูลอยู่มากที่สุดและขอบกว้างที่สุด โดยค่าความผิดพลาดน้อยที่สุดโดยค่าความผิดพลาดจะพิจารณาเฉพาะข้อมูลที่อยู่นอกระยะขอบ ดังนั้นสมการที่ได้จะเป็นการคาดการณ์น้ำท่าโดยดูจากพฤติกรรมโดยรวมของข้อมูลส่วนใหญ่ จากนั้นจะนำค่าคาดการณ์น้ำท่าจาก GLM และ SVR ซึ่งเป็นค่าที่มีความใกล้เคียงกับการจริงมาใช้ในการสร้างโมเดลที่ชื่อว่า ANN-GS โดยนำค่าคาดการณ์จากโมเดลทั้งสองมาใช้ในการเรียนรู้และปรับปรุงโมเดลจากความผิดพลาดด้วยอัลกอริทึม ANN ทั้งนี้งานวิจัยที่เสนอขึ้นมาใหม่นี้จะใช้ข้อมูลปริมาณน้ำฝน จำนวนวันที่ฝนตก ปริมาณน้ำท่า ค่าดัชนีผลต่างพืชพรรณ และตัวเลขของเดือน ซึ่งการดำเนินงานจะเปรียบเทียบกับวิธีการแบบดั้งเดิม เพื่อแสดงให้เห็นว่าวิธีการที่นำเสนอสามารถคาดการณ์ปริมาณน้ำท่าได้อย่างมีประสิทธิภาพ

ตารางที่ 2.15 สรุปเปรียบเทียบงานวิจัยที่เกี่ยวข้องกับการคาดการณ์น้ำท่า

กระบวนการทำงาน	งานวิจัยที่เกี่ยวข้อง						
	ก	ข	ค	ง	จ	ฉ	ช*
ข้อมูลที่ใช้พิจารณาเพื่อคาดการณ์น้ำท่า							
ปริมาณน้ำฝน	✓	✓	✓		✓	✓	✓
ปริมาณน้ำท่า	✓	✓		✓	✓	✓	✓
อุณหภูมิ			✓			✓	
ค่าจากดาวเทียม			✓				✓
ตัวเลขของเดือน							✓
ความชื้นสัมพัทธ์						✓	
การระเหย						✓	
ความเร็วลม						✓	
เทคนิคที่ใช้ในการคาดการณ์น้ำท่า							
Artificial Neural Network	✓	✓	✓		✓		✓
Linear Regression Analysis		✓			✓		
Support Vector Machines	✓				✓	✓	✓
Generalized Linear Model							✓
Autoregressive Integrated Moving Average model				✓			
Adaptive Neuro-Fuzzy Inference System (ANFIS)					✓		
Particle Swarm Optimization						✓	
ปรับปรุงจากวิธีการที่มีอยู่ หรือเสนอเทคนิคใหม่	✓	✓	✓		✓	✓	✓
การเปรียบเทียบประสิทธิภาพ							
Correlation Coefficient	✓		✓		✓	✓	✓
Coefficient of Determination		✓					
Root Mean Square error	✓		✓		✓	✓	✓
Nash–Sutcliffe Efficiency	✓				✓		
Akaike Information Criterion	✓			✓			
Mean Error	✓		✓				

ตารางที่ 2.15 สรุปเปรียบเทียบงานวิจัยที่เกี่ยวข้องกับการคาดการณ์น้ำท่า (ต่อ)

กระบวนการทำงาน	งานวิจัยที่เกี่ยวข้อง						
	ก	ข	ค	ง	จ	ฉ	ช*
การเปรียบเทียบประสิทธิภาพ (ต่อ)							
Mean Absolute Percentage Error	✓						
Average Squared of Error, Mean Absolute Relative Error		✓					
Mean Absolute Error			✓				
วัตถุประสงค์งานวิจัย							
เพื่อทดสอบประสิทธิภาพโมเดล	✓	✓	✓	✓	✓	✓	✓
เพื่อเสนอแนวคิดใหม่	✓		✓		✓	✓	✓
เพื่อประยุกต์ใช้กับข้อมูลจริง	✓	✓	✓	✓	✓	✓	✓

หมายเหตุ งานวิจัยที่เกี่ยวข้องประกอบด้วย

ก แทนงานวิจัยของ Heesung Yoon และคณะ (2011)

ข แทนงานวิจัยของ Imen Aichouri และคณะ (2015)

ค แทนงานวิจัยของ Gökçen Uysal และคณะ (2016)

ง แทนงานวิจัยของ Ayob Katimon และคณะ (2017)

จ แทนงานวิจัยของ Alireza Sharif และคณะ (2017)

ฉ แทนงานวิจัยของ N. Zaini และคณะ (2018)

ช* แทนงานวิจัยของวิทยานิพนธ์ฉบับนี้

บทที่ 3
วิธีดำเนินงานวิจัย

งานวิจัยนี้มีวัตถุประสงค์พัฒนาเทคนิคสำหรับการคาดการณ์น้ำท่า โดยใช้ความสามารถของ 3 อักกอริทึม ได้แก่ โมเดลเชิงเส้นโดยนัยทั่วไป ซัพพอร์ตเวกเตอร์รีเกรสชัน และโครงข่ายประสาทเทียม การผสมผสานสามอัลกอริทึมเพื่อให้การพยากรณ์น้ำท่ามีประสิทธิภาพดี ในบทนี้จะนำเสนอ วิธีการวิจัย เครื่องมือที่ใช้ และกระบวนการต่าง ๆ ของการวิจัย โดยมีรายละเอียดดังนี้

3.1 ข้อมูลที่ใช้ในการสร้างโมเดล

ตารางที่ 3.1 ตัวอย่างข้อมูลที่ใช้ในงานวิจัย

ข้อมูลที่	ปี	เดือน	ตัวเลขของเดือน	น้ำฝน	น้ำท่า	จำนวนวันที่ฝนตก	ดัชนีผลต่างพืชพรรณ
1	2540	JAN	1	52.9	27.7	9	0.415
2	2540	FEB	2	0	22.37	0	0.363
3	2540	MAR	3	11.8	24.63	5	0.278
4	2540	APR	4	57.6	37.47	10	0.274
5	2540	MAY	5	52.9	55.87	9	0.291
6	2540	JUN	6	26.5	49.52	9	0.207
7	2540	JUL	7	178.7	113.35	20	0.141
8	2540	AUG	8	236.4	114.82	22	0.2
9	2540	SEP	9	138	198.53	16	0.276
10	2540	OCT	10	162.9	233.12	8	0.362
11	2540	NOV	11	19.2	83.4	3	0.426
12	2540	DEC	12	0	38.98	0	0.463

งานวิจัยนี้ใช้ข้อมูลภาคพื้นดินจากศูนย์อุทกวิทยาชลประทาน (http://www.hydro-1.net และ http://www.hydro-4.net) ได้แก่ ปริมาณน้ำท่า ปริมาณน้ำฝน จำนวนวันที่ฝนตก ตัวเลขของเดือน และใช้ข้อมูลจากการสำรวจระยะไกลโดยใช้ค่าดัชนีผลต่างพืชพรรณ (NDVI) จากดาวเทียม NOAA STAR - Global Vegetation Health Products (http://www.star.nesdis.noaa.gov) แสดงตัวอย่างข้อมูลที่ใช้ในงานวิจัยได้ในตารางที่ 3.1

- **ข้อมูลปริมาณน้ำฝน** เป็นค่าปริมาณน้ำฝนรายเดือนมีหน่วยเป็นมิลลิเมตร
- **จำนวนวันที่ฝนตก** คือจำนวนวันที่ฝนตกในแต่ละเดือน
- **ปริมาณน้ำท่า** เป็นปริมาณน้ำท่ารายเดือนมีหน่วยเป็นล้านลูกบาศก์เมตร
- **ข้อมูลตัวเลขของเดือน** เป็นตัวเลขที่แทนลำดับของเดือน เช่น มกราคม = 1 กุมภาพันธ์ = 2 เป็นต้น
- **ดัชนีผลต่างพืชพรรณ** (Normalized Difference Vegetation Index: NDVI) เป็นภาพกราฟิกที่แสดงถึงค่าการสะท้อนของคลื่นแม่เหล็ก (Electromagnetic Spectrum) ของช่วงคลื่นใกล้อินฟาเรด (Near Infrared Reflectance) กับ ช่วงคลื่นตามองเห็นสีแดง (Visible Red Reflectance) โดยสามารถนำมาใช้ในการวิเคราะห์การวัดระยะไกล (Remote Sensing Analysis) NDVI (Kriegler et al.,1969) NDVI นิยมใช้ในการตรวจวัดความสมบูรณ์ของพืชพรรณ หรือป่าไม้ เพื่อประเมินว่าพื้นที่ที่ทำการวิเคราะห์มีพืชพรรณสีเขียวหนาแน่นมากน้อยเพียงใด โดย NDVI เป็นการคำนวณหาสัดส่วนของช่วงคลื่นที่เกี่ยวข้องกับพืชพรรณ โดยนำค่าความแตกต่างของการสะท้อนของพื้นผิว ระหว่างช่วงคลื่นใกล้อินฟาเรด กับช่วงคลื่นตามองเห็นสีแดง มาทำสัดส่วนกับค่าผลรวมของทั้งสองช่วงคลื่น เพื่อปรับให้เป็นลักษณะการกระจายแบบปกติ (Normal Distribution) ดังสมการที่ 3-1 โดยทั่วไปแล้วพืชพรรณที่มีความสมบูรณ์จะทำการดูดซับ (Absorb) ช่วงคลื่นใกล้อินฟาเรดและสะท้อนกลับไป (Reflect) ได้เป็นจำนวนมาก แต่จะดูดซับและสะท้อนช่วงคลื่นตามองเห็นสีแดงได้น้อย ในทางตรงกันข้ามพืชพรรณที่ไม่สมบูรณ์จะทำการดูดซับช่วงคลื่นใกล้อินฟาเรดและช่วงคลื่นตามองเห็นสีแดง และทำการสะท้อนกลับปริมาณปานกลางใกล้เคียงกัน ดังแสดงในรูปที่ 3.1 โดยค่า NDVI ที่คำนวณได้นั้นสามารถนำมาแสดงด้วยค่าสีต่าง ๆ ในรูปแผนที่ของบริเวณพื้นที่ที่ต้องการวิเคราะห์

$$NDVI = \frac{(NIR-RED)}{(NIR+RED)} \qquad (3\text{-}1)$$

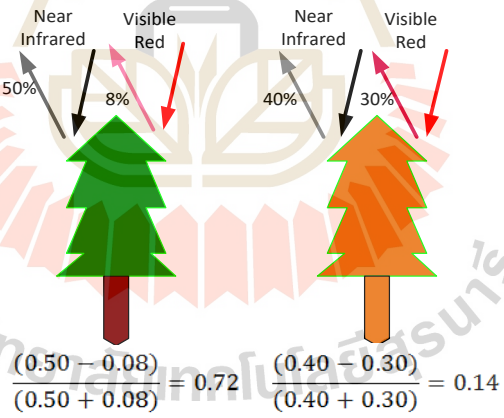
โดย NIR คือ ค่าการสะท้อนแสงช่วงความยาวคลื่นของแสงใกล้อินฟราเรด (%)

RED คือ ค่าการสะท้อนแสงช่วงความยาวคลื่นตามองเห็นของแสงสีแดง (%)

เพื่อให้การแปรผลทำได้ง่ายขึ้น ดัชนีผลต่างพืชพรรณจึงมีค่าอยู่ระหว่าง -1.0 ถึง +1.0 บริเวณที่ค่า NDVI อยู่ในช่วงค่าลบพื้นที่จะเป็นผืนน้ำ หรือทะเล ในบริเวณที่มีค่า NDVI เข้าใกล้ค่า 0 แสดงถึงพื้นที่ที่มีพืชพรรณสีเขียวน้อย และหากค่า NDVI มีค่าเข้าใกล้ +1.0 แสดงถึงพื้นที่ที่พืชสีเขียวปกคลุมมาก ตารางที่ 3.2 แสดงการแปรผลดัชนีผลต่างพืชพรรณ

ตารางที่ 3.2 การแปลผลดัชนีผลต่างพืชพรรณแบบนอมัลไลซ์

ค่า NDVI	ความหมาย
+0.60 ถึง +1.00	มีพืชพรรณอยู่หนาแน่นมาก เช่น พื้นที่ป่าไม้
+0.30 ถึง +0.59	มีพืชพรรณอยู่น้อย เช่น พื้นที่เกษตรกรรม
-1.00 ถึง +0.29	มีพืชพรรณปกคลุมน้อยมากหรือไม่มีอยู่เลย เช่น ทะเล



รูปที่ 3.1 การดูดซับและสะท้อนของช่วงคลื่นใกล้อินฟาเรดและช่วงคลื่นตามองเห็นสีแดงระหว่างพืชพรรณที่สมบูรณ์และพืชพรรณที่ไม่สมบูรณ์

3.2 กรอบแนวคิดของการวิจัย

งานวิจัยนี้เสนอวิธีการทำนายน้ำท่าด้วยวิธีการ ANN-GS ซึ่งเป็นโครงข่ายประสาทเทียมที่อาศัยการเรียนรู้จากการผสมผสานข้อมูลการคาดการณ์น้ำท่าของอัลกอริทึมโมเดลเชิงเส้นโดยนัยทั่วไป และซัพพอร์ตเวกเตอร์รีเกรสชัน กรอบแนวคิดโดยรวมของขั้นตอนการวิจัยแสดงได้ดังรูปที่ 3.2

การแบ่งข้อมูล: งานวิจัยนี้จะแบ่งข้อมูลออกเป็น 2 ชุด ได้แก่ ชุดข้อมูลฝึกสอนและชุดข้อมูลทดสอบ โดยแบ่งข้อมูลออกเป็น 70% และ 30% ตามลำดับ ข้อมูลชุดฝึกสอนจะใช้สำหรับการเรียนรู้ของอัลกอริทึมเพื่อสร้างโมเดลในการคาดการณ์น้ำท่า ส่วนข้อมูลชุดทดสอบจะใช้สำหรับการทดสอบประสิทธิภาพโมเดลที่สร้างขึ้น

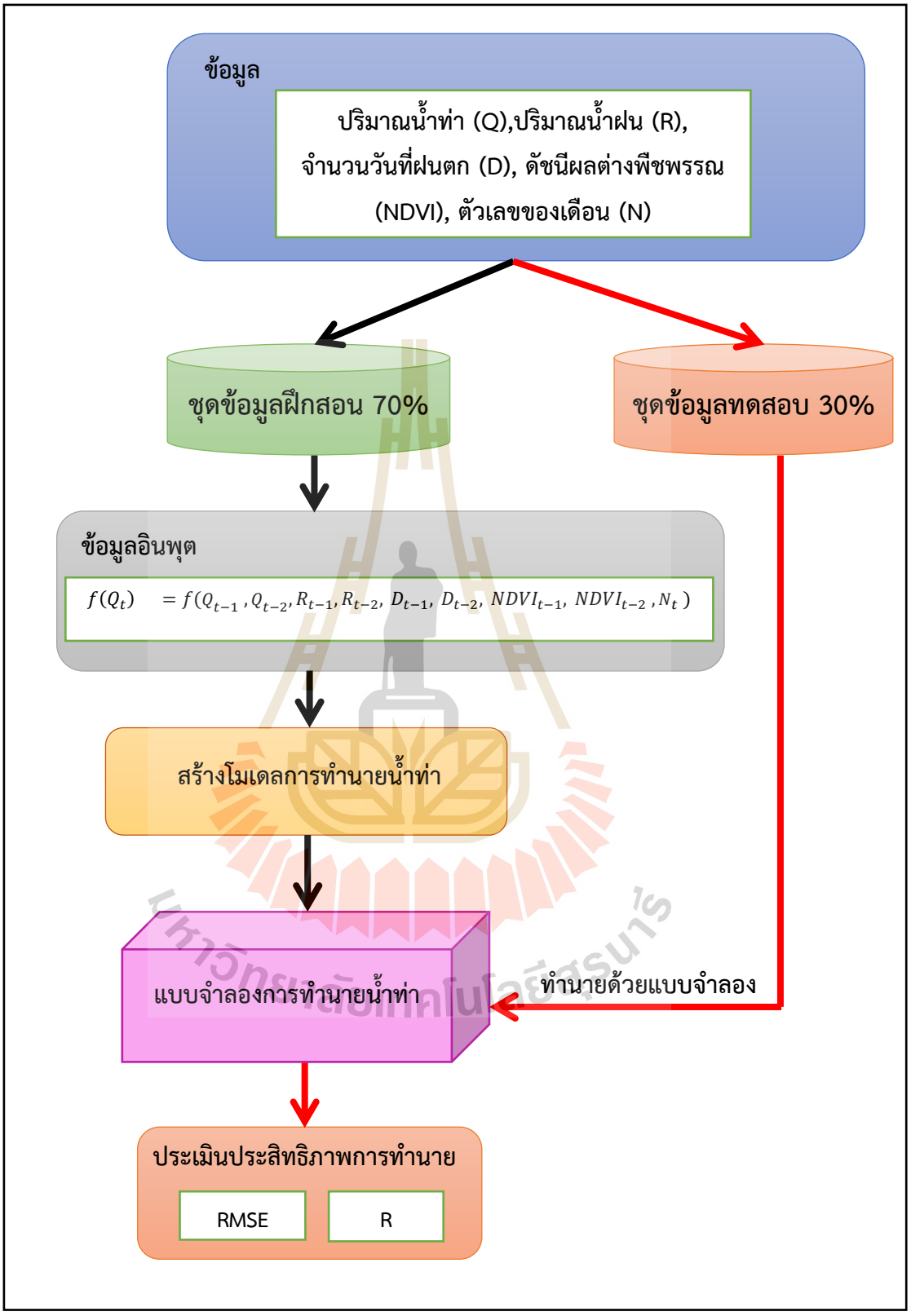
ข้อมูลอินพุต: ชุดข้อมูลฝึกสอนและชุดข้อมูลทดสอบจะประกอบไปด้วยค่าปริมาณน้ำฝน (R) จำนวนวันที่ฝนตก (D) ปริมาณน้ำท่า (Q) ดัชนีผลต่างพืชพรรณ (NDVI) และตัวเลขของเดือน (N) โดย

ในการกำหนดอินพุตเพื่อเข้าสู่การเรียนรู้ของเครื่องเพื่อทำการเรียนรู้ข้อมูลแล้วสร้างโมเดลที่เอาไว้ใช้ในการคาดการณ์น้ำท่า จะกำหนดให้ t คือเวลาที่สนใจ t-1 คือเวลาย้อนหลัง 1 เดือน t-2 คือเวลาย้อนหลัง 2 เดือน สามารถเขียนให้อยู่ในรูปสมการเพื่อให้เข้าใจง่ายดังนี้

$$f(Q_t) \quad = f(Q_{t-1}, Q_{t-2}, R_{t-1}, R_{t-2}, D_{t-1}, D_{t-2}, NDVI_{t-1}, NDVI_{t-1}, N_t) \qquad (3\text{-}2)$$

ถ้าต้องการพยากรณ์ค่าน้ำท่า ที่เดือนมีนาคม จะใช้ค่าปริมาณน้ำท่า ปริมาณน้ำฝน ดัชนีผลต่างพืชพรรณ จำนวนวันที่ฝนตก ในเดือนที่ t-1 และ t-2 ซึ่งคือเดือนมกราคมและกุมภาพันธ์ และตัวเลขของเดือนปัจจุบันคือ เดือนมีนาคม

$$f(Q_{\text{มี.ค.}}) = f(Q_{\text{ก.พ.}}, Q_{\text{ม.ค.}}, R_{\text{ก.พ.}}, R_{\text{ม.ค.}}, D_{\text{ก.พ.}}, D_{\text{ม.ค.}}, NDVI_{\text{ก.พ.}}, NDVI_{\text{ม.ค.}}, N_{\text{มี.ค.}}) \qquad (3\text{-}3)$$



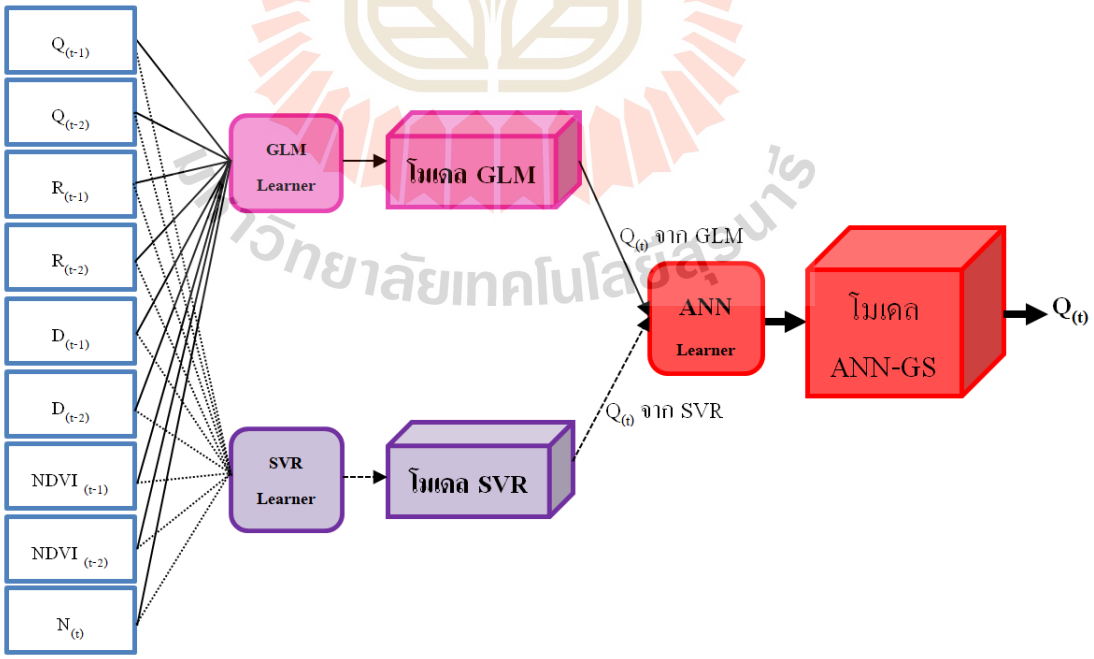
รูปที่ 3.2 กรอบแนวคิดโดยรวมของขั้นตอนการวิจัย

การสร้างโมเดล: การสร้างโมเดล ANN-GS ซึ่งเป็นโมเดลที่งานวิจัยนี้พัฒนาขึ้นเป็นการใช้โครงข่ายประสาทเทียมที่อาศัยการเรียนรู้จากการผสมผสานข้อมูลการคาดการณ์น้ำท่าของอัลกอริทึมโมเดลเชิงเส้นโดยนัยทั่วไป และซัพพอร์ตเวกเตอร์รีเกรสชัน ขั้นตอนการสร้างโมเดล ANN-GS จะเริ่มจากการนำข้อมูลอินพุตที่เตรียมไว้นำเข้าอัลกอริทึมโมเดลเชิงเส้นโดยนัยทั่วไป และซัพพอร์ตเวกเตอร์รีเกรสชัน จากนั้นเมื่อโมเดลทั้งสองให้ค่าการพยากรณ์ จะนำค่าพยากรณ์เหล่านั้นมาเป็นอินพุตให้แก่โครงข่ายประสาทเทียม ซึ่งโมเดลที่ได้จากโครงข่ายประสาทเทียมนี้จะใช้ในการพยากรณ์ค่าน้ำท่า สามารถแสดงรายละเอียดข้อมูลอินพุต และขั้นตอนทั้งหมดของ ANN-GS ได้ดังรูปที่ 3.3 และ 3.4 ตามลำดับ รูปแบบของข้อมูลอินพุตของ ANN-GS สามารถเขียนได้ดังสมการที่ 3.4

$$f(Q_t) = f(Q(GLM)_t, Q(SVR)_t) \tag{3-4}$$

ถ้าต้องการพยากรณ์ค่าน้ำท่าจากโมเดล ANN-GS ที่เดือนมีนาคม จะใช้ค่าปริมาณน้ำท่าเดือนมีนาคมจากโมเดลโมเดลเชิงเส้นโดยนัยทั่วไป และซัพพอร์ตเวกเตอร์รีเกรสชัน เขียนให้อยู่ในรูปสมการดังต่อไปนี้

$$f(Q_{มี.ค.}) = f(Q(GLM)_{มี.ค.}, Q(SVR)_{มี.ค.}) \tag{3-5}$$



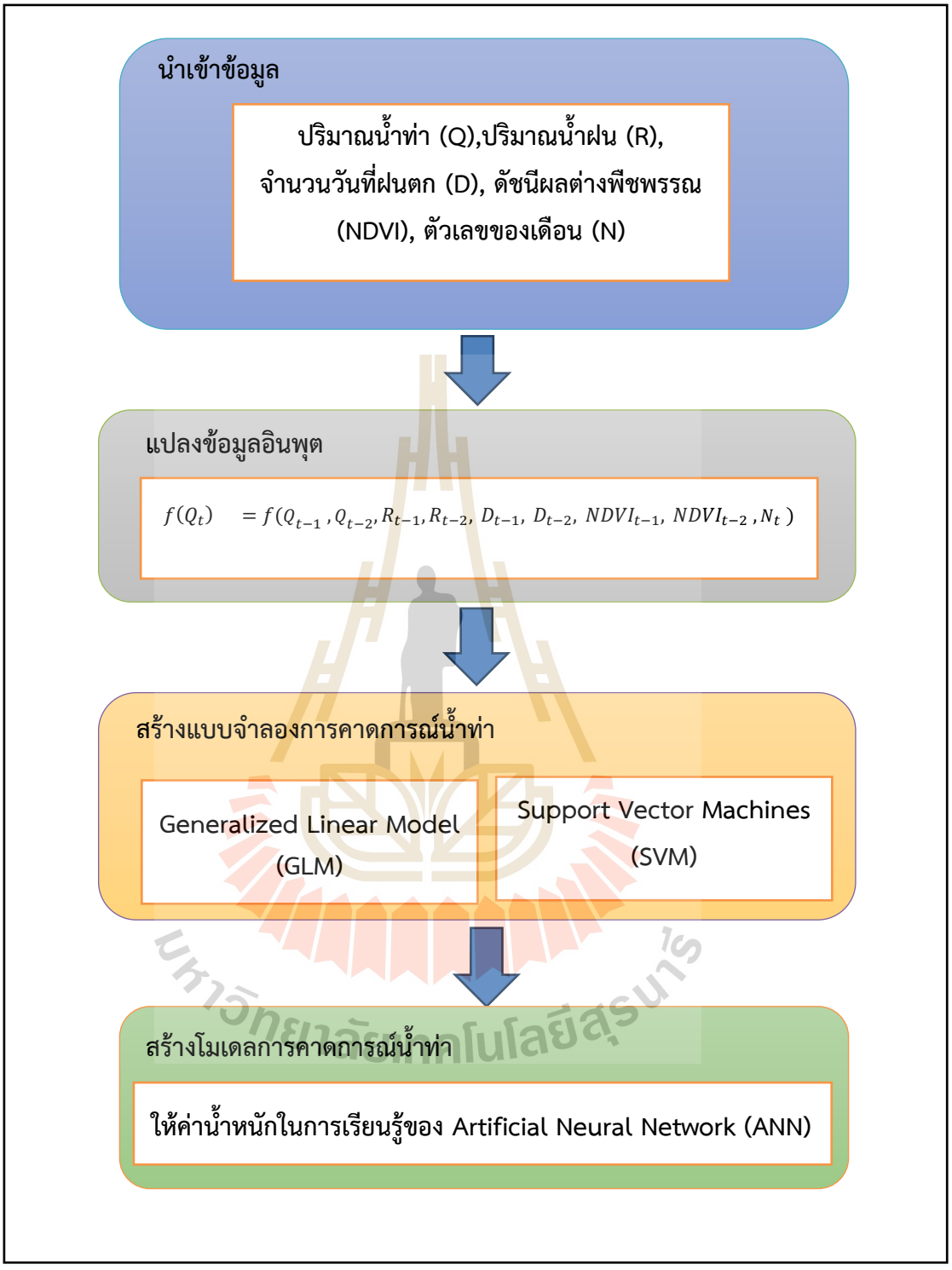
รูปที่ 3.3 แนวคิดของโมเดล ANN-GS

จากรูปที่ 3.3 โมเดล ANN-GS จะประกอบไปด้วยอัลกอริทึม 3 ชนิดได้แก่ ได้แก่ โมเดลเชิงเส้นโดยนัยทั่วไป ซัพพอร์ตเวกเตอร์รีเกรสชัน และโครงข่ายประสาทเทียม โดยจะกำหนดค่าการเรียนรู้ของแต่ละอัลกอริทึมดังนี้

- โมเดลเชิงเส้นโดยนัยทั่วไปจะกำหนดฟังก์ชันเชื่อมโยงตามการแจกแจงของค่าเป้าหมาย คือน้ำท่าที่เวลา t ของข้อมูลชุดฝึกสอน รูปแบบการแจกแจงอาจจะเป็นแบบ Normal, Binomial, Poisson หรือแบบ Gamma ก็ได้
- ซัพพอร์ตเวกเตอร์รีเกรสชันจะใช้เคอร์เนลฟังก์ชันเส้นตรง
- โครงข่ายประสาทเทียม เมื่อกำหนดให้ n เป็นจำนวนอินพุตและ m คือจำนวนเอาท์พุต การออกแบบสถาปัตยกรรมโครงข่ายจะกำหนดจำนวนโหนดในชั้นซ่อนโดยใช้สมการ 3.3 (Masters, 1993)

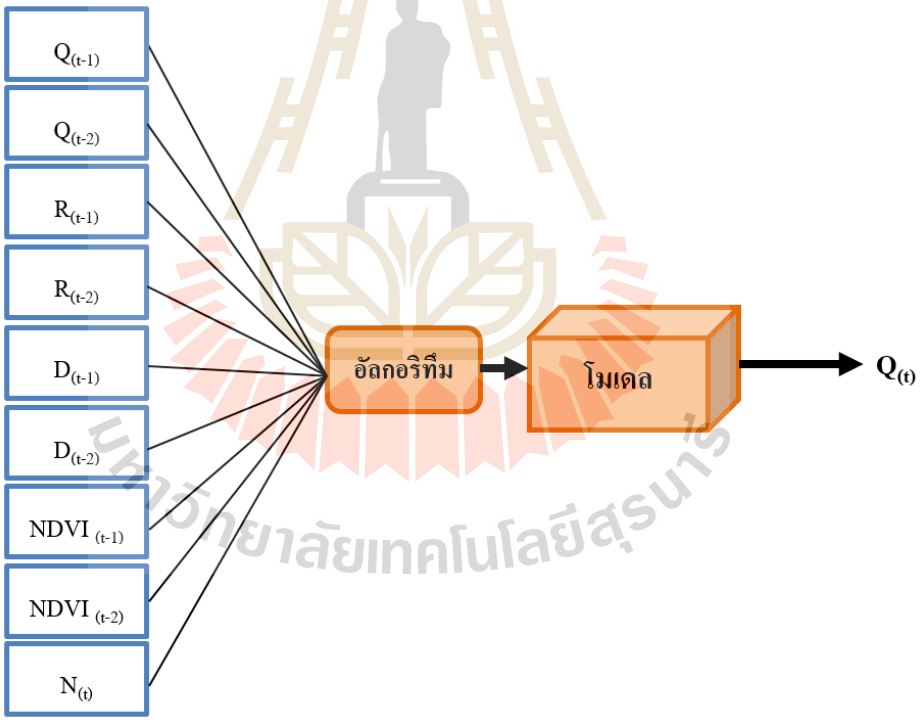
$$\sqrt{n * m} \tag{3.4}$$

ดังนั้นเมื่อแทนค่าจำนวนอินพุตมีค่าเท่ากับ 9 ($Q_{(t-1)}$, $Q_{(t-2)}$, $R_{(t-1)}$, $R_{(t-2)}$, $D_{(t-1)}$, $D_{(t-2)}$, $NDVI_{(t-1)}$, $NDVI_{(t-1)}$, N_t) และจำนวนเอาท์พุตเท่ากับ 1 ($Q_{(t)}$) ในสมการ 3.4 จะได้จำนวนโหนดในชั้นซ่อนคือ 3 โหนด



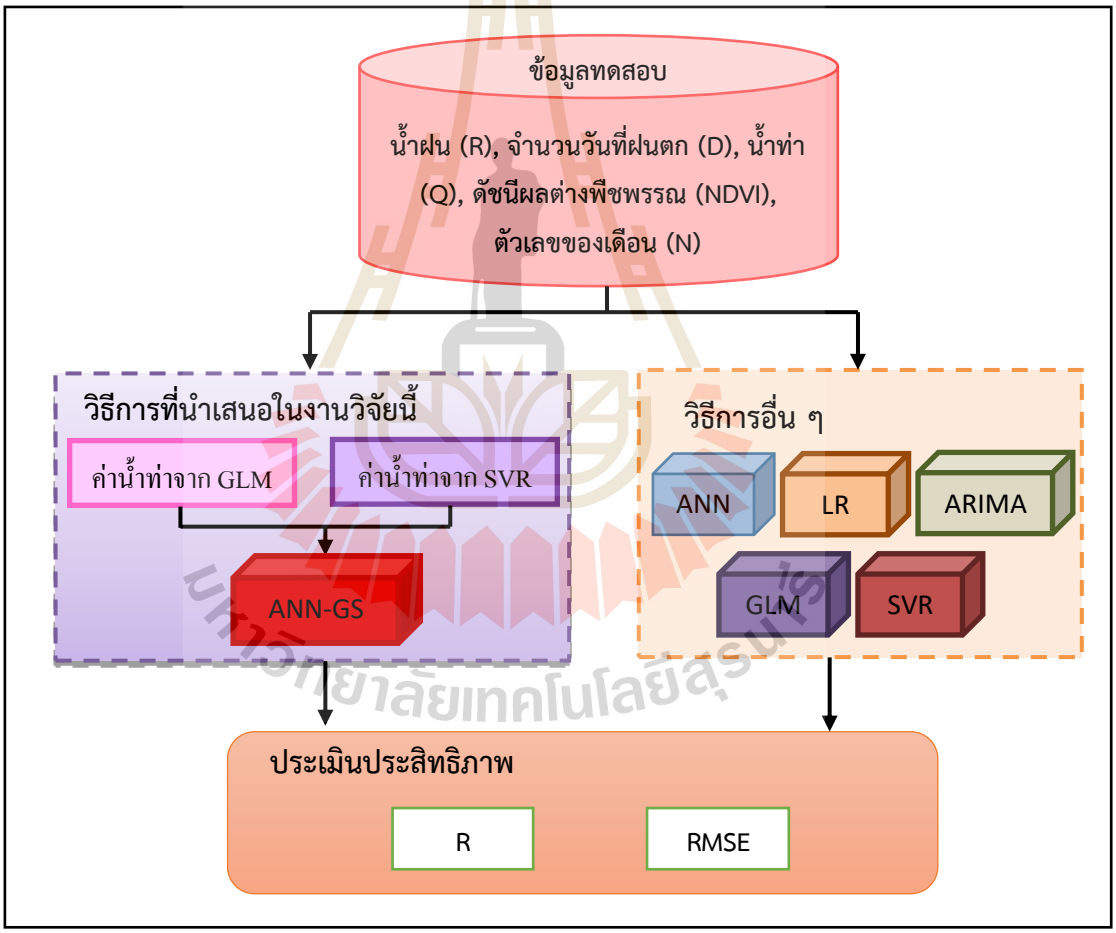
รูปที่ 3.4 สรุปขั้นตอนของการสร้างโมเดล ANN-GS

การสร้างโมเดลอื่น ๆ เพื่อนำมาใช้ในการเปรียบเทียบประสิทธิภาพกับวิธีการที่นำเสนอจะประกอบด้วยการใช้อัลกอริทึมที่ปรากฏในงานวิจัยอื่น ๆ ดังนี้ โครงข่ายประสาทเทียม การถดถอยเชิงเส้น โมเดลอาริม่า โมเดลเชิงเส้นโดยนัยทั่วไป และซัพพอร์ตเวกเตอร์รีเกรสชัน โดยที่อัลกอริทึมโครงข่ายประสาทเทียม โมเดลเชิงเส้นโดยนัยทั่วไป และซัพพอร์ตเวกเตอร์รีเกรสชันจะกำหนดค่าต่าง ๆ เหมือน ANN-GS ขั้นตอนการสร้างโมเดลจะเริ่มต้นด้วยการนำเข้าข้อมูลอินพุตที่เตรียมไว้ซึ่งประกอบไปด้วย ปริมาณน้ำฝน จำนวนวันที่ฝนตก ปริมาณน้ำท่า ดัชนีผลต่างพืชพรรณที่เวลาย้อนหลัง 1 และ 2 เดือนและใช้ค่าตัวเลขของเดือนที่สนใจ นำเข้าอัลกอริทึมต่าง ๆ จะได้โมเดลที่ใช้ในการพยากรณ์น้ำท่า การสร้างโมเดลอื่นที่นำมาเปรียบเทียบกับ ANN-GS สรุปได้ดังรูปที่ 3.5



รูป 3.5 ขั้นตอนการสร้างโมเดลอื่นที่นำมาเปรียบเทียบกับ ANN-GS

การประเมินประสิทธิภาพ: ประเมินประสิทธิภาพการคาดการณ์น้ำท่าของโมเดลต่าง ๆ จะทำโดยการใช้ค่าทางสถิติ 2 ชนิด ได้แก่ ค่าสัมประสิทธิ์สหสัมพันธ์ (R) และค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) ในการวัดประสิทธิภาพการพยากรณ์ค่าน้ำท่า การประเมินประสิทธิภาพการพยากรณ์น้ำท่าในงานวิจัยนี้จะทำได้โดยนำข้อมูลทดสอบนำเข้าโมเดลทั่วไปได้แก่ ANN MLR ARIMA GLM SVR และโมเดลที่นำเสนอในงานวิจัยนี้คือ ANN-GS เมื่อโมเดลต่าง ๆ ให้ค่าการคาดการณ์น้ำท่า จะนำค่าการคาดการณ์เหล่านั้นมาคำนวณ ค่า R และ RMSE จากนั้นเปรียบเทียบค่าทั้งสองเพื่อหาโมเดลที่คาดการณ์น้ำท่าได้ดีที่สุด วิธีการประเมินประสิทธิภาพโมเดลแสดงในรูป 3.6



รูปที่ 3.6 การทดสอบประสิทธิภาพของโมเดล

3.3 เครื่องมือที่ใช้ในงานวิจัย

เครื่องมือที่ใช้ในงานวิจัยนี้ ประกอบด้วยฮาร์ดแวร์และซอฟต์แวร์ ดังนี้

1) เครื่องคอมพิวเตอร์ โดยมีรายละเอียดดังนี้
   - หน่วยประมวลผลกลาง : Intel® Core i7
   - หน่วยความจำสำรอง : 1 TB
   - หน่วยความจำหลัก : 8 GB
   - อุปกรณ์เสริมอื่น ๆ เช่น เมาส์ แป้นพิมพ์ เป็นต้น
2) ระบบปฏิบัติการและโปรแกรมประยุกต์สำหรับสร้างโมเดลการพยากรณ์น้ำท่า ประกอบไปด้วย
   - ระบบปฏิบัติการ : Windows 10 Pro
   - เครื่องมือในการพัฒนาโปรแกรม : IBM SPSS Modeler เวอร์ชัน 18.0

บทที่ 4
การทดสอบและอภิปรายผล

การทดสอบประสิทธิภาพของโมเดลการคาดการณ์น้ำท่า ANN-GS นั้นจะทดสอบประสิทธิภาพของการคาดการณ์น้ำท่าด้วยชุดข้อมูลทดสอบ เปรียบเทียบประสิทธิภาพการคาดการณ์น้ำท่าด้วยมาตรวัดทางสถิติ 2 ค่า ได้แก่ ค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย และค่าสหสัมพันธ์ การทดสอบประสิทธิภาพการคาดการณ์น้ำท่าของงานวิจัยนี้เป็นดังนี้

4.1 ข้อมูลที่ใช้ในการทดสอบ

ข้อมูลที่ใช้เป็นปัจจัยสำหรับคาดการณ์น้ำท่าในงานวิจัยนี้ประกอบด้วย ปริมาณน้ำท่า ปริมาณน้ำฝน จำนวนวันที่ฝนตก ค่าดัชนีผลต่างพืชพรรณ และตัวเลขของเดือน ในการทดลองจะใช้ข้อมูลน้ำท่าของลุ่มน้ำในประเทศไทย (รูปที่ 4.1) โดยใช้ข้อมูลทั้งหมด 4 ชุด จาก 2 ลุ่มน้ำ ได้แก่ ลุ่มน้ำมูล (รูปที่ 4.2) และลุ่มน้ำปิง (รูปที่ 4.3) และแสดงมีรายละเอียดดังนี้

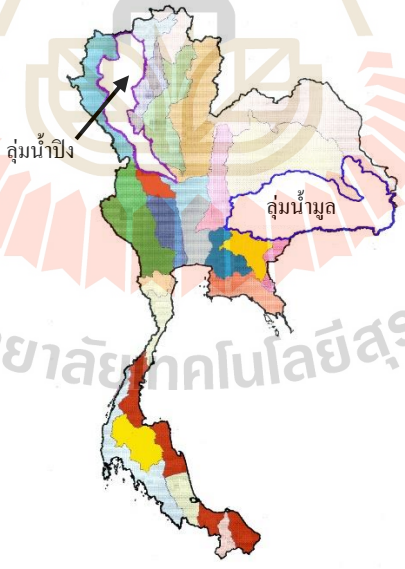
ข้อมูลน้ำท่าพื้นที่ของลุ่มน้ำมูลตอนบนของสถานี M.145 ลำพระเพลิง บ.วังตะเคียนทอง ต.วังกะทะ อ.ปากช่อง จ.นครราชสีมา ใช้ข้อมูลน้ำฝนและจำนวนวันที่ฝนตกที่อยู่ใกล้เคียงกับสถานีน้ำท่าคือ สถานี M.145 ลำพระเพลิง อ.ปากช่อง จ.นครราชสีมา ประกอบด้วยข้อมูลดัชนีผลต่างพืชพรรณ จ. นครราชสีมา เป็นข้อมูลรายเดือนที่เก็บรวบรวม 18 ปี ตั้งแต่ปี พ.ศ. 2541 ถึง 2558 โดยแบ่งข้อมูลเป็นข้อมูลชุดฝึกสอนตั้งแต่ปี พ.ศ. 2541 ถึง 2553 และใช้ข้อมูลช่วงปี พ.ศ. 2554 ถึง 2558 เป็นข้อมูลทดสอบ

ข้อมูลน้ำท่าพื้นที่ของลุ่มน้ำมูลตอนบนของสถานี M.173 แม่น้ำมูล บ.โนนสะอาด ต.ท่าเยี่ยม อ.โชคชัย จ.นครราชสีมา ใช้ข้อมูลน้ำฝนและจำนวนวันที่ฝนตกที่ตั้งอยู่ใกล้เคียงกับสถานีน้ำท่าคือ สถานีศูนย์อุทกวิทยาและบริหารน้ำภาคตะวันออกเฉียงเหนือตอนล่าง บ้านดอน ต.โคกกรวด อ.เมือง จ.นครราชสีมา ประกอบด้วยข้อมูลดัชนีผลต่างพืชพรรณ จ.นครราชสีมา ที่เก็บรวบรวม 14 ปี ตั้งแต่ปี พ.ศ. 2545 ถึง 2558 โดยแบ่งเป็นข้อมูลชุดฝึกสอนตั้งแต่ปี พ.ศ. 2545 ถึง 2554 และข้อมูลทดสอบปี พ.ศ. 2555 ถึง 2558

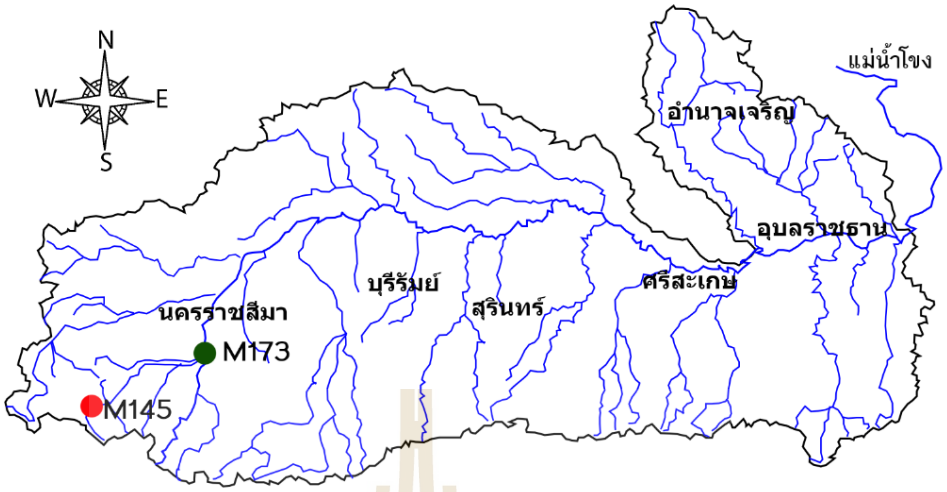
ข้อมูลน้ำท่าพื้นที่ของลุ่มน้ำปิง สถานี P.1 สะพานนวรัฐ อ.เมือง จ.เชียงใหม่ ใช้ข้อมูลน้ำฝนและจำนวนวันที่ฝนตกที่ตั้งอยู่ใกล้เคียงกับสถานีน้ำท่าคือ สำนักงานชลประทานที่ 1 อ.เมือง จ.เชียงใหม่ และใช้ข้อมูลดัชนีผลต่างพืชพรรณ จ. เชียงใหม่ ที่เก็บรวบรวม 19 ปี ตั้งแต่ปี พ.ศ. 2540 ถึง 2558 โดยแบ่งเป็นข้อมูลชุดฝึกสอนตั้งแต่ปี พ.ศ. 2540 ถึง 2553 และชุดข้อมูลทดสอบปี พ.ศ. 2554 ถึง 2558

ข้อมูลน้ำท่าพื้นที่ของลุ่มน้ำปิง สถานี P.4A อ.แม่แตง จ.เชียงใหม่ ใช้ข้อมูลน้ำฝนและจำนวนวันที่ฝนตกที่ตั้งอยู่ใกล้เคียงกับสถานีน้ำท่าคือ สถานีเขื่อนแม่งัด อ.แม่แตง จ.เชียงใหม่ ข้อมูลดัชนีผลต่างพืชพรรณ จ. เชียงใหม่ ที่เก็บรวบรวม 18 ปี ตั้งแต่ปี พ.ศ. 2541 ถึง 2558 โดยแบ่งเป็นข้อมูลชุดฝึกสอนตั้งแต่ปี พ.ศ. 2541 ถึง 2553 และข้อมูลทดสอบปี พ.ศ. 2555 ถึง 2558

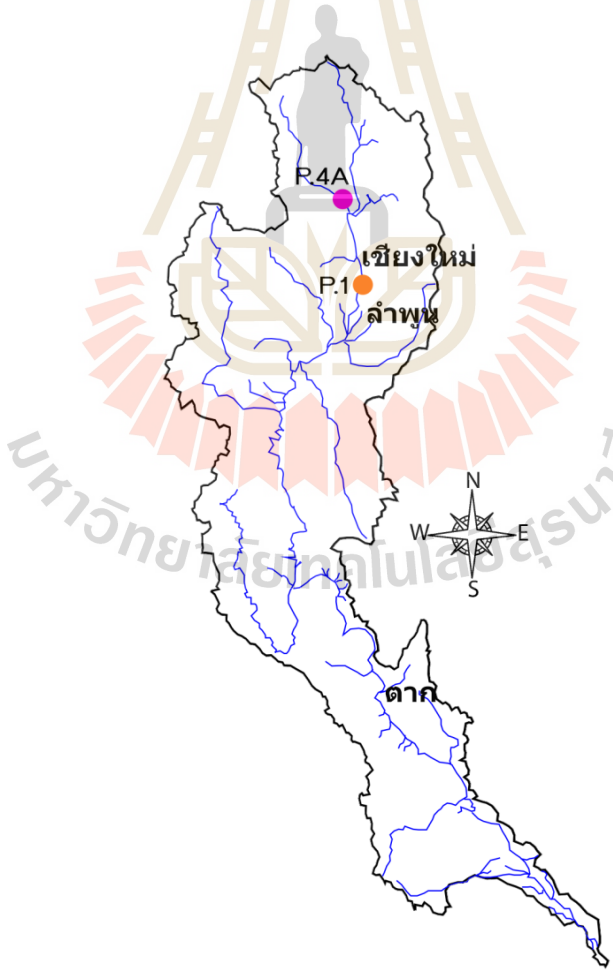
รายละเอียดของข้อมูลในแต่ละสถานีในตาราง 4.1 ข้อมูลทั้งหมดจะถูกรวบรวมให้อยู่ในรูปแบบของข้อมูลน้ำท่า (Q) และตัวเลข (N) ของเดือนที่สนใจ (t) น้ำท่า น้ำฝน (R) จำนวนวันที่ฝนตก (N) ดัชนีผลต่างพืชพรรณ (NDVI) ของเดือนย้อนหลังจากเดือนที่สนใจ 1 เดือน (t-1) และ 2 เดือน (t-2) ดังตารางที่ 4.2 และแสดงค่าสัมประสิทธิ์สหสัมพันธ์ในแต่ละแอตทริบิวต์ของสถานี M145 M173 P1 P4a ในตารางที่ 4.3 4.4 4.5 และ 4.6 ตามลำดับ



รูปที่ 4.1 แผนที่แสดงลุ่มน้ำมูลและลุ่มน้ำปิง



รูปที่ 4.2 แผนที่แสดงสถานีน้ำท่า M145 และ M173 บริเวณลุ่มน้ำมูล



รูปที่ 4.3 แผนที่แสดงสถานีน้ำท่า P1 และ P4a บริเวณลุ่มน้ำปิง

ตารางที่ 4.1 รายละเอียดของข้อมูลในแต่ละสถานี

ลุ่มน้ำ	สถานี	ข้อมูลฝึกสอน	ข้อมูลทดสอบ
ลุ่มน้ำมูลตอนบน	สถานี M.145 (ลำพระเพลิง) บ.วังตะเคียนทอง ต.วังกะทะ อ.ปากช่อง จ.นครราชสีมา	ปี พ.ศ. 2541 ถึง 2553 (13 ปี)	ปี พ.ศ. 2554 ถึง 2558 (5 ปี)
ลุ่มน้ำมูลตอนบน	สถานี M.173 (แม่น้ำมูล) บ.โนนสะอาด ต.ท่าเยี่ยม อ.โชคชัย จ.นครราชสีมา	ปี พ.ศ. 2545 ถึง 2554 (10 ปี)	ปี พ.ศ. 2555 ถึง 2558 (4 ปี)
ลุ่มน้ำปิง	สถานี P.1 (สะพานนวรัฐ) อ.เมือง จ.เชียงใหม่	ปี พ.ศ. 2540 ถึง 2553 (14 ปี)	ปี พ.ศ. 2554 ถึง 2558 (5 ปี)
ลุ่มน้ำปิง	สถานี P.4a (แม่แตง) อ.แม่แตง จ.เชียงใหม่	ปี พ.ศ. 2540 ถึง 2553 (14 ปี)	ปี พ.ศ. 2554 ถึง 2558 (5 ปี)

ตารางที่ 4.2 ตัวอย่างข้อมูลอินพุต

year	month	R_{t-1}	R_{t-2}	Q_{t-1}	Q_{t-2}	D_{t-1}	D_{t-2}	$NDVI_{t-1}$	$NDVI_{t-2}$	N_t	Q_t
2540	JAN	0	79	59.54	123.61	0	6	0.385	0.372	1	27.7
2540	FEB	52.9	0	27.7	59.54	9	0	0.415	0.385	2	22.37
2540	MAR	0	52.9	22.37	27.7	0	9	0.363	0.415	3	24.63
2540	APR	11.8	0	24.63	22.37	5	0	0.278	0.363	4	37.47
2540	MAY	57.6	11.8	37.47	24.63	10	5	0.274	0.278	5	55.87
2540	JUN	52.9	57.6	55.87	37.47	9	10	0.291	0.274	6	49.52
2540	JUL	26.5	52.9	49.52	55.87	9	9	0.207	0.291	7	113.35
2540	AUG	178.7	26.5	113.35	49.52	20	9	0.141	0.207	8	114.82
2540	SEP	236.4	178.7	114.82	113.35	22	20	0.2	0.141	9	198.53
2540	OCT	138	236.4	198.53	114.82	16	22	0.276	0.2	10	233.12
2540	NOV	162.9	138	233.12	198.53	8	16	0.362	0.276	11	83.4
2540	DEC	19.2	162.9	83.4	233.12	3	8	0.426	0.362	12	38.98

ตารางที่ 4.3 ค่าสัมประสิทธิ์สหสัมพันธ์ในแต่ละแอตทริบิวต์ของสถานี M145

แอตทริบิวต์	$NDVI_t$	D_t	N_t	R_t	Q_t
$NDVI_t$	1	0.079	0.698	0.107	0.326
D_t	0.79	1	0.172	0.843	0.502
N_t	0.698	0.172	1	0.145	0.316
R_t	0.107	0843	0.145	1	0.67
Q_t	0.326	0.502	0.316	0.67	1

ตารางที่ 4.4 ค่าสัมประสิทธิ์สหสัมพันธ์ในแต่ละแอตทริบิวต์ของสถานี M173

แอตทริบิวต์	$NDVI_t$	D_t	N_t	R_t	Q_t
$NDVI_t$	1	0.101	0.717	0.098	0.408
D_t	0.101	1	0.164	0.296	0.081
N_t	0.717	0.164	1	0.221	0.335
R_t	0.098	0.296	0.145	1	0.536
Q_t	0.408	0.081	0.316	0.536	1

ตารางที่ 4.5 ค่าสัมประสิทธิ์สหสัมพันธ์ในแต่ละแอตทริบิวต์ของสถานี P1

แอตทริบิวต์	$NDVI_t$	D_t	N_t	R_t	Q_t
$NDVI_t$	1	-0.631	0.95	-0.485	-0.146
D_t	-0.631	1	0.245	0.858	0.534
N_t	0.195	0.245	1	0.234	0.432
R_t	-0.485	0.858	0.234	1	0.666
Q_t	-0.146	0.534	0.432	0.666	1

ตารางที่ 4.6 ค่าสัมประสิทธิ์สหสัมพันธ์ในแต่ละแอตทริบิวต์ของสถานี P4a

แอตทริบิวต์	$NDVI_t$	D_t	N_t	R_t	Q_t
$NDVI_t$	1	-0.651	0.197	-0.551	-0.087
D_t	-0.651	1	0.210	0.896	0.463
N_t	0.197	0.210	1	0.213	0.436
R_t	-0.551	0.896	0.213	1	0.568
Q_t	-0.087	0.463	0.436	0.568	1

4.2 การทดสอบประสิทธิภาพ

การทดสอบประสิทธิภาพของการคาดการณ์น้ำท่าโดยใช้อัลกอริทึมที่นำเสนอคือ ANN-GS เปรียบเทียบกับอัลกอริทึมโครงข่ายประสาทเทียม การถดถอยเชิงเส้น โมเดลอาริม่า โมเดลเชิงเส้นโดยนัยทั่วไป และซัพพอร์ตเวกเตอร์รีเกรสชัน โดยแสดงผลการทดสอบตามเกณฑ์การพิจารณาดังนี้

- ค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Squared Error: RMSE) เพื่อแสดงความผิดพลาดในการคาดการณ์ของโมเดล ค่าที่มีค่า RMSE ที่ต่ำกว่าจะหมายถึงประสิทธิภาพที่ดีกว่าของโมเดล

- ค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient: R) เพื่อแสดงความสัมพันธ์ระหว่างปัจจัยของการคาดการณ์กับค่าเป้าหมายที่ต้องการคาดการณ์ ว่ามีความสัมพันธ์กันมากหรือน้อย ค่า R ที่สูงกว่าแสดงถึงประสิทธิภาพที่ดีกว่าของโมเดล

- ค่าการคาดการณ์น้ำท่าได้ดีขึ้นกว่าวิธีการแบบเดิม (ANN-GS outperform) เป็นการนำค่า RMSE ของวิธีการที่นำเสนอในงานวิจัยนี้คือ ANN-GS เพื่อดูเปอร์เซ็นของความผิดพลาดในการคาดการณ์น้ำท่าที่น้อยลงโดยเปรียบเทียบกับโมเดลจากอัลกอริทึมอื่น โดยคำนวณจากสูตรดังนี้

$$\text{ANN} - \text{GS outperform} = \frac{|\text{RMSE}_{other} - \text{RMSE}_{\text{ANN-GS}}|}{\text{RMSE}_{other}} \times 100 \qquad (4\text{-}1)$$

เมื่อ $\text{RMSE}_{\text{other}}$ คือ ค่า RMSE ของโมเดลจากอัลกอริทึมอื่น ๆ

$\text{RMSE}_{\text{ANN-GS}}$ คือ ค่า RMSE ของโมเดลจากอัลกอริทึม ANN-GS

- การเกิดเหตุการณ์ Overfitting เป็นการแสดงประสิทธิภาพของโมเดลว่าไม่สามารถใช้ได้จริงกับข้อมูลที่ไม่เคยเจอมาก่อนเพราะการเกิดเหตุการณ์ Overfitting คือเหตุการณ์ที่โมเดลทำงานกับข้อมูลชุดฝึกสอนได้ดี แต่เมื่อทำงานกับชุดข้อมูลทดสอบซึ่งเป็นข้อมูลที่ไม่เคยพบมาก่อนโมเดลไม่สามารถทำงานได้ ดังนั้นปัญหาการเกิด overfitting จึงเป็นปัญหาที่สำคัญ นักวิจัยจึงให้ความสนใจกับสร้างโมเดลเพื่อป้องกันการเกิด overfitting หลากหลายวิธีการ เช่น วิธีการป้องกันการเกิด overfitting ด้วยการโมเดลโดยการใช้ชุดข้อมูลตรวจสอบในระหว่างการฝึกสอนโมเดล เมื่อข้อผิดพลาดเริ่มมากขึ้นจะหยุดการฝึกสอนโมเดล (Lang et al., 1990; Doan et al., 2004) การแบ่งออกเป็นหลาย ๆ ชุดเพื่อทำหน้าที่สลับกันระหว่างการฝึกสอนและการทดสอบโมเดล (Haykin, 1999; Piotrowski et al., 2013) การใช้เทคนิคการลดข้อมูลฝึกสอนเพื่อหาข้อมูลฝึกสอนที่ดีที่สุดมาใช้ในการสร้างโมเดลเพื่อป้องกันการเกิด overfitting (Hindi et al., 2011; Sun et al., 2014) เป็นต้น แต่ในปัจจุบันยังไม่มีเกณฑ์ตัดสินการเกิด overfitting อย่างชัดเจน ดังนั้นงานวิจัยนี้จะใช้

การวัดการเกิด overfitting จากประสิทธิภาพของโมเดลในข้อมูลฝึกสอนและข้อมูลทดสอบต้องมีค่า RMSE ไม่แตกต่างกันเกิน 35% ถ้าแตกต่างกันเกิน 35% คือการเกิด Overfitting

$$Overfitting = \left(\frac{|RMSE_{train} - RMSE_{test}|}{RMSE_{train}} \times 100\right) > 35 \quad (4\text{-}2)$$

เมื่อ $RMSE_{train}$ คือ ค่า RMSE ของโมเดลจากชุดข้อมูลฝึกสอน

$RMSE_{test}$ คือ ค่า RMSE ของโมเดลจากชุดข้อมูลทดสอบ

- การประเมินประสิทธิภาพโดยรวมที่พิจารณาจากกราฟเส้นโดยในการพล็อตกราฟจะแสดงค่าที่คาดการณ์จากอัลกอริทึมต่าง ๆ เปรียบเทียบกับค่าจริงในข้อมูลชุดทดสอบ เพื่อดูภาพรวมในการคาดการณ์ค่าน้ำท่าของโมเดลว่าสามารถคาดการณ์น้ำท่าได้ครอบคลุมค่าสูง กลาง ต่ำ ได้หรือไม่เมื่อเทียบกับค่าจริง

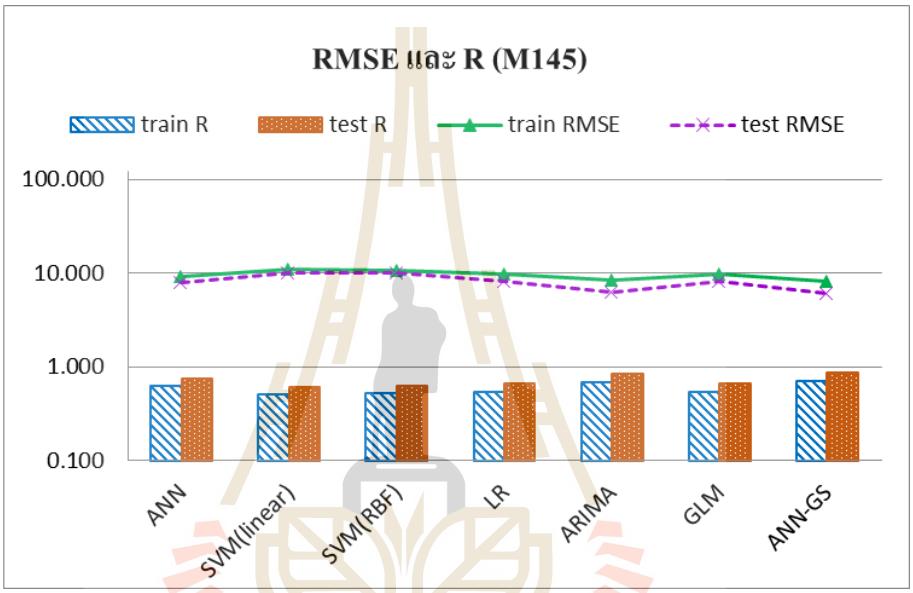
4.2.1 ผลการทดลองที่ข้อมูลสถานี M145

จากตารางที่ 4.7 แสดงผลการทดสอบที่สถานี M145 จะเห็นว่าวิธีการที่นำเสนอ ANN-GS ไม่เกิดเหตุการณ์ Overfitting และให้ประสิทธิภาพดีที่สุดในข้อมูลทดสอบเมื่อเทียบกับอัลกอริทึมอื่น ๆ โดยให้ค่า R มีค่า 0.874 และ RMSE มีค่า 0.599 โมเดล ANN-GS แสดงให้เห็นว่าสามารถการคาดการณ์น้ำท่าดีขึ้นจากโมเดลอื่น ๆ ได้สูงสุด 40.9% เมื่อเปรียบเทียบกับ SVR-Linear โมเดล ARIMA ให้ประสิทธิภาพรองลงมาจากวิธี ANN-GS ที่นำเสนอ ผลการทดลองที่สถานี M145 แสดงเป็นกราฟได้ดังรูปที่ 4.4 และแสดงค่าน้ำหนักของ ANN-GS ได้ดังรูปที่ 4.5

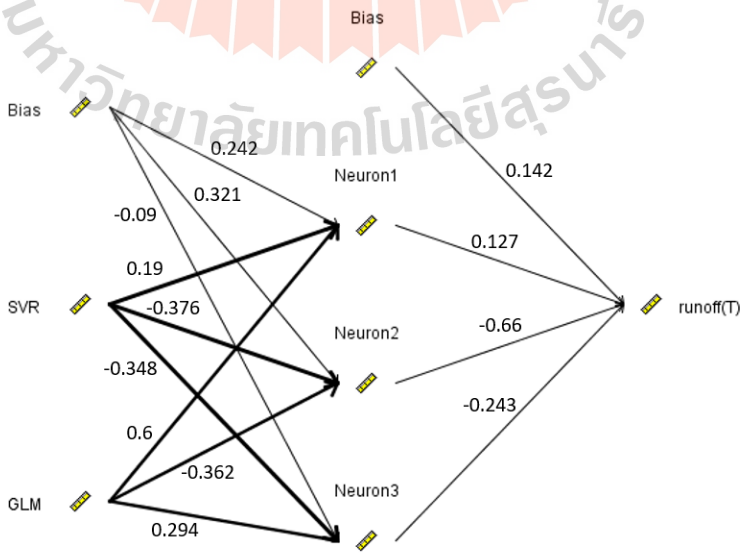
ตารางที่ 4.7 ผลการทดสอบที่สถานี M145

อัลกอริทึม	ข้อมูลฝึกสอน		ข้อมูลทดสอบ		ANN-GS outperform	Overfitting
	R	RMSE	R	RMSE		
ANN	0.623	9.083	0.750	7.878	23.9%	13.3%
SVR (linear)	0.515	10.856	0.612	10.143	40.9%	6.6%
SVR (RBF)	0.528	10.783	0.629	10.088	40.5%	6.4%
LR	0.534	9.785	0.667	8.166	26.5%	16.5%
ARIMA	0.684	8.460	0.832	6.184	3.0%	26.9%
GLM (identity)	0.534	9.785	0.667	8.166	26.5%	16.5%
ANN-GS	0.713	8.224	0.874	5.999	-	27.1%

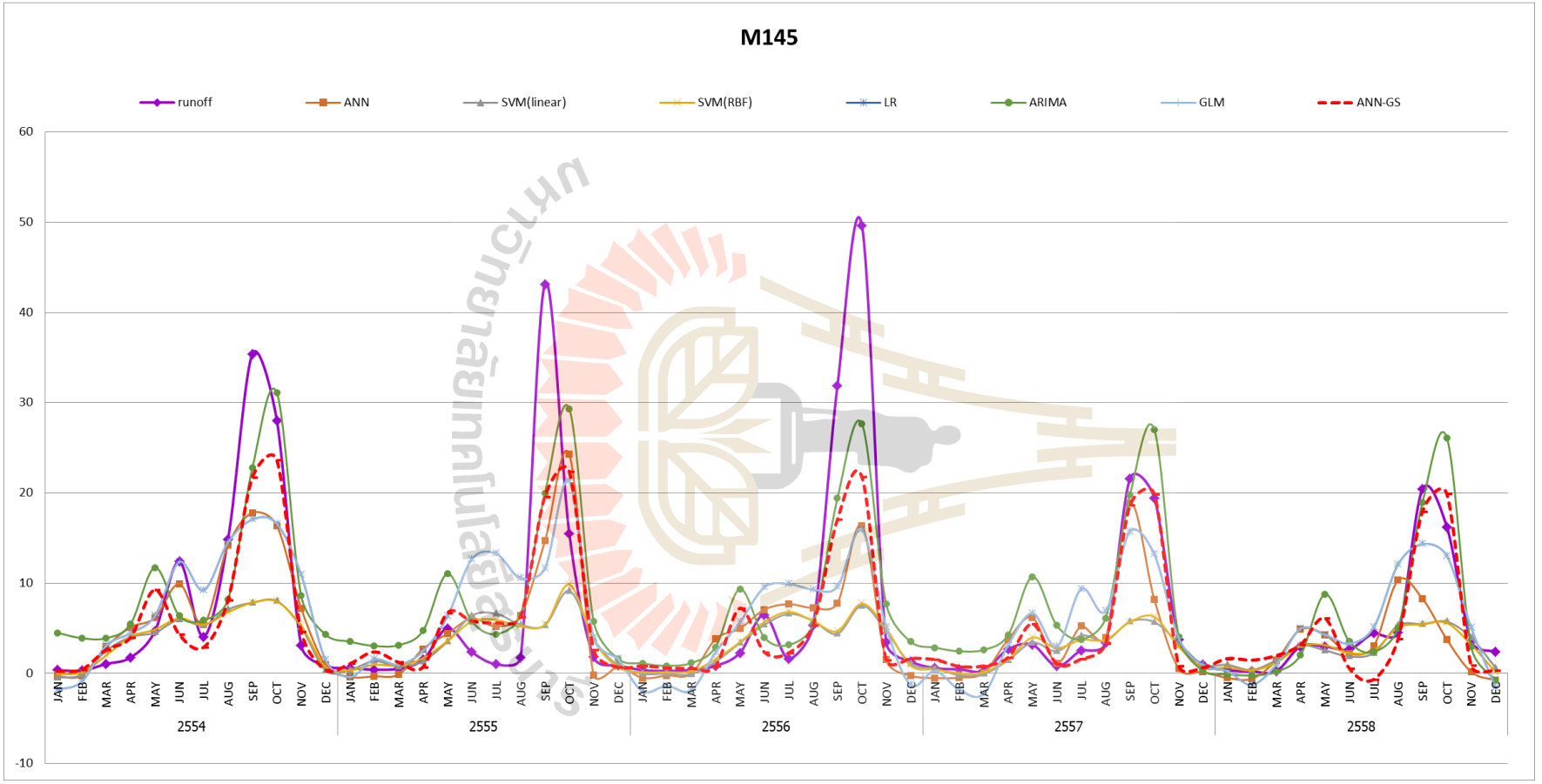
การเปรียบเทียบค่าคาดการณ์และค่าจริงดังรูปที่ 4.6 จะเห็นว่า ARIMA และ ANN-GS สามารถคาดการณ์นำท่าที่ปริมาณสูงได้ดีกว่าโมเดลอื่น แต่ ARIMA จะการคาดการณ์น้ำท่าปริมาณต่ำได้ไม่ดี โมเดล SVR ANN และ ANN-GS สามารถการคาดการณ์น้ำท่าปริมาณต่ำได้ดี แต่ SVR ไม่สามารถคาดการณ์น้ำท่าที่จุดสูง ๆ ได้ โมเดล LR และ GLM มีการคาดการณ์น้ำท่าไม่ใกล้เคียงกับค่าจริงทั้งค่าสูงและต่ำ จะเห็นได้ว่าวิธีการที่นำเสนอ ANN-GS มีภาพรวมในคาดการณ์ค่าน้ำท่าได้ที่ปริมาณสูงและต่ำ อีกทั้งยังเห็นได้ชัดว่า ได้ดีเมื่อเทียบกับอัลกอริทึมอื่น ๆ แต่ในปี



รูปที่ 4.4 กราฟแสดงค่า R และ RMSE ของข้อมูลฝึกสอนและข้อมูลทดสอบที่สถานี M145



รูปที่ 4.5 ค่าน้ำหนักของ ANN-GS ที่สถานี M145



รูปที่ 4.6 กราฟแสดงค่าคาดการณ์น้ำท่าเทียบกับค่าน้ำท่าจริงที่สถานี M145

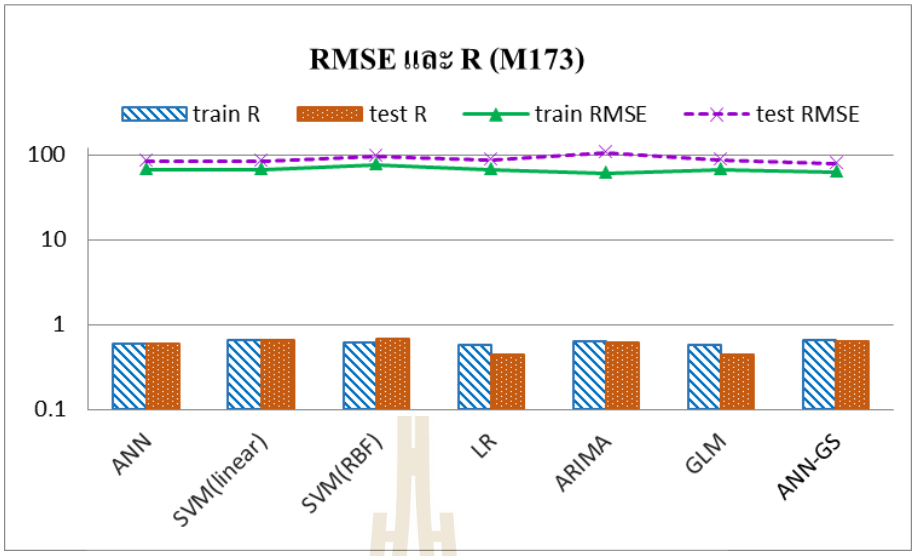
ตารางที่ 4.8 ผลการทดสอบที่สถานี M173

อัลกอริทึม	ข้อมูลฝึกสอน		ข้อมูลทดสอบ		ANN-GS outperform	Overfitting
	R	RMSE	R	RMSE		
ANN	0.578	66.271	0.579	85.412	8.7%	39.9%
SVR (linear)	0.654	66.636	0.638	84.999	8.2%	27.7%
SVR (RBF)	0.613	76.163	0.667	95.06	17.9%	24.1%
LR	0.565	66.334	0.438	87.161	10.5%	31.4%
ARIMA	0.628	60.591	0.599	106.606	26.8%	75.9%
GLM (identity)	0.565	66.334	0.438	87.161	10.5%	31.4%
ANN-GS	0.643	62.981	0.633	78.023	-	29.5%

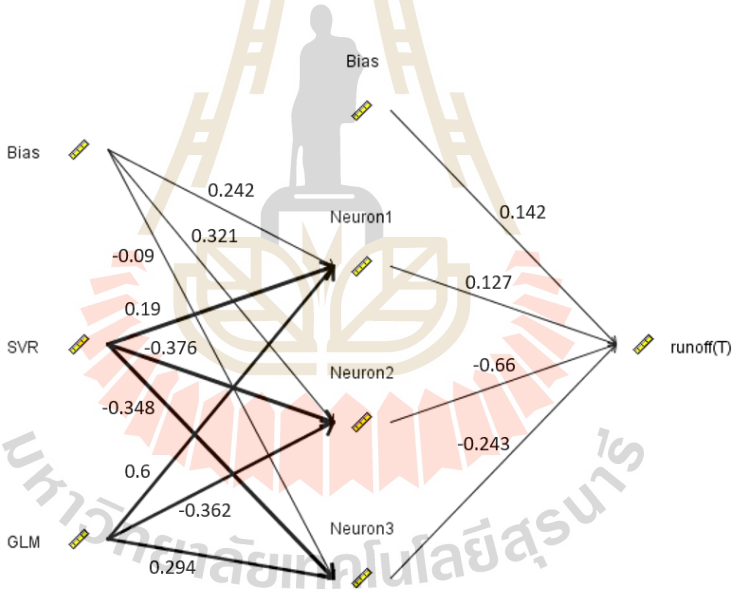
4.2.2 ผลการทดลองที่ข้อมูลสถานี M173

จากตารางที่ 4.8 แสดงผลการทดลองที่สถานี M173 เมื่อพิจารณาค่า R มีค่าสูงสุดคือโมเดล SVR-RBF โดยที่ R มีค่าเท่ากับ 0.667 และโมเดล ANN-GS ให้ค่าคาดการณ์มีความสัมพันธ์กับค่าจริงรองลงมา เมื่อพิจารณา RMSE มีค่าต่ำสุดที่โมเดล ANN-GS โดยที่ RMSE มีค่าเท่ากับ 78.032 โมเดล LR และ GLM ให้ประสิทธิภาพรองลง วิธีการ ANN-GS ที่นำเสนอแสดงให้เห็นว่าสามารถการคาดการณ์น้ำท่าได้ดีขึ้นจากโมเดลอื่น ๆ ได้สูงสุด 26.8% ที่ ARIMA โมเดลของทุกอัลกอริทึมไม่เกิดเหตุการณ์ Overfitting ยกเว้น ANN ผลการทดลองที่สถานี M173 แสดงเป็นกราฟได้ดังรูปที่ 4.7 และแสดงค่าน้ำหนักของ ANN-GSได้ดังรูปที่ 4.8

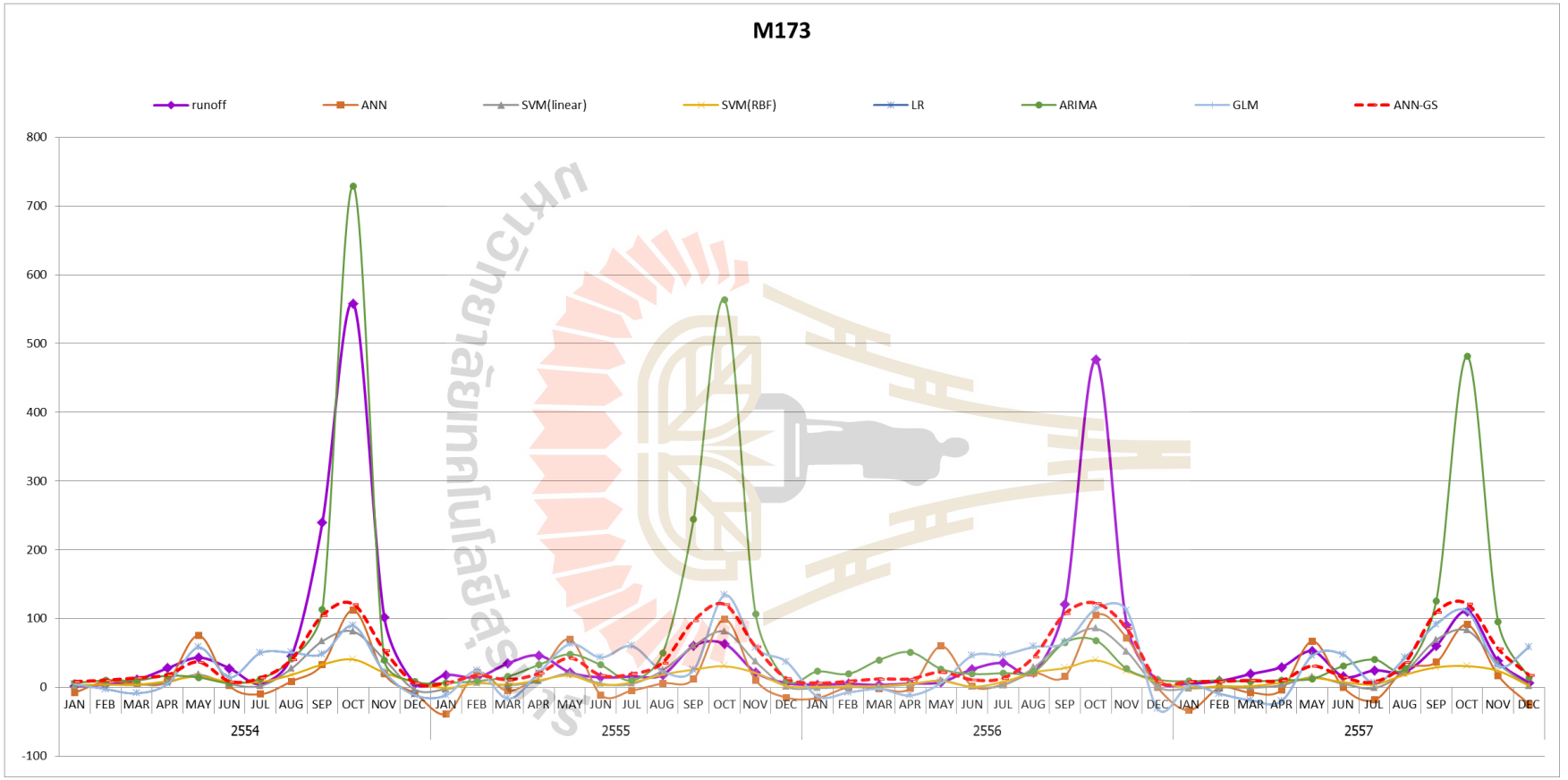
จากรูปที่ 4.9 จะเห็นได้ว่าปริมาณน้ำท่าที่จุดสูงสุดในแต่ละปีมีค่าสูงและต่ำสลับกัน ทำให้โมเดลจากทุกอัลกอริทึมไม่สามารถคาดการณ์น้ำท่าบริเวณนี้ได้โดยเฉพาะ ANN และ SVR วิธีการ ANN-GS ที่นำเสนอมีเส้นแนวโน้มใกล้เคียงกับค่าน้ำท่าจริง โมเดล LR และ GLM มีความผิดพลาดในการคาดการณ์จากค่าจริงมากเพราะแนวโน้มเส้นส่วนใหญ่ไม่เป็นแนวทางเดียวกับค่าน้ำท่าจริง



รูปที่ 4.7 กราฟแสดงค่า R และ RMSE ของข้อมูลฝึกสอนและข้อมูลทดสอบที่สถานี M173



รูปที่ 4.8 ค่าน้ำหนักของ ANN-GS ที่สถานี M173



รูปที่ 4.9 กราฟแสดงค่าคาดการณ์น้ำท่าเทียบกับค่าน้ำท่าจริงที่สถานี M173

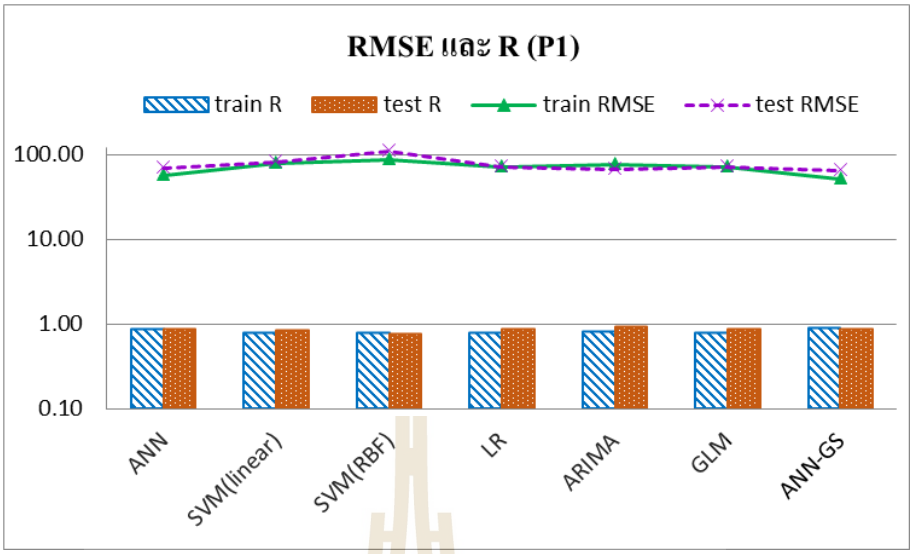
ตารางที่ 4.9 ผลการทดสอบที่สถานี P1

อัลกอริทึม	ข้อมูลฝึกสอน		ข้อมูลทดสอบ		ANN-GS outperform	Overfitting
	R	RMSE	R	RMSE		
ANN	0.88	56.467	0.87	68.876	4.3%	22.0%
SVR (linear)	0.78	78.06	0.84	82.882	20.5%	6.2%
SVR (RBF)	0.80	87.045	0.76	108.325	39.2%	24.4%
LR	0.80	70.932	0.86	72.047	8.6%	1.6%
ARIMA	0.82	76.253	0.92	66.864	1.5%	12.3%
GLM (identity)	0.80	70.932	0.86	72.047	8.6%	1.6%
ANN-GS	0.90	51.004	0.88	65.88	-	29.2%

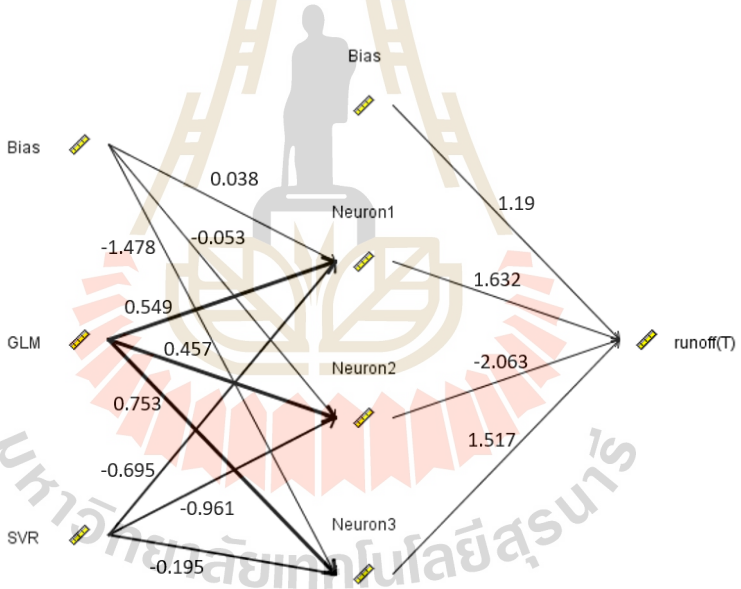
4.2.3 ผลการทดลองที่ข้อมูลสถานี P1

จากตารางที่ 4.9 แสดงผลการทดลองที่สถานี P1 เมื่อพิจารณาค่า R มีสูงสุดที่โมเดล ARIMA โดยมีค่าสูงสุดเท่ากับ 0.88 และเมื่อพิจารณาค่า RMSE มีค่าต่ำสุดที่ ANN-GS โดยมีค่าต่ำสุดที่ 65.88 วิธีการ ANN-GS ที่นำเสนอแสดงให้เห็นว่าสามารถการคาดการณ์น้ำท่าได้ดีขึ้นจากโมเดลอื่น ๆ ได้สูงสุด 39.2% ที่ SVR-RBF และโมเดลในทุกอัลกอริทึมไม่เกิดเหตุการณ์ Overfitting ผลการทดลองที่สถานี P1 แสดงเป็นกราฟในรูปที่ 4.10 และแสดงค่าน้ำหนักของ ANN-GSได้ดังรูปที่ 4.11

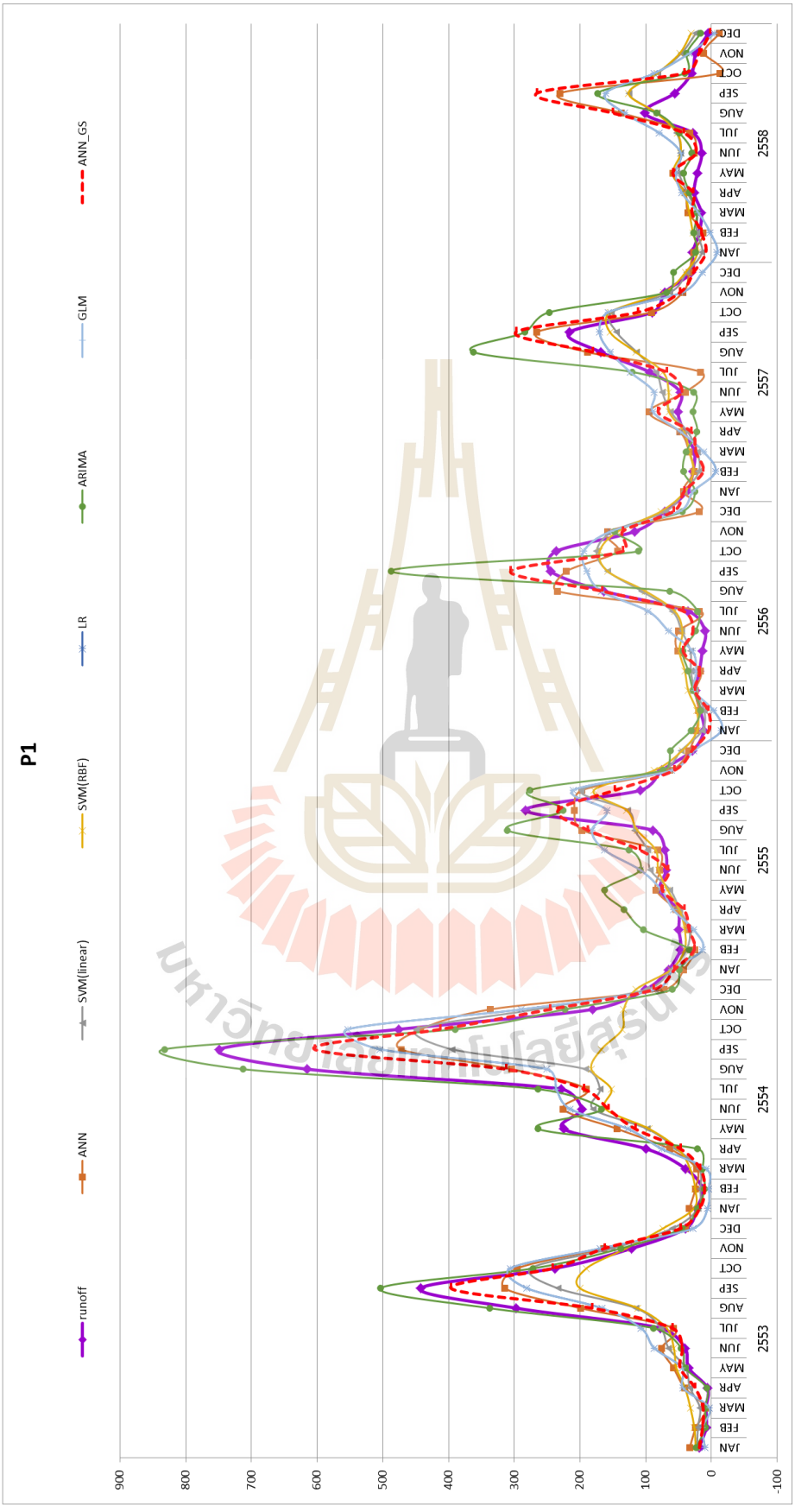
จากรูปที่ 4.12 โดยรวมแล้วทุดโมเดลมีแนวโน้มของเส้นกราฟใกล้เคียงกับค่าน้ำท่าจริง แต่ในปี พ.ศ. 2555 โมเดล ARIMA ไม่มีประสิทธิภาพในการคาดการณ์น้ำท่าโมเดล เพราะแนวโน้มของเส้นไม่เป็นไปตามค่าจริงนั้น จะเห็นได้ว่าซัพพอร์ตเวกเตอร์รีเกรสชันที่เคอร์เนล RBF สามารถคาดการณ์น้ำท่าปกติได้แต่ไม่สามารถคาดการณ์น้ำท่าที่มีปริมาณสูงได้ดีกว่าโมเดลอื่น ๆ อัลกอริทึมที่เสนอ ANN-GS มีภาพรวมในคาดการณ์ค่าน้ำท่าได้ใกล้เคียงกับค่าน้ำท่าจริงเพราะมีเส้นแนวโน้มใกล้เคียงกับค่าจริง



รูปที่ 4.10 กราฟแสดงค่า R และ RMSE ของข้อมูลฝึกสอนและข้อมูลทดสอบที่สถานี P1



รูปที่ 4.11 ค่าน้ำหนักของ ANN-GS ที่สถานี P1



รูปที่ 4.12 กราฟแสดงค่าคาดการณ์น้ำท่าเทียบกับค่าน้ำท่าจริงที่สถานี P1

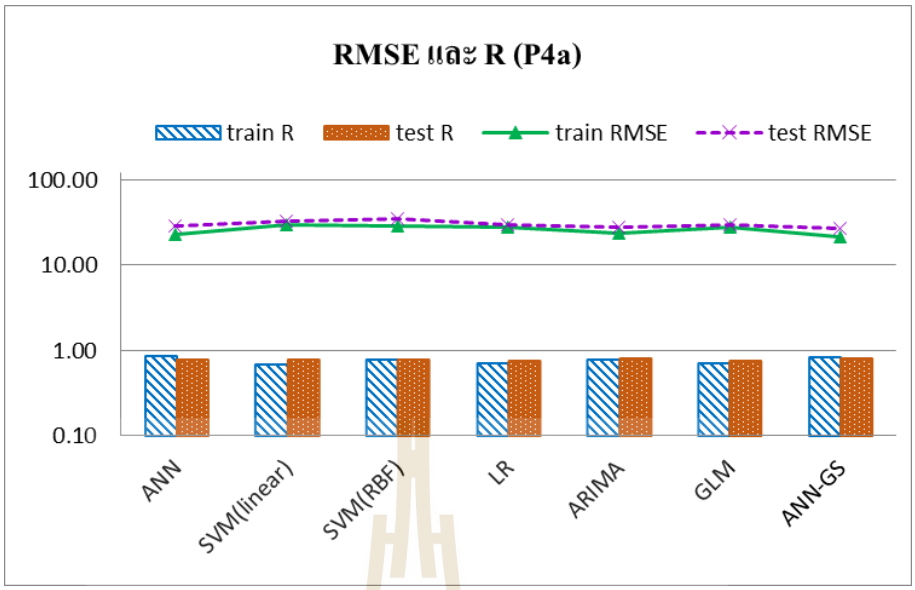
ตารางที่ 4.10 ผลการทดสอบที่สถานี P4a

อัลกอริทึม	ข้อมูลฝึกสอน		ข้อมูลทดสอบ		ANN-GS outperform	Overfitting
	R	RMSE	R	RMSE		
ANN	0.86	22.809	0.79	28.698	5.2%	25.8%
SVR (linear)	0.69	29.504	0.77	32.904	17.3%	11.5%
SVR (RBF)	0.78	28.328	0.77	34.437	21.0%	21.6%
LR	0.70	27.365	0.77	29.684	8.4%	8.5%
ARIMA	0.80	23.673	0.81	27.39	0.7%	15.7%
GLM (identity)	0.70	27.365	0.77	29.706	8.4%	8.6%
ANN-GS	0.83	21.441	0.81	27.197	-	26.8%

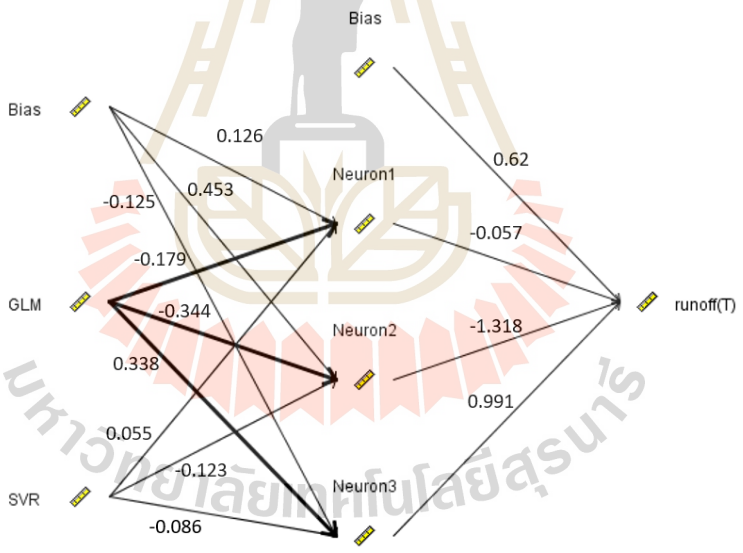
4.2.4 ผลการทดลองที่ข้อมูลสถานี P4a

จากตารางที่ 4.10 แสดงผลการทดสอบที่สถานี P4a จะเห็นว่าเมื่อพิจารณาค่า R มีค่าสูงที่สุดที่โมเดล ARIMA และวิธีการ ANN-GS ที่นำเสนอ โดยมีค่า R สูงสุดเท่ากับ 0.81 และค่า RMSE มีค่าต่ำสุดที่โมเดล ANN-GS โดยมีค่า RMSE เท่ากับ 27.197 ทุกโมเดลไม่เกิดเหตุการณ์ Overfitting เมื่อพิจารณาที่ค่า R โมเดล ANN ให้ค่าคาดการณ์น้ำท่ามีความสัมพันธ์ของกับน้ำท่าจริงมากรองลงมาจากวิธีการ ANN-GS และพิจารณาค่า RMSE โมเดล ARIMA มีความผิดพลาดในการคาดการณ์มากกว่า ANN-GS แต่น้อยกว่าวิธีการอื่น ๆ วิธีการ ANN-GS แสดงให้เห็นว่าสามารถการคาดการณ์น้ำท่าได้ดีขึ้นจากโมเดลอื่น ๆ ได้สูงสุด 21.0% ที่ SVR-RBF ผลการทดลองที่สถานี P4a แสดงเป็นกราฟในรูปที่ 4.13 และแสดงค่าน้ำหนักของ ANN-GS ได้ดังรูปที่ 4.14

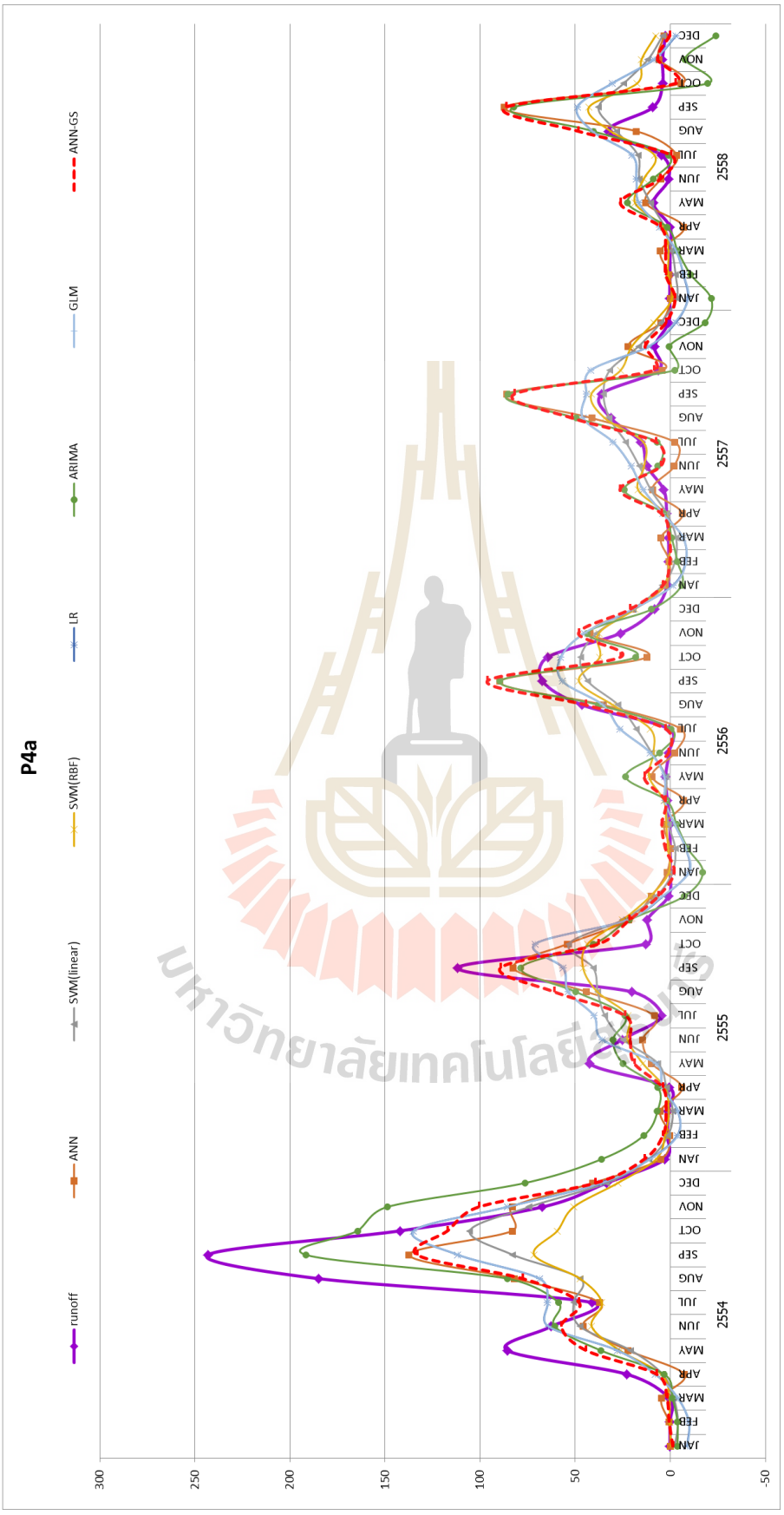
จากรูปที่ 4.15 จะเห็นว่าข้อมูลน้ำท่าของสถานี P4a ที่จุดสูงสุดในแต่ละปีมีค่าน้อยลงเรื่อย ๆ โมเดลทำให้วิธีการ ARIMA และ ANN-GS คาดการณ์ได้ใกล้เคียงเฉพาะจุดที่สูงแต่มีบางบริเวณที่ควรมีค่าสูงกลับมีค่าต่ำทำให้โมเดลมีความผิดพลาด วิธีการที่นำเสนอ ANN-GS มีภาพรวมในคาดการณ์ค่าน้ำท่าได้ใกล้เคียงกับค่าน้ำท่าจริงในช่วงเดือน ธันวาคม ถึง เมษายน



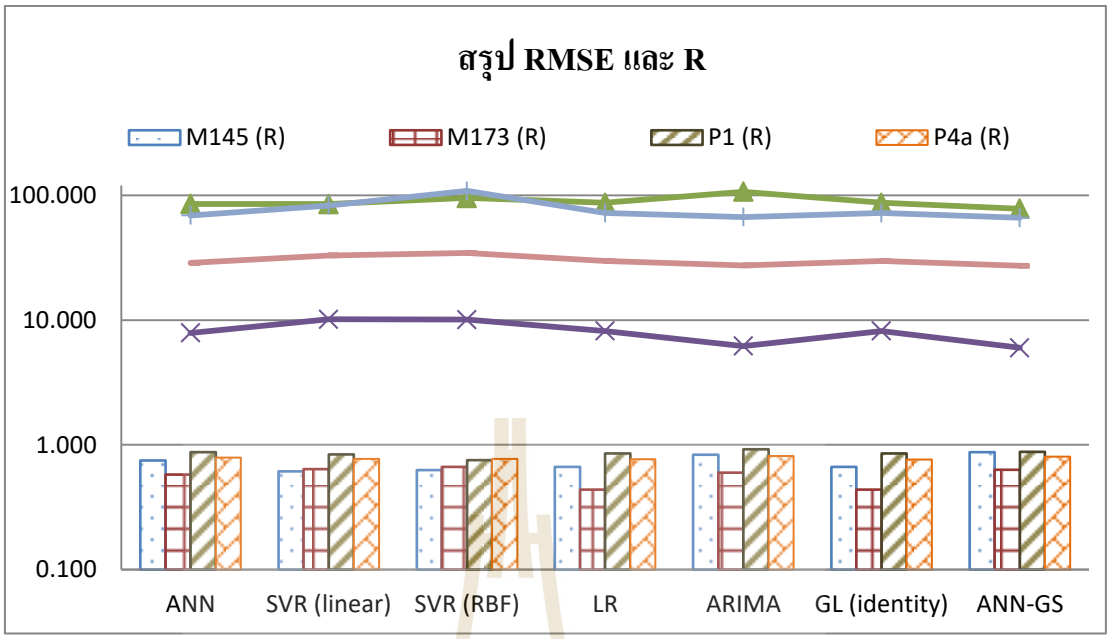
รูปที่ 4.13 กราฟแสดงค่า R และ RMSE ของข้อมูลฝึกสอนและข้อมูลทดสอบที่สถานี P4a



รูปที่ 4.14 ค่าน้ำหนักของ ANN-GS ที่สถานี P4a



รูปที่ 4.15 กราฟแสดงค่าคาดการณ์น้ำท่าเทียบกับค่าน้ำท่าจริงที่สถานี P4a



รูปที่ 4.16 กราฟแสดงค่า R และ RMSE ในข้อมูลทดสอบของทุกสถานี

4.3 อภิปรายผล

งานวิจัยนี้เสนอการคาดการณ์น้ำท่าด้วยวิธีการ ANN-GS ซึ่งเป็นโครงข่ายประสาทเทียมที่เกิดจากการเรียนรู้ของค่าคาดการณ์ปริมาณน้ำท่าของสองโมเดลได้แก่ โมเดลเชิงเส้นโดยนัยทั่วไป และซัพพอร์ตเวกเตอร์รีเกรสชัน โดยทำการเปรียบเทียบกับโมเดลจากอัลกอริทึมอื่น ๆ ได้แก่ โครงข่ายประสาทเทียม การถดถอยเชิงเส้น อาริม่า โมเดลเชิงเส้นโดยนัยทั่วไป และซัพพอร์ตเวกเตอร์รีเกรสชัน ข้อมูลที่ใช้ในการทดสอบมีทั้งหมด 4 ชุดข้อมูล เป็นข้อมูลจากลุ่มน้ำมูล 2 ชุดข้อมูล และมาจากลุ่มน้ำปิง 2 ชุดข้อมูล ผลการทดลองแสดงให้เห็นว่าข้อมูลทั้งหมดไม่เกิดเหตุการณ์ Overfitting แสดงว่าโมเดลที่สร้างขึ้นจากอัลกอริทึม ANN-GS สามารถใช้ในการคาดการณ์น้ำท่าได้ถึงแม้ว่าจะเป็นข้อมูลที่ไม่เคยเห็น

เมื่อพิจารณาที่ค่าสหสัมพันธ์จะเห็นว่าที่สถานี M145 วิธีการ ANN-GS ที่นำเสนอมีความสัมพันธ์ของค่าคาดการณ์น้ำท่ากับค่าน้ำท่าจริงสูงสุดเมื่อเทียบกับโมเดลอื่น ๆ สถานี M173 โมเดล SVR-RBF มีค่าสหสัมพันธ์สูงสุดรองลงมาคือ SVM-linear และ ANN-GS ตามลำดับ สถานี P1 โมเดล ARIMA ให้ค่าคาดการณ์น้ำท่ามีความสัมพันธ์กับค่าจริงสูงสุด รองลงมาคือวิธีการANN-GS สถานี P4a โมเดล ANN-GS และARIMA มีความสัมพันธ์สูงสุด รองลงมาคือโมเดล ANN

เมื่อพิจารณาค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ยวิธีการ ANN-GS ที่นำเสนอมีค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ยน้อยที่สุด แสดงว่า ANN-GS มีประสิทธิภาพในการคาดการณ์น้ำท่าได้ดีที่สุดในทุกชุดข้อมูลเมื่อเปรียบเทียบกับประสิทธิภาพกับโมเดลอื่น ๆ

จะเห็นได้ว่าเมื่อพิจารณาค่า R วิธีการ ANN-GS ที่นำเสนอไม่ได้มีค่ามากที่สุดในทุกสถานี โดยค่า R แสดงให้เห็นว่าวิธีการ ANN-GS ให้ค่าการคาดการณ์น้ำท่ามีความสัมพันธ์มากและไปในทางเดียวกันกับค่าจริงในทุกสถานี ถึงแม้วิธีการที่นำเสนอไม่ได้มี R ดีที่สุดแต่แสดงค่า RMSE น้อยที่สุดในทุกสถานี ซึ่งแสดงได้ว่า ANN-GS มีความผิดพลาดในการคาดการณ์น้ำท่าน้อยที่สุดในทุกสถานี นอกจากนี้เมื่อนำค่าคาดการณ์มาพล็อตกราฟจะเห็นว่าวิธีที่นำเสนอโดยรวมแล้วสามารถนำมาใช้ในการคาดการณ์น้ำท่าได้ดีที่ข้อมูลน้ำท่าที่มีปริมาณต่ำโดยเฉพาะช่วงต้นปีที่เดือนมกราคมถึงมิถุนายนและช่วงปลายปีที่เดือนตุลาคมถึงธันวาคมเพราะข้อมูลที่มีปริมาณต่ำมีจำนวนมากจึงทำให้ ANN-GS เรียนรู้และปรับค่าน้ำหนักจากความผิดพลาดจากข้อมูลน้ำท่าที่มีปริมาณต่ำมากกว่าน้ำท่าที่ปริมาณสูง ดังนั้นจึงทำให้โครงข่ายที่ได้จะสามารถคาดการณ์น้ำท่าปริมาณต่ำได้ดีกว่าปริมาณน้ำท่าสูง

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

ในปัจจุบันปัญหาที่เกิดจากน้ำ ได้แก่ การขาดแคลนน้ำและการเกิดน้ำท่วม มีความรุนแรงทวีคูณขึ้นทุกปี วิธีการเรียนรู้ของเครื่องสามารถนำมาใช้ในการคาดการณ์น้ำท่าเพื่อช่วยในการตั้งรับหรือบรรเทาการเกิดปัญหาที่เกิดจากน้ำได้ งานวิจัยนี้จึงมุ่งเน้นการพัฒนาอัลกอริทึมสำหรับใช้ในการคาดการณ์น้ำท่าให้มีประสิทธิภาพ อีกทั้งยังใช้ข้อมูลในการสร้างโมเดลที่หาได้สะดวกและใช้ข้อมูลจำนวนน้อย

5.1 สรุปขั้นตอนการดำเนินงานวิจัย

งานวิจัยนี้เสนออัลกอริทึมที่ชื่อว่า ANN-GS เป็นอัลกอริทึมที่ใช้ในการคาดการณ์น้ำท่ารายเดือน การพัฒนาอัลกอริทึม ANN-GS เกิดจากการนำโครงข่ายประสาทเทียมมาเรียนรู้ด้วยข้อมูลการคาดการณ์น้ำท่าของอัลกอริทึมโมเดลเชิงเส้นโดยนัยทั่วไป และซัพพอร์ตเวกเตอร์รีเกรสชัน โดยขั้นตอนการดำเนินงานวิจัยมีดังนี้

1) ศึกษาข้อมูลเกี่ยวกับกระบวนการเกิดรวมไปถึงความหมายของน้ำท่า และศึกษาอัลกอริทึมต่าง ๆ ที่ใช้ในการคาดการณ์น้ำท่า

2) ศึกษาปัญหาที่เกิดจากการใช้การเรียนรู้ของเครื่องมาใช้ในการคาดการณ์น้ำท่า และศึกษาการประยุกต์การใช้งานของอัลกอริทึมต่าง ๆ

3) ออกแบบอัลกอริทึม ANN-GS ซึ่งจะเกิดจากการทำงาน 2 ส่วนโดยส่วนแรกจำเป็นจะต้องสร้างโมเดลที่เกิดจากการเรียนรู้ของข้อมูลน้ำท่า น้ำฝน จำนวนวันที่ฝนตก ดัชนีผลต่างพืชพรรณ และตัวเลขของเดือน ด้วยอัลกอริทึมโมเดลเชิงเส้นโดยนัยทั่วไป และอัลกอริทึมซัพพอร์ตเวกเตอร์รีเกรสชัน จากนั้นจะนำค่าการคาดการณ์น้ำท่าจากโมเดลที่เกิดจากส่วนแรก ซึ่งจะมีค่าการคาดการณ์น้ำท่า 2 ค่า คือค่าที่เกิดจากโมเดลเชิงเส้นโดยนัยทั่วไปและค่าที่เกิดจากโมเดลจากซัพพอร์ตเวกเตอร์รีแกรสชัน นำค่าทั้งสองเหล่านี้มาฝึกสอนให้โครงข่ายประสาทเทียมเรียนรู้ จะทำให้ได้โมเดล ANN-GS

4) เปรียบเทียบประสิทธิภาพของวิธีการที่นำเสนอและวิธีการอื่น ๆ งานวิจัยนี้ใช้ มาตรวัดทางสถิติ 2 ค่า ได้แก่ ค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) และค่าสัมประสิทธิสหสัมพันธ์ (R) ค่า RMSE ที่ต่ำกว่าจะบอกถึงประสิทธิภาพที่ดีกว่า ในขณะที่ R ที่สูงกว่าจะบอกถึงประสิทธิภาพการคาดการที่ดีกว่า นอกจากนี้งานวิจัยยังนำเสนอการใช้มาตรวัด Overfitting

เพื่อพิจารณาว่าโมเดลมีความยืดหยุ่นสูงเหมาะสมที่จะนำไปใช้คาดการณ์ข้อมูลใหม่ที่เกิดขึ้นในอนาคตหรือไม่ เกณฑ์ในการตัดสิน Overfitting ทางงานวิจัยพิจารณาจากค่า RMSE ของข้อมูลฝึกสอนและข้อมูลทดสอบว่าจะต้องมีความแตกต่างกันไม่เกิน 35% โดยใช้ RMSE ของข้อมูลฝึกสอนเป็นฐานในการเปรียบเทียบ

5.2 สรุปผลการวิจัย

การทดสอบประสิทธิภาพของ ANN-GS จะเปรียบเทียบกับโครงข่ายประสาทเทียม การถดถอยเชิงเส้น อาริม่า โมเดลเชิงเส้นโดยนัยทั่วไป และซัพพอร์ตเวกเตอร์รีเกรสชัน โดยข้อมูลที่ใช้ในการทดลองนี้มีทั้งหมด 4 ชุดข้อมูล ได้แก่ข้อมูลสถานี M145 และ M173 เป็นข้อมูลจากลุ่มน้ำมูล และข้อมูลสถานี P1 และ P4a จากลุ่มน้ำปิง ข้อมูลที่ใช้สร้างโมเดลประกอบด้วย ปริมาณน้ำท่า ตัวเลขของเดือน ปริมาณน้ำฝน จำนวนวันที่ฝนตก ดัชนีผลต่างพืชพรรณ ประสิทธิภาพในการคาดการณ์น้ำท่าโดยใช้มาตรวัดค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ยแสดงให้เห็นว่า ANN-GS มีความผิดพลาดต่ำสุดเมื่อเทียบกับอัลกอริทึมอื่น ๆ เมื่อวัดประสิทธิภาพด้วยค่าสัมประสิทธิ์สหสัมพันธ์ซึ่งเป็นค่าที่แสดงถึงค่าที่คาดการณ์จากโมเดลมีความสัมพันธ์กับค่าปัจจัยที่ใช้ในการคาดการณ์หรือไม่ ผลปรากฏว่า ANN-GS มีความสัมพันธ์ตั้งแต่ 0.66-0.88 ซึ่งแสดงความสัมพันธ์ที่ระดับปานกลางถึงสูงมาก อีกทั้งวิธีการที่นำเสนอเป็นโมเดลที่สามารถใช้งานกับค่าที่ไม่เคยเห็นได้เนื่องจากไม่เกิดเหตุการณ์ Overfitting นอกจากนี้เมื่อนำค่าที่คาดการณ์จากโมเดลมาพล็อตกราฟเทียบกับค่าจริงยังแสดงให้เห็นอย่างชัดเจนว่า ANN-GS มีการคาดการณ์ค่าน้ำท่าได้ใกล้เคียงกับค่าจริงที่ข้อมูลน้ำท่าที่มีปริมาณต่ำหรือในช่วงฝนแล้งโดยเฉพาะช่วงต้นปีที่เดือนมกราคมถึงมิถุนายนและช่วงปลายปีที่เดือนตุลาคมถึงธันวาคม ส่วนในน้ำท่าที่มีปริมาณสูงหรือช่วงหน้าฝน ANN-GS ยังมีประสิทธิภาพในการคาดการณ์ได้ไม่ดีเท่า ARIMA

5.3 ปัญหาและข้อเสนอแนะ

การคาดการณ์ปริมาณน้ำท่าเป็นวิธีการที่ค่อนข้างยาก เนื่องจากค่าที่คาดการณ์เป็นตัวเลขจำนวนจริงดังนั้นค่าที่คาดการณ์จากโมเดลจะมีความเป็นไปได้มากมาย ดังนั้นข้อมูลในการใช้ฝึกสอนเป็นสิ่งจำเป็นต่อประสิทธิภาพในการคาดการณ์ เพราะเป็นข้อมูลที่ใช้ในการเรียนรู้ของอัลกอริทึมเพื่อสร้างโมเดล เนื่องจากข้อมูลเป็นน้ำท่ารายเดือนดังนั้นข้อมูลที่เป็นค่าน้ำท่าปริมาณสูงจะมีตัวอย่างให้อัลกอริทึมเรียนรู้น้อย ดังนั้นการที่โมเดลจะคาดการณ์น้ำท่าปริมาณสูงจึงเป็นเรื่อง

ยาก ซึ่งจะเห็นว่าวิธีการที่นำเสนอค่าน้ำท่าปริมาณสูงบางข้อมูลก็คาดการณ์ได้ไม่ดีนัก ถ้ามีปริมาณข้อมูลเพิ่มขึ้นอาจจะทำให้โมเดลคาดการณ์น้ำท่าได้มีประสิทธิภาพดียิ่งขึ้น

ในอนาคตจะพิจารณาสภาพภูมิประเทศ และข้อมูลอื่น ๆ จากสถานีใกล้เคียงเพื่อใช้พัฒนาการสร้างโมเดลการคาดการณ์น้ำท่าที่มีประสิทธิภาพที่ดียิ่งขึ้น

รายการอ้างอิง

กมลวรรณ สารพานิช. (2555). การพยากรณ์ราคาน้ำมันดิบล่วงหน้าในตลาดฟิวเจอร์ไนเม็กซ์โดยวิธีอารีมาและอารีแมกซ์. **วิทยานิพนธ์เศรษฐศาสตร์มหาบัณฑิต สาขาวิชาเศรษฐศาสตร์ธุรกิจบัณฑิตวิทยาลัย มหาวิทยาลัยธุรกิจบัณฑิตย์.**

จินดามาส สุทธิชัยเมธี. (2544). การประยุกต์ใช้ ARIMA Model เพื่อการวิจัย. **วารสารสุทธิปริทัศน์**, ฉบับ 76 หน้า 101-120.

พรสิน สุภวาลย์. (2556). **การวิเคราะห์การถดถอย.** กรุงเทพฯ: มหาวิทยาลัยราชภัฏพระนคร.

มนต์ชัย เทียนทอง. (2548). **สถิติและวิธีการวิจัยทางเทคโนโลยีสารสนเทศ**. กรุงเทพฯ: สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

ศูนย์วิจัยกสิกรไทย. (2547). **ภาวะขาดแคลนน้ำ : ภัยร้ายที่กำลังมาเยือน**. [ระบบออนไลน์]. แหล่งที่มา https://mgronline.com/business/detail/9470000089753 (26 กุมภาพันธ์ 2559).

Agarwal A. and Singh R.D. (2004). Runoff modeling through back propagation artificial neural network with variable rainfall–runoff data. **Water Resources Management**.18: 285–300.

Aichouri I., Hani A., Bougherira N., Djabri L., Chaffai H., and Lallahem S. (2015). River flow model using artificial neural networks. **International Conference on Technologies and Materials for Renewable Energy**. 74: 1004–1014.

Castellano-Mendez, M., Gonzalez-Manteiga, W., Febrero-Bande, M., PradaSanchez, J.M., and Lozano-Calderon, R. (2004). Modelling of the monthly and daily behaviour of the runoff of the Xallas River using Box-Jenkins and neural networks methods, **Journal of Hydrology**. 296: 38–58.

Chen S. M., Wang Y. M. and Tsou I. (2013). Using artificial neural network approach for modeling rainfall–runoff due to typhoon, **Journal of Earth System Science**. 122(2): 399–405.

Choubey, S. M. V ., Pandey, S., and Shukla, J. (2014). An efficient approach of support vector machine for runoff forecasting. **International Journal of Scientific & Engineering Research.** 5(3): 158-166.

Daniel T. L. and Chantal D. L. (2015). **Data mining and predictive analytics**. John Wiley & Sons.

David T. G. (2003). Rainfall-runoff Processes. A workbook to accompany the Rainfall-Runoff Processes Web [On-line]. Available: http://hydrology.usu.edu/RRP/userdata/4/87/RainfallRunoffProcesses.pdf

Demuth H, and Beale M. (2000). **Neural Network Toolbox User's Guide**. Mathworks, Inc. Natick, MA .

Demuth H. B., and Hagan M. T. (2000). **Neural network design 2nd edtion**, MHB Inc.

Doan C. D., and Liong S. Y. (2004). Generalization for multilayer neural network: Bayesian regularization or early stopping, In **Proceedings of the 2nd Conference on Asia Pacific Association of Hydrology and Water Resources**, Singapore.

Dorum A., Yarar A., Sevimli M.F., Onüçyildiz M. (2010). Modelling the rainfall–runoff data of susurluk basin. **Expert Systems with Applications.** 37.9: 6587-6593.

Granata F., Gargano R., and de Marinis G. (2016). Support vector regression for rainfall-runoff modeling in urban drainage: A comparison with the EPA's storm eater management model, **Water**. 8(3):69.

Haykin, S. (1999) **Neural Networks: A Comprehensive Foundation (2nd Edition).** Prentice-Hall Inc., Englewood Cliffs, New Jersey, USA.

Hindi K.E., and AL-Akhras M. (2011). Smoothing decision boundaries to avoid overfitting in neural network training. **Neural Network World**, 21(4): 311-325.

Hinkle, D.E, William ,W. and Stephen G. J. (1998). **Applied statistics for the behavior sciences**. 4th ed. New York : Houghton Mifflin.

Jacobs M.C. (1992). **Regression trees versus stepwise regression.** Master's thesis, University of North Florida, Jacksonville, Florida.

John Fox. (2008). **Applied regression analysis and generalized linear models**. USA: SAGE Publication, Inc.

Katimon A., Shahid S., and Mohsenipour M. (2017). Modeling water quality and hydrological variables using ARIMA: a case study of Johor River, Malaysia, **Sustainable Water Resources Management.** pp. 1-8, 2017.

Kriegler F.J., Malila W.A., Nalepka R.F., and Richardson W. (1969). Preprocessing transformations and their effects on multispectral recognition, **Proceedings of the Sixth International Symposium on Remote Sensing of Environment**. University of Michigan, Ann Arbor, MI, p. 97-131

Kumar S. (2004). Neural network –A classroom Approach. [On-line]. Available: http://cs.jnu.edu.cn/international/soft/pptchapter08.pdf

Lang K., Waibel A., and Hinton, G. (1990) A time-delay neural network architecture for isolated word recognition. **Neural Networks**. 3, 23–43.

Masters T. (1993). **Practical neural network recipes in C++,** Academic Press, New York.

McIntyre N., Al-Qurashi A., and Wheater H. (2007). Regression analysis of rainfall–runoff data from an arid catchment in Oman / Analyse par regression de données pluie–débit d'un bassin aride d'Oman, **Hydrological Sciences Journal/Journal des Sciences Hydrologiques.** 52(6): 1103-1118**.**

Patel S., Hardaha M. K., Mukesh K. S., and Madankar K. K. (2016). Multiple linear regression model for stream flow estimation of wainganga river. **American Journal of Water Science and Engineering 2016**. 2(1): 1-5.

Pilgrim, D., Chapman, T., and Doran, D. (1988). Problems of rainfall-runoff modelling in arid and semiarid regions. **Hydrological Sciences-Journal-des Sciences Hydrologiques**. 4(33): 379-400.

Piotrowski, A.P., Napiorkowski, J.J., 2013. A comparison of methods to avoid overfitting in neural networks training in the case of catchment runoff modelling. **Journal of Hydrology.** 476: 97-111.

Riad S. and Mania J. (2004). Rainfall-runoff model using an artificial meural network approach, **Mathematical and Computer Modelling.** 40: 839-846.

Sajikumar N. and Thandaveswara B.S. (1999). A non-linear rainfall–runoff model using an artificial neural network, **Journal of Hydrology.** 216: 32–55.

Sezin, T. A. and Johnson, P. A. (2000). Precipitation-runoff modeling using artificial neural networks and conceptual models, **Journal of Hydrologic Engineering**. 5(2): 156-161.

Sharifi A., Sharifi Y. and Mirabbasi R. (2017). Daily runoff prediction using the linear and non-linear models**, Water Science and Technology.** 76 (3-4): 793-805.

Sun X., and Chan P.K. (2014). An Analysis of Instance Selection for Neural Networks to Improve Training Speed. In **Proceedings of the 2014 13th International Conference on Machine Learning and Applications (ICMLA)**, Detroit, MI, USA, 3–5 December 2014.

Uysala G., Sorman A. and Sensoy A. (2016). Streamflow forecasting using different neural network models with satellite data for a snow dominated region in Turkey. **12th International Conference on Hydroinformatics, HIC 2016.** Procedia Engineering, 154: 1185 – 1192

Wu C.L. and Chau K.W. (2008). River stage prediction based on a distributed support vector regression. **Journal of Hydrology**. 358(1-2): 96-111.

Yoon H., Jun S. C., Hyun Y., Bea G. O. and Lee K. K. (2011). A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. **Journal of Hydrology**. 396: 128–138.

Zaini N., Malek M. A., Yusoff M., Mardi N. H., and Norhisham S. (2018). Daily river flow forecasting with hybrid support vector machine – particle swarm optimization. **IOP Conference Series: Earth and Environmental Science**. 140: 1-8.

ภาคผนวก ก

การใช้งานโปรแกรม

การใช้งานโปรแกรม

เนื้อหาส่วนนี้อธิบายการใช้งานโปรแกรม IBM SPSS Modeler Version 18.0สำหรับการสร้างอัลกอริทึม ANN-GS เพื่อใช้ในการคาดการณ์น้ำท่าโดยจะมีขั้นตอนการทำงานของโปรแกรมดังต่อไปนี้

1. การเตรียมข้อมูล

การใช้งานโปรแกรมสำหรับสร้างอัลกอริทึม ANN-GS นี้จะสามารถใช้งานได้กับไฟล์ข้อมูล .csv ดังนั้นจึงต้องเตรียมข้อมูลให้อยู่ในรูปแบบไฟล์ที่กำหนด โดยตัวอย่างไฟล์ .csv ที่ใช้ในงานวิจัยแสดงดังรูปที่ ก.1

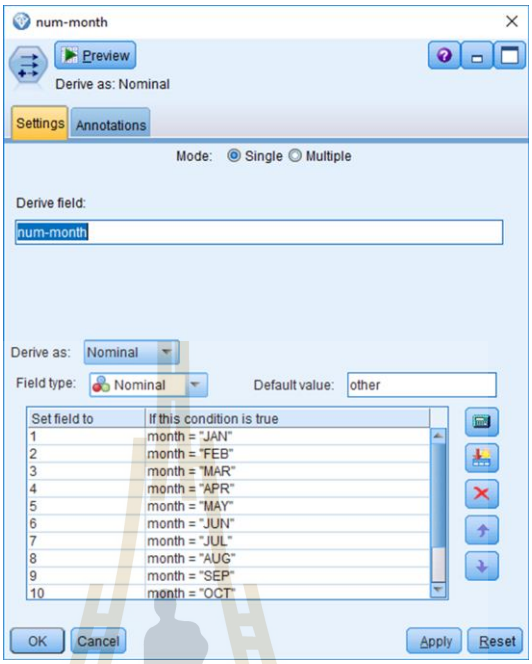
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	row	year	month	rainfall(T)	rainfall(T-1	rainfall(T-2	runoff(T)	runoff(T-1	runoff(T-2	day-rain(T	day-rain(T	day-rain(T	NDVI(T)	NDVI(T-1)	NDVI(T-2)
2	1	2540	JAN	52.9	0	79	27.7	59.54	123.61	9	0	6	0.415	0.385	0.372
3	2	2540	FEB	0	52.9	0	22.37	27.7	59.54	0	9	0	0.363	0.415	0.385
4	3	2540	MAR	11.8	0	52.9	24.63	22.37	27.7	5	0	9	0.278	0.363	0.415
5	4	2540	APR	57.6	11.8	0	37.47	24.63	22.37	10	5	0	0.274	0.278	0.363
6	5	2540	MAY	52.9	57.6	11.8	55.87	37.47	24.63	9	10	5	0.291	0.274	0.278
7	6	2540	JUN	26.5	52.9	57.6	49.52	55.87	37.47	9	9	10	0.207	0.291	0.274
8	7	2540	JUL	178.7	26.5	52.9	113.35	49.52	55.87	20	9	9	0.141	0.207	0.291
9	8	2540	AUG	236.4	178.7	26.5	114.82	113.35	49.52	22	20	9	0.2	0.141	0.207
10	9	2540	SEP	138	236.4	178.7	198.53	114.82	113.35	16	22	20	0.276	0.2	0.141
11	10	2540	OCT	162.9	138	236.4	233.12	198.53	114.82	8	16	22	0.362	0.276	0.2
12	11	2540	NOV	19.2	162.9	138	83.4	233.12	198.53	3	8	16	0.426	0.362	0.276
13	12	2540	DEC	0	19.2	162.9	38.98	83.4	233.12	0	3	8	0.463	0.426	0.362
14	13	2541	JAN	7.2	0	19.2	21.32	38.98	83.4	1	0	3	0.437	0.463	0.426
15	14	2541	FEB	0	7.2	0	13.56	21.32	38.98	0	1	0	0.356	0.437	0.463
16	15	2541	MAR	8.7	0	7.2	26.4	13.56	21.32	1	0	1	0.268	0.356	0.437

รูปที่ ก.1 ตัวอย่างข้อมูล .csv

2. การใช้งานในส่วนโปรแกรม

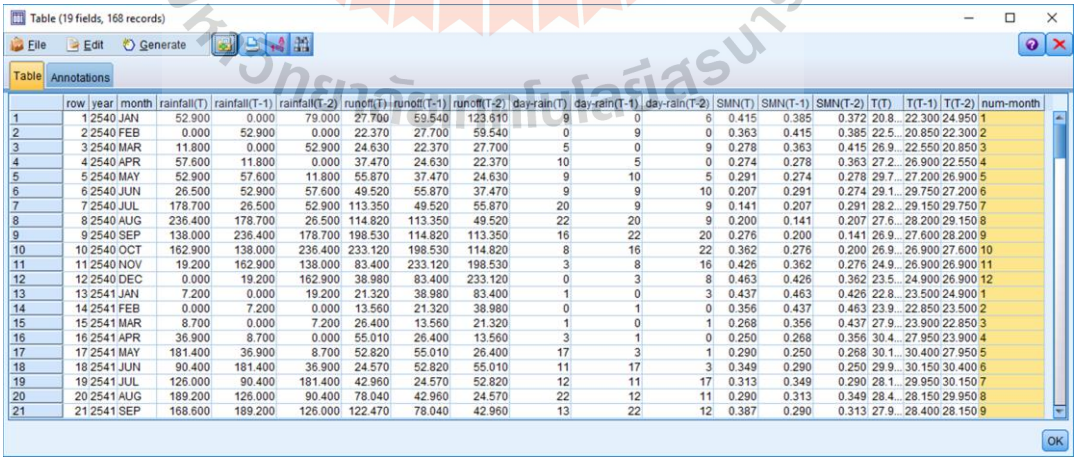
ในขั้นตอนนี้จะใช้ไฟล์ .csv ที่เตรียมไว้ ซึ่งการทำงานมีขั้นตอนการใช้งานดังต่อไปนี้

1) เมื่อกดเริ่มโปรแกรมให้ลากโหนด Var. File เพื่ออ่านไฟล์ .csv
2) ดับเบิลคลิกที่โหนด Var. File เพื่อเลือกไฟล์ที่ต้องการโดยไฟล์นี้จะเป็นข้อมูลชุดฝึกสอน
3) ลากโหนด Derive ทำการเชื่อมโหนด Var. File กับ Derive โดยนำลูกศรไปวางไว้บนโหนด Var. File คลิกปุ่มตรงกลางเมาส์ลากค้างไปที่โหนด Derive จะเกิดลูกศรขึ้น
4) ดับเบิลคลิกที่โหนด Derive เพื่อสร้างคอลัมน์ตัวเลขของเดือน (num-month) จากคอลัมน์ month โดยจะแสดงหน้าต่างการตั้งค่าของโหนด Derive จากนั้นตั้งค่าดังรูปที่ ก.2



รูปที่ ก.2 การตั้งค่าโหนด Derive เพื่อสร้างคอลัมน์ตัวเลขของเดือน (num-month)

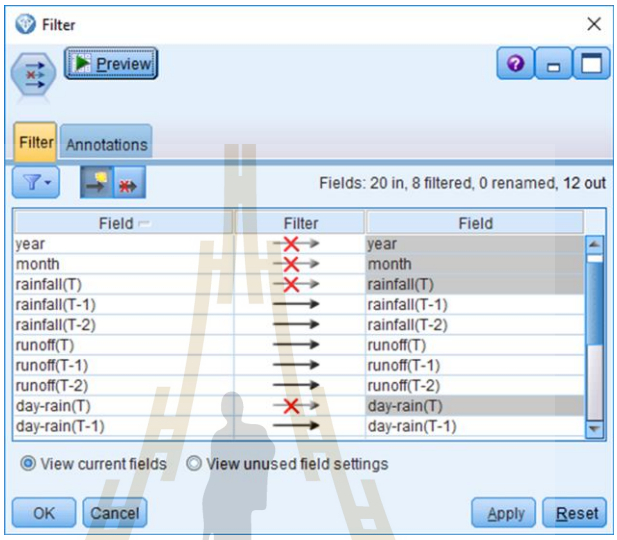
5) สามารถดูคอลัมน์ num-month ที่สร้างขึ้นโดยนำโหนด Table มาต่อโหนด Derive จากนั้นดับเบิลคลิกโหนด num-month กด run เพื่อแสดงตัวอย่างข้อมูลซึ่งจะมีคอลัมน์ num-month เพิ่มขึ้นมาแสดงดังรูป ก.3



	row	year	month	rainfall(T)	rainfall(T-1)	rainfall(T-2)	runoff(T)	runoff(T-1)	runoff(T-2)	day-rain(T)	day-rain(T-1)	day-rain(T-2)	SMN(T)	SMN(T-1)	SMN(T-2)	T(T)	T(T-1)	T(T-2)	num-month
1	1	2540	JAN	52.900	0.000	79.000	27.700	59.540	123.610	9	0	6	0.415	0.385	0.372	20.8...	22.300	24.950	1
2	2	2540	FEB	0.000	52.900	0.000	22.370	27.700	59.540	0	9	0	0.363	0.415	0.385	22.5...	20.850	22.300	2
3	3	2540	MAR	11.800	0.000	52.900	24.630	22.370	27.700	5	0	9	0.278	0.363	0.415	26.9...	22.550	20.850	3
4	4	2540	APR	57.600	11.800	0.000	37.470	24.630	22.370	10	5	0	0.274	0.278	0.363	27.2...	26.900	22.550	4
5	5	2540	MAY	52.900	57.600	11.800	55.870	37.470	24.630	9	10	5	0.291	0.274	0.278	29.7...	27.200	26.900	5
6	6	2540	JUN	26.500	52.900	57.600	49.520	55.870	37.470	9	9	10	0.207	0.291	0.274	29.1...	29.750	27.200	6
7	7	2540	JUL	178.700	26.500	52.900	113.350	49.520	55.870	20	9	9	0.141	0.207	0.291	28.2...	29.150	29.750	7
8	8	2540	AUG	236.400	178.700	26.500	114.820	113.350	49.520	22	20	9	0.200	0.141	0.207	27.6...	28.200	29.150	8
9	9	2540	SEP	138.000	236.400	178.700	198.530	114.820	113.350	16	22	20	0.276	0.200	0.141	26.9...	27.600	28.200	9
10	10	2540	OCT	162.900	138.000	236.400	233.120	198.530	114.820	8	16	22	0.362	0.276	0.200	26.9...	26.900	27.600	10
11	11	2540	NOV	19.200	162.900	138.000	83.400	233.120	198.530	3	8	16	0.426	0.362	0.276	24.9...	26.900	26.900	11
12	12	2540	DEC	0.000	19.200	162.900	38.980	83.400	233.120	0	3	8	0.463	0.426	0.362	23.5...	24.900	26.900	12
13	13	2541	JAN	7.200	0.000	19.200	21.320	38.980	83.400	1	0	3	0.437	0.463	0.426	22.8...	23.500	24.900	1
14	14	2541	FEB	0.000	7.200	0.000	13.560	21.320	38.980	0	1	0	0.356	0.437	0.463	23.9...	22.850	23.500	2
15	15	2541	MAR	8.700	0.000	7.200	26.400	13.560	21.320	1	0	1	0.268	0.356	0.437	27.9...	23.900	22.850	3
16	16	2541	APR	36.900	8.700	0.000	55.010	26.400	13.560	3	1	0	0.250	0.268	0.356	30.4...	27.950	23.900	4
17	17	2541	MAY	181.400	36.900	8.700	52.820	55.010	26.400	17	3	1	0.290	0.250	0.268	30.1...	30.400	27.950	5
18	18	2541	JUN	90.400	181.400	36.900	24.570	52.820	55.010	11	17	3	0.349	0.290	0.250	29.9...	30.150	30.400	6
19	19	2541	JUL	126.000	90.400	181.400	42.960	24.570	52.820	12	11	17	0.313	0.349	0.290	28.1...	29.950	30.150	7
20	20	2541	AUG	189.200	126.000	90.400	78.040	42.960	24.570	22	12	11	0.290	0.313	0.349	28.4...	28.150	29.950	8
21	21	2541	SEP	168.600	189.200	126.000	122.470	78.040	42.960	13	22	12	0.387	0.290	0.313	27.9...	28.400	28.150	9

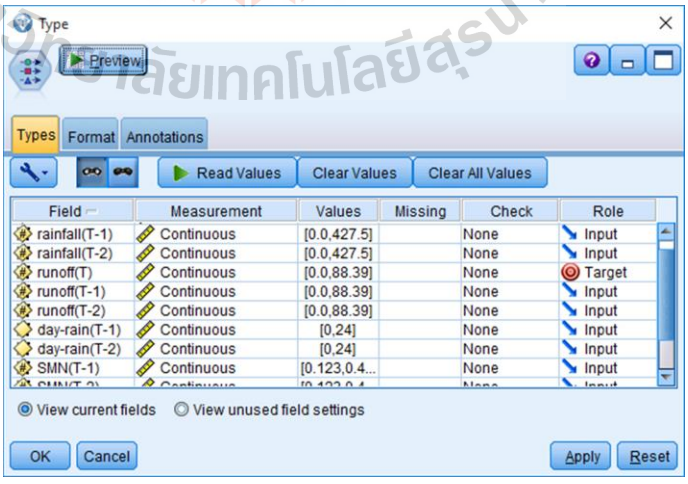
รูปที่ ก.3 ตัวอย่างข้อมูลที่มีคอลัมน์ num-month เพิ่มขึ้น

6) ลากโหนด Filter  มาต่อโหนด Derive เพื่อตัดคอลัมน์ที่ไม่ใช้งานออก โดยงานวิจัยนี้จะใช้ข้อมูล น้ำฝน น้ำท่า จำนวนวันที่ฝนตก NDVI ที่เวลาย้อนหลัง 1 และ2 เดือน (T-1, T-2) ตัวเลขของเดือน และค่าเป้าหมายคือ น้ำท่าที่เวลา T แสดงตัวอย่างการตั้งค่าโหนด Filter ดังรูป ก.4



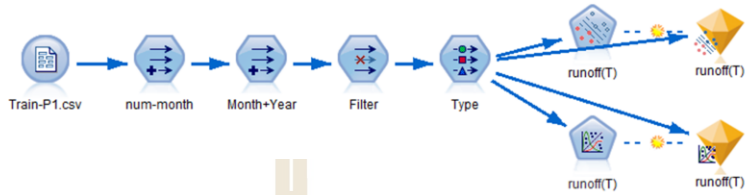
รูป ก.4 ตัวอย่างการตั้งค่าโหนด Filter

7) ลากโหนด Type  มาต่อโหนด Filter เพื่อกำหนดชนิดของข้อมูลที่เลือกว่าเป็นแบบใด กำหนดฟิลด์น้ำท่าที่เวลา T (runoff(T)) เป็น target ข้อมูลอื่นเป็น input และคอลัมน์ row เป็น noneแสดงตัวอย่างการตั้งค่าโหนด Type ดังรูป ก.5



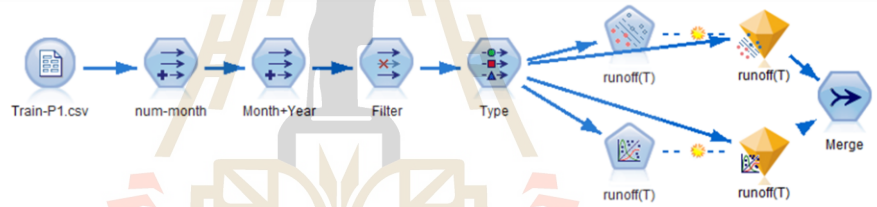
รูป ก.5 ตัวอย่างการตั้งค่าโหนด Type

8) ลากโหนด SVM และ Genlin มาต่อโหนด Type จากนั้นคลิกขวาแล้วเลือก run ในแต่ละโหนดจะเกิดโมเดลที่เป็นรูปเพชรสีทอง แสดงตัวอย่างโมเดลที่เป็นรูปเพชรสีทองในรูป ก.6



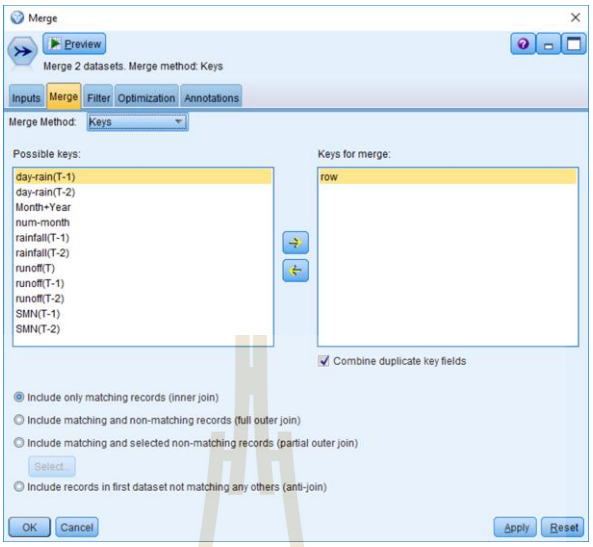
รูป ก.6 ตัวอย่างโมเดลที่เป็นรูปเพชรสีทอง

9) ลากโหนด Merge มาต่อโหนดเพชรสีทองทั้งสองเพื่อจะนำผลการคาดการณ์น้ำท่าของ SVR และ GLM มารวมเป็นข้อมูลชุดเดียวกัน แสดงโหนด Mergeในรูป ก.7

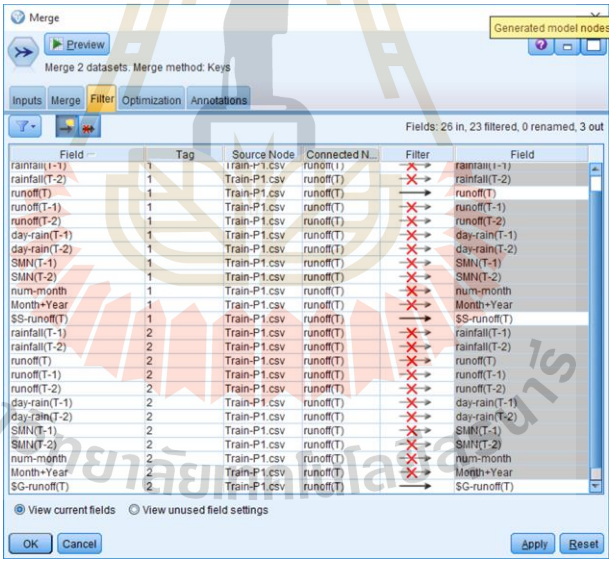


รูป ก.7 การต่อโหนดเพชรสีทองกับโหนด Merge

10) ตั้งค่าโหนด Merge โดยจะกำหนด key คือคอลัมน์ row เพื่อเป็นค่าที่ใช้เปรียบเทียบการรวมข้อมูลถ้าข้อมูลแถวเหมือนกันจะนำข้อมูลมาต่อกัน แสดงตัวอย่างการตั้งค่า key ของโหนด Mergeในรูป ก.8 จากนั้นเลือกคอลัมน์ของข้อมูลที่จะเก็บไว้ได้แก่ ค่าน้ำท่าจริง (runoff(T)) ค่าน้ำท่าที่คาดการณ์จาก SVR ($S-runoff(T)) และจาก GLM ($G-runoff(T)) โดยสามารถตั้งค่าในแท็บ Filler แสดงตัวอย่างการตั้งค่าแท็บ Filler ของโหนด Mergeในรูป ก.9



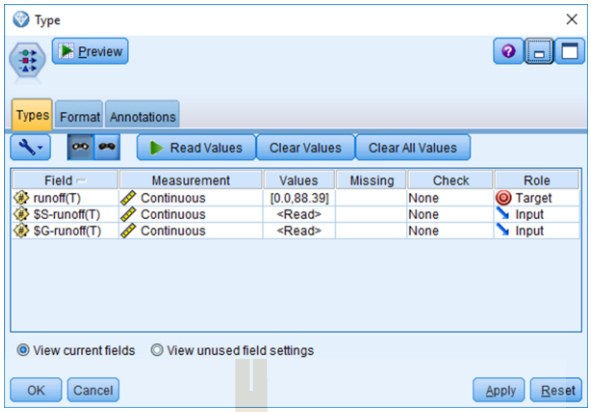
รูป ก.8 ตัวอย่างการตั้งค่า key ของโหนด Merge



รูป ก.9 ตัวอย่างการตั้งค่าแท็บ Filler ของโหนด Merge

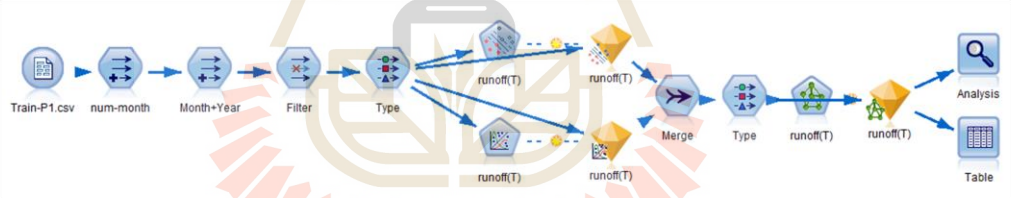
11) ลากโหนด Type มาต่อโหนด Merge เพื่อกำหนดค่า target และ input ก่อนนำเข้า ANN โดยสามารถตั้งค่าโหนด Type ก่อนนำเข้า ANN ในรูป ก.10

12) ลากโหนด Neural Net  ต่อกับโหนด Type คลิกขวาเลือก run ที่โหนด Neural Net จะได้เพชรสีทองซึ่งเป็นโมเดลของ ANN ที่เกิดจากค่าในการคาดการณ์ของ SVR และ GLM



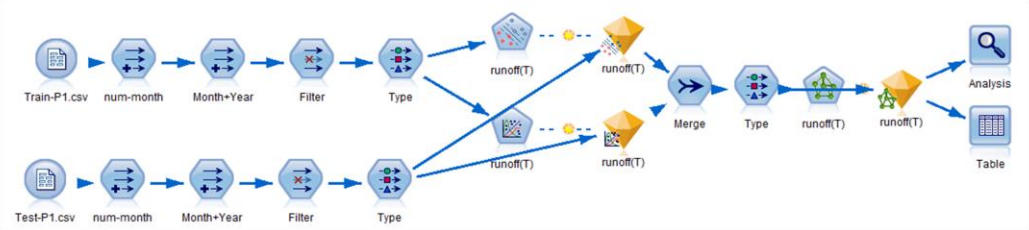
รูป ก.10 ตั้งค่าโหนด Type ก่อนนำเข้า ANN

13) สามารถดูค่าในการคาดการณ์โดยนำโหนด Table มาต่อกับเพชรสีทอง ซึ่งจะแสดงค่าในการคาดการณ์ของข้อมูลชุดฝึกสอน และโหนด Analysis เพื่อดูค่าความผิดพลาดในการคาดการณ์ของโมเดล การต่อโหนดสำหรับการสร้างโมเดล ANN-GS แสดงในรูป ก.11



รูป ก.11 การต่อโหนดสำหรับการสร้างโมเดล ANN-GS

14) สามารถดูมาต่อที่โมเดลโดยข้อมูลชุดทดสอบต้องมีการต่อโหนดต่าง ๆ เพื่อกับหนดค่าให้เหมือนข้อมูลฝึกสอนก่อนนำเข้าโมเดล โดยแสดงในรูป ก.12 สามารถดูค่าในการคาดการณ์และค่าความผิดพลาดโดยคลิกขาวเลือก run ที่โหนด Table และโหนด Analysis



รูป ก.12 การต่อข้อมูลชุดทดสอบเพื่อทดสอบโมเดล ANN-GS

ภาคผนวก ข

บทความวิจัยที่ได้รับการตีพิมพ์เผยแพร่

รายชื่อบทความที่ได้รับการตีพิมพ์เผยแพร่ในระหว่างศึกษา

รติพร จันทร์กลั่น, เกตุกาญจน์ ไชยขันธุ์, กิตติศักดิ์ เกิดประสพ, นิตยา เกิดประสพ. 2560. การพัฒนาแบบจำลองสำหรับคาดการณ์ปริมาณน้ำท่าบริเวณลุ่มน้ำมูล. **วารสารวิชาการและวิจัย มทร.พระนคร**.11: 37-47.(indexing: TCI-tier1)

Ratiporn Chanklan, Keerachart Suksut, Kedkard Chaiyakhan, Nuntawut Kaoungku, Kittisak Kerdprasop, and Nittaya Kerdprasop. (2017). On applying regression and neural network to predict rainfall using satellite based index. In **Proceedings of the International Multi Conference of Engineers and Computer Scientists 2017,** Hong Kong**,** 15-17 March 2017.

Ratiporn Chanklan, Kedkarn Chaiyakhan, Kittisak Kerdprasop, and Nittaya Kerdprasop. (2017). A hybrid modeling technique to predict runoff. **International Journal of Modeling and Optimization,** 7(2): 60-64. (indexing: INSPEC)

Ratiporn Chanklan, Nuntawut Kaoungku, Keerachart Suksut, Kittisak Kerdprasop, and Nittaya Kerdprasop. (2018). Runoff prediction with a combined artificial neural network and support vector regression. **Journal of Machine Learning and Computing,** 8(1): 39-43. (indexing: SCOPUS)

Nuntawut Kaoungku, Keerachart Suksut, Ratiporn Chanklan, Kittisak Kerdprasop, and Nittaya Kerdprasop**.** (2018). The silhouette width criterion for cluster and annociation mining to select image features. **International Journal of Machine Learning and Computing.** 8(1): 69-73. (indexing: SCOPUS)

RMUTP Research Journal, Vol. 11, No. 2, July-December 2017 37

http://journal.rmutp.ac.th/

การพัฒนาแบบจำลองสำหรับคาดการณ์ปริมาณน้ำท่าบริเวณลุ่มน้ำมูล

รติพร จันทร์กลั่น[1*] เกตุกาญจน์ ไชยขันธุ์[2] กิตติศักดิ์ เกิดประสพ[1] และ นิตยา เกิดประสพ[1]

[1] สำนักวิศวกรรมศาสตร์ วิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีสุรนารี

[2] คณะวิศวกรรมศาสตร์และสถาปัตยกรรมศาสตร์ วิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน

[1] 111 ถนนมหาวิทยาลัย อำเภอเมือง จังหวัดนครราชสีมา 30000

[2] 744 ถนนสุรนารายณ์ อำเภอเมือง จังหวัดนครราชสีมา 30000

รับบทความ 4 กุมภาพันธ์ 2017; ตอบรับบทความ 1 สิงหาคม 2017

บทคัดย่อ

บทความนี้นำเสนอการใช้การสำรวจระยะไกลด้วยดัชนีผลต่างพืชพรรณจากดาวเทียม NOAA ค่าการจัดกลุ่มจาก k-means อุณหภูมิ ปริมาณน้ำฝน จำนวนวันที่ฝนตก และปริมาณน้ำท่า เพื่อสร้างโมเดลคาดการณ์ค่าปริมาณน้ำท่ารายเดือนด้วยโครงข่ายประสาทเทียม ประเมินประสิทธิภาพโดยใช้ค่า R^2 และ RMSE เปรียบเทียบประสิทธิภาพด้วยสมการถดถอย ผลการทดลองสรุปว่าเมื่อเพิ่มค่าการจัดกลุ่มร่วมกับพารามิเตอร์อื่น ๆ เพื่อนำมาสร้างโมเดลสามารถเพิ่มประสิทธิภาพการคาดการณ์ได้ เมื่อใช้ค่าดัชนีผลต่างพืชพรรณร่วมกับอุณหภูมิ ที่เวลาย้อนหลัง 1 และ 2 เดือน และค่าการจัดกลุ่ม ที่เวลาย้อนหลัง 1 เดือน สร้างโมเดลให้มีประสิทธิภาพดีที่สุด โดยให้ผล RMSE=0.09 และ R^2=0.743 ผลการทดลองแสดงให้เห็นว่าการใช้ข้อมูลการสำรวจระยะไกล และค่าการจัดกลุ่ม สามารถใช้ในการคาดการณ์ค่าปริมาณน้ำท่าได้อย่างมีประสิทธิภาพ

คำสำคัญ: การสำรวจระยะไกล; น้ำท่า; โครงข่ายประสาทเทียม; สมการถดถอยเชิงเส้น

** ผู้พิพนธ์ประสานงาน โทร: +669 9469 6164, ไปรษณีย์อิเล็กทรอนิกส์: arc_angle@hotmail.com*

38 วารสารวิชาการและวิจัย มทร.พระนคร ปีที่ 11 ฉบับที่ 2 กรกฎาคม-ธันวาคม 2560

http://journal.rmutp.ac.th/

Development Model for Predict Runoff in Mun Basin

Ratiporn Chanklan[1*] Kedkarn Chaiyakhan[2] Kittisak Kerdprasop[1] and Nittaya Kerdprasop[1]

[1] Institute of Engineering, Computer Engineering, Suranaree University of Technology

[2] Faculty of Engineering and Architecture, Computer Engineering, Rajamangala University of Technology Isan

[1] 111 University Avenue, Muang, Nakhon Ratchasima 30000

[2] 744 Suranarai Road, Muang, Nakhon Ratchasima 30000

Received 4 February 2017; accepted 1 August 2017

Abstract

In this paper proposed remote sensing using Normalized Difference Vegetation Index from NOAA STAR, cluster value from k-means, temperature, rainfall, number of rainy days and runoff to create runoff prediction model using Artificial Neural Network (ANN) and evaluated runoff models with the R^2 and RMSE. The results show that the using of cluster value with other parameters to create predictive models can enhance forecasting results. When using Normalized Difference Vegetation Index with temperature value at lag time 1-2 month and cluster value at lag time 1 month to create model with ANN, we have got the best performance which are RMSE=0.09 and R^2=0.743. The experimental results shows that remote sensing data and cluster value from k-means can be used to predictive the runoff effectively.

Keywords: Remote Sensing; Runoff; Artificial Neural Network; Linear Regression

* *Corresponding Author. Tel.: +669 9469 6164, E-mail address: arc_angle@hotmail.com*

1. บทนำ

การคาดการณ์ปริมาณน้ำท่าหรือปริมาณน้ำในแม่น้ำที่เกิดขึ้นจากฝนเป็นการวิเคราะห์ค่อนข้างยาก เพราะมีกระบวนการเกิดที่ซับซ้อนและมีความสัมพันธ์ที่ซับซ้อนไม่เป็นเชิงเส้น การคาดการณ์ปริมาณน้ำท่าที่ได้จะช่วยในการตัดสินใจการวางแผนและจัดการการใช้ทรัพยากรน้ำ กล่าวคือถ้าเราสามารถรู้ปริมาณน้ำท่าหรือปริมาณน้ำในแม่น้ำได้ล่วงหน้า ทำให้สามารถรู้ว่าจะเกิดปัญหาจากน้ำ ได้แก่ น้ำท่วม และการขาดแคลนน้ำ ในอนาคตหรือไม่ ดังนั้นการสร้างโมเดลเพื่อคาดการณ์น้ำท่าจะช่วยให้หลีกเลี่ยงอันตรายที่อาจจะเกิดขึ้นเนื่องจากน้ำท่วมและภัยแล้ง โครงข่ายประสาทเทียม (ANN) เป็นเครื่องมือที่ได้ถูกนำมาใช้ในอุทกวิทยาและสร้างแบบจำลองคาดการณ์ปริมาณน้ำท่า เพราะมีความสามารถในการจำลองทั้งเชิงเส้นและไม่เชิงเส้นโดยไม่จำเป็นต้องตั้งสมมติฐานใด ๆ ตามวิธีการทางสถิติแบบดั้งเดิม มีงานวิจัยหลายชิ้นที่ประสบความสำเร็จในการสร้างแบบจำลองน้ำท่าโดยใช้ ANN

S. Riad et al. [1] ศึกษา ANN เพื่อยืนยันว่าเครือข่ายประสาทเทียมเหมาะที่จะคาดการณ์น้ำท่ามากกว่าวิธีการพื้นฐานด้วยการวิเคราะห์การถดถอย โดยใช้ตัวแปรอินพุท 14 ตัวคือ ปริมาณน้ำฝนและปริมาณน้ำท่าย้อนหลัง 1-7 วัน เพื่อทำนายน้ำท่าที่เวลา t โดยใช้มาตรวัด Squared of Error (ASE), ค่าสัมประสิทธิ์การตัดสินใจ (R^2) ผลการทดลองแสดงให้เห็นว่าวิธี ANN ช่วยให้การคาดการณ์ที่ดีกว่าวิธีการวิเคราะห์การถดถอย

A.R. Ghumman et al. [2] พัฒนา ANN เพื่อให้เหมาะกับชุดข้อมูลที่เก็บรวบรวมที่เป็นระยะสั้น เพื่อทำการคาดการณ์น้ำท่าเดือนถัดไป ค่าพารามิเตอร์ที่ใช้ได้แก่ ค่าน้ำท่าเดือนปัจจุบัน, ค่าน้ำท่าก่อนหน้า, ค่าน้ำฝนเดือนปัจจุบัน, ค่าน้ำฝนก่อนหน้า โดยใช้ค่าจำนวนโหนดในชั้นซ่อนตั้งแต่ 2-15 โดยเลือกค่าจำนวนโหนดในชั้นซ่อนที่ให้ค่าเฉลี่ยความผิดพลาดยกกำลังสอง (MSE) น้อยที่สุด ทดสอบประสิทธิภาพการคาดการณ์น้ำท่าโดยใช้ ค่าสัมประสิทธิ์ของความมีประสิทธิภาพ (CE) และค่าเฉลี่ยกำลังสองข้อผิดพลาด (RMSE) ผลการทดลองยืนยันว่า ANN ให้ประสิทธิภาพดี

R. Modarres et al. [3] ใช้ค่าทางสถิติ 17 ค่าในการทดสอบการคาดการณ์ปริมาณน้ำท่า ใช้พารามิเตอร์ 6 ตัวในการสร้างแม่แบบด้วย ANN ได้แก่ ปริมาณน้ำฝนประจำวันของสถานีที่ 1 ก่อนหน้า 1 และ 2 วัน, น้ำฝนประจำวันของสถานีที่ 2 ก่อนหน้า 1 วัน, สถานีที่ 3 ก่อนหน้า 2 วัน, น้ำท่ารายวันในเวลาก่อนหน้า 1 และ 2 วัน เปรียบเทียบกับการวิเคราะห์การถดถอย โดยใช้ข้อมูลน้ำฝนและน้ำท่ารายวันตั้งแต่ ค.ศ. 1978-2000 ผลการทดลองแสดงให้เห็นว่า ANN ให้ผลที่ดีกว่าการวิเคราะห์การถดถอย

Gunwant Sharma et al. [4] ใช้ข้อมูลปริมาณน้ำฝนและน้ำท่ารายปี ค.ศ. 1987-2012 จาก 6 สถานี เพื่อวิเคราะห์น้ำท่าโดยใช้สมการถดถอยแบบ Linear กับ Parabolic โดยใช้ค่าปริมาณน้ำฝนปีก่อนหน้าเป็นอินพุทเพื่อคาดการณ์ปริมาณน้ำท่า ทำการเปรียบเทียบประสิทธิภาพโดยใช้ค่า R^2 จะเห็นว่าเมื่อใช้ Parabolic มีประสิทธิภาพการทำนายน้ำท่าดีกว่า Linear และแสดงให้เห็นว่าความสัมพันธ์ระหว่างปริมาณน้ำฝนปีก่อนหน้าและน้ำท่าในปีปัจจุบันสอดคล้องกัน

F. Machado et al. [5] สร้างแบบจำลองการทำนายปริมาณน้ำท่าด้วย ANN ด้วยข้อมูลรายเดือนให้มีความแม่นยำ แบ่งข้อมูลออกเป็น 3 ชุด กำหนดค่าโหนดใน Hidden Layer 3, 5, 8, 10 ทดสอบข้อมูลด้วยวิธีการ Nash-Sutcliffe Efficiency Coefficient (NS), R-square Value (R^2) ผลการทดลองที่ข้อมูลชุดที่หนึ่งใช้ อินพุต P(t), EVT(t) และ Q(t−1) ดีที่สุด ชุดที่สองและสามใช้อินพุต P(t−1), P(t), EVT(t−1), EVT(t), Q(t−1) ประสิทธิภาพดีที่สุด

จากที่กล่าวมีข้างต้นจะเห็นว่างานวิจัยส่วนใหญ่นิยมใช้ค่าปริมาณน้ำฝนและปริมาณน้ำท่าย้อนหลังมาใช้ในการสร้างโมเดลคาดการณ์ปริมาณน้ำท่า และ

พยายามปรับปรุงค่าอินพุตที่ใช้กับ ANN เพื่อเพิ่มประสิทธิภาพการคาดการณ์ปริมาณน้ำท่า

ในงานวิจัยนี้เสนอการใช้ข้อมูลรายเดือนได้แก่ ค่าจากการสำรวจระยะไกลคือดัชนีผลต่างพืชพรรณ (NDVI) อุณหภูมิ ค่าคลัสเตอร์หรือค่าการจัดกลุ่ม น้ำฝน น้ำท่า จำนวนวันที่ฝนตก จาก K-means เพื่อสร้างโมเดลในการคาดการณ์ค่าน้ำท่าที่สถานี M.145 ลำพระเพลิง อ.ปากช่อง จ.นครราชสีมา และเปรียบเทียบผลการคาดการณ์น้ำท่าด้วยการใช้การวิเคราะห์การถดถอย โดยทำการประเมินประสิทธิภาพการคาดการณ์โดยใช้ค่า R^2 และ RMSE

2. วิธีการศึกษา

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 การวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression Analysis)

การวิเคราะห์การถดถอยเชิงเส้นอย่างง่ายเป็นการศึกษาความสัมพันธ์ระหว่างสองตัวแปร กำหนดให้ตัวแปร X เป็นตัวแปรอิสระเป็นตัวแปรที่ทราบค่า ตัวแปร Y เป็นตัวแปรตามเป็นตัวแปรที่ไม่ทราบค่า สองตัวแปรนี้มีความสัมพันธ์กันในลักษณะเส้นตรง (Linear) เมื่อนำค่า X และ Y หลาย ๆ ค่ามาพล็อตบนแกน X และ Y แล้วลากเส้นตรงผ่านจุดที่พล็อตลงในกราฟ เส้นตรงที่ลากนั้นจะแสดงความสัมพันธ์ระหว่างค่าเฉลี่ยของตัวแปร X และตัวแปร Y ซึ่งเส้นตรงนี้เรียกว่า เส้นกราฟถดถอย (Regression Line) ค่าของตัวแปร Y จะเปลี่ยนแปลงไปตามตัวแปร X ค่าของ X หนึ่งค่าจะมีค่า Y ที่เป็นคู่ของค่า X เมื่อนำข้อมูลจากตัวแปรมาวิเคราะห์หาความสัมพันธ์ซึ่งสามารถบอกแนวโน้มของความสัมพันธ์โดยใช้แผนภาพเส้นตรง และจะทำการหาเส้นตรงที่ดีที่สุดเพื่อเป็นตัวแทนของรูปแบบความสัมพันธ์ของตัวแปรที่ศึกษา เส้นตรงที่ดีที่สุดจะมีเพียงเส้นเดียว

2.1.2 การวิเคราะห์การถดถอยเชิงเส้นพหุ (Multiple Linear Regression Analysis)

ในการหาความสัมพันธ์ของตัวแปรอิสระกับตัวแปรตามบางครั้งที่จำนวนตัวแปรอิสระที่สนใจในการศึกษามีมากกว่าหนึ่งตัว ความสัมพันธ์นี้ไม่สามารถใช้การถดถอยเชิงเส้นอย่างง่ายในการวิเคราะห์ได้ สำหรับกรณีที่มีตัวแปรอิสระ 2 ตัว (X_1 และ X_2) ที่มีความสัมพันธ์เชิงเส้นกับตัวแปรตาม สมการถดถอยสามารถเขียนในรูปสมการคือ $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$ เมื่อ Y_i เป็นตัวแปรตามที่ในการเก็บข้อมูลครั้งที่ i ตัวแปรอิสระตัวที่ 1 และ 2 ในการเก็บข้อมูลครั้งที่ i แทนด้วย X_{i1} และ X_{i2} ตามลำดับ β_0 คือค่าเฉลี่ยของตัวแปรตามที่ตัวแปรอิสระทั้งสองมีค่าเป็นศูนย์ β_1 และ β_2 ค่าสัมประสิทธิ์การถดถอยของ X_1 และ X_2 ตามลำดับและ ε_i เป็นค่าความคลาดเคลื่อนในการเก็บข้อมูลครั้งที่ i

การประมาณค่าพารามิเตอร์ใช้วิธีการเช่นเดียวกับสมการถดถอยเชิงเส้นอย่างง่ายคือการใช้วิธีกำลังสองน้อยที่สุด ซึ่งสมการถดถอยเชิงเส้นพหุของกลุ่มตัวอย่างคือ $\hat{Y} = b_0 + b_1X_1 + b_2X_2$ ในที่นี้ คือค่าโดยประมาณ โดยสามารถเขียนในรูปของเมตริกซ์ซึ่งสามารถหาเวกเตอร์ b ได้ดังนี้ $b = (X'X)^{-1}. X'Y$ โดย $X'X$ ต้องสามารถหาเมตริกซ์ผกผันได้และเป็นเมตริกซ์สมมาตรที่มีขนาด $p \times p$ โดย p คือจำนวนตัวแปรอิสระบวกหนึ่งและค่าของสมาชิกในแนวเฉียงเป็นผลรวมกำลังสองของค่าในแต่ละหลัก

2.1.3 ดัชนีผลต่างพืชพรรณ (Normalized Differential Vegetation Indices: NDVI)

ดัชนีผลต่างพืชพรรณแบบนอมัลไลซ์เป็นค่า ที่ใช้แสดงการสะท้อนของคลื่นแม่เหล็ก (Electromagnetic Spectrum) ของช่วงคลื่นใกล้อินฟาเรด (Near Infrared Reflectance) กับช่วงคลื่นตามองเห็นสีแดง (Visible Red Reflectance) โดยสามารถนำมาใช้ ในการ

วิเคราะห์การวัดระยะไกล (Remote Sensing Analysis) โดย NDVI เป็นการคำนวณหาสัดส่วนของช่วงคลื่นที่เกี่ยวข้องกับพืชพรรณ โดยนำค่าความแตกต่างของการสะท้อนของพื้นผิว ระหว่างช่วงคลื่นใกล้อินฟาเรด กับช่วงคลื่นตามองเห็นสีแดง มาทำสัดส่วนกับค่าผลรวมของทั้งสองช่วงคลื่น เพื่อปรับให้เป็นลักษณะการกระจายแบบปกติ เพื่อให้การแปรผลทำได้ง่ายขึ้น NDVI นิยมใช้ในการตรวจวัดความสมบูรณ์ของป่าไม้ หรือพืชพรรณ เพื่อประเมินว่าพื้นที่ที่ทำการวิเคราะห์มีพืชพรรณสีเขียวหนาแน่นมากน้อยเพียงใด [6] NDVI มีค่าอยู่ระหว่าง -1.0 ถึง +1.0 บริเวณที่ค่า NDVI อยู่ในช่วงค่าลบพื้นที่จะเป็นผืนน้ำ หรือทะเล ในบริเวณที่มีค่า NDVI เข้าใกล้ค่า 0 แสดงถึงพื้นที่ที่มีพืชพรรณสีเขียวน้อย และหากค่า NDVI มีค่าเข้าใกล้ +1.0 แสดงถึงพื้นที่ที่พืชสีเขียวปกคลุมมาก

2.1.4 โครงข่ายประสาทเทียม (Artificial Neural Network: ANN)

โครงข่ายประสาทเทียมเป็นแนวความคิดที่ต้องการให้คอมพิวเตอร์มีความสามารถในการเรียนรู้เหมือนมนุษย์ โครงข่ายประสาทเทียมประกอบไปด้วยเซตของโหนดและเส้นเชื่อมระหว่างโหนด สามารถแบ่งโหนดเป็น 3 ระดับได้แก่ โหนดในชั้นอินพุต (Input Layer) มีจำนวนเท่ากับจำนวนคุณสมบัติ (Attribute) ที่ใช้อธิบายข้อมูลแต่ละตัว โหนดในชั้นเอาท์พุต (Output Layer) มีจำนวนเท่ากับจำนวนกลุ่มหรือจำนวนประเภทของข้อมูลที่ต้องการจำแนก โหนดในชั้นซ่อน (Hidden Layer) ในแต่ละชั้นซ่อนอาจจะมีได้มากกว่า 1 ชั้น จำนวนชั้นและจำนวนโหนดจะขึ้นอยู่กับผู้ออกแบบโดยต้องทดลองหลาย ๆ แบบแล้วพิจารณาว่าแบบใดให้ประสิทธิภาพที่ดีที่สุด

ในโครงข่ายจะมีเส้นเชื่อมจากทุกโหนดในชั้นอินพุตไปยังทุกโหนดในชั้นซ่อนและมีเส้นเชื่อมจากทุกโหนดในชั้นซ่อนไปยังทุกโหนดในชั้นเอาท์พุต ในแต่ละโหนดจะมีค่า Bias (b) เส้นเชื่อมแต่ละเส้นมีค่าน้ำหนัก (Weight) การทำงานของแต่ละโหนดเทียบได้กับเซลล์ประสาทในสมองมนุษย์ 1 เซลล์ อินพุตที่เข้าสู่โหนดจะเป็นเวกเตอร์ของคุณสมบัติของข้อมูลตัวอย่าง มีค่า $p = [p_1, p_2, ..., p_R]$ มีจำนวน R องค์ประกอบ (คอลัมน์) และเวกเตอร์น้ำหนัก $W = [w_1, w_2, ..., w_R]$ นำอินพุตมาคูณกับน้ำหนักของแต่ละเส้นเชื่อม ผลที่ได้จากอินพุตทุกๆ เส้นเชื่อมของโหนดจะเอามารวมกันแล้วส่งต่อไปยังฟังก์ชันถ่ายโอน (Transfer Function) ซึ่งเกิดเป็นค่าเอาท์พุต a ในที่นี้ f เป็นฟังก์ชันถ่ายทำหน้าที่รับค่าอินพุต n เพื่อเปลี่ยนเป็นค่าเอาท์พุต a ค่าเอาท์พุต a สามารถคำนวณค่าเอาท์พุต a [7] ได้จาก $a = f(Wp + b)$ เมื่อ $n = w_{1,1}p_1 + w_{1,1}p_2 + ... + w_{1,R}p_R + b$ ดังนั้น $n = Wp + b$

การค้นหาค่าน้ำหนักของเส้นเชื่อมแต่ละเส้นที่เหมาะสมที่ทำให้สามารถจำแนกประเภทของข้อมูลตัวอย่างที่ใช้สอน (Training Data) ได้ถูกต้องมากที่สุด เป็นการสอนโครงข่ายประสาทเทียมให้เรียนรู้ ค่าน้ำหนักจะทำการปรับจนกว่าค่าความผิดพลาดจะน้อยลงหรืออยู่ในเกณฑ์ที่ยอมรับ

สิ่งสำคัญคือต้องทราบค่า Weight สำหรับสิ่งที่ต้องการให้โครงข่ายเรียนรู้ ซึ่งเป็นค่าที่ไม่แน่นอน แต่สามารถกำหนดให้โครงข่ายปรับค่าเหล่านั้นได้โดยการสอนให้รู้จักรูปแบบ (Pattern) ของสิ่งที่ต้องการให้รู้จำ เรียกว่า “Back Propagation” ถ้าโครงข่ายประสาทเทียมให้ค่าเอาท์พุตผิด

2.1.5 การแบ่งกลุ่มข้อมูลแบบเคมีน (K-means Clustering)

เป็นการจัดให้ข้อมูลที่มีลักษณะคล้ายกันมาไว้ด้วยกัน ใช้สำหรับการแบ่งข้อมูลจำนวน n เป็น k กลุ่ม [8] ขั้นแรกจะรับค่าพารามิเตอร์ k ซึ่งค่านี้คือจำนวน Cluster ที่ต้องการค้นหา จากนั้นกำหนดจุดศูนย์กลาง (Centroid) ขั้นตอนต่อไปคือสร้างกลุ่มข้อมูลและความสัมพันธ์กับจุดศูนย์กลางที่ใกล้กันมากที่สุด ซึ่งจุดแต่ละจุดแทนด้วยข้อมูลหนึ่งตัว โดยแต่ละจุดจะถูกกำหนดกลุ่มไปยังจุดศูนย์กลางที่ใกล้ที่สุดจนครบหมดทุกจุด และคำนวณจุดศูนย์กลางใหม่ โดยการหา

ค่าเฉลี่ยทุกจุดที่อยู่ในกลุ่ม หากจุดศูนย์กลางในแต่ละกลุ่มถูกเปลี่ยนตำแหน่ง จะได้จุดที่มีความสัมพันธ์กับกลุ่มใหม่และใกล้กับจุดศูนย์กลางใหม่ ทำซ้ำไปเรื่อย ๆ จนกระทั่งจุดศูนย์กลางไม่มีการเปลี่ยนแปลง

2.1.6 ค่าสัมประสิทธิ์การตัดสินใจ (Coefficient of Determination: R^2)

ค่าสัมประสิทธิ์การตัดสินใจ เป็นค่าที่บอกถึงความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระว่ามีความสัมพันธ์กันมากน้อยเพียงใด ค่าสัมประสิทธิ์การตัดสินใจมีค่าตั้งแต่ 0-1 ถ้าค่า R^2 มีค่าเข้าใกล้ 1 แสดงว่าตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันมากในเชิงเส้นตรง คำนวณได้จากสูตร

$$R^2 = \frac{(N\sum XY - (\sum X)(\sum XY))^2}{(N\sum X^2 - (\sum X)^2)(N\sum Y^2 - (\sum Y)^2)} \quad (1)$$

โดยที่

X, Y = ตัวแปรที่พิจารณา

N = จำนวนข้อมูลทั้งหมด

2.1.7 ค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Squared Error: RMSE)

ค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ยใช้ในการประเมินความผิดพลาดในการคาดการณ์ปริมาณน้ำท่า ถ้าค่า RMSE ยิ่งน้อยหมายถึงการคาดการณ์ยิ่งแม่นยำ สามารถคำนวณได้ดังสมการต่อไปนี้

$$RMSE = \sqrt{\left(\frac{\sum(T_i - O_i)^2}{N}\right)} \quad (2)$$

โดยที่

T_i = ค่าของข้อมูลน้ำท่าจริง

O_i = ค่าของข้อมูลน้ำท่าที่คาดการณ์

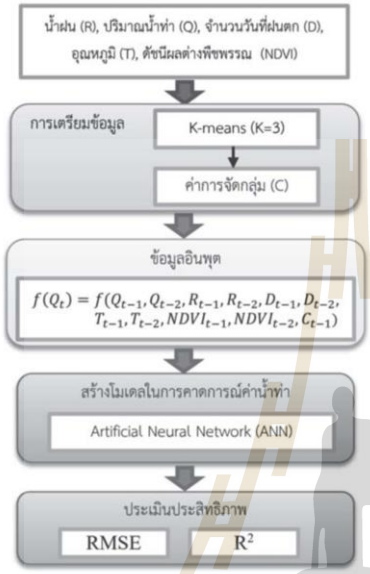
N = จำนวนข้อมูลทั้งหมด

2.2 ขั้นตอนการดำเนินงาน

งานวิจัยนี้ใช้ข้อมูลอุณหภูมิ ปริมาณน้ำฝน และปริมาณน้ำท่าเฉลี่ย ปริมาณน้ำฝน จำนวนวันที่ฝนตก และอุณหภูมิรายเดือน โดยเป็นข้อมูลที่เกี่ยวข้องกับสถานีน้ำท่า M.145 ที่ตั้งอยู่ในจังหวัดนครราชสีมา ซึ่งเป็นบริเวณลุ่มน้ำมูล จากศูนย์อุทกวิทยาชลประทานภาคตะวันออกเฉียงเหนือตอนล่าง (ดาวน์โหลดข้อมูลได้จาก www.hydro-4.com) และใช้ข้อมูลดัชนีผลต่างพืชพรรณ จากดาวเทียม NOAA STAR - Global Vegetation Health Products (ดาว์นโหลดข้อมูลจาก www.star.nesdis.noaa.gov) โดยจะใช้ข้อมูลที่เก็บรวบรวม 15 ปีตั้งแต่ปี พ.ศ. 2543 ถึง พ.ศ. 2557 โดยในขั้นตอนแรกจะนำข้อมูลทั้งหมดได้แก่ ปริมาณน้ำฝน จำนวนวันที่ฝนตก อุณหภูมิ ปริมาณน้ำท่า ดัชนีผลต่างพืชพรรณ ซึ่งเป็นข้อมูลที่ผ่านการลดความซ้ำซ้อนของข้อมูลด้วยวิธีการนอร์มัลไลเซชัน มาทำการจัดกลุ่มด้วยอัลกอริทึม k-means โดยอยู่ในสมมติฐานที่ว่าข้อมูลส่วนมากจะแบ่งเป็นข้อมูลที่มีค่าต่ำ ปานกลาง และสูง จึงจัดกลุ่มเป็น k=3 หลังจากนั้นจะนำค่าดัชนีผลต่างพืชพรรณ ปริมาณน้ำฝน ปริมาณน้ำท่า จำนวนวันที่ฝนตก และอุณหภูมิที่เวลาย้อนหลังหนึ่งและสองเดือนร่วมกับค่าการจัดกลุ่มที่เวลาย้อนหลังหนึ่งเดือน แล้วนำไปสร้างโมเดลการคาดการณ์ด้วยโครงข่ายประสาทเทียม โดยจะสลับการใช้ข้อมูลในการสร้างโมเดลเพื่อพิจารณาว่าควรใช้ข้อมูลใดในการสร้างโมเดลในการทำนายปริมาณน้ำท่าให้ประสิทธิภาพดีที่สุด กล่าวคือ ถ้าต้องการคาดการณ์ค่าน้ำท่าเดือนมีนาคม จะทำการสร้างโมเดลการคาดการณ์โดยใช้ค่าดัชนีผลต่างพืชพรรณ ปริมาณน้ำฝน ปริมาณน้ำท่า จำนวนวันที่ฝนตก และอุณหภูมิที่เดือนมกราคมและเดือนกุมภาพันธ์ ส่วนค่าการจัดกลุ่มจะใช้ที่เดือนกุมภาพันธ์เพื่อนำไปใช้ในการสร้างโมเดลการคาดการณ์ด้วยโครงข่ายประสาทเทียม

ในการวิจัยนี้จะใช้โครงข่ายประสาทเทียมแบบการแพร่ย้อนกลับ (Back-propagation Algorithm) ใช้จำนวนชั้นในชั้นซ่อน 1 ชั้น ทำการเลือกจำนวนโหนดภายในชั้นซ่อนตั้งแต่ 1-10 แล้วเลือกใช้จำนวนโหนด

ที่ให้ค่าเฉลี่ยความผิดพลาดยกกำลังสอง (RMSE) ที่น้อยที่สุดกับการคาดการณ์ค่าน้ำท่าในชุดข้อมูลฝึกสอน



รูปที่ 1 กรอบแนวคิดของงานวิจัย

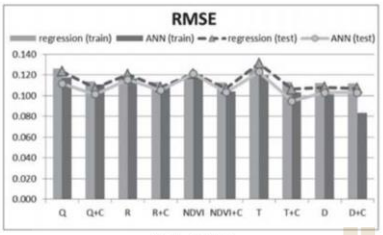
นอกจากนี้เปรียบเทียบประสิทธิภาพด้วยการวิเคราะห์การถดถอยเชิงเส้น และใช้พารามิเตอร์ตัวอื่น ๆ ในการสร้างโมเดลการคาดการณ์ปริมาณน้ำท่า ได้แก่ ปริมาณน้ำฝนรายเดือน (R) จำนวนวันที่ฝนตกในแต่ละเดือน (D) และปริมาณน้ำท่ารายเดือน (Q) ที่เวลาย้อนหลัง 1 เดือนและ 2 เดือน (t-1, t-2) การทดลองนี้ใช้มาตรวัดทางสถิติคือ ค่าสัมประสิทธิ์การตัดสินใจ (R^2) ซึ่งใช้แสดงความสัมพันธ์ระหว่างค่าน้ำท่าจริงกับค่าที่คาดการณ์ และใช้ค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) ซึ่งเป็นค่าที่แสดงความผิดพลาดของการคาดการณ์ โดยกรอบแนวคิดของงานวิจัยนี้สามารถแสดงดังรูปที่ 1

3. ผลการศึกษาและอภิปรายผล

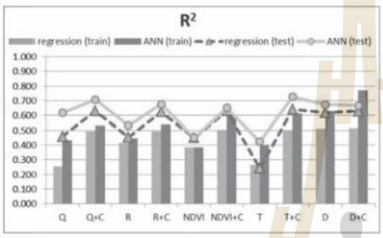
ในขั้นตอนแรกจะใช้อินพุทเป็นพารามิเตอร์ 1 ชนิด เปรียบเทียบกับการใช้ค่าการจัดกลุ่มร่วมด้วย เพื่อจะแสดงการทดสอบว่าค่าการจัดกลุ่มมีผลต่อประสิทธิภาพการคาดการณ์น้ำท่า จากนั้นสร้างแบบจำลองการคาดการณ์น้ำท่าด้วยโครงข่ายประสาทเทียม (ANN) และ การวิเคราะห์การถดถอยเชิงเส้น (Linear Regression: LR) เปรียบเทียบประสิทธิภาพการคาดการณ์ด้วยมาตรวัดทางสถิติด้วยการหาค่า R^2 และ RMSE ผลการทดลองแสดงในตารางที่ 1 และกราฟแสดงในรูปที่ 2

ตารางที่ 1 ใช้พารามิเตอร์ 1 ชนิดและผลการเพิ่มค่าการจัดกลุ่มเพื่อใช้คาดการณ์น้ำท่า (ข้อมูลฝึกสอน)

Parameters	Hidden Nodes	Test Data			
		RMSE		R^2	
		LR	ANN	LR	ANN
Q	9	0.123	0.111	0.459	0.618
Q+C	9	0.107	0.101	0.631	0.707
R	6	0.12	0.115	0.453	0.533
R+C	6	0.108	0.105	0.623	0.677
NDVI	7	0.121	0.121	0.451	0.451
NDVI+C	7	0.107	0.103	0.631	0.651
T	7	0.131	0.123	0.242	0.423
T+C	7	0.106	0.095	0.642	0.727
D	6	0.108	0.103	0.620	0.673
D+C	6	0.107	0.103	0.630	0.667



(ก) ค่า RMSE



(ข) ค่า R^2

รูปที่ 2 กราฟการทดลองใช้พารามิเตอร์ 1 ชนิด และการเพิ่มค่าการจัดกลุ่มของข้อมูลฝึกสอน (กราฟแท่ง) และข้อมูลทดสอบ (กราฟเส้น) ในรูป (ก) แสดงค่า RMSE (ข) แสดงค่า R^2

ตารางที่ 2 ผลการใช้พารามิเตอร์ 2 ชนิดร่วมกับค่าการจัดกลุ่มเพื่อใช้พยากรณ์ (ข้อมูลฝึกสอน)

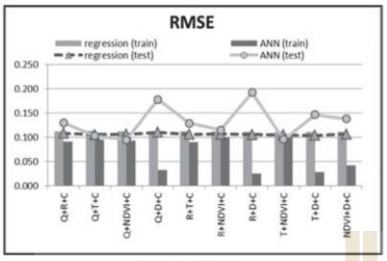
Parameters	Hidden Nodes	Test Data			
		RMSE		R^2	
		LR	ANN	LR	ANN
Q+R+C	8	0.109	0.13	0.615	0.409
Q+T+C	9	0.106	0.103	0.644	0.653
Q+NDVI+C	2	0.107	0.095	0.632	0.711
Q+D+C	10	0.111	0.178	0.595	0.498
R+T+C	9	0.107	0.129	0.638	0.519
R+NDVI+C	6	0.107	0.116	0.629	0.545
R+D+C	10	0.106	0.193	0.642	0.419
T+NDVI+C	3	0.105	0.096	0.66	0.743
T+D+C	10	0.104	0.147	0.655	0.532
NDVI+D+C	9	0.107	0.138	0.629	0.458

จากตารางที่ 1 เมื่อใช้พารามิเตอร์ 1 ชนิดที่เวลาย้อนหลัง 1 เดือนและ 2 เดือน ร่วมกับค่าการจัดกลุ่มที่เวลาย้อนหลัง 1 เดือน ให้ค่า RMSE ลดลง และค่า R^2 เพิ่มขึ้นในทุกพารามิเตอร์ ซึ่งแสดงให้เห็นว่าหากใช้ค่าการจัดกลุ่มร่วมกับพารามิเตอร์อื่น สามารถเพิ่มประสิทธิภาพการพยากรณ์น้ำท่าได้ใน ANN และ Regression จากผลการทดลองให้ประสิทธิภาพดีที่สุดเมื่อใช้อุณหภูมิร่วมกับการจัดกลุ่ม (T+C) โดยที่ ANN ให้ค่า RMSE=0.095 และค่า R^2= 0.727

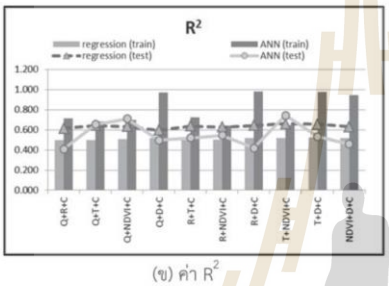
รูปที่ 2 แสดงให้เห็นว่าในข้อมูลฝึกสอนและข้อมูลทดสอบมีค่า RMSE และค่า R^2 ไปในทิศทางเดียวกัน เนื่องจากการทดลองตารางที่ 1 แสดงให้เห็นว่าเมื่อใช้ค่าการจัดกลุ่มสามารถเพิ่มประสิทธิภาพการพยากรณ์ได้ ดังนั้นจึงทดลองการใช้พารามิเตอร์ 2 ชนิดร่วมกับค่าการจัดกลุ่ม ซึ่งแสดงในตารางที่ 2

จากตารางที่ 2 เมื่อใช้ค่าอุณหภูมิ และดัชนีผลต่างพืชพรรณ ร่วมกับค่าการจัดกลุ่ม (T+NDVI+C) ที่ ANN มีประสิทธิภาพดีที่สุดในข้อมูลทดสอบโดยให้ค่า RMSE = 0.096 และ R^2 =0.743 Regression ให้ประสิทธิภาพการพยากรณ์โดยพิจารณาจาก RMSE ไม่แตกต่างจากการใช้พารามิเตอร์ 1 ชนิดแต่ใน R^2 มีค่าสูงตั้งแต่ 0.6 ขึ้นไปในทุกกรณี

เมื่อนำผลการทดลองมาแสดงเป็นกราฟในรูปที่ 3 จะเห็นว่ารูปที่ 3 (ก) ค่า RMSE และรูปที่ 4 (ข) ค่า R^2 ที่ ANN เกิดปัญหาการ Overfitting ชัดเจนในกรณี Q+D+C, R+D+C และ NDVI+D+C กล่าวคือการที่โมเดลที่ได้จากการใช้ชุดข้อมูลฝึกสอนให้ประสิทธิภาพการพยากรณ์สูง แต่เมื่อนำโมเดลไปใช้กับชุดข้อมูลทดสอบให้ประสิทธิภาพการพยากรณ์ต่ำ จะสังเกตเห็นว่าในทุกกรณีเมื่อใช้ D+C จะทำให้เกิดปัญหา Overfitting ดังนั้นผลการทดลองต่อไปจะใช้พารามิเตอร์ Q, T, NDVI, R และ C เป็นพารามิเตอร์ในการทดลองต่อไป เพราะไม่มีแนวโน้มทำให้เกิดปัญหา Overfitting และจะไม่ใช้ D+C ผลการทดลองใช้พารามิเตอร์ 3 ชนิดและ 4 ชนิดร่วมกับค่าการจัดกลุ่มแสดงในตารางที่ 3



(ก) ค่า RMSE



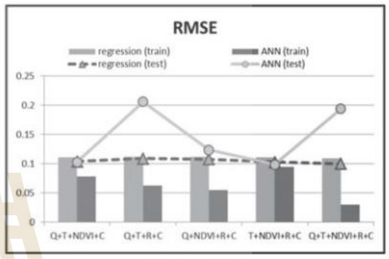
(ข) ค่า R^2

รูปที่ 3 กราฟการใช้พารามิเตอร์ 2 ชนิด ร่วมกับค่าการจัดกลุ่มของข้อมูลฝึกสอน (กราฟแท่ง) และข้อมูลทดสอบ (กราฟเส้น) ในรูป (ก) แสดงค่า RMSE (ข) แสดงค่า R^2

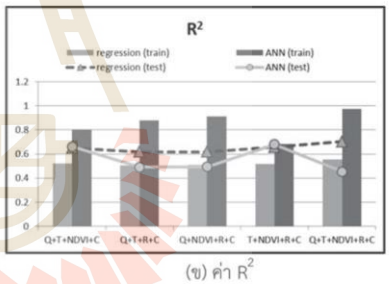
ตารางที่ 3 การใช้พารามิเตอร์ 3 และ 4 ชนิดร่วมกับค่าการจัดกลุ่มเพื่อใช้พยากรณ์ (ข้อมูลฝึกสอน)

Parameters	Hidden Nodes	Test Data			
		RMSE		R^2	
		LR	ANN	LR	ANN
Q+T+NDVI+C	7	0.104	0.103	0.652	0.661
Q+T+R+C	10	0.109	0.206	0.617	0.491
Q+NDVI+R+C	10	0.108	0.124	0.617	0.494
T+NDVI+R+C	8	0.103	<u>0.099</u>	0.658	0.681
Q+T+NDVI+R+C	10	0.100	0.194	<u>0.702</u>	0.453

จากตารางที่ 3 เมื่อใช้ T+NDVI+R+C มีประสิทธิภาพดีที่สุดที่ ANN ในข้อมูลทดสอบให้ค่า RMSE=0.099 และ Q+T+NDVI+R+C ให้ค่า R^2=0.702 มีประสิทธิภาพดีที่สุดที่ ANN และจากรูปที่ 4 จะเห็นว่าเกิดปัญหาการ Overfitting ที่ ANN ยกเว้นกรณี T+NDVI+R+C และประสิทธิภาพการพยากรณ์น้ำท่าไม่ดีกว่าการใช้พารามิเตอร์ 2 ชนิดร่วมกับค่าการจัดกลุ่ม (T+NDVI+C) ที่แสดงในตารางที่ 2



(ก) ค่า RMSE



(ข) ค่า R^2

รูปที่ 4 กราฟการใช้พารามิเตอร์ 3 และ 4 ชนิด ร่วมกับค่าการจัดกลุ่มของข้อมูลฝึกสอน (กราฟแท่ง) และข้อมูลทดสอบ (กราฟเส้น) ในรูป (ก) แสดงค่า RMSE (ข) แสดงค่า R^2

จากการทดลองทั้งหมดแสดงให้เห็นว่าเมื่อใช้ ANN ในการสร้างแบบจำลองมีประสิทธิภาพดีที่สุด โดยใช้อินพุทเป็น T+NDVI+C ที่ชุดข้อมูลทดสอบให้ค่า RMSE = 0.096 และ R^2 =0.743 (แสดงในตาราง ที่ 2) และไม่เกิดปัญหาการ Overfitting ส่วนในการสร้างแบบจำลองด้วย Regression การคาดการณ์น้ำท่าให้ประสิทธิภาพดีที่สุดเมื่อใช้ Q+T+NDVI+R+C ในชุดข้อมูลทดสอบโดยให้ค่า RMSE = 0.100 และ R^2 =0.702 (แสดงในตารางที่ 3) และไม่เกิดการ Overfitting จะเห็นว่าเมื่อใช้ ANN ให้ประสิทธิภาพดีที่สุดในการทดลองนี้และมีประสิทธิภาพการคาดการณ์น้ำท่าดีกว่า Regression นอกจากนี้ยังแสดงให้เห็นว่าการเลือกใช้พารามิเตอร์ในการสร้างโมเดลในการทำนายมีผลต่อความถูกต้องของการทำนาย

4. สรุป

การคาดการณ์น้ำท่าถือเป็นข้อมูลสำคัญที่สามารถช่วยให้สามารถป้องกันปัญหาที่เกิดจากน้ำ ได้แก่ น้ำท่วม และการขาดแคลนน้ำ กล่าวคือถ้าเราสามารถรู้ปริมาณน้ำท่าหรือปริมาณน้ำในแม่น้ำได้ล่วงหน้า ทำให้สามารถรู้ว่าจะเกิดปัญหาเกี่ยวกับน้ำขึ้นในอนาคตหรือไม่ ทำให้มีการวางแผนรับมือกับปัญหาที่จะเกิดขึ้นในอนาคตได้ งานวิจัยนี้เสนอการสร้างโมเดลคาดการณ์ค่าปริมาณน้ำท่าด้วยโครงข่ายประสาทเทียม โดยการทดลองใช้ค่าอุณหภูมิ (T) และดัชนีผลต่างพืชพรรณ (NDVI) ปริมาณน้ำฝน (R) จำนวนวันที่ฝนตก (D) และปริมาณน้ำท่า (Q) ที่เวลาย้อนหลัง 1 และ 2 เดือน (t-1 และ t-2) ร่วมกับ ค่าการจัดกลุ่ม (C) ที่เวลาย้อนหลัง 1 เดือน (t-1) เพื่อแสดงให้เห็นว่าการเลือกใช้พารามิเตอร์ในการสร้างโมเดลมีผลต่อการทำนาย การใช้ค่าการจัดกลุ่มเพื่อเพิ่มประสิทธิ์ภาพของการทำนายค่าน้ำท่า และค่าจากการสำรวจระยะไกลสามารถใช้ในการทำนายน้ำท่าได้ การประเมินประสิทธิภาพโดยใช้วิธีการทางสถิติ 2 ชนิดคือค่า RMSE ที่แสดงความผิดพลาดของการคาดการณ์ปริมาณน้ำท่าและ R^2 เพื่อดูความสัมพันธ์ระหว่างค่าน้ำท่าจริงกับค่าน้ำท่าที่ได้จากการคาดการณ์ นอกจากนี้ยังเปรียบเทียบประสิทธิภาพกับการสร้างแบบจำลองด้วยสมการถดถอย

ผลการทดลองสรุปได้ว่าเมื่อใช้การจัดกลุ่มร่วมกับพารามิเตอร์อื่น ๆ เพื่อสร้างโมเดลคาดการณ์ปริมาณน้ำท่าสามารถเพิ่มประสิทธิภาพการคาดการณ์ได้ และเมื่อใช้ค่าดัชนีผลต่างพืชพรรณ ร่วมกับอุณหภูมิ และค่าการจัดกลุ่ม (NDVI+T+C) สร้างโมเดลด้วยโครงข่ายประสาทเทียม ให้ประสิทธิภาพดีที่สุดโดยให้ค่า RMSE=0.09 และ R^2=0.743 (แสดงในตารางที่ 2) จะเห็นว่าการใช้ข้อมูลการสำรวจระยะไกลด้วยดัชนีผลต่างพืชพรรณ และค่าการจัดกลุ่มจาก k-means สามารถนำมาใช้ในการสร้างโมเดลคาดการณ์ค่าปริมาณน้ำท่าได้ และการสร้างโมเดลด้วยโครงข่ายประสาทเทียมเป็นวิธีการที่มีการคำนวณที่ซับซ้อน การเลือกใช้พารามิเตอร์ในการสร้างโมเดลที่ใช้ในการคาดการณ์ค่าปริมาณน้ำท่ามีประสิทธิภาพ และสามารถป้องกันการเกิดปัญหาการ Overfitting จำนวนพารามิเตอร์และการเลือกใช้พารามิเตอร์ไม่เหมาะสม และพารามิเตอร์ที่เลือกใช้จะต้องมีความสัมพันธ์กับค่าที่คาดการณ์ ถ้าเราเลือกใช้พารามิเตอร์ไม่เหมาะสมจะทำให้โมเดลการคาดการณ์น้ำท่าจากโครงข่ายประสาทเทียมคาดการณ์ปริมาณน้ำท่าผิดพลาด

นอกจากนี้ข้อมูลปริมาณน้ำฝนและน้ำท่าที่ใช้ทำการทดลองในงานวิจัยนี้ใช้แค่หนึ่งสถานี หากทำการทดลองเพิ่มโดยนำข้อมูลปริมาณน้ำฝนและน้ำท่าจากสถานีข้างเคียงมาพิจารณา หรือต้องการวิธีการเลือกพารามิเตอร์ที่เหมาะสมในการสร้างโมเดลคาดการณ์ปริมาณน้ำท่าอาจจะช่วยให้เพิ่มประสิทธิภาพการคาดการณ์ปริมาณน้ำท่าได้ การทดลองต่อไปในอนาคตจะต้องทดลองกับข้อมูลน้ำท่าสถานีอื่นและลุ่มน้ำบริเวณอื่น เพื่อแสดงให้เห็นประสิทธิภาพของวิธีการที่นำเสนอ

5. เอกสารอ้างอิง

[1] S. Riad, J. Mania, L. Bouchaou and Y. Najjar, "Rainfall-runoff model using an artificial neural network approach," *Mathematical and Computer Modelling*, vol. 40, no. 7, pp. 839-846, 2004.

[2] A. R. Ghumman, Y. M. Ghazaw, A. R. Sohail and K. Watanabe, "Runoff forecasting by artificial neural network and conventional mode," *Alexandria Engineering Journal*, vol. 50, no. 4, pp. 345-350, 2011.

[3] R. Modarres, "Multi-criteria validation of artificial neural network rainfall-runoff modeling," *Hydrology and Earth System Sciences*, vol. 13, no.3, pp. 411-421, 2009.

[4] G. Sharm, Y.P. Mathur, S. K. Vyas, & P.K. Navin, "Rainfall-Runoff Regression Model for Meja Catchment," *International Journal of Inventions in Reasearch, Engineering Science and Technology (IJIREST)*, vol. 1, no. 1, pp. 58-62, 2014.

[5] F. Machado, M. Mine, E. Kaviski and H. Fill, "Monthly rainfall–runoff modelling using artificial neural networks," *Hydrological Sciences Journal–Journal des Sciences Hydrologiques*, vol. 56, no.3, pp. 349-361, 2011.

[6] F.N. Kogan, "Vegetation index for a real analysis of crop conditions," in *Proceedings of the 18th Conference on Agricultural and Forest Meteorology*, AMS, W. Lafayette, Indiana, 15–18 September 1987, Indiana, USA, pp. 103–106.

[7] Mark Hudson Beale, Martin T. Hagan and Howard B. Demuth, *Neural Network Toolbox™ User's Guide*: MathWorks Inc, 2015.

[8] Alsabti Khaled, Sanjay Ranka and Vineet Singh, "An efficient k-means clustering algorithm," *Electrical Engineering and Computer Science*, 1997.

Proceedings of the International MultiConference of Engineers and Computer Scientists 2017 Vol I,
IMECS 2017, March 15 - 17, 2017, Hong Kong

On Applying Regression and Neural Network to Predict Rainfall Using Satellite Based Index

Ratiporn Chanklan*, Keerachart Suksut, Kedkard Chaiyakhan, Nuntawut Kaoungku, Kittisak Kerdprasop and Nittaya Kerdprasop

***Abstract*— In this paper, we adopt a statistical method using the linear regression analysis to study relationship between the satellite based vegetation index and the ground based rainfall data, and then apply the data mining method using the neural network to induce a model to predict the amount of annual rainfall. The model is intended to be useful for drought monitoring. Remote sensing data used in our study is the Normalized Difference Vegetation Index (NDVI) obtained from the NOAA STAR. The ground station rainfall data during the years 2005 to 2014 in Nakhon Ratchasima province, Thailand, are obtained from the Meteorological Department. The study of NDVI and ground-based rainfall relationship has been done through the correlation coefficient analysis. The preliminary study results show that vegetation index and rainfall positively correlate with 1-month lagged time. The studied period from June to September shows the strong correlation (r = 0.715, on average). We then induce the rainfall predictive model using neural network with the NDVI and rainfall as the input parameters. The performances of using only a rainfall parameter and a combination of NDVI and rainfall parameters are also compared. The experimental results show that using the remote sensing NDVI data together with the ground based rainfall data can improve accuracy of the neural network model to predict the future annual rainfall.**

***Index Terms*— Remote sensing, Normalized Difference Vegetation Index, Annual rainfall prediction, Neural network**

I. Introduction

A drought is water shortage in an area occurring from precipitation deficiency or the unseasonally lacking of rainfall. The effects of drought are numerous: a lack of water for humans and other living animals, insufficient water for the crops, food inadequate due to crop damage. The agriculture in Thailand depends on natural water, mostly rainfall. Therefore, lacking of enough rainfall for a long period of time has serious effect to agricultural yields. Traditionally, drought monitoring has been done based on meteorological information from ground stations, which does not cover all areas in the country. Due to the limited numbers of ground stations, it results in less, discontinuing, and incomplete data for timely drought warning. Therefore, data analyses are inaccurate and sometime out of date because it takes long time to collect data.

Satellite data can be used to help monitoring drought situation. Environmental observation satellites have been launched to record a continuous, spatial patterns. The satellite data cover the whole global surface area and the data are available for almost real-time access [1]. The remote sensing is thus suitable to track the changes and to monitor the impact of drought on crops [2],[3],[4]. Thus, the intuitive idea of this work is that if we can find the relationship between the remote sensing data and the ground station data from the Meteorological Department, then in the process of model inductive we can use the remotely sensed data incorporating with the ground-based data for predicting the rainfall in the upcoming year. Such model has the advantage of timely monitoring of drought.

In this work, we firstly analyze the relationship between Normalized Difference Vegetation Index obtained from the NOAA satellite and the monthly rainfall data from the Meteorological Department in Nakhon Ratchasima province, Thailand. After confirming its positive correlation, we then build a model to predict the rainfall using artificial neural network. The inputs for our neural network model are the lagged 1-month rainfall and lagged 1-moth NDVI data.

II. Background Theories

A. Normalized Difference Vegetation Index

In this study, we use the remotely sensed data, which is the Global Vegetation Index (GVI)in the area of Nakhon Ratchasima province in the northeast of Thailand. GVI is the basic index for measuring the greenness of the earth surface through the monitoring for density and healthy of land surface vegetation. One specific product of GVI is the Normalized Difference Vegetation Index (NDVI), which is the computation of signals sensed by the channels 1 and 2 of the satellite that aggregates the 4 square km global area coverage daily.

Manuscript received September 26, 2016; revised January 10, 2017. This work was supported in part by grant from Suranaree University of Technology through the funding of Data Engineering Research Unit.

R. Chanklan is a doctoral student with the School of Computer Engineering, Suranaree University of Technology, 111 University Avenue, Muang, Nakhon Ratchasima 30000, Thailand. (corresponding author: +66994696164; e-mail: arc_angle@hotmail.com).

K. Suksut is a doctoral student with the School of Computer Engineering, Institute of Engineering, Suranaree University of Technology, Nakhon Ratchasima, Thailand. (e-mail: mikaiterng@gmail.com).

K. Chaiyakhan is a lecturer with the Computer Engineering Department, Rajamangala University of Technology Isan, Nakhon Ratchasima, Thailand. (e-mail: kedkarnc@hotmail.com).

N. Kaoungku is a lecturer with the School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima, Thailand. (e-mail: nuntawut@sut.ac.th).

K. Kerdprasop is an associate professor with the School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand (e-mail: kerdpras@sut.ac.th).

N. Kerdprasop is an associate professor with the School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima, Thailand. (e-mail: nittaya@sut.ac.th)

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Proceedings of the International MultiConference of Engineers and Computer Scientists 2017 Vol I,
IMECS 2017, March 15 - 17, 2017, Hong Kong

The basic concept of NDVI is based on the fact that internal mesophyll structure of healthy green leaves reflects near-infrared (NIR) radiation, whereas the leaf chlorophyll and other pigments absorb a large proportion of the red visible (VIS) radiation. This function of internal leaf structure becomes reversed in case of unhealthy or water stressed vegetation [5]. The calculation of the NDVI value is thus performed with the Equation1.

$$NDVI = (NIR-VIS) / (NIR+VIS) \quad (1)$$

where, NIR is near infrared and VIS is visible red band of electromagnetic spectrum. The value of NDVI ranges between 1 and +1. It is found below 0.1 in the areas with barren rock, sand and snow cover, whereas it may range from 0.6 to 0.8 in temperate and tropical rainforests. NDVI is suitable for monitoring drought, estimating healthy status of vegetation, crop growth conditions and crop yields [6],[7].

B. Rainfall Data

Rainfall is very important in meteorology because water is a major factor related to the living of people, living creatures, and agriculture. Healthy vegetation and plentiful crop lands are all depending on rainfall. A measure of rainfall is normally done by a rain gauge placed in open space for 24 hours. Observations of daily rainfall are nominally made at 7:00 am to 7:00 pm local clock time each day. The automatic rain gauges can measure rainfall continuously 6, 12, 24 hours or weekly.

C. Correlation Coefficient

The correlation is a statistical measure used to explore relationship between the variables. The degree of correlation is interpreted from the correlation coefficient (R). The correlation coefficient is a numerical value between -1 and 1. It expresses the strength of the linear relationship between two variables (x and y). The direction of the relationship between the two variables can be shown by scatter plot. There are three possible types of relationship: positive correlations, negative correlations, and zero correlations.

Positive correlation is the kind of relationship such that the increase or decrease the value of one variable will cause a corresponding increase or decrease in value of the other variable.

Negative correlation is reverse relationship in which the increase or decrease on a variable's value will cause the other variables change their values in opposite direction.

Zero correlation is the situation in which the two variables have no relationship. The correlation coefficient can be calculated using Equation 2.

$$R = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}} \quad (2)$$

where n is the total number of samples, $x_i = (x_1, x_2, \ldots, x_n)$ are the values of variables x, and y_i is the value of variable y. If the correlation coefficient is closer to 1 or -1, it represents the relationship between the variables at a high level. If the value is close to 0, it represents the relationship between the variables at a low level or no relationship.

D. Artificial Neural Networks

The Artificial Neural Network (ANN) is the most widely used form of neural networks. An ANN is a computational approach inspired by studies of the brain and nervous systems in biological organisms. The powerful functionality of a biological neural system has been attributed to the parallel-distributed processing nature of the biological neurons [8]. Conventionally, the network (fig.1) has three main levels: input layer, hidden layer, and output layer. Nodes in the input layer called the input nodes. The number of nodes in the input layer is equal to the number of features (attributes, independent variables). Nodes in the hidden layer are called the hidden nodes. Number of nodes in the hidden layer is defined by a user. Node in the output layer is called an output node. The number of output nodes is equal to the number of data groups (or target, dependent variable). Nodes in the network are connected with lines; from input nodes to hidden nodes, and from hidden nodes to output nodes. Each line connecting one node to another has weight (w).

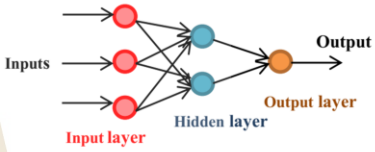


Fig.1. Architecture of the Artificial Neural Network.

Input data is vector of elements: $p = [p_1, p_2, \ldots, p_R]$, R is number of elements (or dimension) in input data. Each line in the network is annotated with weight: $W = [w_1, w_2, \ldots, w_R]$. The network works by multiplying weight on each edge to the input data, summing results from each incoming edge of the node, and finally summing a bias (b) of that node. The summation result of each node is denoted by n. The result (n) has to be transformed through a transfer function to obtain the final computation result of each node, denoted as a. The calculation of the output value from a neuron as shown in fig.2 is summarized in Equation 3.

$$a = f(n) = f(Wp + b) \quad (3)$$

where $n = w_{1,1}p_1 + w_{1,2}p_2 + \cdots + w_{1,R}p_R + b$

$n = Wp + b$

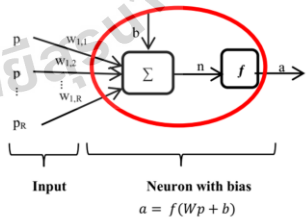


Fig.2. Neural unit with incoming input edges and a bias.

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Proceedings of the International MultiConference of Engineers and Computer Scientists 2017 Vol I,
IMECS 2017, March 15 - 17, 2017, Hong Kong

For the weight and bias, it is can be adjusted from the learned data. If the output value from neural network is false as compared to the true value in the training data, the weight will be update using the error as a guidance. The process continues until the predictive error of the neural network model is less than the acceptable threshold.

III. MATERIALS AND METHODS

In this work, the study area is located in Nakhon Ratchasima, a largest province in the northeast of Thailand (fig.3). We used remote sensing data, which is the Normalized Difference Vegetation Index or NDVI, obtained from the global vegetation health products of NOAA STAR (http://www.star.nesdis.noaa.gov/smcd/emb/vci/VH/vh_browseByCountry_province.php?country=THA&provinceID=28&year1=1981&year2=2015X), and monthly rainfall data in Nakhon Ratchasima area during the years 2005 to 2014. The data are obtained from the Meteorological Department (http://www.dnp.go.th/statistics/dnpstatmain.asp).

Our main objective of study is to find the relationship between Normalized Difference Vegetation Index and rainfall data through the analysis of correlation. The focus of our study is the correlation of the two variables during the southwest monsoon season. To explore the relationship during monsoon period, we use data during the months of June to September. This choice is because the southwest monsoon prevails over Thailand between mid-May to mid-October.

We then use linear regression to determine the relationship between two variables; the independent variable (X) is the Normalized Difference Vegetation Index and the dependent variable (Y) is the rainfall. The NDVI and rainfall data have been lagged from 1 to 6 months. We find the linear correlation between the independent variables and the dependent variable. The preliminary result shows that the lagged 1-month is the discriminative factor, as shown in Equation 4.

$$f(Rainfall_t) = f(Rainfall_{t-1}, NDVI_{t-1}) \quad (4)$$

Fig.3. The study area: Nakhon Ratchasima, Thailand. (www.maphill.com)

In the subsequent step of rainfall prediction model induction with neural network, we thus use rainfall lagged 1-month and NDVI lagged 1-month as input parameters of the neural network. The output node of the network is a node to predict rainfall in current time (t).

IV. EXPERIMENTAL RESULTS

We plot the linear regression to find the relationship between two variables (NDVI and rainfall data). The results are shown in fig 4.

When we analyzed yearly relationship between NDVI and annual rainfall from 2005 to 2014, the results show the correlation coefficient as both positive and negative values. This means that the relationships of the NDVI and annual rainfall during the past ten years are not in the same direction each year. The correlation coefficient range is -0.134 to 0.438. This implies that NDVI and rainfall give quite poor correlation. By merging the ten-year data as one group, we find the correlation coefficient to be 0.159. A single-group NDVI and rainfall relationship is show in fig 5.

To further analyze NDVI-rainfall relationship, we plot NDVI and rainfall values to see a fine-grain relationship in the monthly period, and the result is shown in fig 6. In fig.6, the monthly NDVI and rainfall during the years 2005 to 2014 are plotted as separate graphs.

From fig. 6, we can now notice the trend of NDVI to increase from June to September, every year from 2005-2014. There are two increase trend in the rainfall plot; that is from June to August, and another peak from August to September. June to September is actually the raining season that has some effect from the monsoon. Thailand is under the influence of two monsoon types: southwest monsoon and northeast monsoon. Southwest monsoon prevails over in mid-May to mid-October. During these months, weather are cloudy and rainy. Northeast monsoon prevails over in mid-October to mid-February, with the clear, cold, and dry weather. Fig 6 shows rainfall in June to September with increasing amount of rainfall. The increasing trend also appears in the NDVI plot. We therefore find the correlation coefficient of NDVI with rainfall during June to September, and the results are shown in Table 1.

TABLE I
THE CORRELATION COEFFICIENT RESULTS

Year	January to December	June to September
2005	0.223	0.917
2006	0.320	0.940
2007	-0.098	0.726
2008	0.114	0.983
2009	-0.076	0.709
2010	0.438	0.880
2011	-0.134	-0.114
2012	0.239	0.600
2013	0.552	0.690
2014	0.225	0.817
2005 to 2014	0.159	0.518

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Proceedings of the International MultiConference of Engineers and Computer Scientists 2017 Vol I,
IMECS 2017, March 15 - 17, 2017, Hong Kong

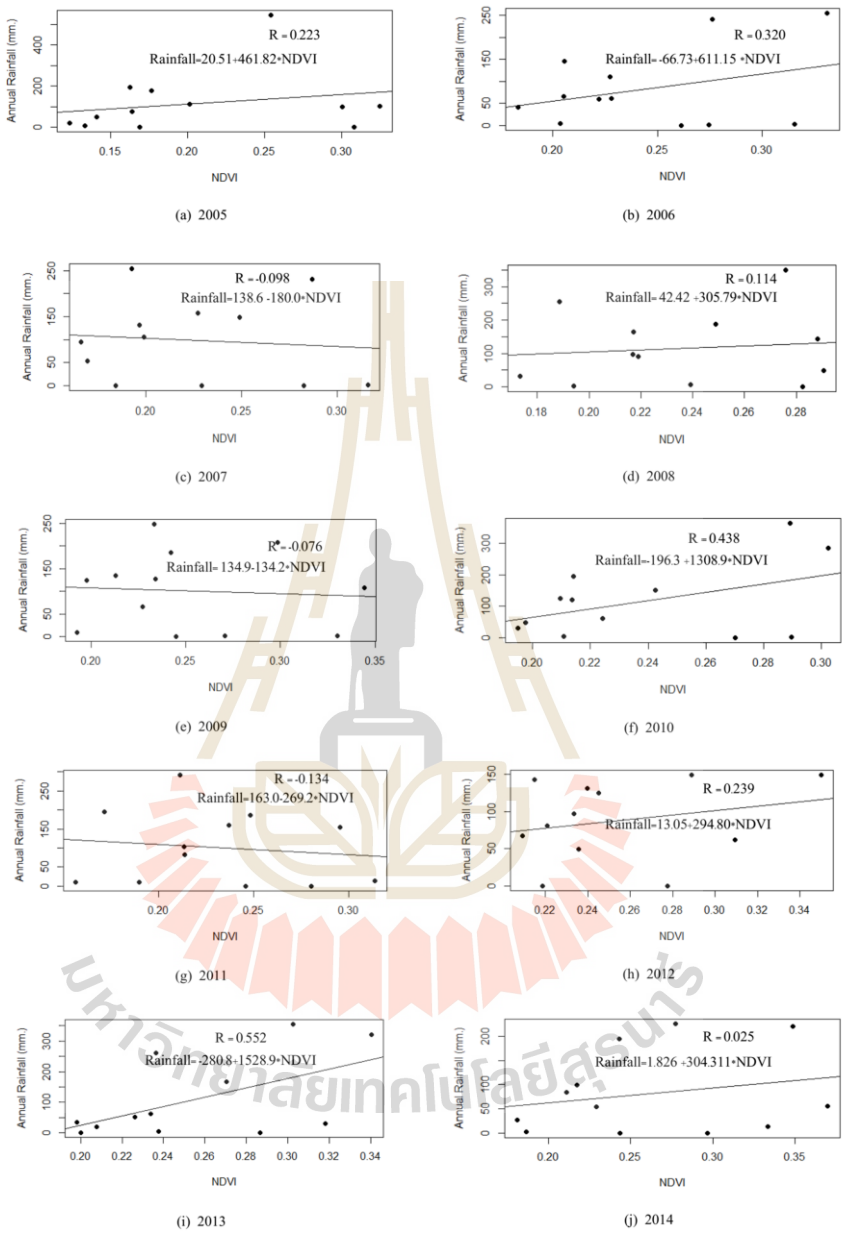


Fig.4. The correlation coefficient (R) of NDVI and rainfall.

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Proceedings of the International MultiConference of Engineers and Computer Scientists 2017 Vol I,
IMECS 2017, March 15 - 17, 2017, Hong Kong

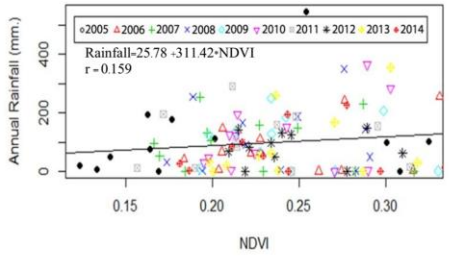
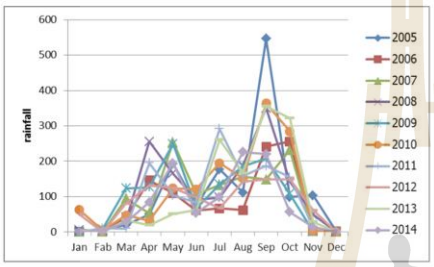
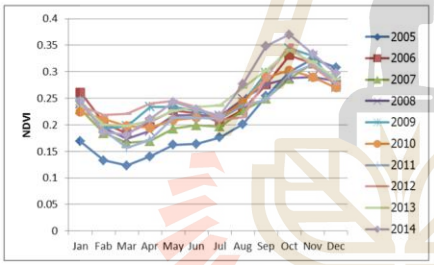


Fig.5. The correlation coefficient of NDVI and rainfall from 2005 to 2014.



(a) Rainfall



(b) NDVI

Fig.6. The graphs show monthly (a) rainfall and (b) NDVI.

From Table 1, it can be noticed that during the Southwest monsoon season, the correlation coefficient of NDVI and rainfall during the months of June to September shows positive direction with strong correlation in the range 0.6 to 0.9. There is only one exception in the year 2011; the NDVI-rainfall relationship shows negative sign. In the year 2011, the rainfall was very unusual affected by the Tropical Storm Nock-ten [9]. It caused big flood in Thailand. NDVI is the index for measuring the 'greenness' of the earth's surface. When flooding covered crops, the relationship between NDVI and rainfall showed negative direction, which is different from other years.

TABLE II
THE RESULTS IN PREDICT RAINFALL USING ARTIFICIAL NEURAL NETWORKS

Data	Input Parameter to ANN	R	RMSE
January to December	Rainfall	0.42	87.91
January to December	Rainfall, NDVI	0.62	74.02

After performing correlation analyses, we then apply artificial neural networks to build a model for predicting rainfall. For experimentation, we use the data from 2005 to 2010 as training data for a neural network, and the data from 2012 to 2014 are used as test data. We use data from January to December in both training and test data.

We also compare the model performance using two different set of input parameters: set one having only rainfall data as input parameter into ANN, and set two having rainfall and NDVI as input parameters for ANN. The results are shown in Table 2. The model that uses both rainfall and NDVI data as input parameters gives a less root-mean-square error (RMSE) than the model that has only rainfall as its input parameter. The correlation coefficient also shows good relationship. The RMSE metric used for model evaluation can be calculated as in equation (4).

$$\text{RMSE} = \sqrt{\left(\frac{\sum(T_i - O_i)^2}{N}\right)} \quad (4)$$

where T_i is the actual data, O_i is the value of predict, and N is the number of all data. The less value of RMSE means the more accurate prediction of the model.

V. CONCLUSION

In this study, we use remote sensing data of Normalized Difference Vegetation Index from the NOAA satellite and rainfall data from the Meteorological Department of Nakhon Ratchasima province in Thailand. The collected data are from the years 2005 to 2014. The analysis results of regression method reveal that a relationship during the months of June to September, which is the period of southwest monsoon, is the best relationships. NDVI and rainfall during this period have positive relationship. The regression model also shows that current rainfall can be estimated from the lagged 1-month rainfall and lagged 1-month NDVI.

To build a model for rainfall prediction with artificial neural network, we apply NDVI and rainfall as input parameters to the network. The model yields prediction result with lower RMSE as compared to the prediction with only rainfall data as input vector of ANN.

The results observed from our experiment show that rainfall data from the ground station and NDVI data from remote sensing are supplement each other for predicting rainfall and monitoring drought, which is the period with less rainfall than the usual. This emphasis the advantage of remote sensing data that can help timely prediction and can cover broad area of land surface.

REFERENCES

[1] Gu, Y., Brown, J. F., Verdin, J. P., & Wardlow, B. (2007). "A five-year analysis of MODIS NDVI and NDWI for grassland drought assessment over the central Great Plains of the United States," *Geophysical Research Letters*, 34(6).

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

[2] Kogan, F., & Guo, W., "Early Detection and Monitoring Droughts From NOAA Environmental Satellites. In Use of Satellite and In-Situ Data to Improve Sustainability," *Springer Netherlands*, pp.11-18, 2011.
[3] Kuri, F., Murwira, A., Murwira, K. S., & Masocha, M, "Predicting maize yield in Zimbabwe using dry dekads derived from remotely sensed Vegetation Condition Index," *International Journal of Applied Earth Observation and Geoinformation*, vol 33, pp. 39-46, 2014.
[4] Dutta, D., Kundu, A., Patel, N. R., Saha, S. K., & Siddiqui, A. R., "Assessment of agricultural drought in Rajasthan (India) using remote sensing derived Vegetation Condition Index (VCI) and Standardized Precipitation Index (SPI)," The Egyptian Journal of Remote Sensing and Space Science. vol. 18, pp. 53-63, 2015.
[5] Dipanwita Dutta, Arnab Kundu, N.R. Patel, S.K. Saha and A.R. Siddiqui, "Assessment of agricultural drought in Rajasthan (India) using remote sensing derived Vegetation Condition Index (VCI) and Standardized Precipitation Index (SPI)," *The Egyptian Journal of Remote Sensing and Space Sciences,* 2015, vol.18, pp.53–63.
[6] Kogan, F.N., "Vegetation index for a real analysis of crop conditions," *In Proceedings of the 18th Conference on Agricultural and Forest Meteorology, AMS, W. Lafayette, Indiana, 15–18 September 1987*, Indiana, USA, pp. 103–106, 1987.
[7] Dabrowska-Zielinska, K., Kogan, F.N., Ciolkoszs, A., Gruszczynska, M., Kowalik, W., "Modelling of crop growth conditions and crop yield in Poland using AVHRR-based indices," *Int. J. Remote Sens*, vol. 23, pp. 1109–1123, 2002.
[8] K.C. Luk, J.E. Ball and A. Sharma, "A study of optimal model lag and spatial inputs to artificial neural network for rainfall forecasting," *Journal of Hydrology*, 2000, vol 226, pp. 56–65.
[9] Watanasak Sornrunk, Tawatchai Kamoltam. "Ministry of Public Health:The operation Flood crisis year 2011," *Journal of Preventive Medicine Association of Thailand*, vol. 2, no. 2, 112-115, 2011.

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

International Journal of Modeling and Optimization, Vol. 7, No. 2, April 2017

A Hybrid Modeling Technique to Predict Runoff

Ratiporn Chanklan, Kedkarn Chaiyakhan, Kittisak Kerdprasop, and Nittaya Kerdprasop

***Abstract*—The management of water resources is important to prevent water problems: floods and water shortages. The foreknowledge allows time for officials to sufficient preparation to deal with the problem. This study aims to determine the appropriate weight for predicting runoff from the merge of runoff prediction results from two algorithms: Artificial Neural Network and Support Vector Regression with linear regression modeling. In this paper, we compare the runoff predictive performance of the three algorithms: Linear Regression, Artificial Neural Network, and Support Vector Regression. We use remote sensing data, which are the Normalized Difference Vegetation Index (NDVI) obtained from the NOAA STAR. The ground station rainfall, runoff, the number of rainy days and temperature data in Mun basin, Thailand, are obtained from the Meteorological Department. We evaluate the model performance using two statistical values: Correlation Coefficient and Root Mean Squared Error. Experimental results confirm the best performance of our proposed method.**

***Index Terms*—Runoff, artificial neural network, support vector regression.**

I. Introduction

Currently, people are experiencing a natural disaster such as floods and water shortages. The strength of such disaster increases every year and causes much damage. A proper management of water resources is one way to protect flood and water shortage problems. Predicting runoff can help to make decision, planning and management of water resources. Runoff is amount of water in the river caused by the rain that fell in the catchment area then flows into the river. The foreknowledge regarding amount of runoff as either excessive or shortage can be useful for estimating the demand for use and planning to fix or deal with floods and water shortages.

Predicting runoff is a very complex process and it also needs an appropriate modeling technique for accurate prediction. Artificial Neural Network (ANN) is a tool that has been used to create model to predict runoff. It has the ability to simulate both linear and non-linear relationships, without any prior assumptions as most traditional statistical methods. ANN has been successfully used for predicting runoff and the method was widely adopted in hydrology [1], [2]. In addition, some researches have suggested the support vector regression (SVR) as an alternative algorithm for predicting runoff effectively. SVR showed the best performance as reported in [3], [4]. Rainfall lag time values are also used to consider for building a model to predict runoff [5], [6]. Runoff lag time values are also proposed to predict runoff [7]. A comparison on efficiency in the literature normally uses statistical values such as Correlation Coefficient, Coefficient of Determination, Root Mean Squared Error (RMSE), Mean Absolute Percentage Error. In this work, we employ the two measures: Correlation Coefficient and RMSE.

In the model building process, we use a hybrid modeling technique from Artificial Neural Network and Support Vector Regression. At the first step, we find cluster among rainfall and runoff data using k-means clustering. Then, we compute predictor importance to select data (runoff, rainfall, the number of rainy days, temperature and cluster data) that are appropriate for creating a predictive model. Then, we use runoff prediction results from Artificial Neural Network and Support Vector Regression as inputs for Linear Regression to make a final runoff prediction.

Manuscript received February 15, 2017; revised April 19, 2017. This work was supported by grant from Suranaree University of Technology through the funding of Knowledge and Data Engineering Research Units.

R Chanklan is with the School of Computer Engineering, Suranaree University of Technology (SUT), 111 University Avenue, Muang, Nakhon Ratchasima 30000, Thailand (e-mail: arc_angle@hotmail.com).

K. Chaiyakhan is with the Computer Engineering Department, Rajamangala University of Technology Isan, Nakhon Ratchasima, Thailand (e-mail: kedkarnc@hotmail.com).

K. Kerdprasop is with the School of Computer Engineering and the Head of Knowledge Engineering Research Unit, SUT, Thailand (e-mail: kerdpras@sut.ac.th).

N. Kerdprasop is with the School of Computer Engineering and the Head of Data Engineering Research Unit, SUT, Thailand (e-mail: nittaya@sut.ac.th).

II. Background Theories

A. Artificial Neural Network

Artificial Neural Network (ANN) is a mathematical method with the basic idea to make a computer machine having the ability to think like humans. ANN has connected multiple computational nodes to form a network. The network has three main levels: input layer, hidden layer, and output layer. The input nodes are node in the input layer. In the input layer, the number of nodes is equal to the number of attributes (or features, independent variables). The hidden nodes means nodes in the hidden layer. The number of nodes in the hidden is defined by a user. There can be more than one layer in the hidden layer of the network. The output nodes are nodes in the output layer. In the output layer, the number of nodes is equal to the number of data groups (or target, dependent variable). The hidden nodes, output nodes, and the structure of the internal calculation are shown in Fig. 1.

In the network, each line has specified weight: $W = [w_1, w_2, ..., w_R]$. Input data are vector of elements: $p = [p_1, p_2, ..., p_R]$, in which R is number of attributes in input data. The network works by multiplying the input data to weight on each edge, summing results from each incoming edge of the node, and summing a bias (b) of that node. The result (n) has to be transformed through a transfer function to obtain the final computation result of each node, denoted as a.

The main objective of ANN learning is the search for proper weight on each edge in the network such that the

DOI: 10.7763/IJMO.2017.V7.559

network is most accurate on separating training data to a correct class. The weight of the edge is uncertain but the network can adjust weight values by teaching it to recognize a pattern of data. It adjusts weight if the output from the neural network is incorrect. The weight adjustment is iterated in a back propagation manner until the network has a small error or it reaches an acceptable threshold.

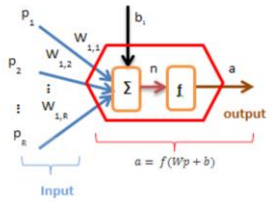


Fig. 1. Neural unit with incoming input edges and a bias.

B. Support Vector Regression

Support vector regression (SVR) has been developed from support vector machines (SVM). The algorithm can estimate the target by a linear equation [8], as shown in equation 1.

$$f(\vec{x}) = \vec{w} \cdot \vec{x} + b \quad (1)$$

When b is a threshold value and w is a weight vector. It finds hyperplane which has small margin and tries to keep all the data within the margin and allows some data to be outside the margin. The margin can be calculated as in equation 2.

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \quad (2)$$

When ξ is slack variable such that $\xi_i, \xi_i^* \geq 0$ and N is number of train data. The margin within the scope can be computed with equations 3 and 4.

$$y_i - (\vec{w} \cdot \vec{x} + b) \leq \varepsilon + \xi_i \quad (3)$$

$$y_i - (\vec{w} \cdot \vec{x} + b) \leq -\varepsilon - \xi_i^* \quad (4)$$

Then linear Support Vector Regression is as shown in equation 5.

$$y = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*) \cdot \langle x_i, x \rangle + b \quad (5)$$

where $\langle x_i, x \rangle$ is the inner product of two vectors in the feature space.

C. K-means Clustering

K-means clustering is a method to partition n objects into k clusters in which each object belongs to the cluster with the nearest central point, or centroid [9]. The number of *k* cluster can be defined by user. K-means steps are as follows:

1) Set number of clusters, denoted as *k*
2) Random cluster centers (centroid) in each *k* group
   Measure the distance between the data D = (x_1,y_1) and Centroid = (cx_1,cy_1) according to the Euclidean distance function, computed as in equation 6.

$$\text{Euclidean distance} = \sqrt{(x_1 - cx_1)^2 + (y_1 - cy_1)^2} \quad (6)$$

3) Assign data to their group based on the closest distance between the data and the cluster center.
4) Calculate the mean of all data in each cluster and set it to be the new centroid. Suppose there are two data points in group A = ((x_1,y_1), (x_2,y_2)), the new centroid can be calculate using the mean of all data in the group, as shown in equation 7.

$$\text{centroid} = \left(\frac{x_1+x_2}{2}, \frac{y_1+y_2}{2}\right) \quad (7)$$

5) Repeat steps 3, 4 and 5 until all the centroids do not change.

D. Predictor Importance

Predictor importance can be determined by computing the reduction in variance of the target attributable to each predictor using a sensitivity analysis. Predictors are ranked according to the sensitivity measure [10] defined as in equation 8.

$$S_i = \frac{V_i}{V(Y)} = \frac{V(E(Y|X_i))}{V(Y)} \quad (8)$$

where Y is target, X_i is predictor ranging from 1,...,k. The number of predictors is k. Model for Y based on predictors X_1 through X_k and $V(Y)$ is the unconditional output variance. Predictor importance is then computed as the normalized sensitivity using equation 9.

$$VI_i = \frac{S_i}{\sum_{j=1}^{k} S_j} \quad (9)$$

where S_i is the proper measure of sensitivity to rank the predictors in order of importance for any combination of interaction and non-orthogonality among predictors. VI_i is the estimate of the conditional variances computing by the multidimensional integrals in the space of the input factors using Monte Carlo method.

E. Correlation Coefficient

The Correlation Coefficient is denoted by R. It is a statistical value to find relationships between two variables (x and y). The coefficient is a numerical value between -1 and 1. There are three possible types of relationship: zero correlations means no relationship, positive correlations is same direction relationship, and negative correlation is the kind of inverse relationship. The correlation coefficient can be calculated using equation 10.

$$R = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}} \quad (10)$$

where $x_i = (x_1, x_2, \ldots x_n)$ are the values of variable x, y_i is the value of variable y, n is the total number of samples.

The correlation coefficient that is closer to 1 or -1 represents high level of relationship. The coefficient that is close to 0 represents low level or even no relationship.

F. Root Mean Squared Error

Root Mean Squared Error (RMSE) is used to check errors in the prediction of the model. If the RMSE has a small value that means the prediction model is efficient. The RMSE metric used for model evaluation can be calculated as in equation 11.

$$\text{RMSE} = \sqrt{\left(\frac{\sum(T_i - O_i)^2}{N}\right)} \quad (11)$$

where N is the number of all data, O_i is the value of prediction, T_i is the actual data.

III. Materials and Methods

The study area in this work is Mun basin, a largest basin in the North-eastern region of Thailand (Fig. 2). We use Normalized Difference Vegetation Index (NDVI), which is a remote sensing data obtained from the NOAA STAR (http://www.star.nesdis.noaa.gov), monthly rainfall, runoff and the number of rainy days data from the Meteorological Department (http://www.hydro-4.com), and temperature data from National Statistical Office (http://www.nso.go.th). This research use IBM SPSS Modeler 14.1 as analysis tool in our experiments. We use two data sources from the Mun basin: M145 and M173 stations.

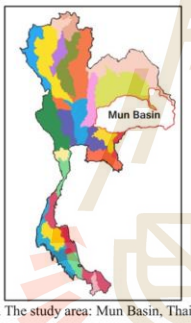


Fig. 2. The study area: Mun Basin, Thailand.

The M145 (Lamphra Phloeng) station locates at Ban Wang Takhian, Tambun Wang Katha, Amphoe Pak Chong, in Nakhon Ratchasima province, Thailand. The 13-year data during 1998 to 2010 have been used as training data. The 5-year data from 2011 to 2015 are to be used as test data.

The M.173 (Moon River) station locates at Ban Non Sa-at, Tambun Tha Yiam, Amphoe Chokchai, in Nakhon Ratchasima province, Thailand. The 10-year training data are those during the periods 2002 to 2011. The 5-year data between 2012 to 2015 are used as test data.

Prepare step: We apply k-means to find clusters from rainfall and runoff data. The appropriate k clusters have been judged from the Silhouette Coefficient. Then, we use predictor importance to select a subset of potential features from the initial set of features including rainfall, runoff, the number of rainy day, temperature, NDVI, and cluster identifier. Data are also lagged 1-month and 2-month. These data are input into two learning algorithms: Artificial Neural Network and Support Vector Regression. Because we use the data from different sources, we thus have to find different sets of parameters that are appropriate to predict runoff in different locations. We later use runoff prediction results from Artificial Neural Network and Support Vector Regression as inputs in Linear Regression, which is a final model to predict runoff in this work.

In the subsequent step, we use test data to test model performance. We evaluate performance using two statistical values: Correlation Coefficient and Root Mean Squared Error. The modeling process of this paper is shown in Fig. 3.

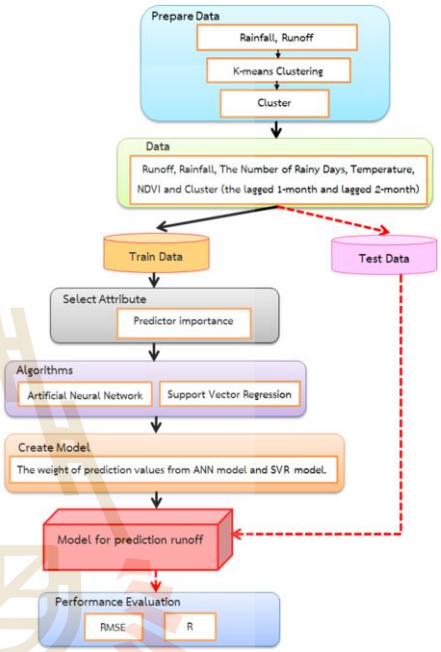


Fig. 3. The modeling process for runoff prediction.

IV. Experimental Results

We predict runoff with test data using the model proposed in Figure 3 and compare the results against the other three models including Artificial Neural Network, Support Vector Regression, and Linear Regression. We experiment with two data sets from the Mun basin: M145 and M.173 stations.

We firstly explore correlation between runoff and other attributes (rainfall, the number of rainy days, NDVI, and temperature) and the computed coefficients are shown in Table 1. The results show that runoff and rainfall has quite strong positive relationship in both stations. Number of rainy days at the M145 station shows positive relationship to the runoff, but it shows no relationship to runoff at M175 station. The NDVI shows low relationship to runoff at both stations (0.326 and 0.41). Temperature show negative relationship to runoff at both locations (-0.014 and -0.08).

In our proposed method to runoff prediction modeling, we introduce new attribute, that is the cluster identifier, to improve the performance of prediction. From the experimental results as shown in Table 1, we thus use only the runoff and rainfall attributes for clustering with k-means and select the appropriate *k* clusters based on the Silhouette Coefficient values. After that, we use predictor importance as a metric to select top important features and use Artificial Neural Network and Support Vector Regression to predict runoff. From our proposed method, we combine the results from these two models and generate a final prediction using linear regression. The runoff prediction results are presented in Table II.

The models to predict runoff at the M145 station have been built based on the k-means clustering process with k = 5 and Silhouette Coefficient = 0.64. The R values in all models is about 0.6. Our proposed method yields the best R value at 0.67. RMSE values of all models are in the range from 8 to10. The best model based on RMSE metric is our proposed model (RMSE = 8.34). Note that for R measure, the higher is the better. But for the RMSE metric, the lower is the better.

TABLE I: THE RESULTS OF CORRELATION COEFFICIENT ANALYSIS

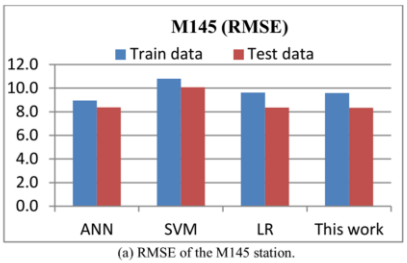
Attribute	M145	M173
Rainfall	0.670	0.54
Day of Rainy	0.502	0.08
NDVI	0.326	0.41
Temperature	-0.014	-0.08

TABLE II: RUNOFF PREDICTION PERFORMANCE AT M145 AND M173 STATIONS

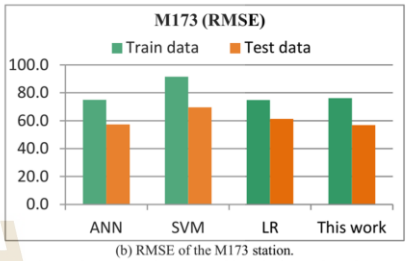
Station	Model	R	RMSE
M145	Artificial Neural Network	0.66	8.38
	Support Vector Regression	0.64	10.09
	Linear Regression	0.66	8.35
	This work	0.67	8.34
M173	Artificial Neural Network	0.58	57.41
	Support Vector Regression	0.42	69.61
	Linear Regression	0.58	61.33
	This work	0.59	56.94

The models to predict runoff at the M173 station have been also built based on the k-means clustering process with k = 5 and Silhouette Coefficient = 0.692. The R values in all models is ranging from 0.4 to almost 0.6. Our proposed method yields the highest R value at 0.59. RMSE values of all models are in the range from 56 to70. Our proposed method performs the best with the lowest RMSE value at 56.94.

The performances of all models based on RMSE and R values are also graphically compared and shown in Figures 4 and 5. In the comparison figures, we show RMSE and R values evaluated on both training data and test data. This is for assessing the over-fitting problem of the models. A model is called over-fitting if it performs well on the training data, but poorly perform on the separate set of test data. Currently, there is no agreement regarding how difference between train-test performance should be considered over-fitting. We, therefore set temporary train-test performance not exceeding 35% as non over-fitting. The models based on ANN, SVM, LR, and our proposed one are non over-fitting.

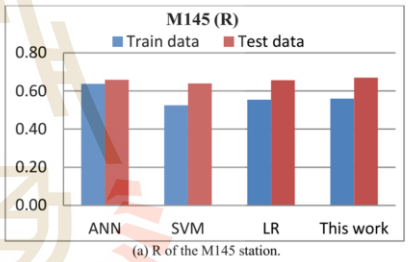


(a) RMSE of the M145 station.

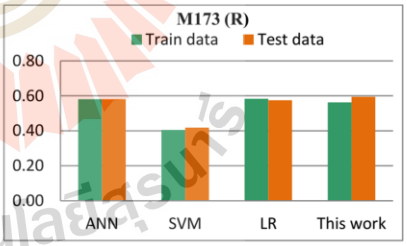


(b) RMSE of the M173 station.

Fig. 4. The RMSE values assessed in train data and test data.



(a) R of the M145 station.



(b) R of the M173 station

Fig. 5. The R values assessed in train data and test data.

It can be noticed from Fig. 5 that ANN model tends to fairly fit model to both train and test data. But our proposed method yields higher R values on the test data than the train data. This may be the reason for our approach being better than others on predicting runoff in the test dataset.

V. Conclusion

In this work, we propose a hybrid modeling technique to predict runoff from combining results from Artificial Neural Network and Support Vector Regression models. On the combination step we adjust weights from the two models by means of Linear Regression. It is the final output from linear regression that to be used as our runoff prediction. From the experimental results the proposed hybrid model shows good efficiency to predict runoff when it is compared with other single learner technique including Artificial Neural Network, Support Vector Regression, and Linear Regression. The comparison is based on the correlation coefficient and RMSE metrics using data from the two stations in the Mun basin of Thailand. Based on the two metrics, our proposed model show the best performance over other three techniques. We also perform experimentation on both training data and test data to check over-fitting problem. The results confirm that all models are non over-fit to the training data.

References

[1] M. A. Kaltech, "Rainfall-runoff modeling using artificial neural network modeling and understanding," *Caspian Journal of Environmental Sciences*, vol. 6, pp. 153-158, 2008.

[2] P. S. Kumar, T. V. Praveen, and M. A. Prasad, "Artificial neural network model for rainfall-runoff-a case study," *International Journal of Hybrid Information Technology*, vol. 9, no. 3, pp. 263-272, 2016.

[3] H. Chu, J. Wei, T. Li, and K. Jia, "Application of support vector regression for mid- and long-term runoff forecasting in 'yellow river headwater' Region," *Procedia Engineering*, vol. 154, pp. 1251–1257, 2016.

[4] F. Granata, R. Gargano, and G. D. Marinis, "Support vector regression for rainfall-runoff modeling in urban drainage: A comparison with the EPA's storm water management model," *Water*, vol. 8, no. 69, 2016.

[5] N. Sajikumar and B. S. Thandaveswara, "A non-linear rainfall–runoff model using an artificial neural network," *Journal of Hydrology*, vol. 216, pp.32–55, 1999.

[6] A. Agarwal and R. D. Singh, "Runoff modeling through back propagation artificial neural network with variable rainfall–runoff data," *Water Resources Management*, vol. 18, pp.285–300, 2004.

[7] A. Dorum, A. Yarar, M. F. Sevimli, and M. Onüçyildiz, "Modelling the rainfall–runoff data of susurluk basin," *Expert Systems with Applications*, vol. 37, no. 9, pp. 6587-6593, 2010.

[8] A. Farag and R. M Mohamed, *Regression Using Support Vector Machines: Basic Foundations*, Technical Report, University of Louisville, Louisville, 2004.

[9] J. A. Hartigan, A. Manchek, and A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100-108, 1979.

[10] A. Saltelli, "Making best use of model evaluations to compute sensitivity indices," *Computer Physics Communications*, vol. 145, no. 2, pp. 280–297, 2002.

R. Chanklan is currently a doctoral student with the School of computer engineering, Suranaree University of Technology, Thailand. She received his bachelor degree in Computer Engineering from Suranaree University of Technology, Thailand, in 2013, the master degree in computer engineering from Suranaree University of Technology, Thailand, in 2014. Her current research of interest includes classification, data mining, artificial intelligence.

K. Chaiyakhan is a lecturer with the Computer Engineering Department, Rajamangala University of Technology Isan, Nakhon Ratchasima, Thailand. She received her bachelor degree in computer engineering from Rajamangala University of Technology Thanyaburi in 1998, the master degree in computer engineering from King Mongkut's University of Technology Thonbuti in 2007 and doctoral degree in computer engineering from Suranaree University of Technology, Thailand in 2016. Her current research includes image classification and image clustering.

K. Kerdprasop is an associate professor and chair of the School of Computer Engineering, Suranaree University of Technology, Thailand. He received his bachelor degree in mathematics from Srinakarinwirot University, Thailand, in 1986, the master degree in computer science from the Prince of Songkla University, Thailand, in 1991 and doctoral degree in Computer Science from Nova Southeastern University, U.S.A., in 1999. His current research includes data mining, artificial intelligence, functional and logic programming languages, computational statistics.

N. Kerdprasop is an associate professor at the School of Computer Engineering, Suranaree University of Technology, Thailand. She received her bachelor degree in radiation techniques from Mahidol University, Thailand, in 1985, the master degree in computer science from the Prince of Songkla University, Thailand, in 1991 and doctoral degree in computer science from Nova Southeastern University, U.S.A, in 1999. She is a member of ACM and IEEE Computer Society. Her research of interest includes knowledge discovery in databases, artificial intelligence, and intelligent databases.

International Journal of Machine Learning and Computing, Vol. 8, No. 1, February 2018

Runoff Prediction with a Combined Artificial Neural Network and Support Vector Regression

Ratiporn Chanklan, Nuntawut Kaoungku, Keerachart Suksut, Kittisak Kerdprasop, and Nittaya Kerdprasop

***Abstract*—Water is an important part of our daily lives: food, manufacture, agriculture, etc. When water is not enough for all population, it leads to many undesirable impacts including drought, famine and death. The solution to this problem is the good management of water resources. The management of water resources is planning and designing of projects related to water. The runoff prediction is one major part of planning. It is a complex process and it also needs an adequate modeling technique for accurate prediction. Therefore, we propose to use combined algorithms to improve prediction performance. Our combination includes the two powerful methods: Artificial Neural Network (ANN) and Support Vector Regression (SVR). The root mean square error (RMSE) and the correlation coefficient (R) are two criteria that we use to evaluate the model performance regarding the comparison between actual runoff and the prediction made by our model. We also compare performance of our model against the other algorithms: Linear Regression, ANN, and Support Vector Machines. The comparison results show that our proposed method shows the best performance and the combined model is also quite accurate on predicting the peak runoff values during heavy rain season.**

***Index Terms*—Runoff prediction, artificial neural network, support vector regression, Mun Basin.**

I. Introduction

Currently, people are experiencing both direct and indirect impacts from droughts such that it results in water usage restrictions, more frequent forest fires, reduced crop yields, loss of livestock, and many others [1]. On the contrary, people also experience floods in larger areas year by year. The efficient management of water resources is one way to protect the two water problems: flood and droughts. A knowledge to make accurate runoff prediction can obviously help planners and water management policy makers to know in advance the water volume as either enough or not enough for the demand of people use and also can improve efficiency on flood and drought control.

Artificial neural network (ANN) shows good performance on predicting outcomes from many complex processes and recognizing patterns [2]. It has also been used to create computational model to predict several kinds of outcomes in the hydrology research, such as stream flow forecasting, rainfall-runoff modeling, water quality and management modeling [3], [4]. Besides ANN, Support vector regression (SVR) is another accurate technique that has been recently applied to predict runoff and [5], [6]. SVR has been known to show high performance on learning solution in complex problems [7].

In this work, we propose a new method to predict monthly runoff. In our method, the better performance can be achieved through the combination of ANN and SVR. The results turn out that our model can predict both low and high runoff values. A literature normally uses statistical values such as mean absolute percentage error (MAPE), coefficient of determination (R^2), correlation coefficient (R), and root mean squared error (RMSE) to compare performance of a prediction algorithm. In our work, we adopt two metrics: R and RMSE.

II. Background Theories

A. Artificial Neural Network

ANN has been developed base on operations of the human brain. It is the most widely used tool in hydrology. The network has three layers: input layer, hidden layer and output layer. The input nodes are name node in input layer. The number of attribute in data is equal to the number of input nodes. The hidden nodes are name node in hidden layer, the number of nodes is defined by a user, and this layer can be more than one layer. The output nodes are name node in output layer, the number of nodes is equal to the number of target on data. The network are connected between the nodes with line, and each line has weight. ANN learning is to find proper weight on each line in the network. The proper weight is the one that can best separating training data into the corrected target groups. There exist several architectures of ANN. In this work, we use feed-forward neural networks. The transfer function in the hidden layer is a linear function.

B. Support Vector Regression

SVR works by transforming the input data into a high-dimensional feature space by linear or nonlinear mapping. SVR is the most popular application form of Support Vector Machine (SVM). The intuitive idea of SVR on predicting future data is illustrated in Fig. 1.

The goal of SVR is to find a function $f(x)$ that has at most ε deviation from the actual target value y_i for all the training data [8]. This linear function f is shown in equation 1. A training data is a set of input-target pairs, $\{(x_1, y_1), \ldots, (x_i, y_i)\}$

Manuscript received September 19, 2017; revised January 20, 2018. This work was supported by grant from Suranaree University of Technology through the funding of Knowledge and Data Engineering Research Units.

The authors are with the School of Computer Engineering, Suranaree University of Technology (SUT), 111 University Avenue, Muang, Nakhon Ratchasima 30000, Thailand. (corresponding author: R. Chanklan; Tel: +66994696164; e-mail: arc_angle@hotmail.com, nuntawut@sut.ac.th, mikaiterng@gmail.com, kerdpras@sut.ac.th, nittaya@sut.ac.th).

doi: 10.18178/ijmlc.2018.8.1.660

$\subset X \times R$.

$$f(x) = \langle w, x \rangle + b \quad (1)$$

When a function $f(x)$ is hyperplane, the size of ε is margin, the symbol $\langle \cdot, \cdot \rangle$ is the dot product in X, $b \in R$ and $w \in X$. The hyperplane has small margin in which the SVR has to find it. This small margin tries to keep all the data lying inside as much as possible. The margin can be calculated as in equation 2.

$$minimize \ \frac{1}{2}\|w\|^2 \ subject\ to \begin{cases} y_i - \langle w, x_i \rangle + b \leq \varepsilon \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon \end{cases} \quad (2)$$

Through equation 2, it is implicitly assumed that this is a function f that approximates all the pairs (x_i, y_i) with precision. If it is not possible to keep all data inside the margin, the slack variables ξ_i, ξ_i^* can be introduced to solve the problem. This can be stated according to the equation 3.

$$minimize \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)$$
$$subject\ to \begin{cases} y_i - \langle w, x_i \rangle + b \leq \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \end{cases} \quad (3)$$

The constant $C > 0$ directs the choice between the flatness of f and the amount up to which deviations larger than ε are tolerated.

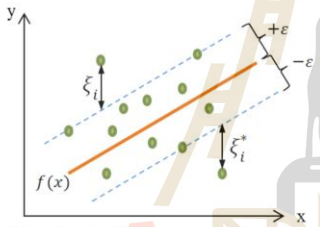


Fig. 1. Example of linear support vector regression.

C. Correlation Coefficient

The correlation coefficient is denoted by R and its numerical value is between -1 and 1. The plus sign means the correlation of the two variables (x and y) moves in the same direction, whereas the minus sign infers opposite direction. The magnitude expresses the strength of the relationship; the higher is the stronger regardless of the sign. No relationship is the magnitude that closes to 0. The correlation coefficient can be computed using equation 4.

$$R = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}} \quad (4)$$

where x_i is the value of variable x or predicted value (when i = 1, 2, … , n), y_i is the value of variable y or actual value, and n is the total number of samples.

D. Root Mean Squared Error

The Root Mean Squared Error is denoted by RMSE and used to measure performance of the model. The RMSE can be calculated as in equation 5.

$$\text{RMSE} = \sqrt{\left(\frac{\sum(T_i - O_i)^2}{N}\right)} \quad (5)$$

where N is the number of all data, O_i is the predicted value by the model, and T_i is the actual value, i = 1,2, …, N.

III. Materials and Methods

Our study area is Mun Basin (Fig. 2), the largest basin in the North-eastern region of Thailand.To build a predictive model, we use temperature data from the National Statistical Office (http://www.nso.go.th), monthly rainfall, runoff and the number of rainy days data from the Meteorological Department (http://www.hydro-4.com) and Normalized Difference Vegetation Index (NDVI), from the NOAA STAR (http://www.star.nesdis.noaa.gov). This research use RStudio as the analysis tools in our experiments. The runoff data are from two sources: M145 and M173 station.

Fig. 2. The study area: Mun Basin, Thailand.

M145 station locates at Ban Wang Takhian, Amphur Pak Chong, in Nakhon Ratchasima Province, Thailand. We use the 18-year data during 1998 to 2015. We spilt the 13-year data (1998-2015) to be training data and the remaining 5-year data (2011-2015) are for testing the model performance.

M.173 station locates at Ban Non Sa-at, Amphur Chokchai, in Nakhon Ratchasima Province, Thailand. The training data is a ten-year period (2002-2011) and the test data is the four-year period (2012-2015).

Then, we use rainfall, runoff, the number of rainy days, temperature and NDVI with the lagging time 1-month and 2-month. These data are input into two learning algorithms: ANN and SVR. Then, the average rainfall is lagged 1-month (average rainfall$_{t-1}$) to select the best subset of data from the initial training set for appropriate algorithm. The average rainfall can be calculated as in equation 6.

$$\text{average rainfall}_{t-1} = \frac{\sum_{i=1}^{N} a}{n} \quad (6)$$

Where a is the average monthly rainfall in lagged 1-month over the training years, N is the number of training data (when i=1, 2, .., N), and n is the number of years from training data.

Our combined prediction model has a flow as shown in Fig. 3. The selection of either ANN or SVR model is based on the amount of rain. If the rainfall in lagged 1-month (rainfall$_{t-1}$) has value less than average rainfall$_{t-1}$, then apply the SVR

model because based on our observation this kind of model is good at prediction on normal or drought situation. But if the rainfall in lagged 1-month ($rainfall_{t-1}$), has a value higher than the average $rainfall_{t-1}$, then apply the ANN model that is good in predicting the near-flooding situation.

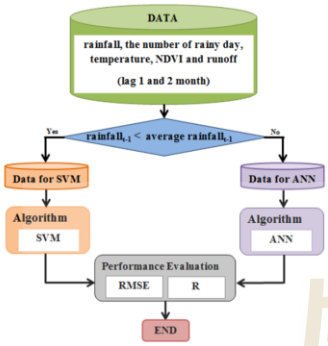


Fig. 3. The modeling process for runoff prediction.

The final step is the performance evaluation using the two statistical values: Correlation Coefficient (R) and Root Mean Squared Error (RMSE). The R and RMSE are computed from experimentation with the separate test data.

IV. Experimental Results

TABLE I: Runoff Prediction Performance from ANN Model

Station	ANN topology	R	RMSE
M145	**10-1-1**	**0.67**	**8.17**
	10-2-1	0.59	8.79
	10-3-1	0.66	8.20
	10-4-1	0.43	10.66
	10-5-1	0.51	9.68
	10-6-1	0.66	8.28
	10-7-1	0.62	8.78
	10-8-1	0.48	10.48
	10-9-1	0.64	8.53
	10-10-1	0.59	9.39
M173	10-1-1	0.51	65.40
	10-2-1	0.48	66.13
	10-3-1	0.47	87.51
	10-4-1	0.43	68.66
	10-5-1	0.37	114.24
	10-6-1	**0.66**	**58.73**
	10-7-1	0.55	79.42
	10-8-1	0.52	80.62
	10-9-1	0.47	97.30
	10-10-1	0.65	69.74

The train and test data have ten input variables (rainfall, runoff, the number of rainy day, temperature and NDVI with lagged 1-month and 2-month). We use the initial training set input into ANN and SVR to create a combined model and then test this model with the test data. The results of applying ANN and SVR alone are shown in Tables I and II, respectively. The result from our proposed combined model is presented in Table III.

TABLE II: Runoff Prediction Performance from SVR Model

Station	Kernel	R	RMSE
M145	linear	0.58	10.17
	sigmoid	0.14	29.20
	polynomial	0.45	10.37
	radial basis function	**0.67**	**8.98**
M173	**linear**	**0.54**	**61.66**
	sigmoid	0.04	135.05
	polynomial	0.36	69.62
	radial basis function	**0.55**	**62.84**

TABLE III: Runoff Prediction Performance from Proposed Model

Station	ANN topology	Kernel	R	RMSE
M145	10-6-1	linear	0.68	8.09
	10-6-1	sigmoid	0.63	8.61
	10-6-1	polynomial	0.68	8.11
	10-6-1	**RBF**	**0.69**	**8.01**
M173	10-8-1	linear	0.71	52.09
	10-8-1	sigmoid	0.71	52.42
	10-8-1	polynomial	0.71	52.38
	10-8-1	**RBF**	**0.71**	**51.99**

Our design of ANN architecture is based on the suggestion "the optimal size of the hidden layer is usually between the size of the input and size of the output layers" [9]. We thus set the hidden layer to be in the range 1-10 as shown in Table I. The results reveal that at the M145 station the best network topology is 10-1-1 (R=0.67, RMSE=8.17). At the M173 station, the best topology is 10-6-1 (R=0.66, RMSE=58.73).

On building the SVR model, we set parameters as follows: cost=1, gamma=1/(data dimension)), degree=3 and coefficients of the support vector =0 for each kernel. The best runoff prediction at the M145 station is the radial basis kernel (R=0.67, RMSE=8.98). At the station M173, the linear kernel (R=0.54, RMSE=61.66) performs almost as good as the radial basis function (R=0.55, RMSE=62.84).

When we combine the power of both ANN and SVR algorithms, it shows clearly good performance. In our proposed method, we use either ANN or SVR depending on the amount of rain in each data instance. We therefore report both the ANN architecture and the kernel function in Table III. The best results are the one highlighted in red bold font. We also show graphical comparisons of ANN and SVR methods for the station M145 (Fig. 4) and the station M173 (Fig. 5). The actual versus predicted runoff values using our proposed combination ANN-SVR method in both stations is presented in Fig. 6.

Figs. 4 and 5 clearly reveal strength of the ANN and SVR models on predicting runoff values. It can be noticed that the ANN performs well on some peak values, but perform poorly on some low runoff values at the M145 and M173. On the contrary, the SVR is good at predicting runoff amount during the water shortage situation. But when runoff is excessive, the SVR model performs quite poor at both the M145 and M173 stations.

It is actually based on these experimental observation that we thus design the proposed model combining the strength from each model using raining amount as a decision criterion.

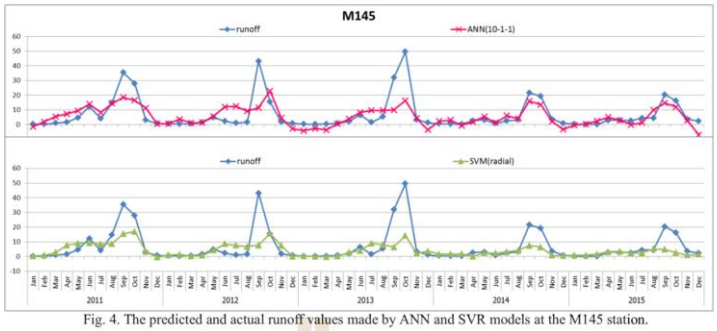


Fig. 4. The predicted and actual runoff values made by ANN and SVR models at the M145 station.

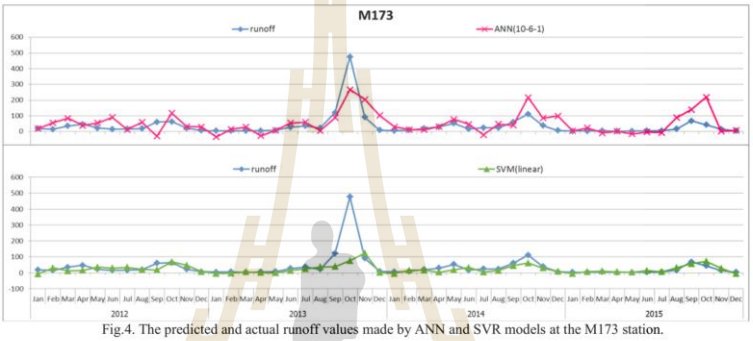


Fig.4. The predicted and actual runoff values made by ANN and SVR models at the M173 station.

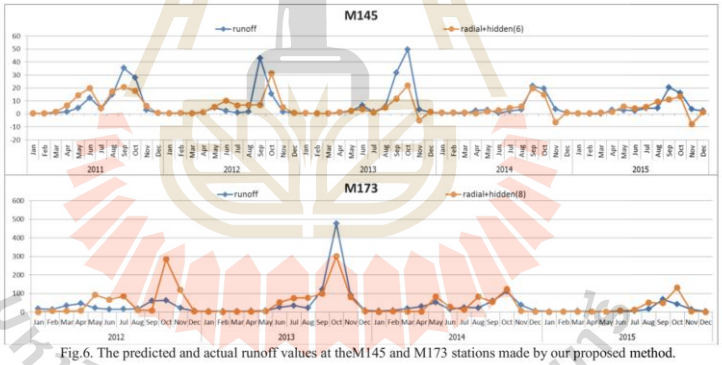


Fig.6. The predicted and actual runoff values at theM145 and M173 stations made by our proposed method.

Our proposed method use average monthly rainfall to select a subset of data for each algorithm with the intuitive idea that rainfall causes runoff. It turns out that our model performs the best in both shortage and excessive rainfall. The proposed method can also find a good architecture for hidden node layer of the ANN.

In addition, we also create model base on Linear Regression (LR). The prediction results are however poor (at the M145 station, R=0.67, RMSE=8.17; at the M173 station, R=0.51, RMSE=65.46). Therefore, we can conclude from these experimental results that our proposed method is the best combined model to predict runoff (at the M145 station, R=0.69, RMSE=8.01; at the M173 station, R=0.71, RMSE=51.99).

V. Conclusion

In this work, we propose a runoff prediction method that combine the advantages of ANN and SVR to predict runoff (rainfall,). The strength of ANN is its high accuracy on predicting runoff when the amount of rainfall is high. The strength of SVR is on the contrary that it is good at the low

amount of rainfall. We thus combine these observed strengths.

On the combination step, we use average accumulative rainfall with lagged 1-month (average rainfall$_{t-1}$) to select a subset of data from the initial training set, then split the initial training set for ANN and SVR to create models. To use our combined method to predict runoff, the decision criteria for choosing either ANN or SVR model is the amount of rainfall on the previous month. If this amount is higher than our threshold value, choose the ANN model; otherwise, choose the SVR model.

From the experimental results, the proposed method shows good efficiency to predict runoff when it is compared against other technique including ANN, SVM and LR. These comparisons are based on R and RMSE metrics using test data from the two stations in the Mun Basin of Thailand.

REFERENCES

[1] S. Subak, "Climate change adaptation in the U.K. water industry: Managers' perceptions of past variability and future scenarios," *Water Resources Management*, vol. 14, pp.137–156, January 2000.

[2] T. A. Sezin and P. A. Johnson., "Precipitation-Runoff Modeling using Artificial Neural Networks and conceptual models," *Journal of Hydrologic Engineering*, vol. 5, no. 2, pp. 156-161, April 2000.

[3] A. W. Minns and M. J. Hall, "Artificial neural networks as rainfall-runoff models," *Hydrological Sciences Journal*, vol. 41, no. 3, pp. 399-417, June 1996.

[4] J. Morshed and J. J. Kaluarachchi, "Application of artificial neural network and generic algorithm in flow and transport simulations," *Advances in Water Resouces,* vol. 22, no. 2, pp. 145-158, 1998.

[5] S. M. V Choubey, S. Pandey, & Shukla, J. "An Efficient Approach of Support Vector Machine for Runoff Forecasting," *International Journal of Scientific & Engineering Research,* vol. 5, no. 3, pp. 158-166, March 2014.

[6] C. L. Wu, K. W. Chau, and Y. S. Li., "River stage prediction based on a distributed support vector regression," *Journal of hydrology,* vol. 358, no. 1-2, pp. 96-111, 2008.

[7] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing,* vol. 1, no. 3, pp. 199–222, 2004.

[8] F. Granata, R. Gargano and Giovanni de Marinis, "Support vector regression for rainfall-runoff modeling in urban drainage: A comparison with the EPA's storm water management model," *Water*, vol.8, no. 3, p. 69, 2016.

[9] *Introduction to Neural Networks with Java*, Heaton Research, Heaton Research, Inc., St. Louis, 2008.

R. Chanklan is currently a doctoral student with the School of Computer Engineering, Suranaree University of Technology (SUT), Thailand. She received her bachelor degree in computer engineering from SUT in 2013, master degree in computer engineering from SUT in 2014. Her current research of interest includes data classification, data mining application, and artificial intelligence.

N. Kaoungku is currently a lecturer at School of Computer Engineering, SUT, Thailand. He received his doctoral degree, master dgree, and bachelor degree in computer engineering from SUT, in 2015, 2013, and 2012, respectively. His current research includes data mining, knowledge engineering, and semantic web.

K. Suksut is currently a doctoral student with the School of Computer Engineering, SUT, Thailand. He received his bachelor degree in computer engineering from SUT in 2011, master degree in computer engineering from SUT in 2013. His current research of interest includes data mining, genetic algorithm, and imbalanced data classification.

K. Kerdprasop is an associate professor and chair of the School of Computer Engineering, SUT. He received his bachelor degree in mathematics from Srinakarinwirot University, Thailand, in 1986, master degree in computer science from the Prince of Songkla University, Thailand, in 1991, and doctoral degree in computer science from Nova Southeastern University, U.S.A., in 1999. His current research includes data mining, artificial intelligence, and computational statistics.

N. Kerdprasop is an associate professor at the School of Computer Engineering, SUT. She received her bachelor degree in radiation techniques from Mahidol University, Thailand, in 1985, master degree in computer science from the Prince of Songkla University, Thailand, in 1991, and doctoral degree in Computer Science from Nova Southeastern University, U.S.A, in 1999. Her research of interest includes knowledge discovery in databases, artificial intelligence, and intelligent databases.

International Journal of Machine Learning and Computing, Vol. 8, No. 1, February 2018

The Silhouette Width Criterion for Clustering and Association Mining to Select Image Features

Nuntawut Kaoungku, Keerachart Suksut, Ratiporn Chanklan, Kittisak Kerdprasop, and Nittaya Kerdprasop

***Abstract*—Image data are normally unstructured and high dimensional due to the photography technology advancement such that an image can be taken at a wide range of resolution levels. To overcome such problem, data miners may consider selecting only a minimal set of features that are really important for classifying their images. Feature selection is a popular method for reducing dimensions in data. However, most feature selection algorithms return results in form of score for each feature. It is still difficult for data miners to choose features based on such scoring scheme because they may not know which score range is the best for their data classification at hand. Therefore, in this research, we aim to assist data miners and novice data analysts on solving dimensionality problem by finding for them the best optimal set of features, instead of just reporting the scores of all features and leaving the selection step to be the burden of miners. We select optimal set of features by firstly apply clustering technique to group similar features based on their scores. We thus propose the silhouette width criterion for selecting the optimal number of clusters during the cluster analysis step. After that we perform association mining to analyze relationships that may exist among different subsets of features toward the target attribute. Our method finally reports user the best subset of features to be potentially used further for data classification. We demonstrate performance of our proposed method on the satellite forest image data in Japan.**

***Index Terms*—Image data, feature selection, clustering, silhouette criterion, forrest type classification.**

I. Introduction

With the rapid development of current electronic devices such as sensors and cameras, the outputs from these devices are of high quality and also high dimensions. Unfortunately, high dimensionality is still an unsolvable issue for many existing data mining and machine learning algorithms. Data with overwhelming attributes or dimensions can be a major cause of low computational performance. It can be even worse when such data may cause a creation of classifying model with low predictive accuracy due to the search for discriminative set of features is obscured by so many irrelevant features. Most classification algorithms are not designed to efficiently handle such high dimensionality problem.

Therefore, the numerous feature selection techniques have been proposed as a pre-classification step for solving the high dimensionality problem. Several research teams introduce many ways to reduce the number of features. The reduced set of features has been proven experimentally increasing the performance of learning process and also being able to build an accurate classification model. Generally, feature selection techniques can be divided into three classes [1]. The first class is called filter method, such as CfsSubsetEval [2], Information Gain, and Chi-Square [3]. The second class is the wrapper method [4], [5]. The filter method introduces some form of scoring computation without actually building a model, whereas the wrapper approach scoring the selected set of features by observing the error made by the classifying model. The last class is called the embedded method; it combines the advantages from both the filter and wrapper methods [5].

Xie *et al.* [6] proposed the association rule mining technique to calculate weight for find the optimal features that are closely correlative with the class attribute, but the proposed technique is quite complex and performance test with cross validation. Nuntawut *et al.* [7] proposed the filter method for feature selection based on association rule mining such that the specific set of association rules that the rules' consequence is the target class. But this feature selection algorithm does not work automatically because human is the one who select the features one by one based on the feature scores reported from the algorithm. Therefore, Nuntawut *et al.* [8] improved the algorithm by proposing clustering technique to cluster the feature scores to assist users on finding an appropriate groups of features. The clustering process is supposed to be automatic in the sense that the number of clusters should be judged by the process itself. However, the clustering algorithm is still semi-automatic in the sense that users must specify the suitable number of feature clusters.

This research, thus, aims at extending the previous work of Nuntawut *et al.* [7], [8] by proposing a silhouette width criterion for automatic setting of initial cluster numbers. We also add confidence criteria into feature selection based on association rule mining technique to increase performance. Experimental results confirm the efficacy of our proposed method that can extract only relevant set of image features from ASTER satellite resulting in better recognition for each forest type.

II. Materials and Methods

A. Feature Selection Based on Association Rule Mining

Association rule mining is finding the frequent patterns in

Manuscript received September 20, 2017; revised January 10, 2017. This work was supported by grant from Suranaree University of Technology through the funding of Knowledge Engineering Research Unit.

The authors are with the School of Computer Engineering, Suranaree University of Technology (SUT), Nakhon Ratchasima 30000, Thailand (corresponding author: N. Kaoungku; Tel: +66872155059; e-mail: nuntawut@sut.ac.th, mikaiterng@gmail.com, arc_angle@hotmail.com, kittisakThailand@gmail.com, nittaya@sut.ac.th).

doi: 10.18178/ijmlc.2018.8.1.665

database and present them in the form of association rules [9]. Generally, there can be so many possible association rules from this technique. Therefore, some constraints are necessary for reducing such exponential growth. There are two popular criteria: support and confidence. Support is the frequency of the occurring event, as shown in (1). Confidence is the proportion of frequency of co-occurring events to the frequency of antecedent event, as shown in (2).

$$\text{Support}, \text{Supp}(X \to Y) = P(X \wedge Y) \quad (1)$$

$$\text{Confidence}, \text{Conf}(X \to Y) = \frac{\text{Supp}(X \to Y)}{\text{Supp}(X)} \quad (2)$$

This technique had been successfully applied to multiple disciplines such as marketing to increase sales. Nuntawut *et al*. [7] applied this technique to find optimal feature set from high dimensional dataset by finding association rules that the target class appears in the consequence of the rule. Then, consider the features or attributes that are most influencing the target class. The algorithm consists of 4 steps:

Step 1: define minimum frequency threshold, support, and confidence. Find frequent patterns and then generate association rules based on the Apriori algorithm [10].

Step 2: select only association rules that their consequence is target class.

Step 3: count features that appear on association rules.

Step 4: calculate frequent features in percentage, as in (3). Then, remove any feature having percentage of frequency appearance in the set of association rules lower than the specified minimum frequency threshold.

$$\text{FrequentFeature} = \frac{\text{AppearFrequency}}{\#\text{Rules}} \times 100 \quad (3)$$

B. k-Means Clustering

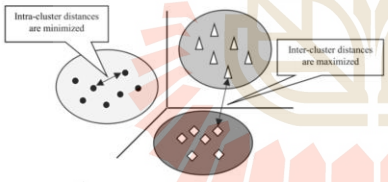


Fig. 1. Forming three clusters with minimized intra-cluster distance but maximized the inter-cluster distance.

k-Means algorithm is unsupervised learning method to form data into clusters based on data similarity regardless of the target class information. Fig. 1 depicts the idea of clustering such that distance between data in the same cluster (intra-cluster) is low, whereas the distance between data in one cluster to data in another cluster (inter-cluster) is high [11], [12]. The k-means algorithm can be explained by the following 4 steps:

Step 1: define the number of clusters (K) and randomly pick k instances as the initial cluster centroids.

Step 2: assign all data points to the closest centroid by measuring the distance such as the Euclidean distance.

Step 3: re-compute the centroid of each cluster by calculating mean value of all the data points in the cluster.

Step 4: repeat steps 2 and 3 until the centroid does not change.

C. Silhouette Coefficient

The shortcoming of k-means clustering is the appropriate choice of *k*, which is the number of clusters. Silhouette coefficient is a popular measure for considering such parameter. The silhouette coefficient can be computed by using average distance between data points in the same cluster compared against average distances between data points in other clusters. Fig. 2 shows main concept of the silhouette coefficient to calculate the silhouette average of all cluster.

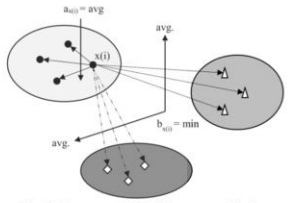


Fig. 2. Concept of the silhouette coefficient.

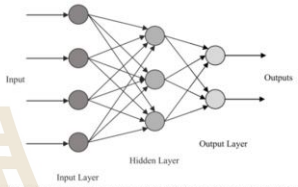


Fig. 3. The architecture of artificial neural networks.

Define K to be cluster composing of data $x(i)$ and $a_{x(i)}$ is average distance between $x_{(i)}$ to every data point in the cluster K. The notation b_x is minimum average distance between $x(i)$ and every data point in other clusters that are not a member in K. The calculation [13] of the silhouette coefficient of $x(i)$, the silhouette average of each cluster, and the silhouette average of all cluster can be shown as in (4), (5), and (6), respectively.

$$S_{x(i)} = \frac{b_{x(i)} - a_{x(i)}}{\max(a_{x(i)}, b_{x(i)})} \quad (4)$$

where

$x(i)$ = data point in the cluster, i = 1, 2, 3, …, n,

$a_{x(i)}$ = average distance between x_i and every data point in the same cluster, and

$b_{x(i)}$ = minimum average distance between x_i and every data point in other clusters.

$$S_k = \frac{1}{n}\sum_{i=1}^{n} S_{x(i)} \quad (5)$$

where k = number of clusters, and

n = number of data points in the same cluster.

$$S_{avg} = \frac{1}{m}\sum_{k=1}^{m} S_k \quad (6)$$

where m is number of all clusters.

D. Artificial Neural Networks

Artificial neural networks is a simulation of human brain with computer program that can self-adjusting from learning the input values. The remarkable feature of this technique is that it consists of many nodes in the hidden layer in which parallel connections are effective for data classification [14]. Fig. 3 shows general architecture of artificial neural networks consisting of nodes and edges between nodes. Form the figure, the network can be partitioned based on node layout into 3 layers. The first layer is input layer; the second is hidden layer (this layer can have more than 1 layer), and the final layer is output layer.

III. Proposed Work

In this section, we present the proposed process of silhouette width criterion consideration for automatic clustering of feature sets with the main focus of finding optimal feature to be discovered by association rule mining. The idea is that we use the silhouette coefficient to find the appropriate number of clusters for clustering the feature scores from feature selection obtained from the association rule mining. The objectives are to increase the predictive accuracy and to reduce the data dimensions of forest type dataset.

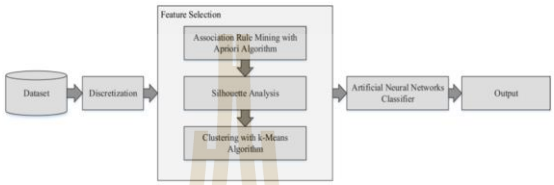


Fig. 4. The concept of feature selection based on association rule mining.

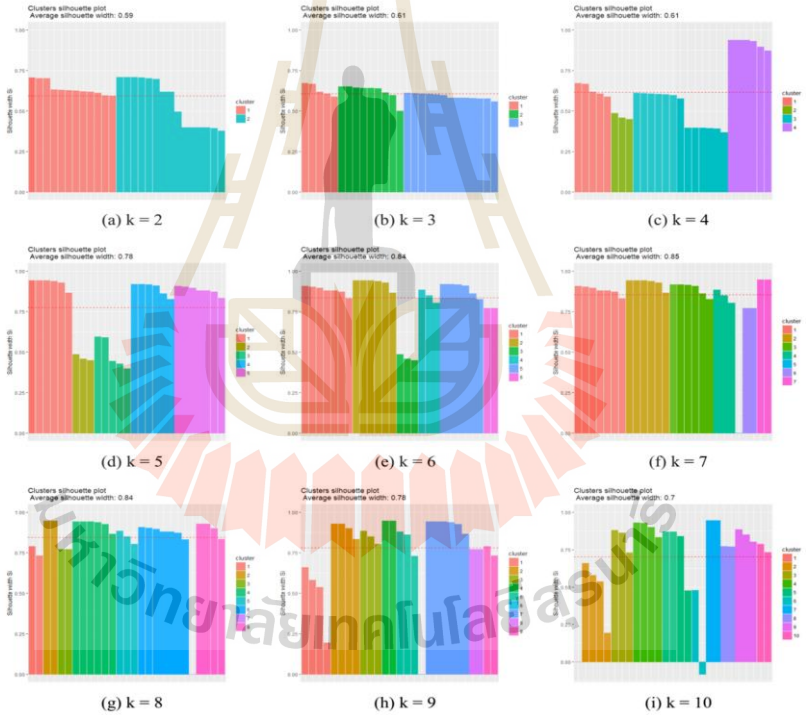


Fig. 5. Comparative graphs showing average silhouette widths of different cluster numbers.

Fig 4 shows the steps of the proposed process, which consists of three phases: phase 1, read the dataset from file or database and then perform discretization with chi-square algorithm if the data type is numeric. This discretization step is necessary for association rule mining that can handle only categorical values. Phase 2 is the feature selection method that

consists of three steps:

Step 1: find frequent patterns and generate from these patterns association rules in a format "*IF condition THEN consequence*." This step is done through Apriori algorithm with initial 2 thresholds: support and confidence. We also constrain the algorithm to generate rules with target class appeared in the consequence part of the rule. The result from this step is a set of features with scores computed as feature frequency in association rules and average confidence of each feature.

Step 2: cluster the features based on their scores with different number of clusters. For each number of clusters, calculate the silhouette coefficient to find the best number of clusters. The higher silhouette coefficient means the better formation of clusters. The result in this step is the optimal parameter *k* to be used in the k-means clustering on step 3.

Step 3: perform k-means clustering with the initial number of cluster (k) according to the recommend value from step 2. We then select a set of features from a cluster showing mean confidence higher than other clusters. The result in this step is optimal feature set to be used for classification.

Finally, phase 3 is the building of classifier using artificial neural networks.

TABLE I: COMPARATIVE RESULTS OF CLASSIFICATION ACCURACY, NUMBER OF FEATURES, AND AVERAGE SILHOUETTE WIDTH

Number of Clusters (k)	Accuracy	Number of Features	Average Silhouette Width
2	80.31%	20	0.59
3	81.54%	14	0.61
4	82.77%	11	0.61
5	82.77%	11	0.78
6	82.77%	11	0.84
7	**84.62%**	10	**0.85**
8	80.00%	7	0.84
9	79.69%	5	0.78
10	78.46%	**3**	0.70

IV. EXPERIMENTAL RESULTS

To test performance of the proposed method of feature selection based on the silhouette width criterion for clustering relevant featured discovered by association rule mining, we use the forest type with high-resolution imaging from ASTER satellite that has been publicly available at the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml). The data are divided into training dataset (198 instances) and test dataset (325 instances). We initialize Apriori algorithm to discover feature sets with support = 0.1 and confidence = 0.1. We experiment with number of clusters (k) between 2 to 10 clusters.

Table I shows comparative results of classification accuracy, number of features, and average silhouette. Fig. 5 shows comparative average silhouette widths of different clusters. It can be seen that when the number of cluster = 7, the average silhouette coefficient is maximized (0.85). At this maximum coefficient value, the predictive accuracy is as high as 84.62%. Moreover, the number of features can be reduced from 26 down to 10. Characteristic of number of features according to the changing number of clusters has been captured and shown in Fig. 6.

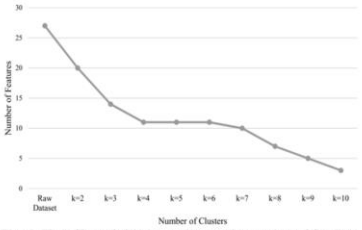


Fig. 6. The effect of cluster numbers to the number of features.

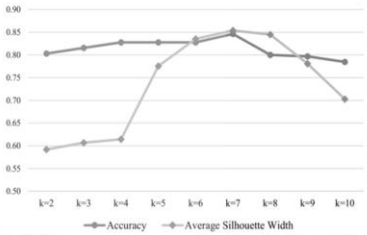


Fig. 7. The accuracy and average silhouette width characteristics.

Fig. 7 shows the comparisons of predictive accuracy and average silhouette width as the number of clusters has been varied from 2 to 10. It can be seen that the average silhouette has direct and positive impact to the classification accuracy. This is observed from the graph that when the silhouette width is low, the accuracy is also low. When the silhouette width is high, the accuracy is high as well.

V. CONCLUSION

This research aims at studying a novel method to use silhouette width criterion for cluster analysis with the main focus of finding optimal feature set to be used for building classification model. Set of features are discovered with the association rule mining method. The proposed feature subset selection method is to be applied on classifying data with high dimensionality such as satellite image data. Our proposed method works with three main phases. Firstly, find and score relevant set of features based on association rule mining technique. Secondly, apply silhouette width criterion to find optimal parameter *k* for the next phase of feature clustering and add average confidence threshold of each cluster to feature score for increasing clustering performance. From the experimental results, we can conclude that the proposed method can select a discriminative set of features resulting in a highly accurate classification model.

REFERENCES

[1] M. Hilario and A. Kalousis, "Approaches to dimensionality reduction in proteomic biomarker studies," *Briefings in Bioinformatics*, vol. 9, no. 2, pp. 102-118, 2008.

[2] Z. N. Hamilton, "Correlation-based feature subset selection for machine learning," Ph.D. Dissertation, Department of Computer Science, Waikato University, 1998.

[3] X. Jin, A. Xu, R. Bie, and P. Guo, "Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles," in *Proc. International Workshop on Data Mining for Biomedical Applications*, 2006, pp. 106-115.

[4] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of Machine Learning Research*, pp. 1205-1224, 2004.

[5] Y. Saeys, P. Rouzé, and Y. Van de Peer, "In search of the small ones: Improved prediction of short exons in vertebrates, plants, fungi and protists," *Bioinformatics*, vol. 23, no. 4, pp. 414-420, 2007.

[6] J. Xie, J. Wu, and Q. Qian, "Feature selection algorithm based on association rules mining method," *In Proc. the ACIS International Conference on Computer and Information Science*, pp. 357-3622009.

[7] N. Kaoungku, K. Suksut, R. Chanklan, K. Kerdprasop, and N. Kerdprasop, "Data classification based on feature selection with association rule mining," *The 25th Int. MultiConference of Engineers and Computer Scientists (IMECS2017)*, Hong Kong, China, 15-17 March 2017, pp. 321-326.

[8] N. Kaoungku, K. Kerdprasop, and N. Kerdprasop, "A method to clustering the feature ranking on data classification using an ensemble feature selection," *International Journal of Future Computer and Communication*, vol. 6, no. 3, September, pp. 81-85, 2017.

[9] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *ACM SIGMOD Record*, vol. 22, no. 2, pp. 207-216, 1993.

[10] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. The 20th Int. Conf. on Very Large Data Bases (VLDB)*, 1994, vol. 1215, pp. 487-499.

[11] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281-297.

[12] M. R. Anderberg, *Cluster Analysis for Applications*, Academic Press, 1973.

[13] S. Aranganayagi and K. Thangavel, "Clustering categorical data using silhouette coefficient as a relocating measure," in *Proc. IEEE Int. Conf. on Computational Intelligence and Multimedia Applications*, 2007, vol. 2, pp. 13-17.

[14] B. Yegnanarayana, *Artificial Neural Networks*, PHI Learning Pvt. Ltd., 2009.

N. Kaoungku is currently a lecturer at School of Computer Engineering, Suranaree University of Technology (SUT), Thailand. He received his doctoral degree, master degree, and bachelor degree in computer engineering from SUT, in 2015, 2013, and 2012, respectively. His current research includes data mining, knowledge engineering, and semantic web.

K. Suksut is currently a doctoral student with the School of Computer Engineering, SUT, Thailand. He received his bachelor degree in computer engineering from SUT in 2011, master degree in computer engineering from SUT in 2013. His current research of interest includes data mining, genetic algorithm, and imbalanced data classification.

R. Chanklan is currently a doctoral student with the School of Computer Engineering, SUT. She received her bachelor degree in computer engineering from SUT in 2013, master degree in Computer Engineering from SUT in 2014. Her current research of interest is data mining and artificial intelligence.

K. Kerdprasop is an associate professor and chair of Computer Engineering School, SUT. He received bachelor degree in mathematics from Srinakarinwirot University, Thailand, in 1986, MS in Computer Science from the Prince of Songkla University, in 1991, and PhD in computer science from Nova Southeastern University, U.S.A., in 1999.

N. Kerdprasop is an associate professor at the School of Computer Engineering, SUT. She received her bachelor degree in radiation techniques from Mahidol University, Thailand, in 1985, MS in Computer Science from the Prince of Songkla University in 1991, and PhD in computer science from Nova Southeastern University, U.S.A, in 1999.

ประวัติผู้เขียน

นางสาวรติพร จันทร์กลั่น เกิดเมื่อวันที่ 30 มีนาคม พ.ศ. 2533 ที่ อำเภอเมือง จังหวัดเพชรบูรณ์ เริ่มเข้าศึกษาระดับชั้นอนุบาล 1 ถึงชั้นมัธยมศึกษาปีที่ 3 ที่โรงเรียนเซนต์โยเซฟศรี-เพชรบูรณ์ อำเภอเมือง จังหวัดเพชรบูรณ์ จากนั้นได้เข้าศึกษาต่อในระดับมัธยมศึกษาตอนปลาย ที่โรงเรียนเพชรพิทยาคม อำเภอเมือง จังหวัดเพชรบูรณ์ ปีการศึกษา 2553 ได้เข้าศึกษาต่อระดับปริญญาตรีสำเร็จการศึกษาเมื่อปี พ.ศ. 2556 สำเร็จการศึกษาในระดับปริญญาโทเมื่อปี พ.ศ. 2557 และภายหลังสำเร็จการศึกษาในระดับปริญญาโท ได้ต่อการศึกษาระดับปริญญาเอก สาขาวิชาวิศวกรรมคอมพิวเตอร์ สำนักวิชาวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี ในปี 2558

ในระหว่างการศึกษาได้รับความอนุเคราะห์อย่างดียิ่งจากอาจารย์โดยได้รับความไว้วางใจให้เป็นผู้ช่วยสอน ผู้ช่วยสอนปฏิบัติการ และได้รับการตีพิมพ์เผยแพร่บทความวิชาการซึ่งรายละเอียดสามารถดูได้ที่ภาคผนวก ข