ข้อเสนอโครงงานวิศวกรรมไฟฟ้า วิชา 2102490 ปีการศึกษา 2562

การเปรียบเทียบวิธีการพยากรณ์กำลังผลิตไฟฟ้าจากเซลล์แสงอาทิตย์ในระยะสั้นมาก

A comparison of intraday solar power forecasting methods

นายชญานนท์ โพธาวนานนท์ ID 5930084921

นายสรารุต พรานนท์สถิตย์ ID 5930515021

อาจารย์ที่ปรึกษา ผศ.ดร. จิตโกมุท ส่งศิริ

ภาควิชาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์

จุฬาลงกรณ์มหาวิทยาลัย

สารบัญ

1 บทนำ

ในปัจจุบันประเทศไทยมีนโยบายที่จะลดปริมาณการใช้พลังงานจากก๊าซธรรมชาติ และส่งเสริมการผลิตไฟฟ้าจากพลังงานพลังงานทางเลือกอาทิ พลังงานจากเซลล์แสงอาทิตย์ ตามแผนพัฒนาพลังงานทดแทนและพลังงานทางเลือก ในปี พ.ศ.2558 (AEDP 2015) โดยกรมพัฒนาพลังงานทดแทนและอนุรักษ์พลังงานมีนโยบายที่จะเพิ่มสัดส่วนการใช้พลังงานทดแทนภายในประเทศ และตามการประเมินภายในสิ้นปี พ.ศ.2579 สัดส่วนของพลังงานไฟฟ้าที่ผลิตได้จากพลังงานแสงอาทิตย์จะคิดเป็นสัดส่วนถึง 30.5 % ของพลังงานในกลุ่มพลังงานทดแทนทั้งหมด ซึ่งสอดคล้องกับการที่ ต้นทุนในการลงทุนติดตั้งระบบผลิตกำลังไฟฟ้าจากเซลล์แสงอาทิตย์ที่มีแนวโน้มที่ลดลงอย่างต่อเนื่อง จึงทำให้การผลิตไฟฟ้าจากพลังงานแสงอาทิตย์ เข้ามามีบทบาทสำคัญและเป็นที่น่าสนใจสำหรับผู้ประกอบการ อย่างไรก็ตามค่าความเข้มรังสีดวงอาทิตย์ที่เป็นตัวแปรสำคัญในการผลิตกำลังไฟฟ้าจากพลังงานแสงอาทิตย์ มีความแปรปรวนซึ่งขึ้นอยู่กับปัจจัยสำคัญ คือ สภาพภูมิอากาศ ทำให้ค่ากำลังไฟฟ้าที่ผลิตได้ในแต่ละช่วงเวลามีความไม่แน่นอน และก่อให้เกิดปัญหาในการบริหารจัดการกำลังผลิตไฟฟ้าให้สอดคล้องกับความต้องการของผู้ใช้ในแต่ละช่วงเวลา จากปัญหาข้างต้น การพัฒนาประสิทธิภาพของการพยากรณ์กำลังไฟฟ้าที่ผลิตได้จากเซลล์แสงอาทิตย์จึงมีความสำคัญ ทั้งในด้านการรักษาความมั่นคงของระบบ ตลอดจนลดต้นทุนอันเนื่องมาจากการสำรองกำลังผลิตไฟฟ้าโดยทั่วไป การพยากรณ์กำลังไฟฟ้าที่ผลิตได้จากเซลล์แสงอาทิตย์สามารถแบ่งได้เป็น 4 ประเภท

1. การพยากรณ์ในระยะ**สั้นมาก** (very short-term forecast หรือ intra-day) เป็นการพยากรณ์ในระยะ 1-6 ชั่วโมงล่วงหน้ามีประโยชน์ในการการรักษาความมั่นคงของระบบโครงข่ายไฟฟ้า รวมไปถึงถึงการใช้งานระบบกักเก็บพลังงานสำรองพร้อมจ่ายทันทีและเพื่อบริหารจัด การพลังงานไฟฟ้า (จากพลังงานหมุนเวียน) ส่วนเกินในบางช่วงเวลา

2. การพยากรณ์ในระยะ**สั้น** (short-term forecast หรือ day-ahead) เป็นการพยากรณ์ในระยะ 1-3 วันล่วงหน้า มีประโยชน์ในการบริหารจัดการ ความต้องการใช้ไฟฟ้า เพื่อเตรียมการสั่งเดินเครื่องในโรงงานที่สามารถควบคุมกำลังผลิตไฟฟ้าได้ เพื่อให้กำลังผลิตไฟฟ้าในแต่ละช่วงเวลาเหมาะสมและ เป็นไปตามกลไกตลาดซื้อขายไฟฟ้าไว้ล่วงหน้า ทั้งนี้เพื่อให้ต้นทุนการจัดหาไฟฟ้าโดยรวมของพื้นที่มีความคุ้มค่าที่สุดในเชิงเศรษฐศาสตร์ และการใช้งานเชื้อเพลิงแต่ละชนิดเป็นไปอย่างเพียงพอและมีประสิทธิภาพ

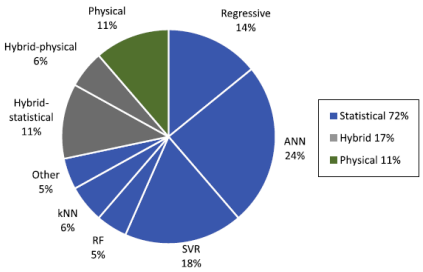
3. การพยากรณ์ในระยะ**กลาง** (medium-term forecast) เป็นการพยากรณ์ในระยะ 1 สัปดาห์ - 1 เดือนล่วงหน้า มีประโยชน์ในการวางแผนกำหนดบำรุงรักษาโดยการทำนายความพร้อมใช้งานของกำลังผลิตไฟฟ้าในอนาคต

4. การพยากรณ์ในระยะ**ยาว** (long-term forecast) เป็นการพยากรณ์ในระยะ 1 เดือน - 1 ปีล่วงหน้า มีประโยชน์ในการบริหารจัดการระบบผลิตกำลังไฟฟ้าในระยะยาว เช่น การสร้างโรงงานผลิตไฟฟ้าแห่งใหม่ หรือการจัดทำแผนประมาณการกำลังไฟฟ้าที่จะผลิตได้ในอนาคต

การศึกษาและพัฒนาความแม่นยำของการพยากรณ์กำลังผลิตไฟฟ้าจากเซลล์แสงอาทิตย์เป็นที่น่าสนใจในวงกว้าง โดยแบ่งออกได้เป็นหลักๆ 2 วิธี คือ วิธีการพยากรณ์ทางตรงและวิธีการพยากรณ์ทางอ้อม วิธีการพยากรณ์ทางอ้อมจะเริ่มจากการพยากรณ์ค่าความเข้มรังสีดวงอาทิตย์ก่อน จากนั้นใช้แบบจำลองของระบบผลิตไฟฟ้าในการแปลงค่าความเข้มรังสีดวงอาทิตย์จากการพยากรณ์ไปเป็นค่ากำลังไฟฟ้าที่คาดว่าจะผลิตได้ ในขณะที่วิธีการพยากรณ์ทางตรงเป็นการพยากรณ์ค่ากำลังผลิตไฟฟ้าที่ได้จากระบบผลิตไฟฟ้าโดยตรง ทั้งนี้หลากหลายงานวิจัยในอดีตจะให้ความสนใจเฉพาะการพยากรณ์ค่าความเข้มรังสีดวงอาทิตย์ เนื่องจากเป็นส่วนที่ยากในการพยากรณ์ และมีการประยุกต์ใช้ที่หลากหลายนอกเหนือจากการพยากรณ์กำลังผลิตไฟฟ้า อย่างไรก็ตามทั้งการพยากรณ์ทางตรงและทางอ้อมต่างมีขั้นตอนวิธีการและเทคนิคที่คล้ายคลึงกัน [2]

หลากหลายงานวิจัยในอดีต ได้นำเสนอวิธีที่หลากหลายในการพยากรณ์รังสีดวงอาทิตย์และกำลังผลิตไฟฟ้าจากเซลล์แสงอาทิตย์ [2, 11] โดยสามารถแบ่งวิธีการได้ออกเป็นหลักๆ 3 ประเภท คือ 1) วิธีการทางสถิติ (statistical methods) 2) วิธีการทางกายภาพ (physical methods) 3) วิธีการแบบผสมผสาน (hybrid methods) [2] วิธีการทางสถิติเป็นการใช้ข้อมูลในอดีตที่วัดได้ เช่น ข้อมูลสภาพอากาศ ค่ากำลังผลิตไฟฟ้าในอดีต ในการพยากรณ์ โดยไม่จำเป็นต้องใช้ข้อมูลความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตาม ตัวอย่างที่นิยมเช่น วิธีการในกลุ่มการเรียนรู้ด้วยเครื่อง วิธีการทางกายภาพ (physical methods) เป็นวิธีที่อาศัยการคำนวณโดยใช้สมการความสัมพันธ์ทางฟิสิกส์ระหว่างตัวแปรต้นและตัวแปรตาม โดยวิธีที่เป็นที่นิยมได้แก่ การพยากรณ์โดยการคำนวณ ค่าพยากรณ์สภาพอากาศเชิงเลข (Numerical Weather Prediction) และ การพยากรณ์โดยใช้วิธีข้างต้นร่วมกันเรียกว่าวิธีการแบบผสมผสาน

แผนภูมิดัง*รูปที่ 1* แสดงให้เห็นว่าในอดีตมีการนำเสนอวิธีที่หลากหลายในการพยากรณ์ค่ากำลังผลิตไฟฟ้าจากเซลล์แสงอาทิตย์อาทิ 1) Regressive methods 2) Artificial neural network (ANN) 3) Support vector regression (SVR) 4) k-Nearest neighbors (k-NN) 5) Random forest (RF) ทั้งนี้ในโครงงานฉบับนี้จะเลือกพิจารณาการพยากรณ์ในระยะสั้นมาก (very short-term forecast หรือ intra-day) เพื่อประโยชน์ในการบริหารและรักษาความมั่นคงในระบบโครงข่ายไฟฟ้า โดยวิธีการพยากรณ์ที่เป็นที่นิยมแพร่หลายในระยะนี้ คือ วิธีการทางสถิติ (statistical methods) ซึ่งมีหลากหลายวิธี ตั้งแต่ การใช้แบบจำลองเชิงเส้น ไปจนถึงวิธีที่มีความซับซ้อนสูงเช่น โครงข่ายประสาทเทียม (neural network) ทั้งนี้การใช้แบบจำลองเชิงเส้นซึ่งที่มีความซับซ้อนต่ำในการพยากรณ์กำลังผลิตไฟฟ้าจากเซลล์แสงอาทิตย์มีประสิทธิภาพต่ำ [15] ซึ่งอาจเกิดมาจากปัจจัยที่มีความสัมพันธ์แบบไม่เป็นเชิงเส้นกับค่ากำลังผลิตไฟฟ้า ด้วยปัญหาข้างต้นในปัจจุบันวิธีการพยากรณ์ในกลุ่มการเรียนรู้ด้วยเครื่องถูกพัฒนา และนำมาใช้ในการพยากรณ์กำลังผลิตไฟฟ้าจากพลังงานแสงอาทิตย์อาทิ ANN, SVR, RF, KNN ซึ่งมีงานวิจัยในอดีตที่เกี่ยวข้องดังนี้



รูป 1: สัดส่วนของงานวิจัยจำแนกตามเทคนิคที่ใช้ในการพยากรณ์ (ที่มา: [2])

M.Rana [14] เปรียบเทียบการใช้วิธี SVR และวิธี NN-ensemble ในการพยากรณ์กำลังผลิตไฟฟ้าในระยะ 5 ถึง 60 นาที โดยข้อมูลกำลังผลิตไฟฟ้าในอดีตเพียงอย่างเดียว และการใช้ข้อมูลกำลังผลิตไฟฟ้าในอดีตร่วมกับข้อมูลสภาพอากาศ นอกจากนี้ยังมีการประยุกต์ใช้วิธี Correlation-based Feature selection (CFS) ในการคัดเลือกคุณลักษณะของข้อมูลที่ใช้ในการทำนาย จากผลลัพธ์พบว่าในการพยากรณ์ระยะใกล้ NN-ensemble และ SVR ให้ความแม่นยำในการพยากรณ์ใกล้เคียงกัน ส่วนในระยะไกลออกไป NN-ensemble จะพยากรณ์แม่นยำกว่า SVR

S.Vagropoulos [17] ใช้วิธี SARIMA ในการพยากรณ์กำลังผลิตไฟฟ้าในระยะ 1 ชั่วโมงโดยใช้ข้อมูล ความเข้มรังสีแสงอาทิตย์ในอดีตร่วมกับข้อมูลสภาพอากาศ ผลลัพธ์การพยากรณ์มีค่า NRMSE เท่ากับ 8.12% โดยเป็นค่าที่ถูกปรับเทียบด้วยค่ากำลังติดตั้งขนาด 0.15-MW (ในที่นี้ Wp ย่อมากจาก Watt peak หมายถึงกำลังผลิตไฟฟ้าสูงสุดของระบบ)

M.Bouzerdoum [6] เปรียบเทียบการใช้วิธี seasonal auto-regressive integrated moving average (SARIMA) , SVR และ การผสมผสานของ SARIMA และ SVR เพื่อพยากรณ์ค่ากำลังผลิตไฟฟ้าในระยะ 1 ชั่วโมง โดยใช้ข้อมูลจากกำลังผลิตไฟฟ้าในอดีตและค่าอุณหภูมิ พบว่า SARIMA-SVR ให้ผลลัพธ์ที่แม่นยำที่สุดมีค่า NRMSE เท่ากับ 9.40 % โดยเป็นค่าที่ถูกปรับเทียบด้วยค่ากำลังติดตั้งขนาด 20-kW

R. Xu และคณะ [19] ประยุกต์ใช้วิธี SVR ร่วมกับการวิเคราะห์ความคล้ายกันของแต่ละวัน ในการพยากรณ์กำลังผลิตไฟฟ้าในระยะ 2 ชั่วโมง โดยใช้ข้อมูลจากกำลังผลิตไฟฟ้าและค่ารังสีดวงอาทิตย์ในอดีตและค่าอุณหภูมิ พบว่าผลลัพธ์การพยากรณ์ที่ได้มีค่า NRMSE เท่ากับ 9.34% ซึ่งมีความแม่นยำสูงกว่าวิธี NN ที่ได้ค่า NRMSE เท่ากับ 13.19% โดยเป็นค่าที่ถูกปรับเทียบด้วยค่ากำลังติดตั้งขนาด 500-kW

W. Björnและคณะ [4] เปรียบเทียบการใช้วิธี SVR , KNN และ combined weight SVM-kNN ในการพยากรณ์ในระยะ 1 ชั่วโมงและ 6 ชั่วโมง โดยใช้ข้อมูลกำลังผลิตไฟฟ้าในอดีต ข้อมูลสภาพอากาศ เวลาที่ทำการพยากรณ์ และค่าดัชนีฟ้าใส พบว่าผลลัพธ์ที่ดีที่สุดของทั้งสองระยะการพยากรณ์มาจากวิธี combined weight SVR-KNN ในระยะการพยากรณ์ 1 ชั่วโมงให้ค่า RMSE เท่ากับ 6.08% และ ในระยะการพยากรณ์ 6 ชั่วโมง ได้ค่า RMSE เท่ากับ 10.16% โดยผลลัพธ์ที่ได้ถูกประเมินจากข้อมูลจากโรงไฟฟ้า 87 โรงในประเทศเยอรมนี

W.A. Muhammad [1] เปรียบเทียบการใช้วิธี SVR และ RF เพื่อพยากรณ์กำลังผลิตไฟฟ้าในระยะ 1 ชั่วโมง โดยใช้ข้อมูลกำลังผลิตไฟฟ้าในอดีต ความเข้มรังสีแสงอาทิตย์ในอดีต ข้อมูลสภาพอากาศ วันและเดือนที่ทำการพยากรณ์ ได้ผลลัพธ์การพยากรณ์จากวิธี RF และ SVR มีค่า RMSE เท่ากับ 2.2470 kWh และ 2.3973 kWh ตามลำดับ โดยข้อมูลที่ใช้ในการทดลองวัดจากระบบไฟฟ้าซึ่งมีค่ากำลังติดตั้งสูงสุดประมาณ 40-kW

จากงานวิจัยที่เกี่ยวข้องข้างต้น แสดงให้เห็นว่าวิธี SVR เป็นวิธีที่นิยมแพร่หลายและมีสมรรถนะที่ดีในการพยากรณ์ในระยะสั้นมาก และวิธี RF เป็นวิธีที่ [1] นำเสนอว่าให้ผลลัพธ์ที่ดีกว่า SVR โครงงานนี้จึงสนใจที่จะทดลองเปรียบเทียบกลุ่มวิธี แบบจำลองในกลุ่มการเรียนรู้ด้วยเครื่องเพิ่มเติม อันได้แก่ 1) linear regression model , 2) Multivariate adaptive regression spline (MARS) , 3) SVR, 4) RF โดยที่ 2 วิธีในกลุ่มแรกจัดทำขึ้นเพื่อเป็นแบบจำลองฐาน (baseline model) สำหรับผลลัพธ์ที่คาดหวังในการจัดทำโครงงานนี้มีดังนี้

1. ผลลัพธ์การเปรียบเทียบการพยากรณ์ความเข้มแสงอาทิตย์ด้วยแบบจำลอง Linear Regression, MARS, SVR, RF ทั้งในแง่ของสมรรถนะการพยากรณ์ และความซับซ้อนในคำนวณของแต่ละแบบจำลองทั้งในส่วนของขั้นตอนการเรียนรู้ทางสถิติและขั้นตอนการดำเนินการพยากรณ์

2. แบบจำลองการพยากรณ์กำลังผลิตไฟฟ้าในระยะสั้นมากในกลุ่มการเรียนรู้ด้วยเครื่องอันได้แก่ Linear Regression, MARS, SVR, RF และชุดคำสั่งโปรแกรมสำหรับพยากรณ์กำลังผลิตไฟฟ้าจากเซลล์แสงอาทิตย์ในระยะสั้นมาก

2 วัตถุประสงค์ของโครงงาน

1. เพื่อศึกษาและสรุปปัจจัยที่ส่งผลต่อกำลังผลิตไฟฟ้าจากเซลล์แสงอาทิตย์ในระยะสั้นมาก
2. เพื่อเปรียบผลลัพธ์การพยากรณ์ความเข้มแสงจากแบบจำลองในกลุ่ม Linear Regression, Multivariate adaptive regression spline, Support vector regression และ Random forest โดยใช้ตัวชี้วัดสมรรถนะของการพยากรณ์ทางสถิติ
3. เพื่อเปรียบเทียบความซับซ้อนในการคำนวณ (computational complexity) ที่เกิดขึ้นในขั้นตอนการเรียนรู้แบบจำลองและการคำนวณค่าพยากรณ์ของแบบจำลองในข้อ 2.

3 หลักการและทฤษฎีที่เกี่ยวข้อง

กำหนดให้ $X_1, X_2, ..., X_p$ แทนตัวแปรต้น, Y แทนตัวแปรตามคือค่ารังสีดวงอาทิตย์
กำหนดสัญลักษณ์และตัวแปรดังนี้

ตาราง 1: สัญลักษณ์และตัวแปร

ตัวแปร	ความหมาย	หน่วย
I	ความเข้มรังสีดวงอาทิตย์	วัตต์ต่อตารางเมตร
P	กำลังผลิตไฟฟ้า	วัตต์
RH	ความชื้นสัมพัทธ์	เปอร์เซ็นต์
WS	ความเร็วลม	เมตรต่อวินาที
UV	ดัชนีรังสีอัลตราไวโอเลต	-
T	อุณหภูมิภายนอก	องศาเซลเซียส
$\cos(\theta)$	โคไซน์ของมุมของดวงอาทิตย์เทียบกับแนวตั้งฉากพื้นโลก	-

- ตัวแปรที่เขียนในรูป $x(t)$ หมายถึงค่า x ณ เวลา t
- ตัวแปรที่เขียนในรูป $\hat{x}(t)$ หมายถึงค่าพยากรณ์ หากเขียนในรูป $x(t)$ หมายถึงค่าที่วัดได้จริง
- ตัวแปรที่เขียนในรูป $\hat{x}_A(t)$ หมายถึงค่าพยากรณ์ของ x จากวิธี A
- การใช้ลำดับเวลาจะเขียนอยู่ในรูป $x(t)$ หมายถึงตัวแปร x ที่เวลา t ในวันหนึ่งๆ หากอยู่เขียนในรูป $x^{(d)}(t)$ หมายถึงตัวแปร x ที่วันที่ d ในเวลา t

ในการทดลองจะกำหนด index ของเวลาดังนี้

- t แทน index ของเวลาปัจจุบัน
- $t-1, t-2, ..., t-k$ หมายถึงเวลา $30, 60, ..., 30k$ นาทีก่อนหน้า
- $t+1, t+2, ..., t+k$ หมายถึงเวลา $30, 60, ..., 30k$ นาทีข้างหน้า

3.1 การคัดเลือกคุณลักษณะ

3.1.1 สหสัมพันธ์

สหสัมพันธ์ (Correlation) เป็นค่าที่บ่งบอกความสัมพันธ์เชิงเส้นระหว่างตัวแปรตั้งแต่ 2 ตัวขึ้นไป โดยในการพิจารณาความสัมพันธ์เชิงเส้นระหว่างตัวแปรว่ามีมากน้อยเพียงใด สามารถบอกได้จากค่าสัมประสิทธ์สหสัมพันธ์ (Correlation coefficient) ซึ่งสามารถคำนวณจากวิธีการทางสถิติได้หลายวิธีซึ่งขึ้นอยู่กับลักษณะของตัวแปรนั้นๆ ในการวัดความสัมพันธ์แต่ละแบบจะต้องมีการทดสอบนัยสำคัญทางสถิติของตัวแปรคู่นั้นๆ ก่อนจึงจะสามารถสรุปความสัมพันธ์ระหว่างตัวแปรได้ การวิเคราะห์ความสัมพันธ์ในรูปแบบนี้จะสามารถตีความถึงความสอดคล้องไปด้วยกันของตัวแปร แต่ไม่ได้หมายความถึงการเป็นเหตุและผลกันระหว่างตัวแปรนั้นๆ

1) สัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน

สัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน (Pearson correlation Coefficient) เป็นวิธีที่ใช้วัดความสัมพันธ์เชิงเส้นระหว่างตัวแปร 2 ชุดในเซตของตัวแปรสุ่มที่เป็นอิสระต่อกัน โดยสามารถคำนวณได้จากสูตรดังนี้

$$\rho = \frac{\mathbf{cov}(X, Y)}{\sigma_X \sigma_Y} \tag{1}$$

โดยที่ ρ แทนสัมประสิทธิ์สหสัมพันธ์เพียร์สัน, $\mathbf{cov}(X, Y)$ แทนความแปรปรวนร่วมของตัวแปร X และ Y σ_x,σ_y แทนส่วนเบี่ยงเบนมาตรฐานของตัวแปร X และ Y ตามลำดับ

2) สัมประสิทธิ์สหสัมพันธ์แบบแยกส่วน

สัมประสิทธิ์สหสัมพันธ์แบบแยกส่วน (Partial correlation coefficient) เป็นวิธีที่ใช้วัดความสัมพันธ์เชิงเส้นระหว่างตัวแปร 2 ชุด โดยคำนวณจากความคลาดเคลื่อนคงค้างของตัวแปร 2 ชุดนั้นหลังจากกำจัดอิทธิพลเชิงเส้นจากตัวแปรอื่นๆออกดังสมการต่อไปนี้

$$\mathbf{cov}(Y_i, Y_j|X) = \mathbf{cov}(Y_i - \hat{Y}_i(X), Y_j - \hat{Y}_j(X)) \tag{2}$$

โดยที่ $\hat{Y}_i$ คือค่าประมาณของ Y_i จาก การวิเคราะห์การถดถอยแบบเชิงเส้นบนข้อมูล X และ $\hat{Y}_j$ คือค่าประมาณของ Y_j จาก การวิเคราะห์การถดถอยแบบเชิงเส้นบนข้อมูล X

$$\rho_{Y_i,Yj|X} = \frac{\mathbf{cov}(Y_i, Y_j|X)}{\sqrt{\mathbf{var}(Y_i - \hat{Y}_i(X))\,\mathbf{var}(Y_j - \hat{Y}_j(X))}} \tag{3}$$

ค่าสัมประสิทธิ์สหสัมพันธ์ของตัวแปรสุ่มแบบเกาส์เซียน 2 ตัวใดๆสามารถคำนวณจากเมทริกซ์ความแปรปรวนร่วมผกผันได้ดังนี้

$$\rho_{X_i X_j \cdot V\setminus\{X_i,X_j\}} = -\frac{\Sigma^{-1}_{ij}}{\sqrt{\Sigma^{-1}_{ii}\Sigma^{-1}_{jj}}} \tag{4}$$

โดยที่ ρ แทนสัมประสิทธิ์สหสัมพันธ์แบบแยกส่วน V แทนเซตของตัวแปรสุ่ม X_1,X_2,...,X_K, Σ แทนเมทริกซ์ความแปรปรวนร่วมของตัวแปรสุ่ม ในเซต V

3.1.2 วิธีการถดถอยเชิงเส้นแบบขั้นตอน

วิธีการถดถอยเชิงเส้นแบบขั้นตอน (Stepwise linear regression) เป็นวิธีหนึ่งในการหาสมการถดถอยเชิงเส้นแสดงความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรที่พิจารณา ซึ่งแตกต่างจากวิธีการถดถอยเชิงเส้น (linear regression) ตรงที่จะมีการเพิ่ม/ลดตัวแปรต้นที่ใช้ในการสร้างสมการโดยการใช้ค่าสถิติเป็นเกณฑ์ในการเลือกตัวแปรต้นที่จะเพิ่ม/ลด แต่ละขั้นตอน ซึ่งวิธีการถดถอยเชิงเส้นแบบขั้นตอน เป็นวิธีที่เกิดจากการประยุกต์ระหว่างวิธีการเลือกแบบก้าวหน้า (foreward seletion) และวิธีการตัดทิ้งแบบถอยหลัง (backward deletion)

- วิธีการเลือกแบบก้าวหน้า (foreward seletion) จะเริ่มต้นจากการสร้างสมการค่าคงที่สำหรับประมาณค่าตัวแปรตามที่พิจารณา จากนั้นในแต่ละขั้นตอนจะทดลองเพิ่มตัวแปรต้นทีละตัวแปรเข้าไปในกลุ่มตัวแปรที่ใช้ในการสร้างสมการถดถอย จากนั้นตรวจสอบว่าการเพิ่มตัวแปรต้นแต่ละตัวแปรนั้นส่งผลให้ค่า RMSE ในการประมาณค่าตัวแปรตามลดลงอย่างมีนัยสำคัญหรือไม่ โดยการทดสอบนัยสำสัญทางสถิติ จากนั้นจึงตัดสินใจเพิ่มตัวแปรที่มี p-value ต่ำสุดเข้าไปในกลุ่มตัวแปรที่จะใช้ในการสร้างสมการถดถอย และดำเนินกระบวนการต่อจนกระทั่งกระบวนการจะสิ้นสุดเมื่อ p-value จากการทดสอบตัวแปรต้นทุกตัวมีค่ามากกว่าค่าที่กำหนด

- วิธีการตัดทิ้งแบบถอยหลัง (backward deletion) จะเริ่มต้นจากการสร้างสมการถดถอยเชิงเส้นที่ประกอบด้วยตัวแปรต้นทุกตัวในสมการก่อน จากนั้นในแต่ละขั้นตอนจะทดลองตัดตัวแปรต้นออกจากกลุ่มตัวแปรที่ใช้ในการสร้างสมการทีละตัวแปร จากนั้นตรวจสอบว่าการที่มีแปรต้นแต่ละตัวอยู่ในกลุ่มนั้น ส่งผลให้ค่า RMSE ในการประมาณค่าตัวแปรตามลดลงอย่างมีนัยสำคัญหรือไม่ (เมื่อเทียบกับหลังตัดตัวแปรออก) โดยการทดสอบนัยสำสัญทางสถิติ จากนั้นจึงตัดสินใจตัดตัวแปรที่มี p-value สูงสุดออกจากกลุ่มตัวแปรที่จะใช้ในการสร้างสมการถดถอย และดำเนิน กระบวนการต่อจนกระทั่งกระบวนการจะสิ้นสุดเมื่อ p-value จากการทดสอบตัวแปรต้นทุกตัวมีค่าต่ำกว่าค่าที่กำหนด

สำหรับวิธีการถดถอยเชิงเส้นแบบขั้นตอนในแต่ละขั้นตอนจะทำการเพิ่มตัวแปรต้นเข้าไปในในกลุ่มตัวแปรที่ใช้ในการสร้างสมการถดถอยโดยวิธีการเลือกแบบก้าวหน้า และเมื่อสิ้นสุดขั้นตอนการเพิ่มตัวแปรแต่ละรอบ จึงตัดตัวแปรออกโดยวิธีการตัดทิ้งแบบถอยหลัง และดำเนินกระบวนการต่อจนกระทั่งกระบวนการจะสิ้นสุด เมื่อไม่มีตัวแปรต้นตัวใดถูกเพิ่มในวิธีการเลือกแบบก้าวหน้าแล้ว ดังนั้นเราจึงสามารถใช้วิธีการถดถอยเชิงเส้นแบบขั้นตอนในการคัดเลือกตัวแปรต้นที่มีความสัมพันธ์เชิงเส้นกับตัวแปรตาม โดยการพิจารณาตัวแปรที่อยู่ในกลุ่มตัวแปรที่ใช้ในการสร้างสมการถดถอยหลังจากกระบวนการเลือกสิ้นสุด

3.2 แบบจำลองการพยากรณ์

ในโครงงานนี้ เราจะพยากรณ์ค่าความเข้มรังสีดวงอาทิตย์ในช่วงเวลาล่วงหน้า 4 ชั่วโมง (ค่าผลลัพธ์การพยากรณ์มีความละเอียด 30 นาที กล่าวคือจะพยากรณ์ 30, 60, 90, ..., 240 นาทีล่วงหน้า) จากนั้นจึงใช้แบบจำลองอักส่วนหนึ่งในการแปลงค่าความเข้มรังสีดวงอาทิตย์เป็นค่ากำลังผลิตไฟฟ้า โดยวิธีที่ใช้แบ่งออกเป็น 4 วิธีคือ 1) Linear Regression 2) Multivariate Adaptive Regression Splines (MARS) 3) Support Vector Regression (SVR) 4) Random Forest (RF) โดย 2 วิธีแรกจัดทำขึ้นเพื่อเป็นแบบจำลองฐาน (baseline model)

3.2.1 Linear regression

เราจะพยากรณ์ค่าความเข้มรังสีดวงอาทิตย์ $I(t+1), I(t+2), ..., I(t+8)$ โดยใช้วิธีการถดถอยเชิงเส้นซึ่งมีตัวแปรต้นดังนี้

1. ค่ารังสีดวงอาทิตย์ในอดีตประกอบด้วย
   - $I(t), I(t-1), ..., I(t-7)$
   - $I^{(d-1)}(t+1), I^{(d-1)}(t+2), ..., I^{(d-1)}(t+8)$
2. ค่าพยากรณ์สภาพอากาศจาก numerical weather prediction model ที่ชื่อว่า WRF ประกอบด้วย
   - $\hat{I}_{\text{WRF}}(t+1), \hat{I}_{\text{WRF}}(t+2), ..., \hat{I}_{\text{WRF}}(t+8)$
   - $\hat{\text{T}}_{\text{WRF}}(t+1), \hat{\text{T}}_{\text{WRF}}(t+2), ..., \hat{\text{T}}_{\text{WRF}}(t+8)$
   - $\hat{\text{RH}}_{\text{WRF}}(t+1), \hat{\text{RH}}_{\text{WRF}}(t+2), ..., \hat{\text{RH}}_{\text{WRF}}(t+8)$
   - $\hat{\text{UV}}_{\text{WRF}}(t+1), \hat{\text{UV}}_{\text{WRF}}(t+2), ..., \hat{\text{UV}}_{\text{WRF}}(t+8)$

   Weather Research and Forecast (WRF) เป็นแบบจำลองที่ใช้พยากรณ์สภาพภูมิอากาศที่ถูกพัฒนาโดยความร่วมมือกันของ National Oceanic and Atmospheric Administration (NOAA) และ National Centers for Environmental Prediction (NCEP) การพยากรณ์โดยใช้ WRF นั้นสามารถกำหนดได้โดยผู้ใช้ ซึ่งผู้ใช้ต้องเลือกชุดสมการพลศาสตร์สำหรับพารามิเตอร์ต่างๆให้เหมาะสมกับพื้นที่ และในที่นี้ค่าพยากรณ์ที่ได้จะถูกนำไปใช้เพื่อเป็นข้อมูลขาเข้าอีกกลุ่มหนึ่งในแบบจำลอง

3.2.2 Multivariate adaptive regression splines (MARS)

เราจะพยากรณ์ค่าความเข้มรังสีดวงอาทิตย์ $I(t+1), I(t+2), ..., I(t+8)$ โดยใช้แปรต้นและตัวแปรตามเช่นเดียวกับวิธี Linear regression Multivariate Adaptive Regression Splines (MARS) เป็นวิธีหนึ่งในวิธีการเชิงถดถอย ซึ่งเป็นวิธีการในสร้างสมการความสัมพันธ์แบบไม่เป็นเชิงเส้น (เชิงเส้นแบบเป็นช่วง) ระหว่างตัวแปรต้นและตัวแปรตาม ซึ่งความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตามเขียนดังนี้ [10]

$$\hat{Y}(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X) \tag{5}$$

โดยที่ β แทนสัมประสิทธิ์, $h(X)$ แทนฟังก์ชันเชิงเส้นแบบเป็นช่วงในตัวแปร X ซึ่งเรียกว่า basis function และ M แทนจำนวนช่วง โดยจะเลือกค่า β และ $h(X)$ ซึ่งทำให้ค่าผลรวมของค่าเศษเหลือกำลังสอง (residual sum of squares) มีค่าน้อยที่สุดในชุดข้อมูลฝึก
เนื่องจากวิธี MARS เป็นตัวอย่างของวิธีการพยากรณ์แบบไม่เป็นเชิงเส้น (เชิงเส้นแบบเป็นช่วง) ซึ่งเข้าใจได้ง่ายและมีพื้นฐานมาจากวิธี Linear regression ดังนั้นเราจึงเลือกวิธีนี้เป็นหนึ่งในวิธีที่จะใช้เป็นแบบจำลองฐาน (baseline model) สำหรับเปรียบเทียบกับวิธีการอื่นๆ

3.2.3 Support Vector Regression

Support vector regression เป็นเทคนิคการเรียนรู้ทางสถิติที่ได้รับการพัฒนามาจาก Vapnik (1995) [8] ที่ได้รับความนิยมและให้สมรรถนะที่ดีในการจัดการกับปัญหาการพยากรณ์อนุกรมเวลา [9] โดยใช้หลักการวิเคราะห์การถดถอยที่สามารถอธิบายได้ทั้งรูปแบบความสัมพันธ์เชิงเส้น และไม่เชิงเส้น ดังอธิบายได้ดังนี้ [16, 18] ภายใต้ชุดข้อมูลฝึก $T = \{(x_1, y_1), ..., (x_n, y_n)\}$ โดยที่ $x_i \in \mathsf{R}^p$ เป็นข้อมูลฝึกขาเข้า และ $y_i \in \mathsf{R}$ เป็นข้อมูลฝึกขาออก หลักการของ Support vector regression คือการเปลี่ยนปริภูมิของข้อมูลฝึกขาเข้า X ไปยังปริภูมิใหม่ ($\mathcal{H}$) ผ่านฟังก์ชัน $\varphi(x)$ หลังจากนั้นวิเคราะห์การถดถอยในปริภูมิใหม่เพื่อหาฟังก์ชันในการประมาณ y_i ดัง (6) โดยหาก $\varphi(x)$ เป็นฟังก์ชันไม่เชิงเส้นแล้วฟังก์ชันเชิงเส้นที่ได้ในปริภูมิใหม่นั้นจะสมนัยกับฟังก์ชันไม่เชิงเส้นในปริภูมิเดิม

$$f(x) = \langle w, \varphi(x) \rangle + b \tag{6}$$

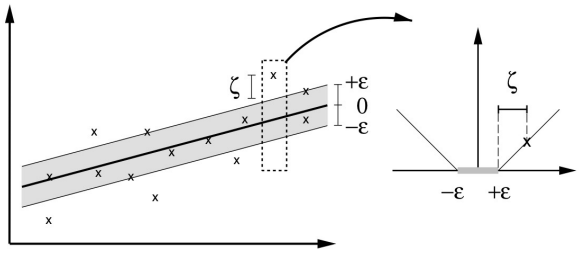
โดยที่ $w \in \mathcal{H}, b \in \mathsf{R}$ แทนเวกเตอร์ค่าน้ำหนัก และค่าคงที่ตามลำดับ ในการหาค่า w, b สามารถทำได้โดยแก้ไขปัญหาการหาค่าที่เหมาะสมภายใต้ข้อจำกัดดังนี้

$$
\begin{aligned}
\underset{w,b,u_i,v_i}{\text{minimize}} \quad & (1/2)\|w\|^2 + C\sum_{i=1}^{n}(u_i+v_i) \\
\text{subject to} \quad & y_i - \langle w, \varphi(x_i)\rangle - b \le \varepsilon + u_i, \;\; i = 1,2,\ldots,n, \\
& \langle w, \varphi(x_i)\rangle + b - y_i \le \varepsilon + v_i, \;\; i = 1,2,\ldots,n, \\
& u_i, v_i \ge 0 \qquad , \;\; i = 1,2,\ldots,n
\end{aligned} \tag{7}
$$

ε แทนพารามิเตอร์ที่กำหนดขนาดของบริเวณค่าความคลาดเคลื่อนที่ยอมรับได้ โดยค่าความคลาดเคลื่อนที่ตกอยู่ภายในบริเวณนี้จะไม่ถูกนำไปคิดในฟังก์ชันสูญเสีย u, v แทนตัวแปรหย่อน (Slack Variable) ซึ่งเป็นค่าที่ยอมให้บางจุดข้อมูลมีค่าความคลาดเคลื่อนมากกว่าค่า ε ที่กำหนดได้ดังแสดงในสมการข้อจำกัด

ฟังก์ชันวัตถุประสงค์ต้องการที่จะหาค่าต่ำสุดของพจน์ $(1/2)\|w\|^2$ ซึ่งเป็นค่าที่ลงโทษความซับซ้อนของแบบจำลองและยังสอดคล้องกับการหาระยะห่างมากที่สุดของระนาบขอบของบริเวณค่าความคลาดเคลื่อนที่ยอมรับได้ และพจน์ $C\sum_{i=1}^{n}(u_i+v_i)$ ซึ่งแสดงถึงฟังก์ชันสูญเสียแบบ ε-incentive ดังแสดงใน (8) พจน์นี้สอดคล้องกับการพิจารณาฟังก์ชันลงโทษ (penalty function) ที่ลงโทษตัวแปรหย่อนที่ยอมให้เกิดความคลาดเคลื่อนมากกว่าค่า ε ในบางจุดข้อมูล ส่วนค่าคงที่ C เป็นค่าน้ำหนักที่ควบคุมความสมดุลในการหาค่าต่ำสุดระหว่าง 2 พจน์ดังกล่าว โดยสรุปการหาค่าต่ำสุดของ (7) สอดคล้องกับหลักการการเรียนรู้ทางสถิติที่ต้องการควบคุมทั้งค่าความคลาดเคลื่อนในชุดข้อมูลฝึกและความซับซ้อนของแบบจำลอง[3]

$$
\left|y - f(x)\right|_{\varepsilon} = \begin{cases} \left|y - f(x)\right| - \varepsilon, & \text{if } \left|y - f(x)\right| > \varepsilon \\ 0, & \text{otherwise} \end{cases} \tag{8}
$$



รูป 2: รูปแสดงฟังก์ชันสูญเสียแบบ ε-incentive ของ linear SVR ([16])

ในการแก้ปัญหา (7) (primal form) เราพบว่ามีการคำนวณหา w ซึ่งอยู่ในปริภูมิ $\mathcal{H}$ ที่อาจมีมิติสูง จึงอาจมีความจำเป็นต้องใช้กำลังในการคำนวณสูง ดังนั้นเราประยุกต์ใช้หลักการ Lagrange duality เปลี่ยนมาพิจารณา dual form ของปัญหานี้ภายใต้เงื่อนไข Karush-Kuhn-Tucker (KKT) แทน ซึ่งเป็นการคำนวณหา λ, ν ซึ่งอยู่ในปริภูมิ R^n ดังนี้

$$
\begin{aligned}
\underset{\lambda,\nu}{\text{maximize}} \quad & -(1/2)(\lambda-\nu)^T Q(\lambda-\nu) - \varepsilon\sum_{i=1}^{n}(\lambda_i+\nu_i) + \sum_{i=1}^{n} y_i(\lambda_i-\nu_i) \\
\text{subject to} \quad & \mathbf{1}^T(\lambda-\nu) = 0 \;\text{ และ }\; \lambda_i, \nu_i \in [0, C], \;\; i = 1,2,,n
\end{aligned} \tag{9}
$$

โดย $Q_{ij} = \varphi(x_i)^T\varphi(x_j)$ และค่าคงที่บวก $\lambda, \nu \in \mathsf{R}^n$ แทนตัวคูณลากรานจ์ ซึ่งจาก (9) ได้ผลลัพธ์ดังนี้

$$
w = \sum_{i=1}^{n}(\lambda_i-\nu_i)\varphi(x_i), \text{ ดังนั้น } \; f(x) = \sum_{i=1}^{n}(\lambda_i-\nu_i)\langle\varphi(x_i),\varphi(x)\rangle + b = \sum_{i=1}^{n}(\lambda_i-\nu_i)k(x_i,x) + b \tag{10}
$$

โดยที่ x_i แทนจุดข้อมูลขาเข้าในชุดข้อมูลฝึก ส่วน x แทนจุดข้อมูลขาเข้าในชุดข้อมูลตรวจสอบหรือชุดข้อมูลทดสอบ จาก (10) การคำนวณผลลัพธ์ที่ได้ขึ้นกับ support vectors โดยไม่ขึ้นกับมิติของปริภูมิ $\mathcal{H}$ นอกจากนี้ยังสามารถประยุกต์ใช้ Kernel Trick โดยการคำนวณเคอร์เนลฟังก์ชันแทนการคำนวณผลคูณแบบจุดของข้อมูลขาเข้าในปริภูมิ $\mathcal{H}$หนึ่งๆ จากข้อได้เปรียบข้างต้นจึงเห็นว่าในการพิจารณา dual problem สามารถลดกำลังการคำนวณในการแก้ปัญหาลงได้มาก ส่วนเงื่อนไข Karush-Kuhn-Tucker (KKT) ที่ทำให้การแก้สมการในรูป dual form ได้ผลลัพธ์เดียวกับการแก้สมการในรูป primal form ได้แก่

$$\lambda_i(\varepsilon + u_i - y_i + \langle w, \varphi(x_i)\rangle + b) = 0 \tag{11}$$
$$\nu_i(\varepsilon + v_i + y_i - \langle w, \varphi(x_i)\rangle - b) = 0 \tag{12}$$
$$(C - \lambda_i)u_i = 0 \tag{13}$$
$$(C - \nu_i)v_i = 0 \tag{14}$$

จากสมการเงื่อนไขดังกล่าว สามารถสรุปได้ดังนี้

1. มีเฉพาะคู่ลำดับ (x_i, y_i) ที่มีค่าตัวคูณลากรานจ์เท่ากับ C เท่านั้นที่ ตกอยู่นอกบริเวณความคลาดเคลื่อนที่ยอมรับได้
2. $\lambda_i\nu_i = 0$ หรือกล่าวได้ว่าคู่ลำดับ (λ_i, ν_i) ใดๆ จะมีค่าใดค่าหนึ่งเท่ากับศูนย์เสมอ
3. ถ้า $\lambda_i \in (0, C)$ แล้ว $u_i = 0$ และถ้า $\nu_i \in (0, C)$ แล้ว $v_i = 0$ ดังนั้นจากเงื่อนไขนี้สามารถนำไปใช้ในการคำนวณหาค่า b จาก (11), (12) ดังนี้

$$\begin{aligned} b &= y_i - \langle w, \varphi(x_i)\rangle - \varepsilon \text{ สำหรับ } \lambda_i \in (0, C) \\ b &= y_i - \langle w, \varphi(x_i)\rangle + \varepsilon \text{ สำหรับ } \nu_i \in (0, C) \end{aligned} \tag{15}$$

เคอร์เนลฟังก์ชันที่เป็นที่นิยมสำหรับ Support Vector Regression มีดังนี้

1. Linear kernel : $k(x, x') = \langle x, x'\rangle$
2. Polynomial kernel : $k(x, x') = (\gamma\langle x, x'\rangle + r)^d$
3. RBF kernel : $k(x, x') = \exp(-\gamma\left\|x - x'\right\|^2)$
4. Sigmoid kernel : $k(x, x') = \tanh(\gamma\langle x, x'\rangle + r)$

โดยที่ r, d, γ เป็นพารามิเตอร์ของเคอร์เนลฟังก์ชัน

เราจะสร้างแบบจำลอง SVR มา 8 แบบจำลองโดย แต่ละแบบจำลองพยากรณ์ค่าความเข้มรังสีดวงอาทิตย์ $I(t + k)$ โดยที่ k = 1,2,...,8 มีตัวแปรต้นดังนี้

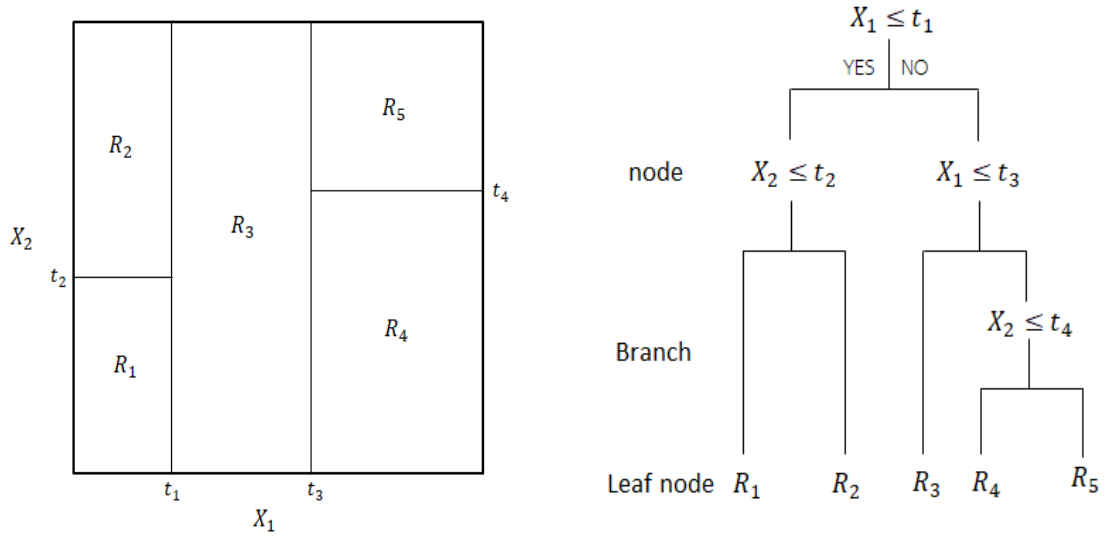
- $I(t), I(t - 1), ..., I(t - 7)$
- $I^{(d-1)}(t + k)$
- $\cos(\theta(t + k))$

โครงงานฉบับนี้ได้ทำการทดลองโดยใช้ the Python package scikit-learn ซึ่งมีพื้นฐานมาจาก LIBSVM library

3.2.4 Random Forest

แบบจำลอง Random forest ถูกนำเสนอครั้งแรกในปี ค.ศ. 1995 โดย Tin Kam Ho เป็นวิธีที่อิงจากแบบจำลองต้นไม้ถดถอย (Regression tree model) ดังที่จะอธิบายต่อไป แบบจำลองต้นไม้ถดถอยเป็นการประยุกต์หลักการของแบบจำลองต้นไม้ตัดสินใจ (Decision tree model) เพื่อใช้ในการพยากรณ์ค่าของตัวแปรที่พิจารณาโดยอาศัยวิธีการแบ่งกลุ่มของตัวแปรต้น
หลักการของแบบจำลองต้นไม้ถดถอย สามารถอธิบายได้เป็น 2 ขั้นตอนดังนี้ [12]

1. แบ่งปริภูมิของตัวแปรต้น $X_1, X_2, ..., X_p$ ออกเป็น J ส่วนที่ไม่มีการซ้อนทับซึ่งกันและกัน, ให้ปริภูมิย่อยนั้นเรียกว่า $R_1, R_2, ..., R_j$
2. สำหรับทุกๆ ข้อมูลของตัวแปรต้นที่อยู่ใน R_j เราจะพยากรณ์ค่าของตัวแปรตามให้มีค่าเท่ากับค่าเฉลี่ยของค่าตัวแปรตามในชุดข้อมูลฝึกทั้งหมด ซึ่งค่าของตัวแปรต้นตกอยู่ใน R_j



(a) ตัวอย่างผลลัพธ์จากการแบ่งปริภูมิของตัวแปรต้นโดยขั้นตอนวิธี recursive binary splitting

(b) แผนภาพต้นไม้ที่สอดคล้องกับการแบ่งปริภูมิในภาพ (a)

รูป 3: ตัวอย่างการแบบจำลองต้นไม้สำหรับปริภูมิตัวแปรต้น 2 มิติ

โดยในขั้นตอนที่ 1 จะทำการเลือกแบ่งปริภูมิของตัวแปรต้นออกเป็น ปริภูมิย่อย $R_1, R_2, ..., R_j$ ซึ่งมีลักษณะเป็น high-dimensional rectangles เพื่อให้ได้ ปริภูมิย่อย $R_1, R_2, ..., R_j$ ซึ่งให้ค่า residual squared error (RSS) ที่น้อยที่สุด
กำหนดโดย

$$\text{RSS} = \sum_{j=1}^{J} \sum_{i \in R_j} \|y_i - \hat{y}_{R_j}\|_2^2 \tag{16}$$

ในทางปฏิบัติขั้นตอนวิธีที่ใช้หาคำตอบของปัญหาข้างต้น เรียกว่าวิธี recursive binary splitting คือ การแบ่งปริภูมิออกเป็นปริภูมิย่อยทีละสองปริภูมิ โดยวิธีการวนซ้ำ ซึ่งมีเงื่อนไขคือในแต่ละรอบจะเลือกการแบ่งปริภูมิที่ทำให้ค่า RSS มีค่าลดลงมากที่สุด *รูปที่ 3* แสดงตัวอย่างในกรณีที่ปริภูมิของตัวแปรต้นเป็นปริภูมิ 2 มิติ และจาก *รูปที่ 3(b)* จะเห็นว่าด้วยลักษณะของขั้นตอนวิธีนี้เองทำให้ลักษณะของแบบจำลองนี้คล้ายการแตกกิ่งของต้นไม้และถูกเรียกว่าแบบจำลองต้นไม้ โดยแต่ละส่วนจะถูกเรียกว่า ปม (node) และแขนง (branch) แบบจำลองต้นไม้ถดถอยดังที่กล่าวไปข้างต้นมักประสบกับปัญหาเรื่องความแปรปรวนที่สูง ซึ่งหมายความว่าหากเราแบ่งชุดข้อมูลฝึกออกเป็นสองส่วน จากนั้นทำการหาแบบจำลองโดยใช้ข้อมูลฝึกแต่ละส่วน ผลลัพธ์ในการพยากรณ์ที่ได้จะแตกต่างกันมาก ซึ่งในทางตรงกันข้ามแบบจำลองที่มีความแปรปรวนต่ำจะให้ผลลัพธ์ที่ใกล้เคียงกันแม้ว่าจะเปลี่ยนชุดข้อมูลฝึก แบบจำลอง Random forest เป็นการรวมผลการพยากรณ์จากแบบจำลองต้นไม้ถดถอยจำนวนมาก โดยผลการพยากรณ์ของแบบจำลอง Random forest จะกำหนดให้เป็นค่าเฉลี่ยของค่าพยากรณ์จากทุกๆแบบจำลองต้นไม้ย่อย โดยในแต่ละปม (node) ของแบบจำลองย่อย จะสุ่มเลือกใช้ จำนวนคุณลักษณะของตัวแปรต้นเพียง m คุณลักษณะจากทั้งหมด p คุณลักษณะ ซึ่งพิจารณาได้ว่ากระบวนการข้างต้นเป็นการลดสหสัมพันธ์ของกลุ่มแบบจำลองต้นไม้ ซึ่งจะทำให้แบบจำลองรวมมีความแปรปรวนลดลง และมีความคงทนต่อการเปลี่ยนแปลงชุดข้อมูล นอกจากนี้แบบจำลอง Random forest ยังสามารถประยุกต์ใช้ร่วมกับวิธี bootstrap ซึ่งมีหลักการคือ ในขั้นตอนฝึกของแต่ละแบบจำลองต้นไม้ย่อย จะมีการสุ่มตัวอย่างชุดข้อมูลฝึกที่จะใช้ในการฝึกแต่ละแบบจำลองจากชุดข้อมูลฝึกทั้งหมด ซึ่งจะทำให้แบบจำลองต้นไม้ย่อยแต่ละแบบจำลองมีความแตกต่างกันมากขึ้น
พารามิเตอร์สำคัญที่เป็นตัวกำหนดเงื่อนไขของแบบจำลอง และยังส่งต่อประสิทธิภาพ/ความซับซ้อนในการคำนวณของการพยากรณ์มีดังนี้

1. จำนวนแบบจำลองต้นไม้ทั้งหมดภายในป่า เขียนแทนด้วย n_{tree}
2. จำนวนระดับหรือความลึกมากสุดของต้นไม้ที่ยอมรับได้ เขียนแทนด้วย d
   คือจำนวนปม (node) ทั้งหมดที่มากที่สุด เมื่อนับตั้งแต่ใบ (leaf node) ไปจนถึงปมบนสุด ดัง*รูปที่ 3(b)*
3. จำนวนตัวอย่างจากชุดข้อมูลฝึกน้อยสุดภายในปริภูมิ ที่ยินยอมให้มีการเริ่มต้นแบ่งปริภูมิ เขียนแทนด้วย $n_{\text{min\_samples\_split}}$
   คือจำนวนตัวอย่างในข้อมูลฝึกที่น้อยที่สุดในแต่ละปม (node) ก่อนเริ่มการแตกใบ (leaf node) ดัง*รูปที่ 3(b)*
4. จำนวนตัวอย่างจากชุดข้อมูลฝึกน้อยสุดที่ยินยอมมีในแต่ละปริภูมิย่อย เขียนแทนด้วย $n_{\text{min\_samples\_leaf}}$
   คือจำนวนตัวอย่างในข้อมูลฝึกที่น้อยที่สุดในแต่ละใบ (leaf node) ดัง*รูปที่ 3(b)*

5. จำนวนคุณลักษณะของตัวแปรต้นใช้ในแต่ละปมของแบบจำลองต้นไม้ เขียนแทนด้วย m

เราจะใช้แบบจำลอง Random forest ในพยากรณ์ค่าความเข้มรังสีดวงอาทิตย์ $I(t+1), I(t+2), ..., I(t+8)$ โดยมีตัวแปรต้นดังนี้

- $I(t), I(t-1), ..., I(t-7)$
- $I^{(d-1)}(t+1), I^{(d-1)}(t+2), ..., I^{(d-1)}(t+8)$
- $\cos(\theta(t+1)), \cos(\theta(t+2)), ..., \cos(\theta(t+8))$
- ตัวเลขชั่วโมงที่บ่งบอกถึงเวลาที่ทำการพยากรณ์ เขียนแทนด้วย $\mathrm{HR}(t)$

3.3 ดัชนีการวัดประสิทธิภาพของการพยากรณ์

การวัดประสิทธิภาพของแบบจำลองการพยากรณ์นั้นมีหลากหลายวิธี โดยดัชนีตัวชี้วัดสมรรถนะที่นิยมใช้ในงานประยุกต์การพยากรณ์พลังงานซึ่งอยู่ในรูปของค่าความผิดพลาดในการพยากรณ์ มีดังนี้
หมายเหตุ : ในที่นี้จะใช้สัญลักษณ์ตัวแปร x และ $\hat{x}$ แทนค่าวัดจริงและค่าพยากรณ์ตามลำดับ โดย x อาจแทนค่าความเข้มแสง หรือค่ากำลังผลิตไฟฟ้า

1. Root Mean Square Error (RMSE): เป็นการหาค่าเฉลี่ยของกำลังสองสัมบูรณ์ซึ่งเทียบได้กับ 2-นอร์มของเวกเตอร์ ค่าความผิดพลาด

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\hat{x}(t) - x(t))^2} \tag{17}$$

2. Mean Bias Error (MBE): เป็นค่าเฉลี่ยของค่าความผิดพลาด ซึ่งอาจมีค่าเป็นบวกหรือลบ เราจะใช้ดัชนีนี้เป็นการบอกว่าแบบจำลองนั้น ประเมินค่าสูงกว่าความเป็นจริง (overestimate) หรือต่ำกว่าความเป็นจริง (underestimate) ได้

$$\mathrm{MBE} = \frac{1}{n}\sum_{t=1}^{n}(\hat{x}(t) - x(t)) \tag{18}$$

3. Normalized Root Mean Square Error (NRMSE): การใช้ดัชนี RMSE นั้นไม่ได้คำนึงถึงขนาดของค่าตัวแปร เมื่อนำดัชนีนี้ไปเปรียบเทียบกับ ข้อมูลชุดอื่นที่มีขนาดต่างกัน จึงอาจจะเปรียบเทียบไม่ได้สมเหตุสมผล ดังนั้นการ normalization แบบต่างๆ จึงได้ถูกเสนอขึ้น เพื่อทำให้สามารถเทียบสมรรถนะกับงานอื่นๆ ที่ทดสอบบนข้อมูลชุดอื่นได้

   a) Normalized by the mean

$$\mathrm{NRMSE} = \frac{\sqrt{\frac{1}{n}\sum_{t=1}^{n}(\hat{x}(t) - x(t))^2}}{\bar{x}} \times 100\% \tag{19}$$

   โดยที่ $\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x(t)$ คือค่าเฉลี่ยของ $x(t)$

   b) Normalized by the capacity
   ในกรณีที่ x เป็นกำลังผลิตไฟฟ้าจากเซลล์แสงอาทิตย์ เราจะนิยมทำให้เป็นปกติด้วยค่ากำลังผลิตที่ติดตั้ง ($\mathrm{Capacity}$)

$$\mathrm{NRMSE} = \frac{\sqrt{\frac{1}{n}\sum_{t=1}^{n}(\hat{x}(t) - x(t))^2}}{\mathrm{Capacity}} \times 100\% \tag{20}$$

4 ผลลัพธ์จากการดำเนินการ

ในส่วนของผลการดำเนินงานในเทอมการศึกษานี้ จะนำเสนอผลลัพธ์ในส่วนของการคัดเลือกลักษณะและแสดงผลลัพธ์การพยากรณ์ค่าความเข้มรังสีดวงอาทิตย์จาก 2 วิธีก่อน คือวิธี SVR และ RF โดยจะนำผลลัพธ์ที่ได้เปรียบเทียบกับผลลัพธ์ จากการพยากรณ์ด้วยวิธี ANN ซึ่งทีมวิจัยระบบสมาร์ทกริด จุฬาลงกรณ์มหาวิทยาลัยได้ทำการทดลองโดยใช้ชุดข้อมูลชุดเดียวกันและมีการแบ่งชุดข้อมูลที่เหมือนกัน
ชุดข้อมูลที่ใช้ในการทดลองคือ ข้อมูลที่วัดได้ ณ ชั้นดาดฟ้าตึกภาควิศวกรรมไฟฟ้า จุฬาลงกรณ์มหาวิทยาลัย ในช่วงเวลาตั้งแต่ เดือนมกราคม พ.ศ. 2560 จนถึงเดือนธันวาคม พ.ศ. 2561 ซึ่งประกอบด้วย ค่ารังสีดวงอาทิตย์ต่อพื้นที่, ค่ากำลังไฟฟ้า, ค่าความชื้นสัมพัทธ์, อุณหภูมิ, ความเร็วลม, ดัชนีรังสีอัลตราไวโอเลต (UV Index) โดยข้อมูลทั้งหมดถูกลดอัตราสุ่มลงเป็น 30 นาที
โดยแบ่งชุดข้อมูลออกเป็น 3 ส่วนคือ ชุดข้อมูลฝึก (training set), ชุดข้อมูลตรวจสอบ (validation set), ชุดข้อมูลทดสอบ (testing set) ในอัตราส่วน $80:10:10$ เรียงลำดับตามเวลา

1. ชุดข้อมูลฝึก ใช้ในขั้นตอนการฝึกของแบบจำลองและการคัดเลือกคุณลักษณะของตัวแปรขาเข้า
2. ชุดข้อมูลตรวจสอบ ใช้ในการปรับค่าพารามิเตอร์ของแบบจำลอง เช่น การเลือกใช้เคอร์เนลฟังก์ชันใน SVR เป็นต้น และการปรับจำนวนคุณลักษณะของตัวแปรขาเข้าในแต่ละแบบจำลองต้นไม้ (m) ใน Random Forest
3. ชุดข้อมูลทดสอบ ใช้ในขั้นตอนทดสอบสมรรถนะในการพยากรณ์ เพื่อนำผลลัพธ์ที่ได้ไปเปรียบเทียบกับแบบจำลองอื่นๆ เพื่อเปรียบเทียบสมรรถนะของแบบจำลอง

ในส่วนของการพยากรณ์จะทดลองเปรียบเทียบผลลัพธ์จากพยากรณ์ความเข้มรังสีดวงอาทิตย์ล่วงหน้า 4 ชั่วโมง (ค่าผลลัพธ์การพยากรณ์มีความละเอียด 30 นาที กล่าวคือจะพยากรณ์ 30, 60, 90, ..., 240 นาทีล่วงหน้า) โดยพยากรณ์ในช่วงเวลา 5:30 น. ถึง 17.30 น. (พยากรณ์ทุกๆ 30 นาทีเพื่อให้ได้ค่าพยากรณ์ในช่วงเวลาตั้งแต่ 6:00 น. ถึง 18.00 น. โดยใช้ผลลัพธ์จากการคัดเลือกคุณลักษณะในการพิจารณาตัวแปรต้นที่จะใช้ในการพยากรณ์ และเพื่อให้ผลลัพธ์ที่ได้สามารถเปรียบเทียบกับวิธี ANN ซึ่งทีมวิจัยระบบสมาร์ทกริดได้จัดทำไว้ จึงเลือกใช้ชุดข้อมูลทดสอบชุดเดียวกัน อย่างไรก็ตามในชุดข้อมูลดังกล่าวมีช่วงเวลาที่ข้อมูลค่า RH, UV, T และ WS สูญหายไปเป็นช่วงเวลานาน ดังนั้นในส่วนของการพยากรณ์จึงเลือกที่จะไม่ใช้ RH, UV, T และ WS เป็นตัวแปรต้น

4.1 ผลลัพธ์การคัดเลือกคุณลักษณะ

ตัวแปรต้นที่พิจารณาคือ $I(t-1), I(t-2), ..., I(t-7), I^{(d-1)}(t+1), \mathrm{T}(t), \mathrm{RH}(t), \mathrm{UV}(t), \mathrm{WS}(t), \cos(\theta(t+1))$ (ในที่นี้ตัวแปรตามคือ $I(t+1)$)

ตาราง 2: ผลลัพธ์การคัดเลือกคุณลักษณะสำหรับพยากรณ์ $I(t+1)$

ตัวแปร	สหสัมพันธ์		สหสัมพันธ์แยกส่วน		การถดถอยเชิงเส้นแบบขั้นตอน	
	สัมประสิทธิ์	p-value	สัมประสิทธิ์	p-value	สัมประสิทธิ์ในสมการถดถอย	p-value
$I(t)$	0.8956	0	0.4366	0	$0..6574$	10^{-34}
$I(t-1)$	0.7789	0	-0.0014	0.8675	$-$	$-$
$I(t-2)$	0.6478	0	0.0101	0.2397	$-$	$-$
$I(t-3)$	0.5018	0	-0.0172	0.0466	-0.0191	10^{-2}
$I(t-4)$	0.3610	0	-0.0099	0.2533	$-$	$-$
$I(t-5)$	0.2260	10^{-155}	-0.0344	0.0001	-0.0510	10^{-7}
$I(t-6)$	0.1039	10^{-33}	-0.0202	0.0192	-0.0288	10^{-2}
$I(t-7)$	-0.0059	0.4955	-0.0720	0	-0.0834	10^{-21}
$I^{(d-1)}(t+1)$	0.7369	0	0.1021	0	0.0876	10^{-50}
$\mathrm{T}(t)$	0.4290	0	0.0035	0.6825	$-$	$-$
$\mathrm{RH}(t)$	-0.1291	10^{-51}	-0.0638	0	-1.2015	10^{-17}
$\mathrm{UV}(t)$	0.8540	0	0.1090	0	1.4957	10^{-39}
$\mathrm{WS}(t)$	0.1388	10^{-59}	-0.0088	0.3090	$-$	$-$
$\cos(\theta(t+1))$	0.7810	0	0.0910	0	109.78	10^{-75}

หมายเหตุ : ตัวแปรที่ไม่ถูกเลือกในการสร้างสมการถดถอยเชิงเส้นแบบขั้นตอนคือ
$I(t-1), I(t-2), I(t-4), WS(t), T(t)$

จากผลลัพธ์ดัง*ตารางที่ 2* จะเห็นว่าผลลัพธ์จากการวิเคราะห์ด้วยสัมประสิทธิ์สหสัมพันธ์และสัมประสิทธิ์สหสัมพันธ์แบบแยกส่วนแตกต่างกัน เพราะสัมประสิทธิ์สหสัมพันธ์แยกส่วนเป็นการวิเคราะห์ความสัมพันธ์เชิงเส้นระหว่างตัวแปร ในขณะที่กำหนดให้ตัวแปรอื่นๆ เป็นค่าคงที่

ในการคัดเลือกคุณลักษณะที่มีนัยสำคัญเพื่อใช้ในการพยากรณ์ค่า $I(t+1)$ เราจึงพิจารณา p-value จากการทดสอบนัยสำคัญทางสถิติเป็นเกณฑ์ สำหรับสัมประสิทธิ์สหสัมพันธ์จะเห็นว่า p-value ของทุกๆตัวแปรมีค่าใกล้เคียงศูนย์ อย่างไรก็ด้วยเหตุผลที่กล่าวไปข้างต้น เราจึงพิจารณา p-value จากสัมประสิทธิ์สหสัมพันธ์แบบแยกส่วนประกอบกัน จะพบว่ากลุ่มตัวแปรที่มี p-value ต่ำ ซึ่งหมายความว่าเป็นกลุ่มตัวแปรที่มีนัยสำคัญในการพยากรณ์ค่า $I(t+1)$ ประกอบด้วย

$$I(t), I(t-3), I(t-5), I(t-6), I(t-7), I^{(d-1)}(t+1), RH(t), UV(t), \cos(\theta(t+1)) \tag{21}$$

ซึ่งสอดคล้องกับผลลัพธ์การสร้างสมการด้วยวิธีการถดถอยเชิงเส้นแบบขั้นตอน ซึ่งในที่นี้กำหนด p-value สูงสุดของตัวแปรต้นที่ยอมรับให้เพิ่มเข้ามาในสมการเท่ากับ 0.05 ในขั้นตอนคัดเลือกคุณลักษณะจึงได้ข้อสรุปว่าตัวแปรที่มีนัยสำคัญสำหรับการพยากรณ์ความเข้มรังสีดวงอาทิตย์ประกอบด้วย ความเข้มรังสีดวงอาทิตย์ย้อนหลังในวันเดียวกัน, ความเข้มรังสีดวงอาทิตย์ย้อนหลังในวันก่อนหน้าที่เวลาเดียวกัน , ความชื้นสัมพัทธ์ และดัชนีรังสีอัลตราไวโอเลต ส่วนตัวแปรที่มีนัยสำคัญต่ำประกอบด้วย ความเร็วลม, อุณหภูมิ และความเข้มรังสีดวงอาทิตย์ย้อนหลังในบางช่วงเวลา

4.2 ผลการปรับค่าพารามิเตอร์ของแบบจำลอง Support Vector Regression

สมรรถนะของ SVR นั้นขึ้นอยู่กับชนิดของเคอร์เนลฟังก์ชันที่เลือกใช้และพารามิเตอร์ของฟังก์ชันเคอร์เนลนั้นๆ โดยในรายงานฉบับนี้เลือกใช้ Radial-basis function (RBF) kernel เนื่องจากการคำนวณเคอร์เนลฟังก์ชันดังกล่าวสมนัยกับการคำนวณผลคูณแบบจุดของข้อมูลขาเข้าที่อยู่ในปริภูมิมิติอนันต์ทำให้สามารถรับมือกับความสัมพันธ์ที่ไม่เป็นเชิงเส้นได้ [16] โดยมีพารามิเตอร์ของฟังก์ชันเคอร์เนลดังนี้

1. สัมประสิทธ์ของฟังก์ชันลงโทษ (C) เป็นค่าที่ควบคุมความสมดุลระหว่าง การยอมรับความคลาดเคลื่อนที่มากกว่า ε ในชุดข้อมูลฝึกและความซับซ้อนของแบบจำลอง [3] การปรับค่า C ให้มีค่าน้อยเป็นการยอมให้เกิดความคลาดเคลื่อนในชุดข้อมูลฝึกได้มากซึ่งสามารถนำไปสู่เกิดปัญหาการเข้ากันระหว่างแบบจำลองและชุดข้อมูลน้อยเกินไป (Under-fitting) ในทางตรงกันข้ามการปรับค่า C ให้มีค่ามากจะเป็นการบังคับให้ฟังก์ชันวัตถุประสงค์มุ่งเน้นที่จะลด emperical risk ให้ต่ำที่สุดในชุดข้อมูลฝึกซึ่งสามารถนำไปสู่แบบเกิดปัญหาการเข้ากันระหว่างแบบจำลองและชุดข้อมูลมากเกินไป (Over-fitting)

2. สัมประสิทธ์เคอร์เนล (γ)
เป็นค่าที่แสดงถึงระยะของอิทธิพลของจุดข้อมูลฝึกหนึ่งๆ โดยถ้ากำหนดให้ γ มีค่าน้อย แสดงถึงอิทธิพลของจุดข้อมูลฝึกหนึ่งๆมีระยะไกล ในทางตรงกันข้ามถ้ากำหนดให้ γ มีค่ามาก แสดงถึงอิทธิพลของจุดข้อมูลฝึกหนึ่งๆมีระยะใกล้

3. ค่าความคลาดเคลื่อนที่ยอมรับได้ (ε)
เป็นพารามิเตอร์ที่กำหนดขนาดของบริเวณที่ยอมให้เกิดความคลาดเคลื่อนระหว่างค่าพยากรณ์จากฟังก์ชันและค่าจริง โดยการปรับค่า (ε) สูงจะเป็นการลดความแม่นยำของการพยากรณ์ในชุดข้อมูลฝึก

ในการปรับค่าพารามิเตอร์ทั้ง 3 จะเริ่มจากกำหนดให้ค่า $\gamma = 1/p$ โดยที่ p แทนจำนวนคุณลักษณะทั้งหมดของตัวแปรต้น ตามที่มีการเสนอใน [7] ซึ่งในที่นี้ $\gamma = 1/11$ และค่า $\varepsilon = 0.1$ จากนั้นปรับค่า C ระหว่างช่วง 2^{-3} ถึง 2^{9} จากผลใน *ตารางที่ 3* พบว่าสำหรับการเพิ่มค่า C ในตอนต้น สมรรถนะของแบบจำลองในชุดข้อมูลตรวจสอบจะเพิ่มขึ้นอย่างมีนัยสำคัญแต่เมื่อเพิ่มค่า C ถึงค่าหนึ่งสมรรถนะแบบจำลองในชุดข้อมูลตรวจสอบจะค่อนข้างคงที่ในขณะที่สมรรถนะแบบจำลองในชุดข้อมูลฝึกยังคงเพิ่มขึ้นอย่างต่อเนื่องจึงสรุปได้ว่าการเพิ่มค่า C ต่อไปจากค่าดังกล่าวสามารถนำไปสู่ปัญหาการเข้ากันระหว่างแบบจำลองและชุดข้อมูลมากเกินไป ดังนั้นจึงเลือกค่า $C = 16$ หลังจากนั้นปรับค่า γ ระหว่างช่วง 2^{-7} ถึง 2^{2} ในขณะที่กำหนดค่า $C = 16$ และ $\varepsilon = 0.1$จากผลใน*ตารางที่ 4* พบว่าค่า γ ที่ทำให้ RMSE ต่ำที่สุดคือ $\gamma = 0.125$ สุดท้ายปรับค่า ε ระหว่างช่วง 2^{-4} ถึง 2^{7} ผลใน*ตารางที่ 5* พบว่าสำหรับค่า ε ที่น้อยกว่า 16 การปรับค่า ε นั้นไม่ส่งผลอย่างมีนัยสำคัญต่อสมรรถนะของแบบจำลองสะท้อนให้เห็นว่าในการประมาณฟังก์ชันใดๆย่อมมีความคลาดเคลื่อนที่ไม่สามารถทำให้ลดลงได้ (irreducible error) เกิดขึ้นเสมอในที่นี้เลือกค่า $\varepsilon = 4$

สรุปค่าพารามิเตอร์ที่เลือกใช้คือ $C = 16, \gamma = 0.125$ และ $\varepsilon = 4$

ตาราง 3: สมรรถนะของแบบจำลอง SVR เมื่อปรับค่า C โดยที่ $\gamma = 1/11$ และ ε = 0.1

C	RMSE	
	training set	validation set
2^{-3}	167.4992	153.5940
2^{-2}	137.4814	131.0199
2^{-1}	121.6232	121.6037
2^{0}	113.5810	116.3607
2^{1}	109.0157	112.6498
2^{2}	106.1898	110.1800
2^{3}	104.5957	108.7586
2^{4}	103.4812	**107.9071**
2^{5}	102.6484	107.3979
2^{6}	101.8962	107.4631
2^{7}	101.0159	107.3815
2^{8}	100.1987	107.4247
2^{9}	99.2640	107.5412

ตาราง 4: สมรรถนะของแบบจำลอง SVR เมื่อปรับค่า γ โดยที่ C = 2^4 และ ε = 0.1

γ	RMSE	
	training set	validation set
2^{2}	197.1516	201.7977
2^{1}	164.1737	165.2336
2^{0}	132.5032	132.7949
2^{-1}	113.8153	115.7224
2^{-2}	106.4343	109.5702
2^{-3}	104.7495	**108.6500**
2^{-4}	104.7007	109.1461
2^{-5}	105.1972	110.0012
2^{-6}	106.2430	111.4005
2^{-7}	108.2354	113.1515

ตาราง 5: สมรรถนะของแบบจำลอง SVR เมื่อปรับค่า ε โดยที่ C = 2^4 และ γ = 2^{-3}

ε	RMSE	
	training set	validation set
2^{7}	115.3209	118.9872
2^{6}	106.2897	110.4693
2^{5}	104.9467	109.0708
2^{4}	104.7150	108.6971
2^{3}	104.7013	108.5600
2^{2}	104.6915	**108.6042**
2^{1}	104.7177	108.6412
2^{0}	104.7319	108.6539
2^{-1}	104.7444	108.6456
2^{-2}	104.7503	108.6537
2^{-3}	104.7491	108.6510
2^{-4}	104.7499	108.6464

4.3 ผลการปรับค่าพารามิเตอร์ของแบบจำลอง Random Forest

พารามิเตอร์สำคัญที่เป็นตัวกำหนดเงื่อนไขของแบบจำลอง และยังส่งต่อประสิทธิภาพ/ความซับซ้อนในการคำนวณของการพยากรณ์มีดังนี้

1. จำนวนแบบจำลองต้นไม้ทั้งหมดภายในป่า เขียนแทนด้วย n_{tree}
   เป็นพารามิเตอร์ที่ส่งผลโดยตรงต่อกำลังการคำนวณที่ใช้ในทั้งการฝึกแบบจำลองและขั้นตอนการพยากรณ์โดยยิ่ง n_{tree} มีค่ามาก แบบจำลองจะมีความแปรปรวนต่อการเปลี่ยนแปลงชุดข้อมูลฝึกลดลง แต่ในขณะเดียวกันในการคำนวณจะใช้กำลังการคำนวณที่มากขึ้น

2. จำนวนระดับหรือความลึกมากสุดของต้นไม้ที่ยอมรับได้ เขียนแทนด้วย d
   เป็นพารามิเตอร์ที่กำหนดความซับซ้อนของแบบจำลองและการเข้ากันได้ในชุดข้อมูลฝึก การปรับค่า d ให้มีค่ามากเกินไปจะนำไปสู่เกิดปัญหาการเข้ากันระหว่างแบบจำลองและชุดข้อมูลฝึกมากเกินไป (Over-fitting)

3. จำนวนตัวอย่างจากชุดข้อมูลฝึกน้อยสุดภายในปริภูมิ ที่ยินยอมให้มีการเริ่มต้นการแบ่งปริภูมิ เขียนแทนด้วย $n_{\text{min\_samples\_split}}$
   เป็นพารามิเตอร์ที่ควบคุมความสมดุลระหว่าง ค่า RMSE ในชุดข้อมูลฝึก และความซับซ้อนของแบบจำลอง การปรับค่า $n_{\text{min\_samples\_split}}$ ให้มีค่าน้อยเกินไปมีโอกาสที่จะทำให้จำนวนระดับหรือความลึกของต้นไม้มีค่ามาก และนำไปสู่เกิดปัญหาการเข้ากันระหว่างแบบจำลองและชุดข้อมูลฝึกมากเกินไป (Over-fitting)

4. จำนวนตัวอย่างจากชุดข้อมูลฝึกน้อยสุดที่ยินยอมมีในแต่ละปริภูมิย่อย เขียนแทนด้วย $n_{\text{min\_samples\_leaf}}$
   เป็นพารามิเตอร์ที่ควบคุมความสมดุลระหว่าง ค่า RMSE ในชุดข้อมูลฝึก และความซับซ้อนของแบบจำลอง การปรับค่า $n_{\text{min\_samples\_leaf}}$ ให้มีค่าน้อยเกินไปมีโอกาสที่จะทำให้จำนวนระดับหรือความลึกของต้นไม้มีค่ามาก และนำไปสู่เกิดปัญหาการเข้ากันระหว่างแบบจำลองและชุดข้อมูลฝึกมากเกินไป (Over-fitting) เป็นพารามิเตอร์ที่มีความสัมพันธ์กันกับ $n_{\text{min\_samples\_split}}$

5. จำนวนคุณลักษณะของตัวแปรต้นใช้ในแต่ละปมของแบบจำลองต้นไม้ เขียนแทนด้วย m
   เป็นพารามิเตอร์ที่มีผลกับความแตกต่างกันระหว่างแบบจำลองต้นไม้ย่อยแต่ละแบบจำลอง เมื่อเราปรับค่า m ให้น้อยลง แบบจำลองต้นไม้ย่อยแต่ละแบบจำลองจะมีความแตกต่างกันมากขึ้น แต่การเข้ากันระหว่างแบบจำลองและชุดข้อมูลฝึกจะลดลง

n_{tree} เป็นพารามิเตอร์ที่ส่งผลโดยตรงต่อกำลังการคำนวณที่ใช้ในทั้งการฝึกแบบจำลองและขั้นตอนการพยากรณ์ ซึ่งในการทดลองนี้เราจะกำหนด $n_{\text{tree}} = 1000$ จากนั้นจะทำคำนวณค่า RMSE ในชุดข้อมูลฝึกและชุดข้อมูลตรวจสอบ เมื่อปรับพารามิเตอร์ $n_{\text{min\_samples\_split}}$ และ $n_{\text{min\_samples\_leaf}}$ โดยไม่พิจารณาเงื่อนไขของ d และกำหนด $m = p = 25$

ตาราง 6: สมรรถนะของแบบจำลอง RF เมื่อปรับค่า $n_{\text{min\_samples\_split}}$ และ $n_{\text{min\_samples\_leaf}}$ โดยที่ $m = 25$

$n_{\text{min\_samples\_split}}$	$n_{\text{min\_samples\_leaf}}$	RMSE	
		training set	validation set
28	16	83.4965	105.3259
32	16	83.4965	105.3259
34	**16**	83.5903	**104.7497**
36	16	84.2358	105.2982
40	16	84.9581	105.2853
44	16	85.6133	105.2779
34	10	80.1608	104.8592
34	12	81.2877	104.8138
34	14	82.2806	104.7892
34	**16**	83.5903	**104.7497**
34	18	84.3997	104.7647
34	20	85.7072	104.6964

จากผลลัพธ์ดัง*ตารางที่ 6* พบว่าการปรับค่า $n_{\text{min\_samples\_split}}$ และ $n_{\text{min\_samples\_leaf}}$ จะส่งผลต่อค่า RMSE ในชุดข้อมูลฝึก คือ เมื่อเราเพิ่มค่า $n_{\text{min\_samples\_leaf}}$ หรือ $n_{\text{min\_samples\_leaf}}$ จะทำให้ค่า RMSE ในชุดข้อมูลฝึกมีค่าสูงขึ้นเพราะเป็นการจำกัดเงื่อนไขในการเข้ากัน (fitting) ของแบบจำลอง อย่างไรก็ตามการเพิ่มค่า $n_{\text{min\_samples\_split}}$ และ $n_{\text{min\_samples\_split}}$ เป็นการป้องกันปัญหาการเข้ากันระหว่างแบบจำลองและชุดข้อมูลมากเกินไป (Over-fitting) ซึ่งในที่นี้เราจะเลือก $n_{\text{min\_samples\_split}}$ และ $n_{\text{min\_samples\_leaf}}$ ซึ่งทำให้ค่า RMSE ในชุดข้อมูลตรวจสอบมีค่าต่ำที่สุด นั่นคือ

$$n_{\text{min\_samples\_split}} = 34, n_{\text{min\_samples\_leaf}} = 16$$

จากนั้นปรับค่า d ซึ่งเป็นพารามิเตอร์ซึ่งกำหนดความซับซ้อนของแบบจำลองได้ผลลัพธ์ตาม*ตารางที่ 7* เลือกค่า d ซึ่งทำให้ RMSE ในชุดข้อมูลตรวจสอบมีค่าต่ำที่สุด นั่นคือ $d = 10$ สุดท้ายปรับค่า m ได้ผลลัพธ์ตาม*ตารางที่ 8* และเลือกค่า m ซึ่งทำให้ RMSE ในชุดข้อมูลตรวจสอบมีค่าต่ำที่สุด นั่นคือ $m = 13$

ตาราง 7: สมรรถนะของแบบจำลอง RF เมื่อปรับค่า d โดยที่ $n_{\text{min\_samples\_split}} = 34, n_{\text{min\_samples\_leaf}} = 16, m = 25$

d	RMSE	
	training set	validation set
20	83.5903	104.7497
15	83.3568	104.7410
13	83.8632	104.7304
12	84.4242	104.7097
11	85.3068	104.6863
10	86.6353	**104.6413**
9	88.5226	104.6944
8	90.9809	104.7087
7	93.8837	104.7113

ตาราง 8: สมรรถนะของแบบจำลอง RF เมื่อปรับค่า m โดยที่ $n_{\text{min\_samples\_split}} = 34, n_{\text{min\_samples\_leaf}} = 16, d = 10$

m	RMSE	
	training set	validation set
25	86.6353	104.6413
20	87.0418	104.6347
15	87.3493	104.5573
14	87.8009	104.5927
13	87.9583	**104.4538**
12	88.1825	104.5702
11	88.4028	104.5692
10	88.6606	104.7341

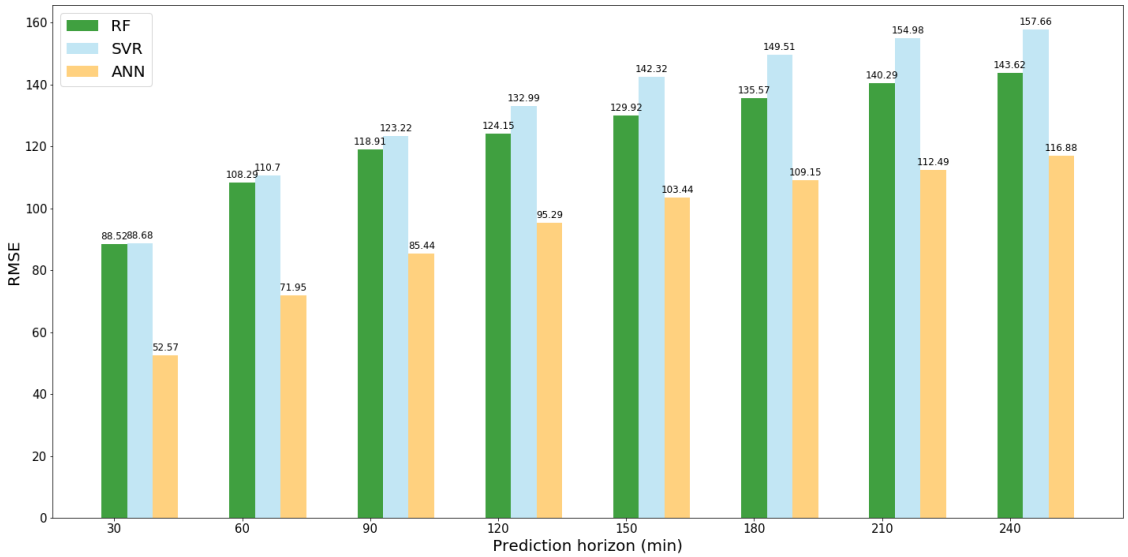
สรุปค่าพารามิเตอร์ที่เลือกคือ

$$n_{\text{tree}} = 1000, n_{\text{min\_samples\_split}} = 34, n_{\text{min\_samples\_leaf}} = 16, d = 10, m = 13 \tag{22}$$

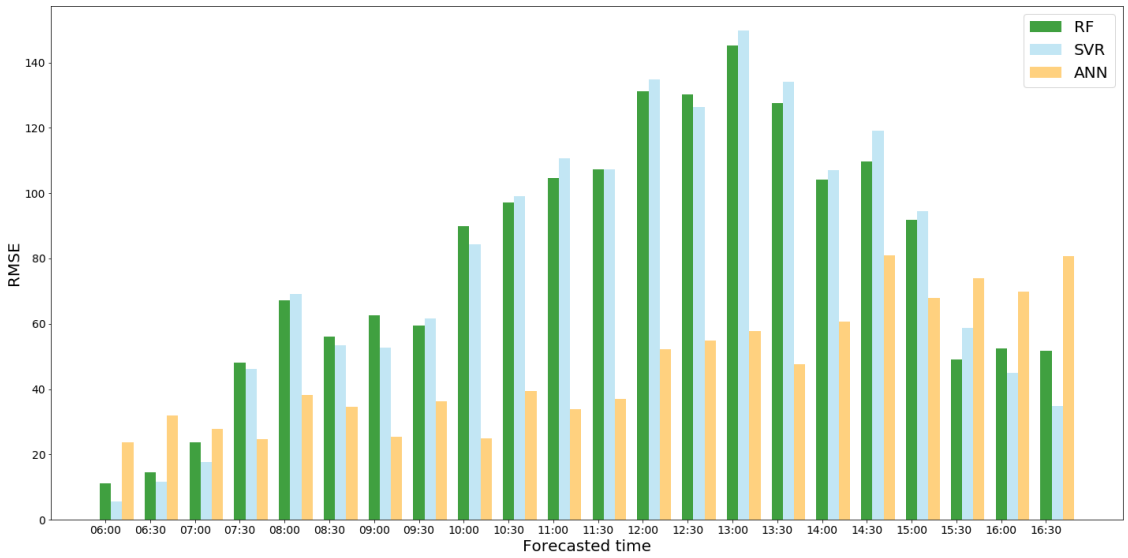
4.4 ผลการพยากรณ์ความเข้มรังสีดวงอาทิตย์

ในส่วนนี้จะนำเสนอผลลัพธ์การพยากรณ์ความเข้มรังสีดวงอาทิตย์โดยวิธี SVR, RF ซึ่งใช้ค่าพารามิเตอร์จากผลลัพธ์ในหัวข้อที่ 4.2และหัวข้อที่ 4.3 โดยจะทำการเปรียบเทียบผลการพยากรณ์ทั้ง 2 วิธี และวิธี ANN ซึ่งเป็นผลลัพธ์ที่ทางทีมวิจัยสมาร์ทกริด จุฬาลงกรณ์มหาวิทยาลัยได้จัดทำไว้โดยใช้ชุดข้อมูลเดียวกัน

จากผลลัพธ์ดัง*รูปที่ 4* จะเห็นว่าสมรรถนะของแบบจำลองทั้งสามแบบจะมีแนวโน้มลดลงเมื่อระยะการพยากรณ์ไกลขึ้นซึ่งเป็นผลลัพธ์ที่สมเหตุสมผล อย่างไรก็ตามจะสังเกตว่าสมรรถนะของแบบจำลอง SVR ลดลงมากกว่าแบบจำลอง RF เมื่อเพิ่มระยะการพยากรณ์ และจะเห็นว่าแบบจำลอง ANN ให้สมรรถนะที่ดีกว่าแบบจำลอง RF และ SVR ในทุกๆ ระยะการพยากรณ์



รูป 4: แผนภูมิแท่งเปรียบเทียบค่า RMSE ในแต่ละระยะการพยากรณ์

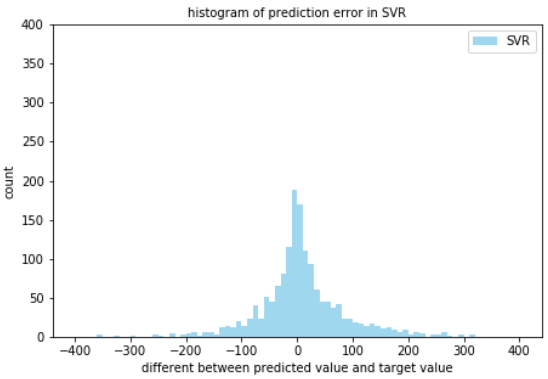


รูป 5: แผนภูมิแท่งเปรียบเทียบค่า RMSE ของการพยากรณ์ที่แต่ละจุดเวลา

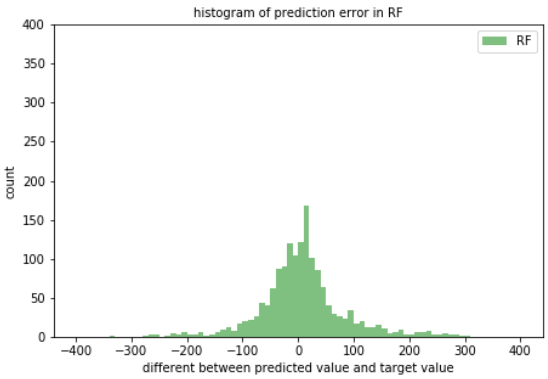
จากผลลัพธ์ดัง*รูปที่ 5* จะเห็นว่าการพยากรณ์ความเข้มรังสีดวงอาทิตย์ด้วยแบบจำลอง RF ให้ค่า RMSE ต่ำกว่าแบบจำลอง SVR ในช่วงเวลา 11:00 น. - 15.30 น. ซึ่งเป็นช่วงเวลาที่ค่าความเข้มรังสีดวงอาทิตย์มีความผันผวนมาก แต่ในทางกลับกัน เมื่อพิจารณาในช่วงเช้ามืดและหัวค่ำแบบจำลอง SVR มีแนวโน้มที่จะให้ค่า RMSE ต่ำกว่าแบบจำลอง RF ทั้งนี้ผลลัพธ์ที่น่าสนใจคือจะสังเกตว่าค่า RMSE จากวิธี ANN ในเวลาเช้ามืดและเวลาเย็นมีค่าสูง ซึ่งในเวลาดังกล่าวควรจะเป็นเวลาที่ค่าความเข้มรังสีดวงอาทิตย์มีความผันผวนน้อย

จากผลลัพธ์ดัง*รูปที่ 6* จะเห็นว่าในแต่ละแบบจำลองมีการกระจายตัวของความคลาดเคลื่อนแตกต่างกันโดยจะเห็นได้ว่าแบบจำลอง ANN มีการกระจายตัวของที่แคบสอดคล้องกับผลลัพธ์จากกราฟก่อนหน้าที่จะเห็นว่าแบบจำลอง ANN เป็นวิธีที่พยากรณ์ได้อย่างแม่นยำที่สุด ส่วนแบบจำลอง RF มีช่วงที่ค่าความคลาดเคลื่อนเกิดขึ้นเยอะที่สุดมีค่ามากกว่า 0 สะท้อนถึงการที่แบบจำลองมักจะพยากรณ์สูงกว่าค่าความเป็นจริงในขณะที่แบบจำลอง ANN และ SVR ช่วงที่ค่าความคลาดเคลื่อนเกิดขึ้นเยอะที่สุดมีค่าประมาณ 0

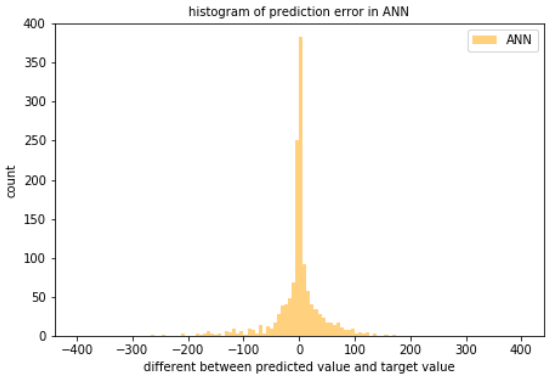
จากผลลัพธ์ดัง*รูปที่ 7* ความคลาดเคลื่อนจากแบบจำลอง RF และ SVR ในช่วงเวลา 11:00 น. - 15.00 น. มีช่วงการกระจายตัวที่กว้างกว่าในเวลาเช้ามืดหรือเวลาเย็น ซึ่งสอดคล้องกับช่วงเวลาที่ค่าความเข้มรังสีดวงอาทิตย์มีความผันผวนมาก นอกจะนี้ยังสังเกตได้ว่าค่าความผิดพลาดสำหรับ แบบจำลอง RF ในช่วงเวลาเช้ามืดและเวลาเย็นจะมีแนวโน้มที่จะมีค่ามากกว่าศูนย์ กล่าวคือแบบจำลองมีโน้วโน้มที่จะประมาณค่ารังสีดวงอาทิตย์เกินกว่าความเป็นจริงในช่วงเวลาดังกล่าว ในขณะที่แบบจำลอง SVR ไม่พบปัญหาดังกล่าว ส่วนผลลัพธ์จากแบบจำลอง ANN ความคลาดเคลื่อนมีแนวโน้มกระจายตัวสูงขึ้นในเวลาบ่ายถึงเย็น ส่วนในช่วงเช้ามีการกระจายตัวที่ต่ำมากเมื่อเทียบกับแบบจำลองที่เหลือ



(a) การกระจายตัวของค่าความคลาดเคลื่อนใน SVR

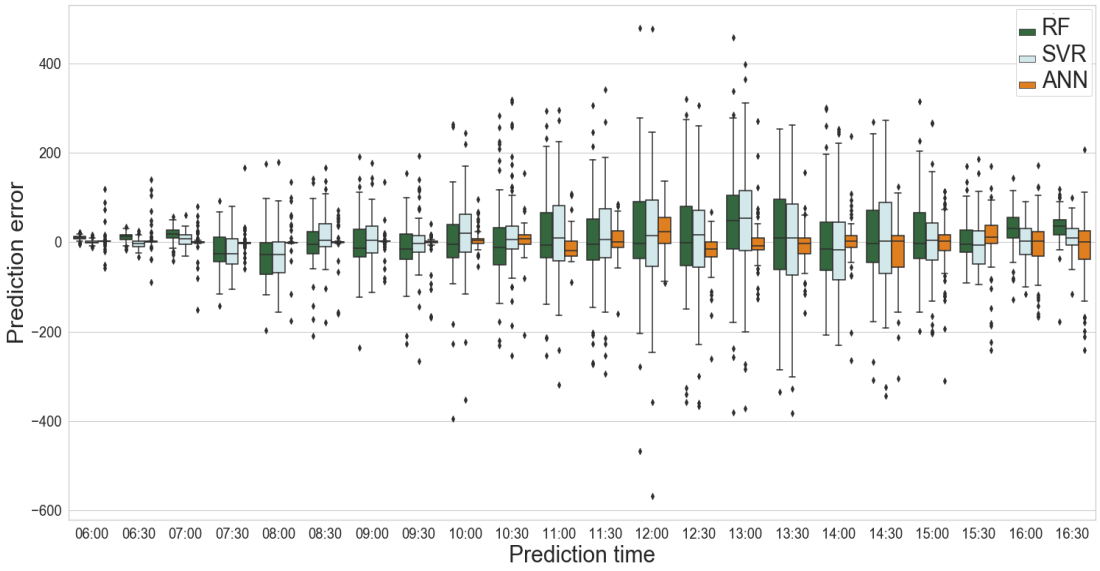


(b) การกระจายตัวของค่าความคลาดเคลื่อนใน RF



(c) การกระจายตัวของค่าความคลาดเคลื่อนใน ANN

รูป 6: แผนภาพฮิสโตแกรมแสดงการกระจายตัวของค่าความผิดพลาดในแต่ละแบบจำลอง



รูป 7: แผนภาพกล่องแสดงการกระจายตัวของค่าความผิดพลาดในการพยากรณ์ที่แต่ละจุดเวลา

4.5 ผลการเปรียบเทียบความซับซ้อนในการคำนวณ

เมื่อกำหนดให้ n แทนจำนวนจุดข้อมูลทั้งหมดของชุดข้อมูลฝึก, p แทนจำนวณคุณลักษณะทั้งหมดของตัวแปรต้น

Support Vector Regression

ความซับซ้อนของการคำนวณในขั้นตอนฝึกของ non-linear kernel SVR มาจาก 2 ส่วนหลัก คือ 1.การแก้สมการเชิงเส้นของอนุพันธ์ของฟังก์ชันจุดประสงค์ซึ่งมีความซับซ้อนในการคำนวณเป็น $\mathcal{O}(R^3)$ 2.การคำนวณ gradient ของ dual เพื่อยืนยันเงื่อนไขค่าสุดขีดโดยมีความซับซ้อนในการคำนวณเป็น $\mathcal{O}(nS)$ เมื่อกำหนดให้ R คือจำนวนของ free support vectors ซึ่งมีค่าเท่ากับจำนวนของสมาชิกในเวกเตอร์ตัวคูณลากรานจ์ที่มีค่าอยู่ในช่วง $(0, C)$ และ S คือจำนวนของ support vectors ซึ่งมีค่าเท่ากับจำนวนของสมาชิกในเวกเตอร์ตัวคูณลากรานจ์ที่มีค่าอยู่ในช่วง $(0, C]$ จะเห็นว่าความซับซ้อนในการคำนวนในขั้นตอนนี้ขึ้นอยู่กับจำนวนของ support vectors และเนื่องจากจำนวนของ support vectors เพิ่มขึ้นอย่างเป็นเชิงเส้นตามจำนวนชุดข้อมูลฝึกที่เพิ่มขึ้น จึงสามารถกล่าวได้ว่าความซับซ้อนของการคำนวณในขั้นตอนฝึกมีค่าอยู่ระหว่าง $\mathcal{O}(n^2)$ ถึง $\mathcal{O}(n^3)$ [5]

สำหรับในการพยากรณ์จาก (10) จะเห็นว่าความซับซ้อนในการคำนวณมาจากจำนวนของ support vectors และการคำนวณค่าของฟังก์ชันเคอร์เนล ดังนั้นความซับซ้อนในการคำนวณในขั้นตอนนี้จึงขึ้นอยู่กับชนิดของฟังก์ชันเคอร์เนลที่เลือกใช้

Random forest

ความซับซ้อนของการคำนวณในขั้นตอนฝึกขึ้นอยู่กับปัจจัยหลักคือ n_{tree} และ n โดยเราสามารถแสดงได้ว่าความซับซ้อนในการคำนวณในขั้นตอนฝึกกรณีที่แย่ที่สุดเท่ากับ $\mathcal{O}(n_{\text{tree}}mn^2 \log n)$ [13] และเมื่อเพิ่มเงื่อนไขจำนวนระดับหรือความลึกมากสุดของต้นไม้ที่ยอมรับได้ (d) ซึ่งเป็นการลดความซับซ้อนของแบบจำลอง ความซับซ้อนของการคำนวณในกรณีที่แย่ที่สุดจะมีค่าลดลงเหลือ $\mathcal{O}(n_{\text{tree}}mdn \log n)$ นอกจากนี้หากใช้วิธี bootstrap ร่วมด้วยในขั้นตอนฝึกความซับซ้อนในการคำนวณจะลดลงเหลือ $\mathcal{O}(n_{\text{tree}}md\tilde{n} \log \tilde{n})$ ($\tilde{n}$ คือจำนวนจุดข้อมูลที่ใช้ในขั้นตอนฝึกสำหรับวิธี bootstarp โดยทั่วไป $\tilde{n} \approx 0.632n$) [13] สำหรับในขั้นตอนการพยากรณ์ดังจะเห็นว่าการคำนวณค่าพยากรณ์ในแบบจำลองต้นไม้แต่ละแบบจำลองเป็นการตรวจสอบเงื่อนไขของตัวแปรต้นในแต่ละระดับความลึกของต้นไม้ (d) ดังนั้นความซับซ้อนของการคำนวณในขั้นตอนทำนาย กรณีที่แย่ที่สุดเท่ากับ $\mathcal{O}(n_{tree}d)$

5 รายละเอียดหัวข้อโครงงาน

5.1 ขอบเขตของโครงงาน

1. การทดลองหลักจะทดลองบนข้อมูลที่วัดได้ ณ ชั้นดาดฟ้าตึกภาควิศวกรรมไฟฟ้า จุฬาลงกรณ์มหาวิทยาลัย ในช่วงเวลาตั้งแต่ เดือนมกราคม พ.ศ. 2560 จนถึงเดือนธันวาคม พ.ศ. 2561 ซึ่งประกอบด้วย ค่ารังสีดวงอาทิตย์ต่อพื้นที่, ค่ากำลังไฟฟ้า, ค่าความชื้นสัมพัทธ์, อุณหภูมิ, ความเร็วลม, ดัชนีรังสีอัลตราไวโอเลต (UV Index) โดยข้อมูลทั้งหมดถูกลดอัตราสุ่มลงเป็น 30 นาที ส่วนข้อมูลสำรองเป็นข้อมูลที่วัดได้จากโรงไฟฟ้าในภาคกลางจำนวน 1 โรง ในช่วงเวลาตั้งแต่ เดือนมกราคม พ.ศ. 2560 จนถึง เดือนธันวาคม พ.ศ. 2561 ซึ่งประกอบด้วย ค่ารังสีดวงอาทิตย์ต่อพื้นที่, ค่ากำลังไฟฟ้า, อุณหภูมิ โดยข้อมูลทั้งหมดถูกลดอัตราสุ่มลงเป็น 30 นาที (จะมีผลการทดลองเมื่อมีข้อมูลเพียงพอ)

2. วิเคราะห์หาตัวแปรต่างๆที่ส่งผลต่อกำลังผลิตไฟฟ้าจากเซลล์แสงอาทิตย์ ด้วยวิธีการคัดเลือกคุณลักษณะ ได้แก่ การวิเคราะห์สหสัมพันธ์, การวิเคราะห์สหสัมพันธ์แบบแยกส่วน และการถดถอยเชิงเส้นแบบขั้นตอน

3. เปรียบเทียบผลลัพธ์จากพยากรณ์ความเข้มรังสีดวงอาทิตย์ล่วงหน้า 4 ชั่วโมง (ค่าผลลัพธ์การพยากรณ์มีความละเอียด 30 นาที กล่าวคือจะพยากรณ์ 30, 60, 90, ..., 240 นาทีล่วงหน้า) โดยพยากรณ์ในช่วงเวลา 5:30 น. ถึง 17.30 น. (พยากรณ์ทุกๆ 30 นาที) เพื่อให้ได้ค่าพยากรณ์ในช่วงเวลาตั้งแต่ 6:00 น. ถึง 18.00 น.

4. การเปรียบเทียบแบบจำลองจะพิจารณากลุ่มแบบจำลองอันได้แก่ 1) Linear Regression 2) Multivariate Adaptive Regression Splines (MARS) 3) Support Vector Regression 4) Random Forest โดย 2 วิธีแรก จัดทำขึ้นเพื่อเป็นแบบจำลองฐาน (baseline model)

5. เปรียบเทียบผลลัพธ์การพยากรณ์กับผลลัพธ์จากแบบจำลอง ANN ซึ่งทีมวิจัยสมาร์ทกริด จุฬาลงกรณ์มหาวิทยาลัยได้จัดทำขึ้น

6. ใช้แบบจำลองการแปลงความเข้มรังสีดวงอาทิตย์ไปเป็นกำลังผลิตไฟฟ้าจากเซลล์แสงอาทิตย์

5.2 แผนการดำเนินงาน

สำหรับแผนการดำเนินงานจะแบ่งออกเป็นสองส่วนใหญ่ๆ คือส่วนของแบบจำลอง SVR และแบบจำลอง RF โดยนายชญานนท์ โพธาวนานนท์ จะเป็นผู้รับผิดชอบการทดลองตลอดจนรายงานผลลัพธ์ในส่วนของแบบจำลอง SVR และนายสรารุต พรานนท์สถิตย์จะเป็นผู้รับผิดชอบในส่วนของแบบจำลอง RF สำหรับการจัดทำแบบจำลองฐาน (baseline model) , ผลลัพธ์การเปรียบเทียบและวิเคราะห์ผลจะเป็นการทำงานร่วมกัน

ขั้นตอนการดำเนินงาน	วิชา 2102490				วิชา 2102499				
	ส.ค.	ก.ย.	ต.ค.	พ.ย.	ธ.ค.	ม.ค.	ก.พ.	มี.ค.	เม.ย.
1.ศึกษาภาพรวมของหลักการการพยากรณ์กำลังผลิตไฟฟ้าจากเซลล์แสงอาทิตย์ และตัวชี้วัดสมรรถนะของการพยากรณ์									
2.ทำการคัดเลือกคุณลักษณะที่จะใช้ในการพยากรณ์โดยการทำ correlation analysis, partial correlation analysis และ stepwise regression									
3.ทำ pre-analysis data ศึกษาสมบัติทางสถิติของข้อมูล และ pre-process data เช่นจัดการกับ outlier และข้อมูลที่ขาดหาย และ data augmentation									
4.ทำการทดลองสร้างแบบจำลองฐานโดยใช้ persistence model									
5. ทดลองใช้แบบจำลอง SVR และ Random Forest และปรับค่า parameter เพื่อหาค่าที่ดีที่สุดในแต่ละแบบจำลอง									
6.นำผลการทดลองจากแบบจำลอง ANN ที่ทางทีมวิจัยสมาร์ทกริดจุฬาลงกรณ์ได้จัดทำไว้มาปรับเทียบเพื่อเปรียบเทียบกับผลการทดลองที่มี									
7.วิเคราะห์และเปรียบเทียบผลการทดลองโดยคำนวณจากตัวชี้วัดสมรรถนะการพยากรณ์									
8.ศึกษาเกี่ยวกับกำลังในการคำนวณของแบบจำลอง SVR และ Random Fores ทั้งในส่วนการเรียนรู้แบบจำลองและการคำนวณค่าพยากรณ์									
9.จัดทำรูปเล่มและเตรียมนำเสนอโครงร่างโครงงาน									
10.ทดลองใช้แบบจำลองแปลงค่าความเข้มรังสีดวงอาทิตย์เป็นค่ากำลังผลิตไฟฟ้า									
11.ทำการทดลองสร้างแบบจำลองฐานโดยใช้วิธี Linear regreesion และ Mutivarate adaptive regression spline									
12.ทดลองปรับเปลี่ยน Configuration ในการทดลองโดยใช้แบบจำลองเฉพาะสำหรับการพยากรณ์ค่าในแต่ละช่วงเวลาและทำ feature extraction									
13.ทดลองการพยากรณ์โดยใช้ชุดข้อมูลจากโรงไฟฟ้าในภาคกลาง									
14.วิเคราะห์สรุปผลเตรียมการนำเสนอโครงงาน									

รูป 8: Gantt Chart แสดงแผนการดำเนินงานตลอดทั้งปีการศึกษา 2562

6 เอกสารอ้างอิง

[1] W.M. Ahmad, M. Mourshed, and Y. Rezgu. Tree-based ensemble methods for predicting PV power generation and their comparison with support vector regression. *Energy*, 164:465–474, 2018.

[2] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F.J. Martinez de Pison, and F. Antonanzas-Torres. Review of photovoltaic power forecasting. *Solar Energy*, 136:78–111, 2016.

[3] S. Bernhard and S. Alexander J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[4] W. Björn, L. Elke, and K. Oliver. Statistical learning for short-term photovoltaic power predictions. In *Computational sustainability*, pages 31–45. Springer, 2016.

[5] L. Bottou and C. Lin. Support vector machine solvers. *Large scale kernel machines*, 3(1):301–320, 2007.

[6] M. Bouzerdoum, A. Mellit, and A.M. Pavan. A hybrid model (sarima–svm) for short-term power forecasting of a small-scale grid-connected photovoltaic plant. *Solar Energy*, 98:226–235, 2013.

[7] C. Chang and C. Lin. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

[8] C. Corinna and V. Vladimir. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[9] M. Francisco, T. Alicia, C. Gualberto, and R. José. A survey on data mining techniques applied to energy time series forecasting. *Energies*, In press, 11 2015.

[10] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[11] R.H. Inman, H. Pedro, and C.F.M. Coimbra. Solar forecasting methods for renewable energy integration. *Progress in energy and combustion science*, 39(6):535–576, 2013.

[12] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to statistical learning*, volume 112. Springer, 2013.

[13] G. Louppe. Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*, 2014.

[14] R. Mashud, K. Irena, and A. Vassilios G. Univariate and multivariate methods for very short-term solar photovoltaic power forecasting. *Energy Conversion and Management*, 121:380–390, 2016.

[15] M. Samanta, B. Srikanth, and J. Yerrapragada. Short-term power forecasting of solar pv systems using machine learning techniques,(nd).

[16] A.J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[17] S. Vagropoulos, G. Chouliarasand . Kardakosand, C. Simoglou, and A. Bakirtzis. Comparison of SARIMAX, SARIMA, modified SARIMA and ANN-based models for short-term PV generation forecasting. In *2016 IEEE International Energy Conference (ENERGYCON)*, pages 1–6. IEEE, 2016.

[18] V. Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[19] R. Xu, H. Chen, and X. Sun. Short-term photovoltaic power forecasting with weighted support vector machine. In *2012 IEEE International Conference on Automation and Logistics*, pages 248–253. IEEE, 2012.