

การพัฒนาแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่องเพื่อพยากรณ์พื้นที่ในประเทศ
ไทยที่มีความเสี่ยงได้รับผลกระทบจากการเกิดวาตภัย

THE MODEL DEVELOPMENT OF A LARGE-SCALE PREDICTION
SYSTEM FOR WINDSTORM IN THAILAND THROUGH MACHINE
LEARNING TECHNIQUE

อนุตริดา กลิ่นหอม¹

เอกสิทธิ์ พัทธวงษ์ศักดิ์²

บทคัดย่อ

การเกิดพายุและปรากฏการณ์สภาพอากาศที่มีความรุนแรงก่อให้เกิดการเสียชีวิต บาดเจ็บ เกิดความเสียหายต่อทรัพย์สินที่สำคัญ และเกิดความล่าช้าในการค้าขายและการคมนาคม งานวิจัยนี้จึงมีวัตถุประสงค์เพื่อพัฒนาแอปพลิเคชันโดยประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่องเพื่อพยากรณ์พื้นที่เสี่ยงได้รับผลกระทบจากวาตภัยที่เกิดในประเทศไทย และแสดงข้อมูลพายุลูกเก่าๆ ที่เหมือนกับพายุลูกล่าสุด จากข้อมูลการแจ้งเตือนวาตภัย ข้อมูลประวัติการเกิดวาตภัย และข้อมูลสภาพอากาศ ผลการวิจัย ณ ตอนนี้นำเสนอให้เห็นว่า การพยากรณ์โดยใช้ข้อมูลสภาพอากาศปัจจุบันร่วมกับข้อมูลสภาพอากาศล่วงหน้าหนึ่งวัน จำนวน 11 Features ที่ Learning Rate 0.01 Training Cycle 160 มีความถูกต้องของการแจ้งเตือนพายุ 66.77% ไม่แจ้งเตือนพายุ 74.24% Accuracy ที่ 70.21 และสามารถเปรียบเทียบข้อมูลประวัติการเกิดวาตภัย กับวาตภัยลูกปัจจุบันที่มีความคล้ายกันมากที่สุด 5 อันดับแรก

คำสำคัญ: การพยากรณ์วาตภัย ความเหมือนของวาตภัย การเรียนรู้ของเครื่อง โครงข่ายประสาทเทียม

ABSTRACT

The occurrence of storms and other severe weather phenomena can result in fatalities, injuries, damage to significant properties, and delays in trade and transportation. This research aims to develop an application using the machine learning techniques to forecast areas which are prone to storm risk in Thailand. The research also attempts to show similarities between previous storms and the latest ones based on the alert data of storm events, historical data of storm events, and weather information. The results revealed that the forecast combining current weather information with weather information of one day in advance, which includes 11 features at Learning Rate 0.01 and Training Cycle 160, generates 66.77% class

recall of storm warnings and 74.24% of no storm warnings at 70.21 accuracy. The model can also compare the historical data of storm events with the top five current ones which bear some similarities.

Keywords: Windstorm prediction, Storm Similarity, Machine learning, Neural Network

บทนำ

วาตภัยหรือพายุ (Windstorm) คือปรากฏการณ์ทางธรรมชาติที่เกิดจากการเปลี่ยนแปลงทางสภาพอากาศที่รุนแรง ที่มักก่อให้เกิดผลกระทบต่อทรัพยากรทางธรรมชาติ สิ่งปลูกสร้าง และสิ่งมีชีวิต เกิดจากการเคลื่อนที่ของลมหรือมวลอากาศ ซึ่งเกิดจากความต่างกันของอุณหภูมิบรรยากาศโดยรอบ โดยกระแสลมจะพัดเข้าหาจุดศูนย์กลางของบริเวณดังกล่าว เนื่องจากมวลอากาศลอยตัวขึ้น ส่งผลให้มวลอากาศด้านข้างที่มีอุณหภูมิต่ำกว่าไหลเข้ามาแทนที่ ทำให้เกิดการหมุนของอากาศ ก่อนพัฒนาไปเป็นพายุ

ในหลายปีที่ผ่านมา นักอุตุนิยมวิทยาและศูนย์เตือนภัยจำนวนมากทุ่มเทให้กับการศึกษานี้และมีความก้าวหน้าในเทคโนโลยีสำหรับใช้สังเกตการณ์ที่ดีขึ้น อย่างไรก็ตามเทคโนโลยีที่ช่วยในการแจ้งเตือนภัยในประเทศไทยมีการใช้งานยังไม่แพร่หลาย หรือยังอยู่ในช่วงที่กำลังพัฒนา

งานวิจัยนี้จึงมีวัตถุประสงค์เพื่อพัฒนาแอปพลิเคชันโดยใช้แบบจำลองทาง Machine Learning ในการทำนายพื้นที่เสี่ยงได้รับผลกระทบจากพายุที่เกิดในประเทศไทย และแสดงข้อมูลพายุลูกเก่าๆ ที่เหมือนกับพายุลูกล่าสุด โดยการทำงานสามารถแบ่งได้เป็น 3 ส่วน ส่วนแรก การเตรียมข้อมูล โดยใช้ข้อมูลจากกรมอุตุนิยมวิทยา และกรมป้องกันและบรรเทาสาธารณภัย ส่วนที่สองการสร้างแบบจำลองโครงข่ายประสาทเทียม (Neural Network) เพื่อทำนายพื้นที่เสี่ยงได้รับผลกระทบและเปรียบเทียบพายุโดยใช้การหาระยะทางแบบยูคลิด (Euclidean Distance) ส่วนสุดท้ายสร้างเว็บแอปพลิเคชันในการแสดงผล

ผลที่ได้จากงานวิจัยทำให้สามารถเฝ้าระวังผลกระทบจากพายุในตำบลต่างๆ ของประเทศไทย เพื่อลดความเสียหายต่อทรัพยากรธรรมชาติ สิ่งปลูกสร้าง ผลผลิตทางการเกษตร ชีวิตและทรัพย์สินของประชาชนในพื้นที่เสี่ยงที่จะได้รับผลกระทบ

วัตถุประสงค์ของงานวิจัย

1. ทำนายพื้นที่ในประเทศไทย ที่มีความเสี่ยงได้รับผลกระทบจากพายุที่เกิด โดยแสดงในรูปแบบการแจ้งเตือนพายุในระดับตำบล
2. เปรียบเทียบเพื่อหาความเหมือนระหว่างพายุลูกล่าสุดกับพายุลูกเก่าที่เคยเกิดขึ้นไปแล้ว

ขอบเขตงานวิจัย

1. ใช้ชุดข้อมูลต่างๆดังนี้
 - 1) ชุดข้อมูลสภาพอากาศจากกรมอุตุนิยมวิทยา ตั้งแต่วันที่ 2 มกราคม พ.ศ. 2561 ถึงวันที่ 22 มกราคม พ.ศ. 2564
 - 2) ชุดข้อมูลแจ้งเตือนภัยจากกรมป้องกันและบรรเทาสาธารณภัย (ปภ.) ตั้งแต่วันที่ 8 มกราคม พ.ศ. 2559 ถึงวันที่ 29 พฤศจิกายน พ.ศ. 2563
 - 3) ชุดข้อมูลประวัติการเกิดวาทภัยจากกรมอุตุนิยมวิทยา ตั้งแต่วันที่ 1 มกราคม พ.ศ. 2561 ถึงวันที่ 15 พฤศจิกายน พ.ศ. 2563
2. การแจ้งเตือนภัยให้กับตำบลในประเทศไทยโดยแสดงเป็นแผนที่ความรุนแรง
3. การพยากรณ์พื้นที่ผลกระทบโดยใช้ข้อมูลสภาพอากาศล่วงหน้าหนึ่งวัน
4. แสดงรายชื่อวาทภัยที่เหมือนกับวาทภัยลูกปัจจุบันที่สุด 5 อันดับแรก

ประโยชน์ที่คาดว่าจะได้รับ

1. เพื่อช่วยให้ตำบลต่างๆ ในประเทศไทยมีการเฝ้าระวังพายุที่กำลังจะเข้ามาในประเทศ
2. เพื่อเป็นแนวทางสำหรับต่อ ยอดการพัฒนาแบบจำลองที่เกี่ยวข้องกับวาทภัยหรือภัยอื่นๆ
3. เพื่อเป็นแนวทางในการศึกษาปัจจัยต่างๆของพายุที่มีความคล้ายกัน

ทฤษฎี และผลงานวิจัยที่เกี่ยวข้อง

1. เทคนิคการเรียนรู้ของเครื่อง (Machine Learning)

คือระบบที่สามารถเรียนรู้จากข้อมูลที่มีอยู่ได้ด้วยตนเอง ประกอบด้วยข้อมูลและเครื่องมือทางสถิติเพื่อช่วยพัฒนากระบวนการแก้ปัญหา และทำนายผลลัพธ์ออกมาอย่างเหมาะสม

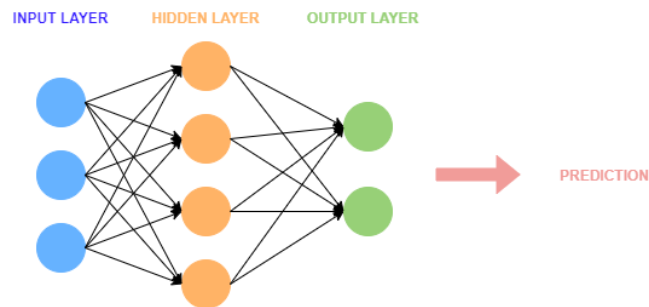
Machine Learning จะทำงานคล้ายกับการเรียนรู้ของมนุษย์ โดยการป้อนชุดข้อมูลและชุดคำสั่งให้คอมพิวเตอร์ “เรียนรู้” เพื่อจำแนกแยกแยะวัตถุต่างๆ รวมถึงบุคคล สิ่งของ ฯลฯ

2. โครงข่ายประสาทเทียม (Neural Network)

โครงข่ายประสาทเทียม (Neural Network) เป็นอัลกอริทึมหลักของการเรียนรู้เชิงลึก (Deep Learning) ซึ่งเป็นแนวทางการเรียนรู้แบบหนึ่งของเทคนิคการเรียนรู้ของเครื่อง (Machine Learning)

การทำงานของโครงข่ายประสาทเทียมจะเป็นการจำลองการทำงานมาจากสมองของมนุษย์โดยตรง ซึ่งสมองของมนุษย์นั้นจะมีหน่วยประมวลผลมากมายที่เชื่อมกันด้วยระบบประสาทช่วยให้มนุษย์สามารถจดจำ วิเคราะห์และตัดสินใจได้อย่างรวดเร็ว แต่ในส่วนของคอมพิวเตอร์มีหน้าที่เพียงแค่รับข้อมูล

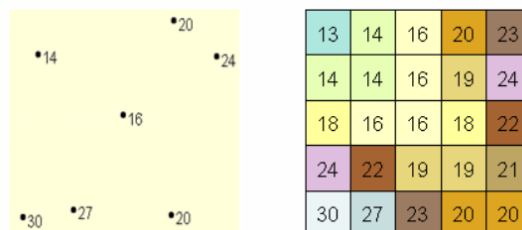
และรันโปรแกรมตามคำสั่งที่มนุษย์ป้อนให้ ดังนั้นจึงเกิดการจำลองแนวทางการเรียนรู้ของมนุษย์ไปสู่คอมพิวเตอร์ด้วย Neural Network ขึ้นมา



ภาพที่ 1: แสดงการทำงานของโครงข่ายประสาทเทียม

3. ทฤษฎีการประมาณค่าในช่วง (Interpolation)

ใช้หลักการของ IDW หรือ Inverse Distance Weight เป็นวิธีการประมาณค่าโดยจะทำการสุ่มจุดตัวอย่างจากข้อมูลที่มี ซึ่งเป็นจุดที่สามารถส่งผลกระทบไปยังเซลล์ที่ต้องการประมาณค่าได้ ซึ่งจุดผลกระทบนี้จะมีผลกระทบน้อยลงตามระยะทางที่อยู่ห่างจากจุดที่ต้องการประมาณค่าไปเรื่อยๆ เหมาะกับตัวแปรที่ระยะทางมีผลต่อการคำนวณ ซึ่งจุดยิ่งใกล้กันจะยิ่งมีอิทธิพลมาก เช่น ความสูง ปริมาณน้ำฝน การกระจายตัวของสารเคมี ระดับเสียงรบกวน เป็นต้น และในงานวิจัยนี้จะเป็นการทำนายค่าสภาพอากาศต่างๆ ตามตำบลในประเทศไทย

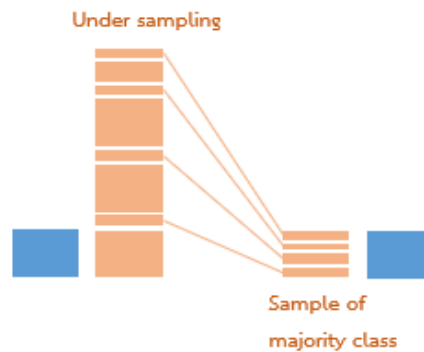


ภาพที่ 2: แสดงการหาค่าจากจุดที่ทราบค่า โดยค่าที่ไม่รู้จักจะถูกทำนายด้วยสูตรทางคณิตศาสตร์

ที่มา: desktop.arcgis.com

4. การสุ่มลด (Under sampling)

เป็นหนึ่งในวิธีการจัดการข้อมูลโดยจะทำการลดจำนวนข้อมูลที่มีจำนวนมากให้มีจำนวนใกล้เคียงกับข้อมูลที่มีจำนวนน้อย ถูกนำมาใช้บ่อยครั้งเมื่อพบข้อมูลที่มีความไม่สมดุล (Imbalance Data) หรือข้อมูลมีจำนวนต่างกันมากๆ



ภาพที่ 3: แสดงการสุ่มลด (Under sampling)

5. Features Selection (Weight by chi-square)

Features selection คือการเลือกคุณลักษณะหรือเลือก Features ที่เหมาะสมเพื่อลดขนาดหรือมิติของข้อมูล กรณีที่มี Features เยอะเกินไปเครื่องอาจจะต้อง Fit Model เทียบกับ Features จำนวนมากทำให้ใช้เวลาานพอสมควร หรือบางครั้งอาจมี Features บางตัวที่ไม่จำเป็นต้องนำมาสร้างแบบจำลอง จึงต้องมีการคัดเลือก Features ให้เหมาะสมเพื่อลดระยะเวลาการประมวลผล และเพื่อไม่ให้ model over fitting จนเกินไป การคัดเลือกคุณลักษณะจึงเป็นหัวใจสำคัญในการทำเหมืองข้อมูล (Data Mining) หรือการเรียนรู้ของเครื่อง (Machine Learning) ซึ่งการคัดเลือกลักษณะนี้มีหลายวิธีมาก

เทคนิค Chi-Square เป็นเทคนิคการคำนวณค่าน้ำหนักของแต่ละคุณลักษณะโดยใช้ค่า Chi-Square โดยจะทำการดูความถี่ที่เกิดขึ้นระหว่างแอตทริบิวต์ต่างๆ เทียบกับแอตทริบิวต์ลาเบล (Label) โดยค่า Chi-Square คำนวณได้จาก

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

χ^2 = ค่าสถิติไคสแควร์

O_i = ความถี่ที่ได้จากการสังเกต (Observed Frequency)

E_i = ความถี่ที่คาดหวัง (Expected Frequency) ซึ่งมีค่าเท่ากับจำนวนข้อมูลคูณด้วยสัดส่วนที่คาดหวัง ซึ่ง ($E_i = np_i$)

6. การวัดประสิทธิภาพแบบจำลองจาก Confusion Matrix

การจะนำแบบจำลองไปใช้จริง จำเป็นต้องมีการวัดประสิทธิภาพของแบบจำลองก่อนว่าแบบจำลองที่สร้างขึ้นมานั้นมีประสิทธิภาพเพียงพอที่จะนำมาพัฒนาต่อหรือใช้งานจริงหรือไม่ ซึ่งส่วนใหญ่นั้นจะใช้ Confusion Matrix ในการวัดประสิทธิภาพ

Confusion Matrix คือตารางที่มีแนวคิดมาจากสิ่งที่แบบจำลองทำนายได้กับสิ่งที่เกิดขึ้นจริงนั้นมีส่วนเป็นอย่างไร

		<u>Actual values</u>	
		Positive	Negative
<u>Predicted values</u>	Positive	TP	FP
	Negative	FN	TN

ภาพที่ 4: แสดงตัวอย่างตาราง Confusion Matrix

7. การหาความเหมือน (Similarity) โดยวิธีการหาระยะทางแบบยูคลิด (Euclidean Distance)

การหาความเหมือนมีหลายวิธี ซึ่งในงานวิจัยฉบับนี้จะใช้วิธีการ **Euclidean Distance** ในการหาค่าความเหมือนของตัวแปร

Euclidean Distance คือระยะทางปกติระหว่างจุดสองจุดในแนวเส้นตรง เป็นค่ารากที่สองของผลรวมของผลต่างกำลังสองระหว่างเวกเตอร์ทั้งสอง โดยที่

q_i, p_i = จุดของข้อมูล

n = จำนวนมิติ

$$d = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$

ระเบียบวิธีวิจัย

1. ข้อมูลที่ใช้ในงานวิจัย

1.1 ข้อมูลสภาพอากาศจากกรมอุตุนิยมวิทยาซึ่งมีลักษณะของข้อมูลดังนี้

- 1) เป็นข้อมูลรายชั่วโมงจำนวน 2,338,065 ตัวอย่าง ซึ่งเก็บข้อมูลตั้งแต่ 2 มกราคม 2561 ถึง 22 มกราคม 2564

- คุณลักษณะที่สนใจได้แก่ ทิศทางลม, ความเร็วลม, อุณหภูมิ, ความกดอากาศและความชื้น

1.2 ข้อมูลแจ้งเตือนภัยจาก กรมป้องกันและบรรเทาสาธารณภัย(ปภ.) โดยมีลักษณะของข้อมูลดังนี้

- เป็นข้อมูลรายวันจำนวนทั้งหมด 17,255 ตัวอย่าง ซึ่งจะเก็บเฉพาะช่วงวันที่มีการแจ้งเตือนภัยในประเทศไทยเท่านั้น
- เก็บข้อมูลตั้งแต่วันที่ 8 มกราคม 2559 จนถึงวันที่ 29 พฤศจิกายน 2563
- รวมตำบลที่มีการแจ้งเตือนภัยแล้วทั้งหมด 5,431 ตำบล

1.3 ข้อมูลประวัติวาทภัยที่เกิดขึ้น มีลักษณะของข้อมูลดังนี้

- เป็นข้อมูลราย 6 ชั่วโมงจำนวนทั้งหมด 2,328 ตัวอย่าง ซึ่งจะเก็บเฉพาะช่วงเวลาที่เกิดพายุเท่านั้น
- เป็นข้อมูลทำการเก็บตั้งแต่วันที่ 1 มกราคม 2561จนถึงวันที่ 15 พฤศจิกายน 2563
- มีพายุรวมทั้งสิ้น 96 ลูก

แหล่งข้อมูล

#	ชุดข้อมูล	ประเภทแหล่งที่มาของข้อมูล		ประเภทข้อมูล		ลักษณะข้อมูล		
		Owner Data	Portal Data	ข้อมูล Static	ข้อมูล Dynamic	ข้อมูลอดีต	ข้อมูลปัจจุบัน	ข้อมูลอนาคต
1	ข้อมูลสภาพอากาศ	กรมอุตุนิยมวิทยา	กรมอุตุนิยมวิทยา		✓	✓	✓	✓
2	ข้อมูลแจ้งเตือนวาทภัย	ปภ.	ปภ.	✓		✓	✓	
3	ข้อมูลประวัติการเกิดวาทภัย	กรมอุตุนิยมวิทยา	กรมอุตุนิยมวิทยา	✓		✓	✓	

Owner Data: หน่วยงานเจ้าของข้อมูล

Portal Data: หน่วยงานที่รวบรวมข้อมูลจากเจ้าของข้อมูลและนำมาเผยแพร่

ข้อมูล Static: ข้อมูลที่มีการ update นาน ๆ ครั้งหรือจัดทำขึ้นเพียงครั้งเดียว

ข้อมูล Dynamic: ข้อมูลที่มีการ update สม่ำเสมอและมีช่วงเวลาที่แน่นอน

รายชั่วโมง

รายวัน

ราย 6 ชั่วโมง

ภาพที่ 5: แสดงชุดข้อมูลที่ใช้ในงานวิจัย

2. การเตรียมข้อมูล (Prepare Data)

- จากข้อมูลสภาพอากาศจำนวน 2,338,065 ตัวอย่าง พบว่ามี Missing values อยู่ ได้ทำการ Remove missing values ออกเหลือตัวอย่างทั้งหมดจำนวน 2,267,368 ตัวอย่าง
- ทำการปรับเปลี่ยนข้อมูลรายชั่วโมงเป็นข้อมูลแบบรายวันโดยค่าของคุณลักษณะจะใช้เป็นค่าเฉลี่ยแทน ได้ข้อมูลรวม 96,968 ตัวอย่าง

- 3) ทำ Interpolation เพื่อให้มีค่าครบก่อนนำไปเชื่อมโยงกับ ข้อมูลการแจ้งเตือนวาทภัย (Label) (จากข้อมูลที่ได้จากการ Remove missing values จำนวน 96,968 ตัวอย่าง มี 1,071 วัน 5 คุณลักษณะ นำมา Interpolate ให้มีข้อมูลครบทุกตำบลในประเทศไทยทั้ง 5 คุณลักษณะ โดยการใช้เทคนิค IDW ในการเติมข้อมูลสภาพอากาศของตำบลที่หายไป ได้ข้อมูลรวม 7,946,820 ตัวอย่าง)
- 4) นำข้อมูลที่ได้จากการ Interpolation มาเชื่อมโยงกับข้อมูลการแจ้งเตือนวาทภัยแบบล่วงหน้า 1 วัน และเพิ่มคุณลักษณะสภาพอากาศที่เป็นวันปัจจุบันเข้าไปด้วย โดยใช้รหัสตำบลและวันที่ของข้อมูลเป็นคีย์
- 5) ข้อมูลที่ได้มีลักษณะเป็น Imbalance data ไม่เหมาะต่อการนำมาสร้างแบบจำลองจึงต้องมีการทำ Under sampling โดยคัดเลือกให้ Label = Y และ Label = N มีจำนวนเท่ากัน ซึ่งผลจากการทำ Under sampling มีดังนี้

ตารางที่ 1: แสดงผลสรุปจากการ Remove missing values และ Under sampling

	Label = Y	Label = N
จำนวนตัวอย่างก่อนกำจัด Missing Values	15,009	7,931,811
จำนวนตัวอย่างหลังกำจัด Missing Values	15,001	7,916,979
จำนวนตัวอย่างหลังทำ Under Sampling	15,001	15,001

3. การคัดเลือกตัวแปร (Features Selection)

โดยใช้ข้อมูลสภาพอากาศล่วงหน้า 1 วัน และข้อมูลสภาพอากาศปัจจุบัน โดยประโยชน์หลักของการทำ Features Selection คือ เพื่อเพิ่มประสิทธิภาพของโมเดลและเพื่อลดจำนวน Features ที่ไม่จำเป็น ในงานวิจัยฉบับนี้ใช้ weight by chi-square ในการหาค่าน้ำหนักความสัมพันธ์ของทั้ง 11 คุณลักษณะ

4. สร้างแบบจำลอง Neural Network (จากข้อมูลล่วงหน้า 1 วัน และข้อมูลปัจจุบัน)

ทำการสร้างแบบจำลองโครงข่ายประสาทเทียมโดยใช้ 11 คุณลักษณะดังนี้

ตารางที่ 2: แสดงคุณลักษณะที่ใช้ในการสร้างแบบจำลองโครงข่ายประสาทเทียม

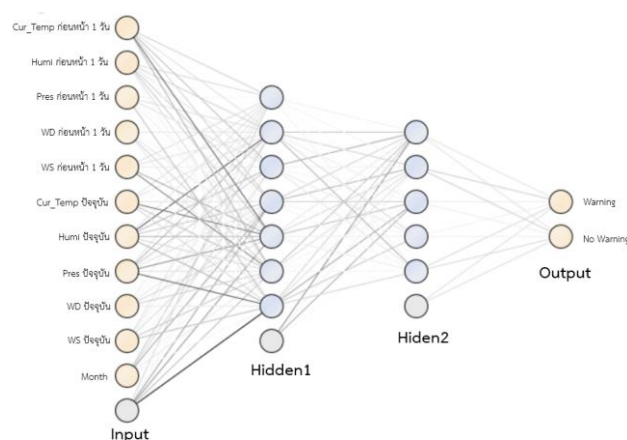
1	ความกดอากาศก่อนหน้า 1 วัน	7	ความชื้น ณ วันปัจจุบัน
2	ความชื้นก่อนหน้า 1 วัน	8	อุณหภูมิ ณ วันปัจจุบัน
3	อุณหภูมิก่อนหน้า 1 วัน	9	ความเร็วลม ณ วันปัจจุบัน
4	ความเร็วลมก่อนหน้า 1 วัน	10	ทิศทางลม ณ วันปัจจุบัน
5	ทิศทางลมก่อนหน้า 1 วัน	11	เดือน (1-12) ที่ทำการเก็บข้อมูล
6	ความกดอากาศ ณ วันปัจจุบัน		

ใช้โอเพอร์เรเตอร์ “Neural Net” ของ RapidMiner Studio เพื่อทำการสร้างแบบจำลองโครงข่ายประสาทเทียม โดยกำหนดพารามิเตอร์ต่างๆดังนี้

ตารางที่ 3: แสดงค่าพารามิเตอร์ที่ใช้ในการสร้างแบบจำลองโครงข่ายประสาทเทียม

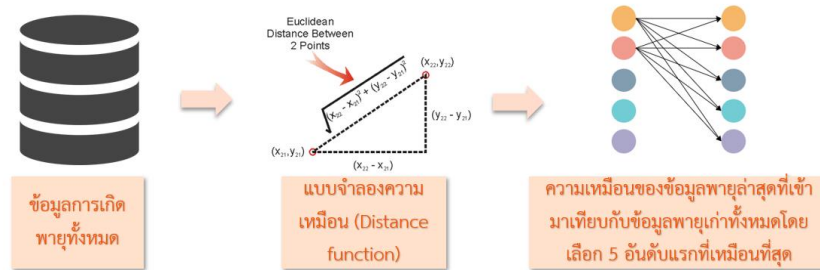
Hidden layer 1	7
Hidden layer 2	5
Training cycle	160
Learning rate	0.01
Momentum	0.9

และมี Output Label ได้แก่ ค่าบ่งชี้ที่มีการแจ้งเตือนวาทภัย (Warning) และค่าบ่งชี้ที่ไม่มีการแจ้งเตือนวาทภัย (No Warning)



ภาพที่ 6: แสดงโครงสร้างของแบบจำลองโครงข่ายประสาทเทียมด้วย RapidMiner Studio

5. หาความเหมือนของพายุโดยเปรียบเทียบกับข้อมูลประวัติวาทภัย



ภาพที่ 7: แสดงวิธีการหาความเหมือนของพายุ

ทำการเตรียมข้อมูลพายุโดยมีขั้นตอนดังนี้

- 1) จัดเรียงข้อมูลการเกิดพายุตามวันและเวลาการเกิด
- 2) ทำการจัดการข้อมูลให้เป็นไปในรูปแบบเดียวกัน ยกตัวอย่างเช่น Wind Direction ซึ่งเป็นข้อมูลที่เกิดจากการบันทึกข้อมูลโดยผู้ใช้ ทำให้เกิดความผิดพลาดที่เกิดจากมนุษย์ (Human Error) จำเป็นต้องมีการทำ Data Cleansing เนื่องจากการใช้ตัวอักษรเล็กใหญ่หรือคำย่อไม่เหมือนกัน
- 3) ใช้ Operator Data to Similarity และ Similarity to Data ในการหาความเหมือนของตัวแปรที่สนใจและเลือกใช้เป็น Euclidean Distance ในการหาระยะทาง
- 4) โดยค่าความเหมือนจะถูกแสดงในช่อง Distance ซึ่งยิ่งค่า Distance มีค่าต่ำแสดงว่าข้อมูลใกล้เคียงหรือเหมือนกันมากนั่นเอง
- 5) ทำการจัดเรียงค่า Distance จากน้อยไปมากตามลำดับ และคัดเลือกมาจำนวน 5 ลำดับแรก

6. เครื่องมือที่ใช้ในการวิจัย

6.1 RapidMiner Studio

ในงานวิจัยนี้จะใช้ RapidMiner Studio ในการเตรียมข้อมูล สร้างแบบจำลองโครงข่ายประสาทเทียม และใช้ในการวัดผลแบบจำลองที่สร้างขึ้น ซึ่งเป็นเครื่องมือที่เข้าใจง่ายเหมาะสำหรับทำ Data Mining

6.2 Microsoft Power BI Desktop

เป็นเครื่องมือสำหรับออกแบบและสร้าง Dashboard เพื่อนำเสนอผลการพยากรณ์ล่วงหน้า นอกจากนี้ Power BI Desktop ยังสามารถ Publish งานของเราขึ้นเป็น Web Application เพื่อแชร์ให้ผู้อื่นได้ใช้งานอีกด้วย

7. การนำแบบจำลองไปประยุกต์ใช้กับข้อมูลจริง

ขั้นตอนการนำแบบจำลองไปใช้งานมีความคล้ายคลึงกับขั้นตอนการเตรียมข้อมูลเพื่อนำมาสร้างแบบจำลองแต่มีความต่างกันอยู่บางส่วน

ข้อมูลใหม่ที่จะนำมาใช้กับแบบจำลองเป็นข้อมูลสภาพอากาศล่วงหน้า 7 วันจากกรมอุตุนิยมวิทยา เป็นข้อมูลแบบรายชั่วโมง ซึ่งจะถูกเปลี่ยนเป็นข้อมูลแบบรายวันเพื่อให้ Input เหมือนกับช่วงสร้างแบบจำลอง

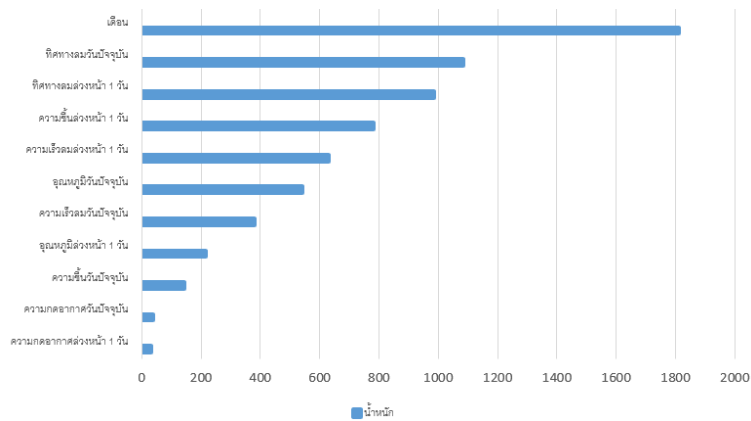
ผลการศึกษา

1. ผลการคัดเลือกตัวแปรโดยหาค่าถ่วงน้ำหนัก

ตารางที่ 4: แสดงค่าถ่วงน้ำหนักของตัวแปรจากน้อยไปมาก จากการทำ Features selection โดยการหาค่า chi-square

ตัวแปร	ถ่วงน้ำหนัก
เดือน	1812.88
ทิศทางลมวันปัจจุบัน	1085.932
ทิศทางลมล่วงหน้า 1 วัน	984.596
ความชื้นล่วงหน้า 1 วัน	783.762
ความเร็วลมล่วงหน้า 1 วัน	632.32
อุณหภูมิวันปัจจุบัน	541.681
ความเร็วลมวันปัจจุบัน	382.449
อุณหภูมิล่วงหน้า 1 วัน	216.659
ความชื้นวันปัจจุบัน	142.635
ความกดอากาศวันปัจจุบัน	40.182
ความกดอากาศล่วงหน้า 1 วัน	31.671

จากผลการทำ Features Selection โดยใช้ข้อมูลสภาพอากาศล่วงหน้า 1 วัน และสภาพอากาศปัจจุบัน พบว่าสภาพอากาศของวันปัจจุบันค่อนข้างมีความสำคัญกับการสร้างแบบจำลองอยู่บ้าง งานวิจัยนี้จึงทำการใส่สภาพอากาศปัจจุบันเป็นคุณลักษณะเพิ่มเข้าไปด้วย เพื่อเพิ่มประสิทธิภาพของแบบจำลองให้พยากรณ์ออกมาได้แม่นยำมากขึ้น



ภาพที่ 8: แสดงกราฟค่าน้ำหนักจากการทำ Features selection โดยการหาค่า chi-square

2. ผลจากการวัดผลแบบจำลอง Neural Network

ตารางที่ 5: ตาราง Confusion Matrix จากการวัดผล

	True No Warning	True Warning	Class Precision
Pred. No Warning	4012	2105	65.59%
Pred. Warning	1392	4230	75.24%
Class Recall	74.24%	66.67%	

จากการวัดประสิทธิภาพได้ค่า Accuracy ที่ 70.21 เปอร์เซ็นต์ และได้ค่า Class Recall และ Class Precision อยู่ในเกณฑ์ที่น่าพอใจ

3. ผลจากการหาความเหมือนของพายุ

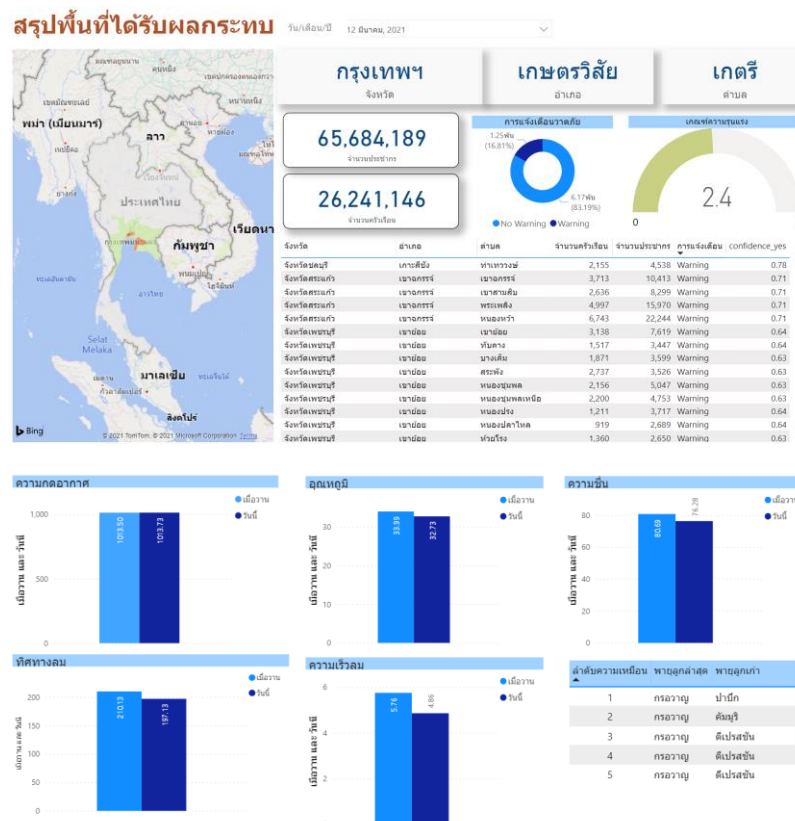
DISTANCE	first_typhoon_id	second_typhoon_id	typhoon_lat	typhoon_long	typhoon_speed	typhoon_date_month	direction	typhoon_date	name...
2.739	285	206	20	105	20	9	W	Sep 13, 2018 ...	บาริจัด
3.082	285	277	16	106	20	10		Oct 12, 2020 ...	หลันฟ้า
3.082	285	278	20	102	20	10		Oct 14, 2020 ...	ซิงกา
3.415	285	280	16	106.600	20	10		Oct 26, 2020 ...	โซเดล
3.937	285	281	15	103	20	10		Oct 29, 2020 ...	โมลาน
285	18.500	104.500	20	11				Nov 15, 2020 6:00:00 PM ICT	หว่ามก่อ

ภาพที่ 9: ตัวอย่างผลการหาความเหมือนด้วยวิธีการหา Euclidean Distance โดย RapidMiner Studio

จากภาพตัวอย่าง จะเห็นว่าพายุลูกที่ 285(หว่ามก้อ) เมื่อเทียบกับพายุลูกที่ 206(บารีจัด), 277(หลั่นฟา), 278(นังกา), 280(โซเคล) และ 281(โมลาเบ) นั้นมีระยะทาง(Distance) น้อยที่สุด 5 ลำดับแรก

4. การแสดงผล

จากการนำแบบจำลองที่ได้ไปประยุกต์ใช้กับข้อมูลสภาพอากาศจริงที่ได้จากกรมอุตุนิยมวิทยา และแสดงผลเป็น Dashboard ผ่านทาง Web Application โดยใช้โปรแกรม Power BI desktop ในการแสดงผล



ภาพที่ 10: ตัวอย่างผลจากแบบจำลองผ่าน Power BI desktop

บทสรุปและข้อเสนอแนะ

1. สรุปผลการศึกษา

แบบจำลองโครงข่ายประสาทเทียมที่พัฒนา เมื่อใช้กับตัวแปรทั้ง 11 ตัว โดยมีค่า Learning rate ที่ 0.01 Training cycle ที่ 160 มีค่า Accuracy อยู่ที่ 70.21% โดยค่าความถูกต้องของการแจ้งเตือนวาทภัย 66.67% และค่าความถูกต้องของการไม่แจ้งเตือนวาทภัย 74.24% ซึ่งเป็นค่าที่ให้ผลดีที่สุดและยอมรับได้

ผลการหาค่าความเหมือนเมื่อเปรียบเทียบค่าตัวแปรต่างๆ ของพายุลูกปัจจุบันกับลูกล่าสุด พบว่ามีความใกล้เคียงกันทั้ง ทิศทางของพายุ, เดือนที่เกิดพายุ, ละติจูด, ลองจิจูด และความเร็วลม ณ จุดศูนย์กลางพายุ ทำให้ได้ค่าระยะทาง(Distance) ที่ต่ำเมื่อเทียบกับพายุลูกปัจจุบัน

2. ข้อเสนอแนะ

2.1 การคัดเลือกตัวแปรที่ใช้ในการสร้างแบบจำลอง รวมถึงการหาตัวแปรเพิ่มเติมจากงานวิจัยนี้ ซึ่งในงานวิจัยนี้มีตัวแปรจำกัดในการสร้างแบบจำลอง

2.2 การทดสอบด้วยแบบจำลองอื่นๆ และวัดผลเปรียบเทียบกับ Neural Network หรือผสมผสานกันระหว่างแบบจำลองแบบต่างๆ เพื่อผลที่ดีขึ้น

2.3 ทดสอบหาค่าความเหมือนของพายุด้วยเทคนิคระยะทางแบบต่างๆ และนำมาเปรียบเทียบผล

2.4 ผลของแบบจำลองที่ทำนายว่าจะไม่มีการแจ้งเตือน แต่ความจริงแล้วมีการแจ้งเตือนเกิดขึ้นจำนวน 2,105 ตัวอย่างในพื้นที่จริงจากแผนที่ประเทศไทย เมื่อนำมาเปรียบเทียบกับพื้นที่จริงพบว่าเกิดขึ้นกับจังหวัด นครสวรรค์ พิจิตรและจังหวัดลพบุรีเป็นส่วนใหญ่ ทั้งนี้จำนวนตำบลในแต่ละจังหวัดอาจมีผลต่อการเปรียบเทียบ และอาจขึ้นกับปัจจัยอื่นๆทางด้านสภาพอากาศที่มีความคล้ายคลึงกันส่งผลให้ผลการทำนายมีความคลาดเคลื่อนไม่มากนัก

2.5 ในการเพิ่มตัวแปรของเดือนเข้ามาช่วยในการสร้างแบบจำลองโดยใช้เป็นเลข 1 ถึง 12 ตามลำดับของเดือนสามารถทดสอบเปลี่ยนเป็นเป็นช่วงของฤดูแทน และเปรียบเทียบผลของแบบจำลองซึ่งคาดว่า การใช้ตัวแทนเป็นตัวเลขอาจจะมีข้อจำกัดเช่น เดือนธันวาคมและเดือนมกราคม ถึงแม้ตัวเลขจะต่างกันมาก คือ 1 และ 12 แต่ยังเป็นฤดูหนาวเหมือนกันทั้งคู่

2.6 ในพื้นที่บางส่วนหากสภาพอากาศเหมือนกัน ส่วนใหญ่แล้วควรจะมีการแจ้งเตือนเหมือนกัน แต่ด้วยข้อจำกัดต่างๆเช่น พื้นที่ที่มักจะมีการแจ้งเตือนควรจะให้มีการแจ้งเตือนต่อไปถึงแม้สภาพอากาศจะแตกต่างออกไป ทั้งนี้ขึ้นอยู่กับเกณฑ์ของหน่วยงานที่เกี่ยวข้องด้วย เช่นพื้นที่ทางภาคใต้ที่มักเกิดมรสุมอยู่บ่อยครั้ง อาจไม่จำเป็นต้องมีการแจ้งเตือนยกเว้นเสียแต่ว่าถกกันนั้นจะมีเกณฑ์ความรุนแรงมาก เป็นต้น

บรรณานุกรม

ภาษาไทย

กิตติพงษ์ เนียเจริญ (2564), *Practical Data Visualization with Power BI*. ไอดีซีฯ

ชวลิต ภู่อธิกุล (2564), *Analyzing Data with Power BI (เอกสารประกอบการอบรม)*. เอเซีย

ดิจิทัลการพิมพ์

มูลนิธิโครงการสารานุกรมไทยสำหรับเยาวชน โดยพระราชประสงค์ในพระบาทสมเด็จพระบรมชนกาธิเบศร มหาภูมิพลอดุลยเดชมหาราช บรมนาถบพิตร, *ข้อมูลที่ได้จากการตรวจอากาศ*.

<https://www.saranukromthai.or.th/sub/book/book.php?book=35&chap=6&page=t35-6-infodetail04.html>.

เอกสิทธิ์ พัทธวงศ์ศักดิ์ (2562), *Introduction to Feature (Attribute) Selection with RapidMiner Studio*

6. Retrieved from <https://www.slideshare.net/sitake/introduction-to-feature-attribute-selection-with-rapidminer-studio-6-55432350>.

เอกสิทธิ์ พัทธวงศ์ศักดิ์ (2562), *Practical Data Mining with RapidMiner Studio 9 @BD5*

(เอกสารประกอบการอบรม). เอเซีย ดิจิทัลการพิมพ์

ภาษาต่างประเทศ

ArcGIS for Desktop. *Understanding interpolation analysis*. Retrieved from

<https://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/understanding-interpolation-analysis.htm>

RapidMiner. (2564). *Neural Net*. Retrieved from

https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/neural_nets/neural_net.html