บทที่ 3
วิธีดำเนินการวิจัย

ในการดำเนินการวิจัยในการพัฒนาแบบจำลองเพื่อพยากรณ์ปริมาณ PM10 ในพื้นที่จังหวัดเชียงใหม่ โดยใช้โครงข่ายเพอร์เซ็ปตรอนหลายชั้น ประกอบด้วยขั้นตอนดังแสดงต่อไปนี้

3.1 การรวบรวมข้อมูล

3.2 การสร้างแบบจำลองโครงข่ายประสาทเทียม

3.3 การทำ Incremental Clustering

3.4 ศึกษาเปรียบเทียบประสิทธิภาพของแบบจำลอง

3.1 การรวบรวมข้อมูล

ทำการรวบรวมข้อมูลปริมาณมลพิษทางอากาศในพื้นที่จังหวัดเชียงใหม่ ข้อมูลที่ใช้ในการทดลองอ้างอิงมาจาก รายงานสถานการณ์หมอกควันในพื้นที่ภาคเหนือของประเทศไทย ซึ่งได้แสดงข้อมูลทางอุตุนิยมวิทยาที่เกี่ยวข้อง ได้แก่ อุณหภูมิ ความกดอากาศ ความชื้นสัมพัทธ์ น้ำระเหย ปริมาณน้ำฝน ทิศทางลม กำลังลม และ และใช้ข้อมูลปริมาณฝุ่นขนาดเล็กกว่า 10 ไมครอน (PM10) จากสถานีตรวจวัดคุณภาพอากาศที่โรงเรียนยุพราชวิทยาลัย ซึ่งเป็นสถานีตรวจวัดคุณภาพอากาศของกรมควบคุมมลพิษ ในช่วงระยะเวลา 4 ปี ตั้งแต่ปี 2555 -2558 มาเป็นข้อมูลในการสร้างรูปแบบการพยากรณ์

3.2 การสร้างแบบจำลองโครงข่ายประสาทเทียม

ดัชนีคุณภาพอากาศที่ใช้อยู่ในประเทศไทย คำนวณโดยเทียบจากมาตรฐานคุณภาพอากาศในบรรยากาศโดยทั่วไปของสารมลพิษทางอากาศ 5 ประเภท ได้แก่ ก๊าซโอโซน (O_3) เฉลี่ย 1 ชั่วโมง ก๊าซไนโตรเจนไดออกไซด์ (NO_2) เฉลี่ย 1 ชั่วโมง ก๊าซคาร์บอนมอนอกไซด์ (CO) เฉลี่ย 8 ชั่วโมง ก๊าซซัลเฟอร์ไดออกไซด์ (SO_2) เฉลี่ย 24 ชั่วโมง และฝุ่นละอองขนาดเล็กกว่า 10 ไมครอน (PM10) เฉลี่ย 24 ชั่วโมง ทั้งนี้ ดัชนีคุณภาพอากาศที่คำนวณได้ของสารมลพิษทางอากาศประเภทใดมีค่าสูงสุด จะใช้เป็นดัชนีคุณภาพอากาศของวันนั้น

ดัชนีคุณภาพอากาศของประเทศไทยแบ่งเป็น 5 ระดับ คือ ตั้งแต่ 0 ถึง มากกว่า 300 ซึ่งแต่ละระดับจะใช้สีเป็นสัญลักษณ์เปรียบเทียบระดับของผลกระทบต่อสุขภาพอนามัย ตามตารางที่ 3.1 โดยดัชนีคุณภาพอากาศ 100 จะมีค่าเทียบเท่ามาตรฐานคุณภาพอากาศในบรรยากาศโดยทั่วไป หากดัชนีคุณภาพอากาศมีค่าสูงเกินกว่า 100 แสดงว่าค่าความเข้มข้นของมลพิษทางอากาศมีค่าเกินมาตรฐานและคุณภาพอากาศในวันนั้นจะเริ่มมีผลกระทบต่อสุขภาพอนามัยของประชาชน

ตารางที่ 3.1 เกณฑ์ของดัชนีคุณภาพอากาศสำหรับประเทศไทย

AQI	ความหมาย	สีที่ใช้	แนวทางการป้องกันผลกระทบ
0-50	คุณภาพดี	ฟ้า	ไม่มีผลกระทบต่อสุขภาพ
51-100	คุณภาพปานกลาง	เขียว	ไม่มีผลกระทบต่อสุขภาพ
101-200	มีผลกระทบต่อสุขภาพ	เหลือง	ผู้ป่วยโรคระบบทางเดินหายใจ ควรหลีกเลี่ยงการออกกำลังภายนอกอาคารบุคคลทั่วไป โดยเฉพาะเด็กและผู้สูงอายุ ไม่ควรทำกิจกรรมภายนอกอาคารเป็นเวลานาน
201-300	มีผลกระทบต่อสุขภาพมาก	ส้ม	ผู้ป่วยโรคระบบทางเดินหายใจ ควรหลีกเลี่ยงกิจกรรมภายนอกอาคารบุคคลทั่วไป โดยเฉพาะเด็กและผู้สูงอายุ ควรจำกัดการออกกำลังภายนอกอาคาร
มากกว่า 300	อันตราย	แดง	บุคคลทั่วไป ควรหลีกเลี่ยงการออกกำลังภายนอกอาคาร สำหรับผู้ป่วยโรคระบบทางเดินหายใจ ควรอยู่ภายในอาคาร

3.2.1 นำข้อมูลที่ใช้ในการทดลองทางอุตุนิยมวิทยาที่เกี่ยวข้อง ได้แก่ อุณหภูมิ ความกดอากาศ ความชื้นสัมพัทธ์ น้ำระเหย ปริมาณน้ำฝน ทิศทางลม กำลังลม และ และใช้ข้อมูลปริมาณฝุ่นขนาดเล็กกว่า 10 ไมครอน (PM10) มาเป็นข้อมูลในการสร้างรูปแบบการพยากรณ์ แล้วแทนค่า Missing Value ด้วยค่า Mean

3.2.2 กำหนด Class อ้างอิงตาม ตารางที่ 3.1 โดยสร้าง Class ตามค่า PM10 จำนวน 3 Class ตามตารางที่ 3.2 ดังนี้

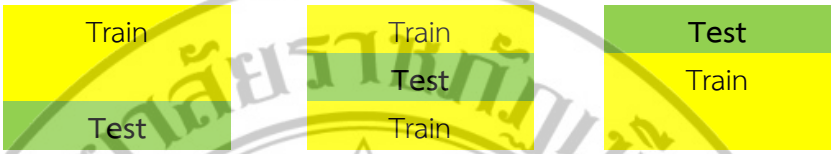
ตารางที่ 3.2 แสดง Class ตามค่า PM10

Class	PM10
1	0 - 100
2	101 – 300
3	> 300

3.2.3 แปลงค่าข้อมูล (Data transformation) เป็นกระบวนการปรับขอบเขตของข้อมูลด้วยวิธีการนอมัลไลซ์ (Normalization) ซึ่งเป็นการแปลงข้อมูลให้อยู่ในช่วงใหม่มีขอบเขตที่น้อยลง เพื่อให้เหมาะสมกับฟังก์ชันของโครงข่ายประสาทเทียม คือให้อยู่ในช่วง 0 -1 เพื่อให้เหมาะสมต่อการนำไปใช้งานในการสอนให้โครงข่ายประสาทเทียมเกิดการเรียนรู้ ตามสมการ

$$v' = \frac{v-min}{max-min}(new_\max - new_min) + new_min$$

3.2.4 ทำการแบ่งชุดข้อมูลเพื่อใช้สำหรับการเรียนรู้ (Training Set) และข้อมูลชุดทดสอบ (Test Set) ออกเป็น 3 กลุ่ม ๆ ละ 150 รายการ แบบ 3 fold cross validation



ภาพที่ 3.1 การแบ่งกลุ่มข้อมูลแบบ 3 fold cross validation

3.2.5 ทำการเขียนโปรแกรม สำหรับการ Train Backpropagation (the generalized delta rule) สำหรับ Multi-Layer Perceptron (MLP) ในส่วนของ MLP จะสามารถกำหนดจำนวน Hidden Node และตัวแปรที่เกี่ยวข้องคือ Momentum Rate และ Learning Rate ได้โดยที่ไม่ต้องไม่แก้โปรแกรม

3.2.6 ทำการสอนโครงข่ายประสาทเทียม (neural network training) โดยใช้ตัวแปรข้อมูลเข้า $X(t)$ และตัวแปรผลลัพธ์ $Y(t+1)$ โดยใช้ข้อมูลชุดสอน และวัดค่าคลาดเคลื่อนเฉลี่ย (Sum square error: SSE) เพื่อวัดความคลาดเคลื่อนของแบบจำลองในการพยากรณ์

3.2.7 ทำการทดลอง แต่ละ fold ของ 3 fold cross validation และวัดความถูกต้องของการพยากรณ์

3.3 การทำ Incremental Clustering

จากการทดลองหารูปแบบของการพยากรณ์ปริมาณฝุ่นละอองขนาดไม่ เกิน 10 ไมครอน (PM10) ที่เป็นสารมลพิษทางอากาศที่เป็นปัญหาหลักในพื้นที่จังหวัดเชียงใหม่ โดยใช้เทคนิคโครงข่ายประสาทเทียม (Neural Networks) ในรูปแบบของโครงข่ายเพอร์เซฟตรอนแบบหลายชั้น (Multi-layer perceptron) โดยมีวิธีการเรียนรู้แบบ Back-propagation นั้น ได้ผลดังนี้

ตารางที่ 3.3 ผลการทดลองกับข้อมูล โดยใช้ 3 fold cross validation

รอบที	เวลา (วินาที)	จำนวนรอบ	Avg ssqerr	ความถูกต้อง (เปอร์เซ็นต์)
1	8.673874	1562	9.49986861	83.33
2	0.077655	13	9.45476340	73.33
3	0.071448	12	9.37973767	72
ค่าเฉลี่ย	2.940992	529	9.44479	76.22

3.3.1 เลือก Model จากข้อมูลในตารางที่ 6 จะเห็นว่าในการทดลองกับข้อมูล โดยใช้ 3 fold cross validation รอบที่มีเปอร์เซ็นต์ความถูกต้องมากที่สุดคือ รอบที่ 1 จึงเลือก model ที่ได้ผลการ classify ที่ดีที่สุด มาใช้ในการทำ clustering โดยมีค่า Weight ของตัวแปร input ตามตารางที่ 3.4 และ ค่า Weight ของ Hidden Node ตามตารางที่ 3.5

ตารางที่ 3.4 ค่าน้ำหนักของตัวแปร input

W1	W2	W3	W4	W5	W6	W7
-12.032	-26.0642	63.43898	1.433411	-1.66743	-5.48449	32.94506
-13.1953	-9.83382	14.55446	-7.65679	-4.13667	9.379011	5.685942
5.905771	6.286288	1.47497	-0.80403	-0.5496	-14.1557	-7.90282

ตารางที่ 3.5 ค่าน้ำหนักของ Hidden Node

W1	W2	W3
-38.2393	-32.4184	-30.4246
-14.8726	11.41909	1.85406
15.44495	-11.9733	-3.19105

3.3.2 เตรียมชุดข้อมูลสภาพอากาศ ที่มี 3 classes ได้แก่ class 1 ค่า PM10 อยู่ในช่วง 0 – 100 class 2 ค่า PM10 อยู่ในช่วง 101 – 300 class 3 ค่า PM10 มากกว่า 300 และ 7 attribute input ได้แก่ อุณหภูมิ ความกดอากาศ ความชื้นสัมพัทธ์ น้ำระเหย ปริมาณน้ำฝน ทิศทางลม กำลังลม ทั้งหมดจำนวน 450 รายการ แล้วทำการลบ class ของ dataset ออก

3.3.3 นำ dataset ทั้งหมด มาผ่าน Neural Network ที่เลือก โปรแกรมจะทำการสร้างไฟล์ข้อมูลผลลัพธ์ ตามตารางที่ 3.6

ตารางที่ 3.6 ตัวอย่างข้อมูลผลลัพธ์

Desired Output			Program Output		
0	0	1	2.24E-30	2.42E-06	0.999995
0	0	1	3.48E-30	2.09E-05	0.999965
0	0	1	7.05E-30	0.006914	0.993445
0	0	1	2.13E-29	0.004375	0.995904
0	0	1	6.81E-30	3.01E-06	0.999994
0	0	1	2.27E-30	2.79E-06	0.999994
0	0	1	7.52E-30	2.26E-06	0.999996
0	0	1	4.65E-30	2.35E-06	0.999995
0	0	1	1.78E-27	0.000621	0.999441

ตารางที่ 3.6 ตัวอย่างข้อมูลผลลัพธ์ (ต่อ)

Desired Output			Program Output		
0	0	1	2.94E-28	0.002002	0.998231
0	0	1	9.03E-30	2.20E-06	0.999996
0	0	1	1.82E-31	0.018035	0.981277
0	0	1	1.40E-31	0.023365	0.976019
0	0	1	3.50E-39	0.082953	0.836333
0	0	1	2.06E-31	0.030641	0.969894
0	0	1	2.00E-31	0.029688	0.9707
0	0	1	2.14E-31	0.030142	0.97039
0	0	1	2.18E-31	0.02977	0.970748
0	0	1	3.64E-30	0.001159	0.998662
0	0	1	4.71E-26	3.67E-05	0.999964
0	0	1	2.40E-31	0.029345	0.971229
0	0	1	3.43E-25	1.11E-06	0.999999
0	0	1	2.10E-30	2.64E-06	0.999995
0	0	1	3.33E-28	0.001005	0.99906

3.3.4 นำค่า Program Output ที่ได้จาก Neural Networks มาทำ Incremental Clustering ได้ผลลัพธ์ตามตารางที่ 3.7

ตารางที่ 3.7 ตัวอย่างการเปรียบเทียบกับผลลัพธ์

Desired Output			Program Output	Result
0	0	1	C1	T
0	0	1	C2	F
0	0	1	C1	T
0	0	1	C2	F
0	0	1	C1	T
0	0	1	C1	T
0	0	1	C1	T
0	0	1	C1	T
0	0	1	C1	T
0	0	1	C1	T
0	0	1	C1	T
0	0	1	C1	T

ตารางที่ 3.7 ตัวอย่างการเปรียบเทียบกับผลลัพธ์ (ต่อ)

Desired Output			Program Output	Result
0	0	1	C1	T
0	0	1	C1	T
0	0	1	C1	T
0	0	1	C1	T
0	0	1	C1	T
0	0	1	C1	T
0	0	1	C1	T
0	0	1	C1	T
0	0	1	C1	T
0	0	1	C1	T
0	0	1	C1	T
0	0	1	C1	T
0	0	1	C1	T
0	0	1	C1	T
0	0	1	C1	T

3.3.5 วิเคราะห์ผลการทำงาน นำค่า Program Output ที่ได้จาก Neural Networks มาทำ Incremental Clustering ต่อนั้นทำให้ได้เปอร์เซ็นต์ความถูกต้องของผลลัพธ์

3.4 ศึกษาเปรียบเทียบประสิทธิภาพของแบบจำลอง

ในขั้นตอนการศึกษานี้เป็นการนำผลของข้อมูลที่ได้จากการทำนายด้วยโครงข่ายเพอร์เซ็ปตรอนหลายชั้น และแบบจำลอง Incremental Clustering มาเปรียบเทียบประสิทธิภาพของแบบจำลอง