# 深層学習手法を用いた光化学オキシダント濃度予測システムの自作

環境衛生科学研究所

○小田祐一、杉山優雅、渡邉崇之、結城茜、金子亜由美、矢嶋雅、太田良和弘

## はじめに

大気汚染物質である光化学オキシダント(以下、Ox)は、大気中の窒素酸化物(以下、NOx)や非メタン炭化水素(以下、NMHC)等の揮発性有機化合物が紫外線による光化学反応を起こすことにより生成し、ヒトへの健康被害等を引き起こす。本県を含めた全国の多くの自治体では、県民へのOx予測情報提供のため、Ox濃度が高くなりやすい初夏から早秋にかけて、当日のOx濃度予測が行われている。今年度の研究で、このOx濃度予測の自動化、予測精度の向上を目的に、「Ox濃度予測」を行う人工知能(AI)を、プログラミング言語のPythonによって内製・自作する試みに取り組んでいる。本稿では、今回開発した試作AIの開発過程について報告する。

## 方法

### 1 開発環境マシンスペック

CPU:Intel(R)Core(TM)i7-12700H(14コア、定格クロック周波数2.30GHz、最大クロック周波数4.70GHz)、メモリ:64GB、GPU:NVIDIA GeForce RTX 3050Ti LaptopGPU、GPUメモリ:4GB、OS:windows11 Home 64bitを使用した。

### 2 AIアルゴリズムの選定

本研究で検討したAIアルゴリズムは、深層ニューラルネットワークを用いた手法のうち、時系列データの扱いに特化したリカレントニューラルネットワーク(Recurrent Neural Networks:RNN)とした。これは、本アルゴリズムの特徴として、ネットワークを構成するニューロンの出力が、未来の当該ニューロン自身の入力にもなる仕組みであるため、「ある時点のデータが、その次の時点のデータに影響を及ぼす」性質を持つ時系列データの特徴を再現していることによる。本研究では、比較的長期間の時系列データの勾配消失を低減させるため、記憶セルを長・短期記憶ユニット(Long Short-Term Memory:LSTM)とした。

### 3 データセット構築

本研究の試作AIは、予報(予測)値が入手可能な気象データのみを用いてOxの時系列データ予測を行えることを目的としたため、Ox生成に重要な日射量データを収集・公開している静岡地方気象台(静岡市)近傍の大気汚染常時監視局舎(静岡市立長田南中学校)のOxデータを用いた(**図1**)。気象データは、日射量、日照時間、気温、湿度、降水量、現地気圧、風速、天気とし、常時監視データはOxのみとした。これらはいずれも1時間値を使用した。評価対象期間は、2021年5月1日1:00から6月10日0:00とし、2021年5月1日1:00から6月9日10:00までを訓練データとし、6月9日11:00からから6月10日0:00までを予測した。


**図1** 評価対象地点の位置関係

### 4 RNN(LSTM)によるOx時系列データ予測

AIプログラムはPython(3.10)を用いてコードし、ディープラーニングのモデル構築、学習、推論(予測)にはオープンソースライブラリであるTensorFlow(2.9.1)及びTensorFlowを計算バックエンドとした高水準APIであるKeras(2.9.0)を用いている。予測の手法としては、まずはOxのみの単変量時系列データを用いて未来を予測し(単変量RNN)、その後に、Ox及びその他の気象データを訓練データとした学習結果を用いて多変量時系列データによる予測を行った(多変量RNN)。この際、単変量RNNによるOx予測結果及びその他の気象データをテストデータとした(**図2**)。この際、過去300時間から次の1時間を予測するプログラムとした[1),2)]。

RNNニューラルネットワークモデルは、ニューロン数300の中間層を1つ、活性化関数は線形関数とし、損失関数は二乗誤差、最適化アルゴリズムにAdamオプティマイザーを用いた。また、学習に用いるデータのバッチサイズは256、学習回数は100回に設定した。

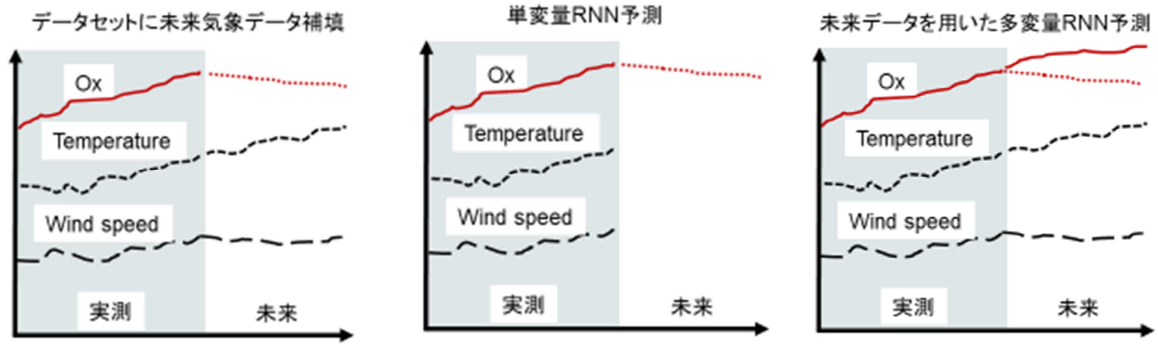


**図 2** 単変量及び多変量 RNN による Ox 時系列データ予測模式図

## 5 気象データ以外の特徴量選定の検討

今回の検討では、気象データのみを用いて Ox の時系列データ予測を行った。今後、さらなる精度の改善を目的に、学習・予測に用いる特徴量の追加を検討するため、気象データに加え、大気汚染常時監視データと Ox 生成の関連性について調べた。方法として、2017 年から 2021 年までの 5 月から 9 月までの間で、1 日の平均気温、最高気温、平均湿度、合計降水量、積算日照時間、合計全天日射量、平均風速、最大風速、現地平均気圧、NOx 日平均値、NMHC 日平均値と Ox 日最高値の相関係数を評価した(ピアソンの積率相関係数)[3]。また、同様のデータセットを用いて、主にクラス分類に用いる機械学習手法であるランダムフォレストによって Ox 予測に対する各特徴量の重要度を評価した[4]。この際、クラス分類のラベルは、Ox 日最高値が 100ppb 以上か否かとした。Ox 予測は、2017 年から 2020 年までの 5 月から 9 月までのデータを訓練データ、2021 年の 5 月から 9 月までのデータをテストデータとした。データは、Ox の他に NOx、NMHC いずれも測定しており、かつ、RNN(LSTM)による時系列データ予測を評価した静岡市立長田南中学校近傍の測定局舎である静岡市常盤公園測定局のデータを用いた(**図 1**)。

本検討で、相関行列の作成には、統計解析用プログラミング言語である R(4.0.3)を用いた。また、ランダムフォレストは Python(3.10)を用いてコードし、モデル構築、学習、推論(予測)、評価に用いたライブラリは Scikit-learn(1.1.2)を使用した。

## 結果

### 1 RNN(LSTM)による Ox 時系列データ予測

Ox 実測値が 6 月 9 日 16:00 に 101ppb のピークがあったのに対し、多変量 RNN による予測(2nd_prediction)では、18:00 に 96ppb の値を出力した。実測値との誤差は 5%未満であった。また、単変量 RNN による予測(1st_prediction)は Ox 実測値の最高濃度到達時刻は合致したものの、予測値は 80ppb と、実測値と比較し約 20%の誤差があった(**表 1**、**図 3**)。

**表 1** 単変量、多変量 RNN による Ox 時系列データ予測結果

| 時刻(2021/6/9) | 11:00 | 12:00 | 13:00 | 14:00 | 15:00 | 16:00 | 17:00 | 18:00 | 19:00 | 20:00 | 21:00 | 22:00 | 23:00 | 0:00 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ox(実測) | 76 | 78 | 80 | 90 | 92 | 101 | 98 | 94 | 88 | 81 | 72 | 66 | 60 | 56 |
| Ox(単変量予測) | 64 | 69 | 76 | 78 | 79 | 80 | 80 | 79 | 77 | 70 | 63 | 63 | 63 | 61 |
| Ox(多変量予測) | 73 | 83 | 84 | 85 | 88 | 91 | 95 | 96 | 90 | 76 | 68 | 62 | 56 | 50 |

(単位：ppb)

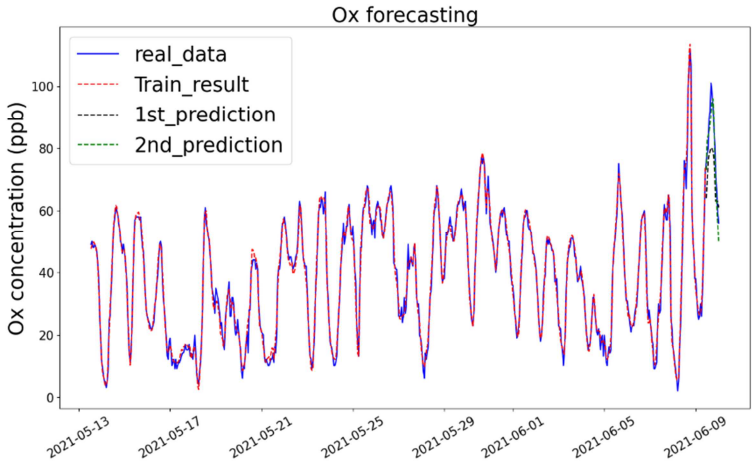


**図 3** 単変量、多変量 RNN による Ox 時系列データ予測グラフ

## 2 気象データ以外の特徴量選定の検討

気象データ、NOx、NMHC の 1 時間値と Ox 日最高値の相関係数を評価した結果、今回の評価対象地点、期間においては、合計全天日射量(相関係数 0.45)、積算日照時間(相関係数 0.42)、平均湿度(相関係数-0.41)の相関係数が比較的大きかった。また、NOx と NMHC では、NOx(相関係数 0.2)、NMHC(相関係数-0.05)と NOx の方が相関係数の値は大きかった(**図 4**)。ランダムフォレストによる Ox 予測に対する各特徴量の重要度の評価では、NOx 日平均値、合計全天日射量、積算日照時間、最大風速、1 日の平均気温、平均湿度、現地平均気圧、NMHC、平均風速、最高気温、合計降水量の順に重要度が高いと判定された(**図 5**)。

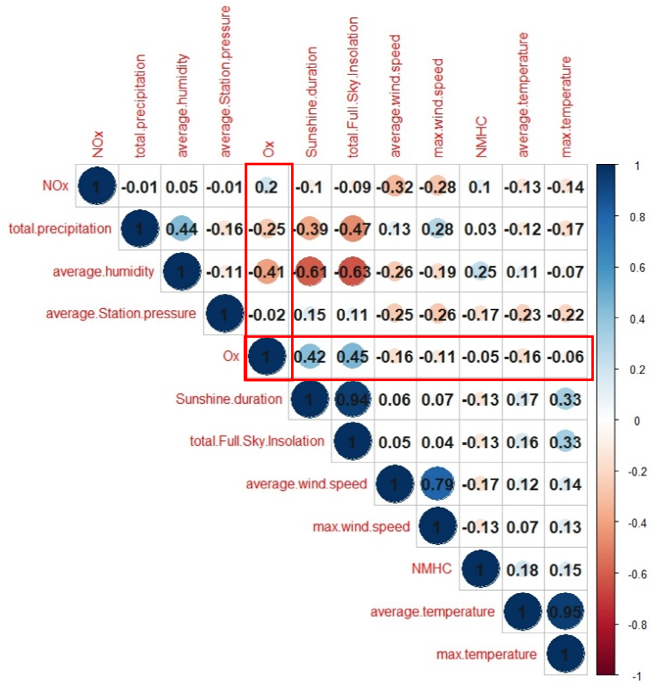


**図 4** Ox 日最高値と各データの相関行列

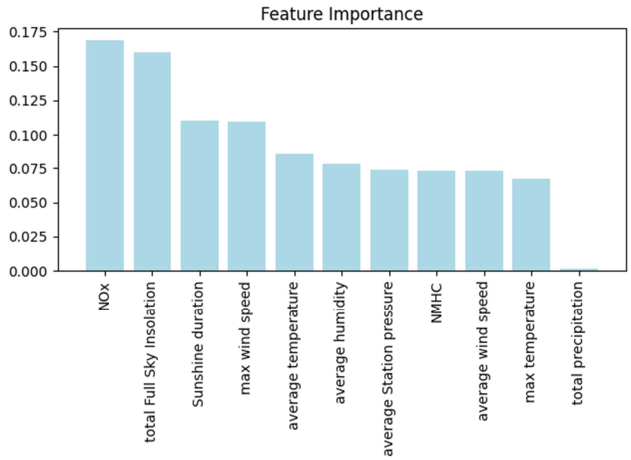


**図 5** ランダムフォレストによる特徴量重要

### 考察

今回の評価対象期間による Ox 時系列データ予測では、今回のケースでは、予測値が実測値のトレンドを外すことなく、かつ、Ox 実測値の最高濃度到達時刻も 2 時間の誤差という結果となった。未来データの 1 回目の予測は単変量 RNN で行ったが、予測は、最後の訓練データから次の未来 1 時間を予測、その予測値をもってまた次の 1 時間を予測、という方法をとった。この方法では、訓練データから時間が遠ざかるにつれ、次第に誤差が大きくなっていく問題が生じ得る。しかし、Ox のように比較的周期性のある時系列データは少なくともトレンドを外さない程度に予測できるようである。また、今回の予測モデルは単変量 RNN による予測値をテストデータにし、8 種類の気象データと Ox データの計 9 変量で再度、学習を行うものであったが、この手法でさらに予測精度が改善した。これは、Ox の時系列予測に気象データの要因を加えたことにより、「補正」がなされたものと思われる。しかし、多変量 RNN によるアルゴリズムを採用した際、多変量 RNN にて順次予測していくことが一般的である。こちらもプログラム自体は完成しているため、今後、精度の検証を進めていきたい。

また、気象データ以外の特徴量選定においては、相関係数の分析、ランダムフォレストによる特徴量の重要度評価ともに NOx は NMHC より Ox 生成に重要であることが示唆された。相関係数の評価及び RNN と別のアルゴリズムであるランダムフォレストによる検証は、RNN の精度との関連は直接的にはないものの、今回、参考として解析した。

今回の「未来予測」は、過去データを用いて試行しているが、実際に未来予測を行う際には、気象データは天気予報中の「予測値」を用いることを想定している。NOx も学習・予測に用いる特徴量の 1 つとして採用する場合、何らかの方法で NOx も時系列データ予測を行う必要がある。今後、あらゆる時系列ポイントで予測を行い、本予測手法が実用に耐え得るか検証を重ねていく必要がある。この際、併せて精度の更なる改善も進めていく予定である。


### (謝辞)

今回の検証では、静岡市様所有の大気常時監視局舎のデータを使用させていただきました。御協力に深く感謝申し上げます。


(参考文献)


1) https://github.com/ishikawa08/tf_multi_LSTM/ 2023 年 3 月 10 日最終確認

2) 巣籠悠輔:詳解ディープラーニング[第 2 版], p270-271, マイナビ出版(2020)

3) Winston Chang:R グラフィックスクックブック-ggplot2 によるグラフ作成のレシピ集, p274, オーム社(2016)

4) Andreas C. Muller, Sarah Guido:Python ではじめる機械学習-scikit-learn で学ぶ特徴量エンジニアリングと機械学習の基礎, p83-85, オーム社(2020)