

感染症例数予測のための時系列解析手法及び機械学習の応用と検証 Application and Validation of Time Series Analysis Methods and Machine Learning for Forecasting the Number of Infection Cases

1255055 土肥 直樹 (Soft Intelligent System On Chip 研究室)
(指導教員 星野 孝総 准教授)

1. はじめに

1.1 研究背景

紀元前より人類は感染症と闘い続けてきた。特に近年では、COVID-19やサル痘が流行している。このような背景から、予測モデルの需要が高まってきている。そこで、本研究ではCOVID-19及びサル痘の症例数予測モデルを構築する。

1.2 関連研究

Wuら[1]は、インフルエンザ様疾患発生数の予測を統計モデル(Auto-Regressive Integrated Moving Average:ARIMA)と機械学習モデル(Long Short-Term Memory:LSTM、Attention機構付き Sequence To Sequence:Seq2Seq、Transformer)の予測精度を比較した。この際、比較の基準としてARIMAのRoot Mean Square Error (RMSE)を設定している。彼らの結果では、TransformerがARIMAのRMSEから最も減少させることができた。Transformerは様々な分野で高い精度を誇るモデルで、時系列予測分野ではZhou[2]らによって時系列予測用に改良されたInformerが提唱されている。通常のモデルは、再帰的に予測値を入力することで次の時刻の予測を行う。Informerは、1度に複数時刻予測することが可能で、これにより精度を上げている。しかし、Transformerベースのモデルの学習には大量のデータが必要であり今回の感染症予測には不向きである。また、Spyrosら[3]は、機械学習モデルが古典的な統計モデルに代替可能か検証した。その結果、統計モデルの方が精度が高かった。Wuら[1]ら、Spyrosら[3]の研究より、統計モデルと機械学習の比較は重要であると我々は考える。そこで、本研究では図1に示す小規模なモデルのRecurrent Neural Network(RNN)を基本としたモデルをZhouら[2]の研究を参考に複数時刻予測できるモデルとして構築し、統計モデル(ARIMA、SARIMA)との予測比較を行った。また、Wuら[1]らの研究を参考にARIMAのRMSEをベースラインとして用い彼らの研究結果とも比較した。

1.3 研究目的

本研究では、最適な症例数予測モデルの発見を最終目的とする。まず、実験1を以降の実験で使用するRNNにおける最適な入出力関係を明らかにする目的で行った。次に、実験2及び実験3は、感染症症例数予測において最適な予測モデルを探索する目的で行った。

2. 研究内容

実験1では、3種類の入出力のRNNを第2波までのCOVID-19のデータを用いて学習し構築した。その後、ARIMAとSARIMAと比較し最適な入出力を選択した。

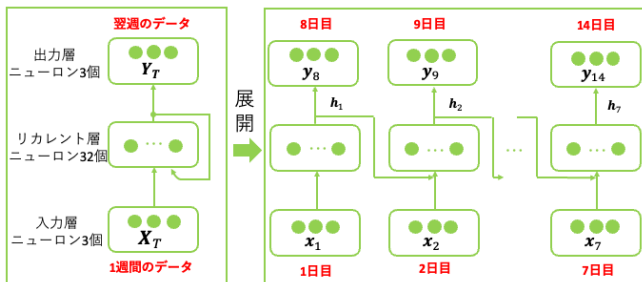


図1:RNN

なお、ARIMAのRMSEとの比較は式1を用いて行った。

$$\left(\frac{RMSE}{ARIMA's RMSE} - 1 \right) \times 100 - (1)$$

次に、実験2では実験1で決定した入出力関係と第7波までCOVID-19のデータを用いてRNN、LSTM、Seq2Seq、Stacked LSTMを構築しARIMA及びSARIMAと比較を行った。最後に、実験3では実験2で決定した入出力関係とサル痘のデータを用いてRNN、LSTM、Seq2Seq、Stacked LSTMを構築しARIMA及びSARIMAと比較を行った。また各実験において、ARIMAとSARIMAの関係においてモデルが季節成分を含むかを確認するため各モデルの残差における自己相関のコレログラムを描画した。

3. 実験結果と考察

実験1の結果を表1に示す。表1中の数値to数値は入力次元to出力次元の意味である。また差(%)は、式1によって算出された。表1より、実験1において最も精度の高いモデルは3次元入力3次元出力であった。この結果は、(Self)Adaptive-Attention[4]によってもたらされたと考ええる。Attentionは、Wuらの実験で最も高い精度であったTransformerにも用いられている機構である。この結果より、以降の実験で用いる入出力関係を3次元入力3次元出力に決定した。

表1: COVID-19のデータを学習したモデルの評価(実験1)

	ARIMA	SARIMA	1to1	3to1	3to3
RMSE	464	67	110	59	47
差(%)	0.0	-85.6	-76.3	-87.3	-89.9

次に、COVID-19のデータを用いた実験2及びサル痘のデータを用いた実験3の結果とWuらのインフルエンザ様疾患発生数予測モデルとの比較を行った。表2に、統計モデル及び機械学習モデルの中で最も精度の高かったモデルとWuらの研究で最も精度の高かったTransformerのRMSEと各症例数用のARIMAのRMSEからの増減率を示す。なお、表中の差(%)は、式1によって算出された。

表2: 実験2及び実験3の結果とWuらのモデル比較

	Transformer (By Wu[3])	SARIMA (COVID)	LSTM (COVID)	SARIMA (サル痘)	RNN (サル痘)
RMSE	0.588	6707	10827	276	360
差(%)	-42.4	-76.2	-61.6	-83.1	-77.9

表2より、COVID-19とサル痘どちらにおいても統計モデルのSARIMAの結果が最も良かった。RNNやLSTMを含め我々の結果はWuらのTransformerを約20%超える結果を示した。また、各実験の統計モデルにおいて残差のコレログラムを描画した結果、ARIMAにおいては残差に周期的な自己相関が観測された。この結果より、SARIMAの精度が高いのはモデルが季節性を含んでいることが起因していると考えた。

参考文献

- [1] Wu, Neo, Green, Bradley, Ben, Xue, O'Banion, Shawn: "Deep transformer models for time series forecasting: The influenza prevalence case", arXiv preprint arXiv:2001.08317, 2020.
- [2] Zhou, Haoyi, Zhang, Shanghang, Peng, Jieqi, Zhang, Shuai, Li, Jianxin, Xiong, Hui, Zhang, Wancai: "Informer: Beyond efficient transformer for long sequence time-series forecasting", Vol. 35, No. 12, pp. 11106-11115, 2021.
- [3] Makridakis, Spyros, Spiliotis, Evangelos, Assimakopoulos, V assilios: "Statistical and Machine Learning forecasting methods: Concerns and ways forward", PloS one, Vol.13, No.3, p.e0194889, 2018.
- [4] Bahdanau, Dzmitry, Cho, Kyunghyun, Bengio, Yoshua: "Neural machine translation by jointly learning to align and translate", arXiv preprint arXiv:1409.0473, 2014.