COVID-19 感染者数予測モデルにおける性能比較と検証

1210113 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

現在、新型コロナウイルスの流行により経済的打撃やリモートワークの推進など大きな社会体制の変化が起きている。そこで本研究は、病床の早期確保、ワクチンの確保、緊急事態宣言等の行政的判断の早期決断のために新型コロナウイルスの感染者数予測を行った。また、先行研究[1]により時系列予測をする際は統計モデルと機械学習モデルを比較するように指摘されていた為、統計モデル(自己回帰(AR)モデル、自己回帰和分移動平均(ARIMA)モデル、季節変動自己回帰和分移動平均(SARIMA)モデル)と機械学習モデル(多層パーセプトロン(MLP)モデル)、再帰型の機械学習モデル(リカレントニューラルネットワーク(RNN)モデル、ロングショートタイムメモリー(LSTM)モデル、ゲート付きリカレントユニット(GRU)モデル)を用いて予測を行った。

2. 実験内容・方法

2.1 累積確認症例数予測

累積確認症例数の予測を統計モデルでは、AR(1)モデルとSARIMA(*p,d,q,P,D,Q*)モデルを用いて行った。SARIMA(*p,d,q,P,D,Q*)モデルにおいては最適なパラメータを見つけるために *p*=1 から *p*=3、*d*=0 から *d*=2、*q*=0 から *q*=3、*P*=0 から *P*=3、*D*=0 から *D*=2、*Q*=0 から *Q*=3 の範囲で総当たり探索を行った。

機械学習モデル及び再帰型の機械学習モデルでは、入力層及び出力層のニューロン数を 1 個とし、中間層のニューロン数またはブロック数を 16 個、32 個、64 個、128 個、256 個、512 個、1024 個で総当たり探索を行った。また、ドロップアウトによりマスクするニューロンの比率を 0%、20%、40%、50%、60%で総当たり探索行った。

2.2 新規感染者数予測

新規感染者数の予測は ARIMA(p,d,q) モデルと SARIMA(p,d,q,P,D,Q)モデルを用いて行った。ARIMA(p,d,q)モデルにおいては最適なパラメータを見つけるために *p*=1 から *p*=3、*d*=0 から *d*=1、*q*=0 から *q*=3 の範囲で、SARIMA(*p,d,q,P,D,Q*)モデルにおいては最適なパラメータを見つけるために *p*=1 から *p*=3、*d*=0 から *d*=1、*q*=0 から *q*=3、*P*=0 から *P*=3、*D*=0 から *D*=2、*Q*=0 から *Q*=3 の範囲で総当たり探索を行った。機械学習モデル及び再帰型の機械学習モデルでは、入力層及び出力層のニューロン数を 1 個とし、中間層のニューロン数またはブロック数を 16 個、32 個、64 個、128 個、256 個、512 個、1024 個で総当たり探索を行った。また、ドロップアウトによりマスクするニューロンの比率を 0%、20%、40%、50%、60%で総当たり探索行った。

なお、両データセットの機械学習モデル及び再帰型の機械学習モデルの学習は 7 日間のデータを入力とし正解ラベルを 8 日目とした。予測時は直前 7 日間を入力とし、8 日目を予測した。また、予測値を次の予測時の入力として用いた。

2.3 評価関数

評価関数として次の式 1 で定式化される二乗平均平方根誤差(RMSE)を用いた。

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \quad (1)$$

式 1 中のy_iは観測値を、$\hat{y}_i$は予測値を、nはデータ数を意味する。二乗平均平方根誤差は誤差を二乗することで負の値をとらないようにしている。また、二乗することで予測が外れた際に大きな値になる。

3. 実験結果

表 1 に各モデルで累積確認症例数を予測した際の評価値を示す。

累積確認症例数予測において最も精度が良かった統計モデルは表 1 より SARIMA モデルであった。また、機械学習モデルでは、GRU モデルであった。また、全モデルで最も良い評価を得たのは SARIMA モデルであった。

表 1:各モデルによる累積確認症例数予測の評価(単位:人)

	AR	SARIMA	MLP	RNN	LSTM	GRU
RMSE	2,8001	127	342	425	441	211

次に、表 2 に各モデルで新規感染者数を予測した際の評価値を示す。

新規感染者数予測において最も精度が良かった統計モデルは表 2 より SARIMA モデルであった。また、機械学習モデルでは、MLP モデルであった。また、全モデルで最も良い評価を得たのは SARIMA モデルであった。

表 2:各モデルによる新規感染者数予測の評価(単位:人)

	ARIMA	SARIMA	MLP	RNN	LSTM	GRU
RMSE	103	70	75	90	115	89

4. 考察

本研究の結果として SARIMA モデルが最も良い結果を得た。しかし、再帰型の機械学習モデルには精度向上の余地があると考える。これは、SARIMA モデルが季節変動成分の相関を含むことでその精度が上がったことが表 1 の AR モデルとの比較及び表 2 の ARIMA モデルとの比較した際に考えられるからである。従って、再帰型の機械学習モデル 3 種では 7 日間の訓練データに対して 8 日目を正解データにするのではなく、次の 7 日間の正解データを学習させる必要があると考える。これは、8 日目だけを正解データにした際は、中間層の最後の隠れ状態ベクトル$\boldsymbol{h}_t$のみ出力層に渡される。それ対し、次の 7 日間を正解データにした場合は、各時刻における中間層の隠れ状態ベクトル$\boldsymbol{h}_t$が出力層に渡されるため 1 日ずつの誤差を最小化することができる。さらに、ある日のデータと 1 週間後のデータから予測を行うという考え方は SARIMA モデルの考え方と似ている為、今回最も精度の良かった SARIMA モデルと同程度もしくはよりも高い精度が期待されるからである。

また、今回-1 から 1 で正規化したデータを活性化関数の ReLU 関数(0 以下の入力の出力を 0、0 より大きい入力の場合はそのまま出力する)に入力したため学習が不十分であったと考えられる。その為、正規化の範囲の変更もしくは活性化関数の変更によって精度が向上すると考えられる。

また、同様に複数のデータを入力することによる精度向上も期待される。

5. 結論

本研究の結果では統計モデルである SARIMA モデルが最も良い予測精度を出した。ただし、再帰型の機械学習モデルにおいては訓練データと正解データの関係や適切な正規化の範囲の設定もしくは活性化関数の設定、複数のデータの入力により SARIMA モデルと同程度もしくはより良い精度がもたらされると期待される。その為、引き続き検証を行っていく必要がある。

参考文献

[1] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. "Statistical and machine learning forecasting methods: Concerns and ways forward," PloS one, Vol.13, No.3, pp.e0194889, 2018.