LSTM を用いた covid-19 の感染者数の予測

坪野谷　友都
Yuto Tsubonoya
法政大学情報科学部コンピュータ科学科
yuto.tsubonoya.8d@stu.hosei.ac.jp

Abstract

The COVID-19 epidemic that occurred in late December 2019 is still spreading rapidly in many countries and regions of the world. Therefore, there is an urgent need to predict the outbreak and spread of the epidemic. Nowadays, machine learning is often used to predict infectious diseases. In this study, I used LSTM and RNN, which are excellent for predicting time series data, to predict the number of COVID-19 infections in Japan. In the process of learning, I also focused on the weekly periodicity from the testing regime of the PCR test for COVID-19. To achieve this, I implemented Conv1D + LSTM by including a 1D convolutional layer in the input layer of machine learning. As a result, I was able to achieve higher accuracy in prediction than that produced by the results of previous studies. As a discussion, we compared three evaluation metrics: mean absolute error, mean squared error, and coefficient of determination.

1　はじめに

新型コロナウイルス (COVID-19) は、感染症である。COVID-19 は、2019 年 12 月に中国の武漢（湖北省の州都）より世界に広がった。そして、ウイルスパンデミックとして現在も世界中で感染が進行している。このウイルスは多くの患者の急性呼吸器症候群を引き起こすことで知られている。そのため、現在では COVID-19 は国際公衆衛生上の緊急事態として考えられている。厚生労働省のオープンデータを確認すると、日本国内では、2020 年 01 月 18 日現在の感染人数は 321,484 例、死亡者数は 4,445 名となっており、被害が拡大している。国連事務総長は、COVID-19 を、人間の健康を直接脅かし、世界中の経済、社会、環境の発展に影響を与える世界的なカタストロフィーと表現している。そのため、感染症流行の予測は公衆衛生における 1 つの課題である。感染症流行は患者数の増加による医療現場の逼迫、経済的損失や混乱を引き起こすことから、流行を予測することは予防や流行の対処といった観点からも重要である。また、医療現場の問題としては、主に 3 つが挙げられる。新型コロナウイルス感染者の受け入れのための病床の確保、人材の確保、救急搬送などである。要因はいろいろ考えられるが、それらが集積した結果が感染という形では簡単に観測できる、もしくは観測データが公開されているものから間接的に推測できるようなニューラルネットワークの構築を目指す。

Supervisor: Prof. Katunobu Itou

2　covid-19 の感染者数の予測

現在では、COVID-19 の研究や予測が世界中で行われている。従来の感染症予測については数理モデルを取り扱うことが一般的であった。主要な数理モデルに関しては、SIR モデルや SEIR モデルなどの感染症の短期的な流行過程を決定論的に記述する古典的なモデル方程式である（2020 年 11 月 17 日に google が発表した COVID-19 の感染予測もこの数理モデルを用いている。）。また、昨今では、機械学習を用いた感染症予測も多い。機械学習の例としては、RNN、LSTM、CNN（Convolutional Neural Network）などが挙げられる。従来研究で使用されている LSTM、ハイパーパラメータを用いて単日の感染者数、死亡者数、回復者数の 3 つの特徴量を用いて予測を実施する。

本研究では従来研究 [6] などを参考に、日本における 1 日当たりの PCR 検査陽性者数などの特徴量を入力データに使用し、機械学習を用いて covid-19 の日々の感染予測を行う。従来研究では、人口の多い国であるアメリカ、中国、インドネシアなどにおける covid-19 の流行予測が行われている。先行研究では入力データとして主に各特徴量の累積数が用いられる。特徴量の累積数における回帰は単調増加であることに加えて、相対的な誤差が少なくなるため比較的、容易に機械学習ができるためだと考えられる。多くの先行研究では、特徴量として累積感染者数、累積死亡者数、累積回復者数などが使用される。そのため、1 日当たりの感染者数などの特徴量を用いて機械学習を行う研究は少なくなっている。従来研究では LSTM を 4 層使っている。ここで、ニュラルネットワークの構成を図 1 に示す。

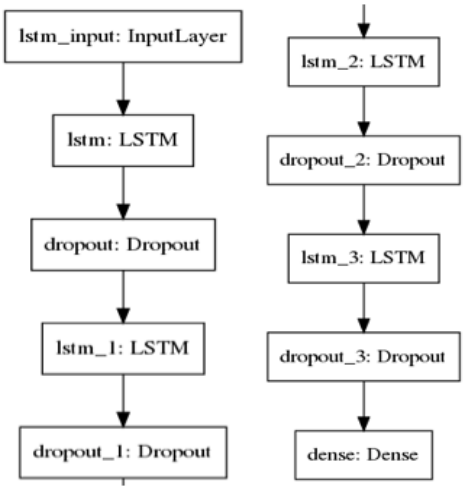

図 1. 従来研究のニュラルネットワーク構成

それぞれのハイパーパラメータについて、LSTM 1 層におけるユニット数は 30 個、ドロップアウトは 0.1、活性化関数はシグモイド関数、最適化アルゴリズムは Adam となっている。こ

れらを用いて、感染者数を予測した結果を図 2,3,4 に示す。

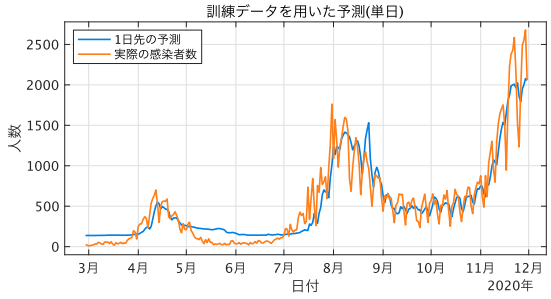

図 2. 訓練データによる 1 日当たりの予測

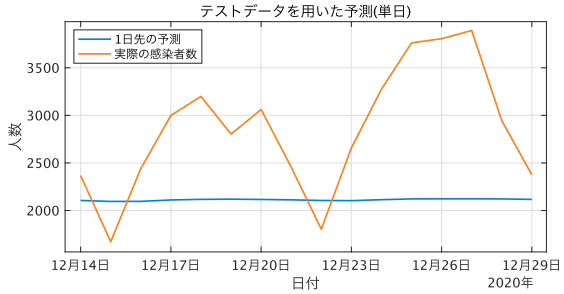

図 3. テストデータによる 1 日当たりの予測

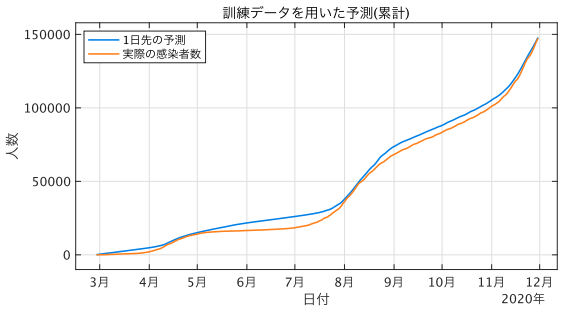

図 4. 訓練データによる累積値の予測

訓練データの期間は 2020 年 02 月 15 日～12 月 01 日、テストデータの期間は 2020 年 12 月 02 日～12 月 29 日である。図 2 は訓練データを用いた 1 日先の感染者数の予測を行ったものである。図 3 はテストデータを用いて同様に予測を行ったものである。図 4 は図 2 の累積を表している。国内の新型コロナウイルスの現状としては、4 月-5 月、8 月-9 月、11 月-12 月にかけて感染のピークが現れている結果となっている。日本の covid-19 にへの対策の一つに緊急事態宣言などが挙げられる。図 2 に示すように、緊急事態宣言 (期間は 2020 年 04 月 07 日～2020 年 05 月 25 日) が発令される前と後では、コロナウイルス陽性者数の数は大きく異なる。これは「人の移動」が著しく制限されたためである。また、緊急事態宣言が解除されてから数ヶ月経過した今では、「人の移動」が盛んになり、コロナウイルス陽性者数は感染が拡大した初期段階の 3 月と比べても約 2 倍以上、緊急事態宣言が発令した場合と同じ状況になっていることが図 2 より見て取れる。このことから、感染症の流行は人の移動量に大きく関係していることが推測できる。また、グラフの傾向としては徐々に増加している。急激に増加しているわけではなく、増加と減少を繰り返しながら、増えている。

訓練データの 1 日先の予測に関しては、ある程度の増加や減少するなどの傾向を追えている。しかし、細かいところの予測、感染の急激な増加（12 月はじめ）などの予測はできていない。また、テストデータの予測に関しても、なるべく誤差が少なくなるような値を取り続けているような予測を行っており、細かい部分の予測できておらず、増減などの傾向が掴めていない。一方、予測の累積のグラフに関しては、100 日～150 日あたりの感染者数が停滞していた時期（緊急事態宣言が発令が解除されたあたり）の予測に 1 万人強の差はでているが、それ以前や以降では、ほぼ実際の曲線と類似している。予測を累積したグラフの方が 1 日あたりの予測した結果よりも直感的には正確な予測が行えているように感じる。 1 日あたりの感染者数の方では、大きく差がでているような予測になっているにもかかわらず、累積の予測では精度が高くなってしまう。これは、累積表示では、感染者数の値が減少することがなく、増え続けてしまう単調増加が現れるので、相対的な誤差が少なくなるように見えてしまうためである。そこで、本研究では、 1 日あたりの予測でも特徴量の累積数を用いた予測と同等の結果が得られることを示す。また、構築したモデルで学習を行い、比較、検討も実施する。

3　1 次元畳み込みと LSTM

先行研究で用いられているデータセットでは主に、covid-19 の PCR 検査の感染者数の累計、死亡者数の累計、回復者数の累計などが特徴量として用いられ、学習で使用される期間としては 2 ヶ月程度である。[4] 特徴量については、機械学習のモデルを作成する際、効率性の観点から、複数の特徴量を選別し、目的に必要な特徴量のみを選択する必要性が生じる。考察として、評価指標の RMSE や MAE の値を算出した。評価指標は、１６日間先を予測したものに相関があるのか、精度が高いのか低いのかなどの類似点や相違点を見つける。

3.1　畳み込みによる移動平均

検査体制の影響から、感染者数には周期的な特徴が存在する。そのため、ある程度の傾向を捉えるため、移動平均（畳み込み）を行う。 7 日間で畳み込みを行った結果を、図 5 に示す。

図 5. 畳み込みの結果

全体的に感染者数の一番多いピークのポイントを取ることができている。特に、5 月終わり～8 月初めや 10 月終わり～12 月初めではピークを正確に捉えることができている。このことから、1 次元畳み込み (Conv1D) ＋ LSTM による感染者数の予測を提案する。Conv1D+LSTM のモデルの例を図 6 に示す。

3.2　学習用データセット

LSTM は RNN の拡張したネットワークであり、株式予測、気象など多様な時系列トピックに用いられている。つまり、時系列データの機械学習に用いられている。RNN は予測精度を示す上で、ベースラインとして考えられている。

本研究では、学習を少しでも容易にするため、データセットの値が最小値が 0 ，最大値が 1 になるように正規化を施す。正規化には Python ライブラリの sklearn.preprocessing.MinMaxScaler() を利用した。MinMaxScaler() における正規化の式は以下の通りである。

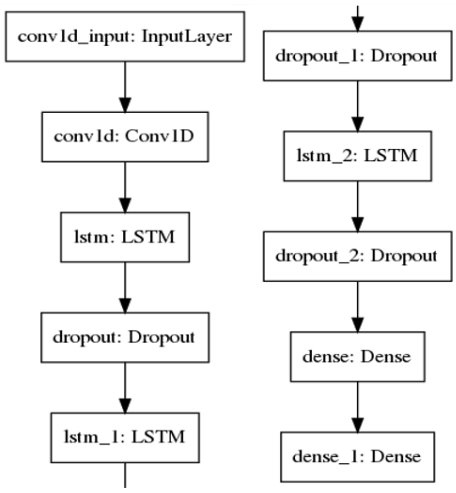

図 6. Conv1D+LSTM

式 1 に示す。ここで、X は元のデータセットを表し、X_{new} は新しく正規化されたデータセットを表している。

$$X_{new} = \frac{X - min(X)}{max(X) - min(X)} \tag{1}$$

LSTM で学習させるためにはデータセットの整形を行う必要がある。LSTM では、過去のデータを参照して未来のデータを予測をする。そのため、モデルを作成する際の入力データと出力データは表 1 のような形に修正する。

表 1. LSTM のためのデータセット

入力データ	出力データ
t, t+1, t+2, ..., t+step	t+step+1
t+1, t+2, t+3, ..., t+step+1	t+step+2
t+2, t+3, t+4, ..., t+step+2	t+step+3
...	...

ここで t は時系列データの特徴量、$step$ は日数を表している。何日間のデータを基に予測をするのかを決めている。本研究では、$step$ は 12 日間である。つまり、12 日間のデータ (感染者数、死亡者数、回復者数) からその次のデータ (t+step+1 日目のデータ) を予測する。

4 評価

4.1 実験

先行研究では covid-19 の RNN と LSTM の比較を示している。その結果では、LSTM の方が精度が高いことが示されている。本研究では、先行研究を参考に評価指標である RMSE（二乗平均平方根誤差）及び MAE（平均絶対誤差）を使用している。式（1）、（2）はそれぞれ RMSE と MAE を表している。また、16 日先までの予測に関しては、2 つの指標に加えて決定係数 R^2 を使用している。

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{n}(f_t - y_t)^2} \tag{2}$$

平均平方二乗誤差（RMSE）は，予測値と実測値との差の二乗和の平均から平方根をとることにより求められるもので，測定値のバラツキ具合を数量的に表すものである。平均二乗誤差が小さいほど，その測定精度は良いと判断する。

$$MAE = \frac{1}{N}\sum_{t=1}^{n}|f_t - y_t| \tag{3}$$

平均絶対誤差（MAE）は予測値と実測値の単純な指標である。実測値と予測値の絶対値を平均したものである。 MAE が小さいほど誤差が少なく、予測モデルが正確に予測できていることを示し、MAE が大きいほど実際の値と予測値に誤差が大きく、予測モデルが正確に予測できていないといえる。（y_t:実測値, f_t:予測値）

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(y_t - f_t)^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2} \tag{4}$$

観測値とモデルから計算した計算値（予測値）がほぼ同じになると、次式の分子が 0 に近づくため、R^2 は 1 に近づく。逆に、観測値と予測値がかけ離れていると、分子が大きな値となり、R^2 は 1 から離れた値となる。先行研究 [4] では、RNN や LSTM における最適なハイパーパラメーターの検討が行われている。先行研究では、3 つのハイパーパラメーターについて検討されていた。本研究では、ユニットの数、隠れ層の数、ドロップアウト、活性化関数、最適化アルゴリズムの計 5 つを検討した。検討するライブラリとしては optuna を導入し、活用した。本研究で使用した最終的なハイパーパラメーターを表 2 に示す。

表 2. ハイパーパラメーター

隠れ層の数	畳み込み層:1 LSTM:3 Dense:1 出力層:1
ユニットの数	畳み込み層:200 LSTM:150-50 Dense:25 出力層:3
ドロップアウト	0.2
活性化関数	relu
最適化アルゴリズム	Adam

これらの設定により、訓練データやテストデータに関しては良好な結果が得られた。本研究では、 1 日あたりの感染者数などの特徴量を用いて学習を実行した。データセットの区間は、2020 年 02 月 15 日～2020 年 12 月 29 日までの 319 日間を使用している。そして、データセットを訓練データとテストデータで 9 ： 1 の割合に分割している。訓練データの区間はデータセットにおける 2020 年 2 月 15 日から 2020 年 12 月 01 日までのデータを使用している。テストデータの区間は 2020 年 12 月 02 日から 2020 年 12 月 29 日までのデータを使用している。それに加えて、学習のデータセットに「 1 日当たりの感染者数」、「 1 日当たりの死亡者数」、「 1 日当たりの回復者数」を使い、 1 6 日間先までの予測を行う。また、データセットの検討や機械学習におけるハイパーパラメータの検討なども実施する。学習結果については、どの程度の予測ができているかを確認するため、評価指標である RMSE（二乗平均平方根誤差）、MAE（平均絶対誤差）、R^2（決定係数）、MRE（平均相対誤差）による精度とその考察を示す。Conv1D+LSTM による学習結果を図 7、図 8、図 9 に示す。また、評価指標による結果を表 4.1 に示す。

それぞれのグラフは、学習結果から 1 日当たりの感染者数のみを抽出しグラフにしたものである。この 3 つのグラフの中で 16 日間先の予測をしているグラフに関しては、訓練データとテストデータの検証の際の予測方法とは異なっている。訓練データとテストデータの検証では、実際のデータ 12 日分を用いて

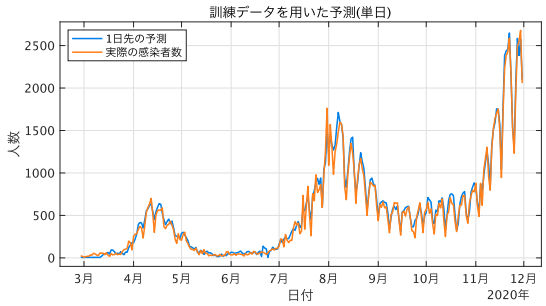

図 7. 訓練データの学習結果

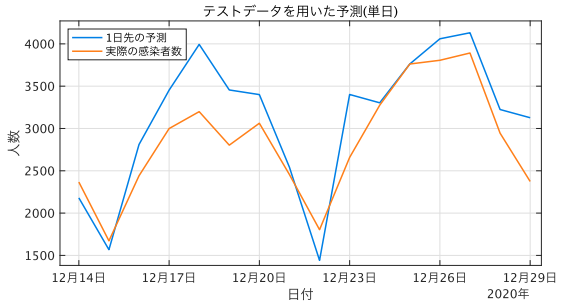

図 8. テストデータの学習結果

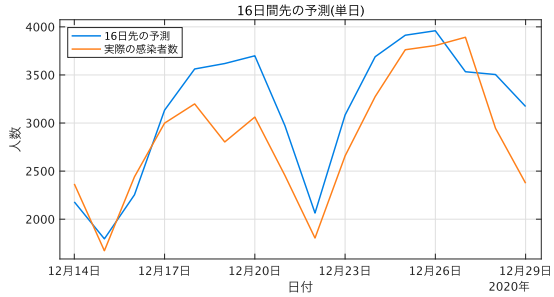

図 9. 16 日間先までの予測

表 3. RMSE, MAE, R^2

	訓練データ	テストデータ	16 日間先の予測
RMSE	70.44	434.40	441.50
MAE	52.01	352.91	380.37
R^2	0.98	0.70	0.571

1 日先を予測している。それに対して、この予測では、実際のデータ 12 日分から 1 日先を予測した後、予測したデータを実際の 12 日分のデータに追加する。13 日分のデータの内、追加された 1 日分を含む後ろの 12 日分を使用して 1 日先を予測する。これを繰り返すことで 16 日先まで予測している。

4.2 考察

まず、訓練データでは、RMSE が 70.44、MAE が 52.01 となっており、図 7 を確認しても予測した曲線は実際の曲線とほぼ同じようなグラフを描いている。予測した曲線は実際の曲線と同様に緩やかに増加している。また、緊急事態宣言が発令した期間の減衰も予測できており、50 日前後や 175 日前後で減少傾向に転じるような、増加や減少を含めて周期性があるような似た特徴を予測曲線でも示せている。次に、テストデータでは、RMSE が 434.40、MAE が 352.91 となっている。実際の曲線と予測曲線はほぼ同じような軌跡を描いている。訓練データのときと同様に大まかな増加や減少 (4 日目まで増加し、8 日まで大まかに減少している傾向やそれ以降の 10 日目から 14 日目まで凸のような曲線) を示せている。しかし、所々で差がでてしまっている。4 日目や 7 日目 16 日目などの予測は最大で約 1000 人以上の差がでており、実際よりも人数が増えすぎていたり、減りすぎているような結果となっている。この差は RMSE や MAE などを確認しても開きがある。この原因としては、訓練データとテストデータの RMSE や MAE を比較すると、MAE はどちらも低い数値ではあるが、RMSE に関しては約 350 程度の差が存在しており、ほんの少しだけ訓練データに学習がフィットしていることが考えられる。最後に、16 日間先の予測に関しては、RMSE が 441.50、MAE が 380.37、R^2 が 0.571 となっている。予測方法が異なるため、2 つの結果よりも RMSE、MAE ともに差がほんの少しだけ（50 程度）大きくなっている。ただ、決定係数の値を見ると、ある程度グラフの概形を取ることができている。　しっかりとグラフに着目すると、予測曲線は全体的に実際の曲線よりも値が大きくなってしまっている。また、予測曲線は実際の曲線と比べてみると 3 日目当たりから 6 日目までの細かい増減の傾向は予測できていない。また、12 日目あたりから増加や減少する傾向が左方向に早くなるようにずれている。結果として、単純な LSTM よりも 1 次元畳み込みを加えた conv1D+LSTM の方がコロナウイルスの感染者数の予測には有用であることが分かった。

5 おわりに

本研究では、先行研究の累積値の予測ではなく、1 日あたりの正確な予測を目的とし、日本における COVID-19 の感染者数を予測した。単純な LSTM 層だけでなく、1 次元畳み込み層を加えることで従来研究よりも良い結果が得られた。具体的には、評価指標の RMSE や MAE の両方の値で、先行研究よりも 100 以上誤差を少なく予測することができている。Conv1D+LSTM では、大まかな傾向の予測は Conv1D で学習でき、細かい部分の予測を LSTM 層で学習するという分業体制ができたため、うまく予測ができたのではないかと考えている。

参考文献

[1] Yuexin Wu, et al. “Deep Learning for Epidemiological Predictions”, In Proc. of SIGIR, pp.1085-1088, 2018

[2] Yaguang Li, at al., “Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting”, In Proc. of ICLR, 2018.

[3] 村山太一, 清水伸幸, 藤田澄男, 若宮翔子, 荒牧英治, 位置関係を考慮した地域ごとのインフルエンザ流行予測, 2019

[4] Novanto Yudistira, ”COVID-19 growth prediction using multivariate long short term memory”, 2020, https://github.com/cbasemaster/lstmcorona

[5] Ratnabali Pal , Arif Ahmed Sekh , Samarjit Kar, Dilip K. Prasad, ”Neural Network Based Country Wise Risk Prediction of COVID-19”, Department of Computer Science, UiT The Arctic University of Norway, August 2020

[6] Sourabh Shastri, Kuljeet Singh, Sachin Kumar, Paramjit Kour, Vibhakar Mansotra. ”Time series forecasting of Covid-19 using deep learning models”, Department of Computer Science and IT, University of Jammu, Jammu, Kashmir, India India-USA comparative case study”, 2020 August