F-023

ヒストグラムの簡略化が時系列予測に与える影響
Effect of histogram simplification on time series forecasting

野口 真[1)]　　徳山 豪[1)]
Makoto Noguchi　Takeshi Tokuyama

1　はじめに

時系列データのような 1 次元のデータ分析の際に用いられるデータの表現方法の 1 つにヒストグラムがあり，それは一意にピークを持つ単峰性ヒストグラムと，その単峰性ヒストグラムが複数結合された多峰性ヒストグラムに分けられる．ピークを多く持つ多峰性ヒストグラムを用いてデータ解析や予測を行う場合，近似ヒストグラムを計算して入力とすることは解析の易化や過学習の抑制を可能とする. 特にヒストグラムの単峰部分の抽出においてはガウス分布を用いた曲線近似が広く用いられているが，実際のデータはピーク数が多く凹凸の激しいヒストグラムデータであることが多く，元データと近似曲線との誤差が非常に大きくなってしまうことがあるため，データを最適に近似する多峰性ヒストグラムを求めることが重要な問題となる．[1] では，入力がヒストグラム関数である場合に，単峰部分毎の L2 二乗距離が任意の閾値以下であるという条件の下，単峰部分の数が最小となる近似ヒストグラムを考えるピーク数最小化近似問題とその最適化手法が提案されている．

また，中国の武漢にて発生した Covid-19 の世界への感染拡大に伴い，その感染者数や死者数等の解析，予測に深層学習モデルや感染症数理モデルを用いた様々な手法が用いられている．その 1 つに，LSTM(Long Short Time Memory) がある．これは RNN(Recurrent Nweural Network) を拡張したモデルとして知られており，株式の予測や，気象をはじめとする様々な時系列データの予測に用いられている．また，Shastri らは Covid-19 の時系列予測を行う深層学習モデルとして，1 次元畳み込みを用いた LSTM(ConvLSTM) を提案した [2]．これによりデータは平滑化され，データ解析の精度や汎化性能の向上が見込まれる．

そこで本研究では，よりデータの揺れが小さい平滑化として，[1] の単峰近似アルゴリズムを応用した k ピーク近似アルゴリズムを実装する．更に，これを入力に加えた LSTM ネットワークを構築し ConvLSTM との比較実験を行うことで，より良いヒストグラムの簡略化が時系列予測の精度向上に有効であることを示す．

2　関連研究

2.1　1 次元畳み込みを用いた LSTM

2015 年に Shi らによって畳み込み LSTM(ConvLSTM) が提案された [3]．これはデータの空間情報を保持しながら，それを時間方向に扱うことを可能にした手法であり，特に動画像の処理において優れた予測性能を達成している．この ConvLSTM を応用し，Shastri らは Covid-19 の時系列予測を行う深層学習モデルとして，1 次元畳み込み (Conv1D) を用いたモデルを提案した [2]．このモデルは，インドとアメリカ合衆国における Covic-19 の 1 ヶ月先の感染者数等の予測において，Stacked LSTM, Directional LSTM を用いた他のモデルよりも高い予測精度を示している．本研究における比較モデルとしてこの ConvLSTM を用い，畳み込みについては Covid-19 の検査体制による感染者数の周期的な特徴を考慮し，7 日間の移動平均を計算する．

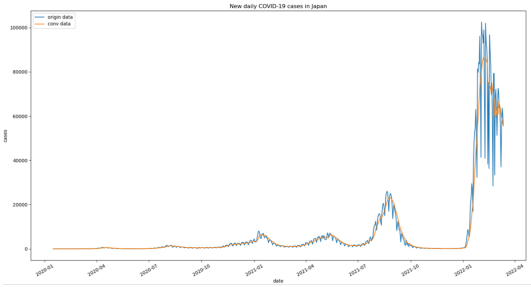

図 1　日本の Covid-19 の新規感染者数と畳み込みデータ

2.2　単峰近似の線形時間アルゴリズム [1]

ヒストグラム関数 $y = f(x)$ が与えられた時，f との L_2 二乗距離を最小化する単峰近似ヒストグラム関数 $y = \phi(x)$ を凸包の平面走査アルゴリズムを用いてデータ数に対して線形時間で計算するアルゴリズムである．本研究では，このアルゴリズムを k ピーク数近似の最適化アルゴリズムのサブルーチンとして用いる．

3　ヒストグラムの k ピーク近似アルゴリズム

2.2 のアルゴリズムを用いて，入力されたヒストグラム f に対して L_2 二乗距離が最小となる k ピーク近似ヒストグラム ϕ を計算する．4 における予測結果の評価指標に RSME を用いるため，ここでは近似誤差に L_2 二乗距離を採用している．ヒストグラムのデータサイズは n とする．

f の極小値のインデックスを抽出しその冒頭に 0 を，末尾に n を追加した配列 MIN を作成し，MIN のサイズを m と定義する．f において，$0 \leq i < j \leq m-1$ を満たす区間 $[MIN_i, MIN_j]$ における単峰近似ヒストグラム ϕ_{MIN_i,MIN_j} を 2.2 のアルゴリズムを用いて計算しする．

次に，辺の重みを $w_{MIN_i,MIN_j} = ||f_{i,j} - \phi_{i,j}||^2 = \sum_{x=i}^{j} |f(x) - \phi(x)|^2$ とする有向非閉路グラフ G を作成し，G について辺の数が k となる最短経路を動的計画法を用いて求める．漸化式は以下の通り．

$$dp_{k,j} = \min_{0 \leq i < j-1} dp_{k-1,i} + w_{i,j} \tag{1}$$

また，最短経路を更新したときのパスを P として保存しておく．辺の数が k となる最短経路を計算した後，P において $0 \leq i < j \leq k$ を満たす区間 $[P_i, P_j]$ における単峰近似ヒストグラムを計算し，それらを結合することで k ピーク近似ヒストグラム ϕ を求める．

尚，このアルゴリズムの計算量は，2.2 のアルゴリズムが $O(n)$，式 (1) より G の k リンク最短経路の探索が $O(kn^2)$ であることから，$O(kn^2)$ であることが保証

1)　関西学院大学大学院 理工学研究科 情報科学専攻 徳山研究室

できる．図 2 に日本の Covid-19 の新規感染者数データ

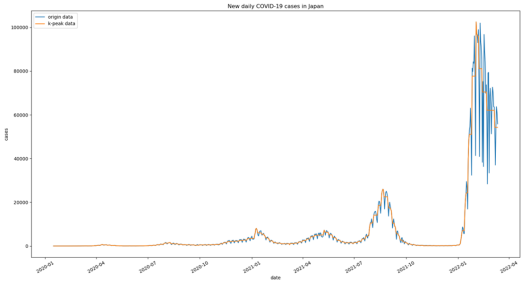

図 2　k ピーク近似の例 ($k = 6$)

に対して k ピーク近似を行った例を示す．ピーク数は AIC(赤池情報量基準) を用いて最適化し，$k = 6$ としている，

4　LSTM による時系列予測

評価指標

本研究では，先行研究を参考に評価指標として RMSE(二乗平均平方根誤差) を用いる．RMSE が小さいほど時系列予測の精度が良いと判断する．

実験

実験に用いる入力データは，日本における Covid-19 の日毎の新規感染者数，死者数，回復者数としている，データセットの区間は 2020 年 01 月 16 日から 2022 年 3 月 12 日までの 786 日間としている．LSTM の学習には 756 日間のデータを用い，最後の 30 日間のデータは予測結果の評価に用いている．

本研究では，学習に際して入力データに最小値が 0，最大値が 1 となるような正規化を施す．また，LSTM で学習を行うためにデータセットの整形を行う必要がある．これは，時刻 t の値を予測するために，時刻 $t-1$ 以前の任意のステップ数のデータを用いるためである．本研究においては，このステップ数を 1 ヶ月の平均日数である 30 日間とする．

LSTM のハイパーパラメータについては，LSTM の層の数，隠れ層のユニット数，ドロップアウト，最適化アルゴリズム，学習率の 5 つについて検討し，より良いパラメータを探索する方法として Optuna を導入した．最終的なパラメータは以下の通り．

表 1　ハイパーパラメータ

層の数	2
隠れ層のユニット数	144
ドロップアウト	0.17
最適化アルゴリズム	Adam
学習率	0.001

結果と考察

まず，Conv1D を用いたモデルと k ピーク近似を用いたモデルの予測結果を図 3，図 4 に示す．次に，評価指標である RMSE による予測精度を表 2 に示し，最後にその考察を示す．表 2 より，Conv1D よりも k ピーク近似の方が良い精度であることが分かった．また，図 3, 4 からは，どちらにおいても実際の感染者数と同じタイミングでの凹凸が確認でき，感染者数の増減に関して良い予測ができていることが分かった．さらに，k ピーク近似アルゴリズムを用いた場合に，より元のデータに近い値での予測を可能にしている．その原因として，より凹凸が小さい，すなわちデータの揺れが小さい k ピーク近似ヒストグラムを入力とすることで，時刻 t の値の予測の際に時刻 $t-1$ 以前のデータの凹凸に対して過学習しないことが考えられる．

表 2　RMSE(小数第 1 位は四捨五入している)

	Conv1D	k ピーク近似
RMSE	18752	15470

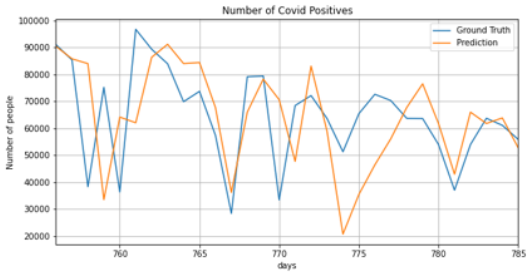

図 3　Conv1D を用いた LSTM による 30 日間の予測

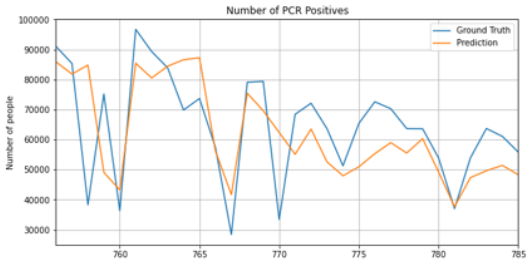

図 4　k ピーク近似を用いた LSTM による 30 日間の予測

5　まとめ

本研究では，Covid-19 の感染者数を予測する LSTM ネットワークを構築し，その入力に，k ピーク近似データを加えた場合に，Conv1D を用いた従来手法よりも良い予測精度を示すことを確認し，データをいくつかの単峰部分に分割し近似する k ピーク近似が，データ解析に有効であることを示した．この k ピーク近似アルゴリズムを他の解析手法等に用いることや，アルゴリズム自体を高次元に拡張することが今後の重要な課題であると考える．

参考文献

[1] Jinhee Chun, Kunihiko Sadakane, and Takeshi Tokuyama. Linear time algorithm for approximating a curve by a single-peaked curve. In *International Symposium on Algorithms and Computation*, pp. 6–15. Springer, 2003.

[2] Sourabh Shastri, Kuljeet Singh, Sachin Kumar, Paramjit Kour, and Vibhakar Mansotra. Time series forecasting of covid-19 using deep learning models: India-usa comparative case study. *Chaos, Solitons & Fractals*, Vol. 140, p. 110227, 2020.

[3] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, Vol. 28, , 2015.