The 37th Annual Conference of the Japanese Society for Artificial Intelligence, 2023

時系列予測における人工データを用いたデータ拡張

Synthetic Data Augmentation for Time Series Forecasting

大野 花純 Kasumi Ohno　牧野 晃平 Kohei Makino　三輪 誠 Makoto Miwa　佐々木 裕 Yutaka Sasaki

豊田工業大学 知能数理研究室
Computational Intelligence Lab, Toyota Technological Institute

Recent years, deep learning models have achieved high prediction performance in Time Series Forecasting (TSF), which predicts future sequences from a given sequence. Training deep learning models requires a large amount of data; it is, however, difficult to prepare enough data since it takes time to collect the data. In this study, we augment the training data by using a variety of synthetic waveforms created with functions and other methods for data expansion. In the experiments, we employed Neural Laplace, which models the dynamics of time series in the Laplace domain, as a model for TSF and evaluated the effect of data augmentation with the synthetic waveforms on the Electricity Transformer Temperature (ETT) m2 dataset, a standard benchmark for TSF. We found that the data augmentation with the synthetic waveforms is effective for TSF on the ETTm2 dataset.

1. はじめに

与えられた時系列に対してその将来の系列を予測する時系列予測は，経済や医療，産業の現場など，様々な場面での活用が見込めることから盛んに研究されている．例えば経済の分野では，DeepAR [Salinas 17] という深層学習モデルが商品の売上予測に活用されている．従来，時系列予測においては，自己回帰モデルと移動平均モデルに和分過程の考え方を組み合わせた自己回帰和分移動平均（Autoregressive Integrated Moving Average; ARIMA）モデル [Box 15] が標準的に用いられてきた．このような統計モデルは，訓練データが少ない場合でもある程度の予測を行うことができるという利点がある反面，周期が一定でない場合や長期的に依存し合うような複雑な時系列データに対しては適用が難しい．

近年，様々なドメインの時系列データの予測において，高い表現力で複雑な時系列パターンを捉えることができる深層学習モデルが高い予測性能を達成している [Lim 21]．しかしながら，深層学習モデルを適切に学習するためには大量の訓練データが必要であり，実際に対象の時系列データを収集し，教師データを作成することはコストが高い．データが不足している場合，モデルが訓練データに対して過適合するなど，十分に学習出来ない場合がある．

このような場合に，訓練データの不足を改善し，深層学習モデルを適切に学習する手法として，データ拡張や転移学習が提案されている．Rotem らは，自動生成した人工的な時系列データ（人工データ）でモデルを事前に学習する**時系列分類**のための転移学習を提案した [Rotem 22]．この研究では，人工データを利用した転移学習手法が，現実の時系列データをソースデータとする場合と比較して，より高い分類精度を達成できる場合があることを示した．しかしながら，**時系列予測**においては，データ拡張に人工データを用いる研究は少なく，その有効性は明らかになっていない．

本研究では，最先端の時系列予測モデル Neural Laplace を用いた単変量時系列予測を対象に，人工データを用いたデータ拡張が，学習時のデータ不足の問題を軽減し，予測性能の向上に寄与するかを検証することを目的とする．本研究の貢献は，以下の 2 点である．

- 時系列予測の標準ベンチマークである ETT (Electricity Transformer Temperature) m2 (ETTm2) [Zhou 21] データセットにおいて，教師データへの人工データの追加による予測性能の向上を確認した．
- Neural Laplace の目的関数の変更により，目的関数が学習に与える影響を調査した．

連絡先: 佐々木裕，豊田工業大学 知能数理研究室，〒 468-8511 名古屋市天白区久方 2 丁目 12 番地 1，052-802-1111，yutaka.sasaki@toyota-ti.ac.jp

2. 関連研究

2.1 時系列予測手法

時系列予測とは，与えられた時系列データから，予測対象の変量の将来の系列を予測するタスクである．時系列予測で扱うデータはセンサデータや音声データなど多岐に渡る [Gemmeke 17, Li 18]．時系列予測は，$0 \leq t \leq T$ の範囲の時点にサンプリングされる系列 x において，ある時点 τ より前の系列 x_t $(0 \leq t < \tau)$ から以降の系列 x_t $(\tau \leq t \leq T)$ を予測するタスクである．時系列予測の問題は，入力する変量の数が単数か複数か，つまり x の次元数が 1 次元か複数次元かで単変量と多変量に分類される．単変量の場合は入力系列からその将来の系列を予測し，多変量の場合は入力する変量同士の相互関係を考慮して対象系列を予測する．

この時系列予測を高性能かつ効率的に行うモデルとして，Neural Laplace [Holt 22] が提案されている．Neural Laplace は時系列を連続時間で扱える連続時間モデルの一つで，原関数 $f(t)$ をラプラス変換して得られる像関数 $F(s)$ をニューラルネットワークでモデル化することで，遅延微分方程式や積分微分方程式のような広範な微分方程式を扱える．Neural Laplace は，図 1 のように，$F(s)$ をモデル化したラプラス表現ネットワークと数値ラプラス逆変換アルゴリズムで構成される．

ラプラス表現ネットワークは，入力系列 x_t が関数 $f(t)$ からサンプリングされたと考えたときに，そのラプラス変換による像関数 $F(s) = \mathcal{L}[f(t)]$ をモデル化したニューラルネットワークであり，s を入力として，$F(s)$ の値を出力する．そして，ラプラス表現ネットワークを利用して計算した $F(s)$ を利用して，

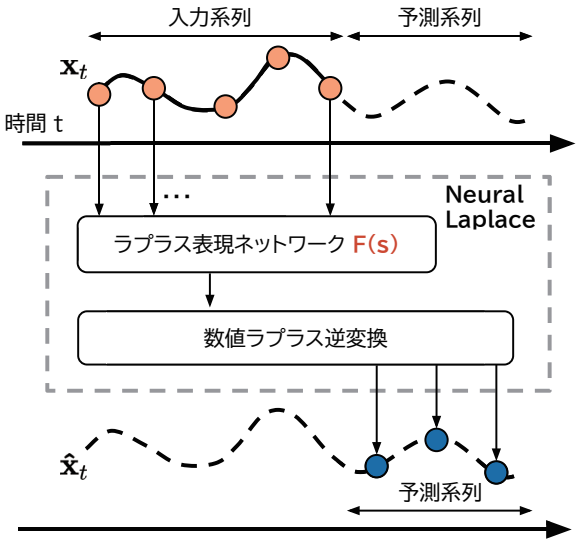

図 1: Neural Laplace の概要図

数値ラプラス逆変換により予測したい時点 t の数値 $\hat{x}_t$ を計算する．Neural Laplace では予測した系列と教師データの系列の二乗誤差を損失 J として，ラプラス表現ネットワークのパラメタを最適化する．

$$J = \sum_{\tau \leq t \leq T} \|\hat{x}_t - x_t\|_2^2 \tag{1}$$

2.2 人工データを用いた時系列データの拡張

仮想的に訓練データ量を増やすデータ拡張は，訓練データが大量に必要な深層学習を用いたモデルの学習に有用であり，広く用いられている [Wen 21]．よく用いられる時系列データの拡張手法として，対象データを時間方向に反転させるなど，元の入力系列を直接操作する方法がある [Le Guennec 16]．他にも，他の目的で収集された既存データをデータ拡張に用いる手法などがあり，データ拡張の時系列タスクへの有効性が示されている [Wen 21]．

人為的に操作可能な人工データを用いて，時系列分類の基本的な変動を学習する手法が提案されている [Rotem 22]．時系列データは，一般的に長期的に持続する変化である「傾向変動」，周期的な変化である「季節変動」，測定誤差などによる変化である「不規則変動」の 3 つの基本的な変動で構成されたものとして扱われる [Cleveland 90]．現実世界で収集されたデータはこれらの変動が混ざり合っているため，分離して個別にモデルに与えることは困難である．そのため，これらの変動に対応する人工的なデータを作成して，時系列分類のモデルを事前に学習する研究がされている [Rotem 22]．このような，人工データを用いて，現実世界のデータで構成するのが困難な教師を与える研究は注目されており，コンピュータビジョン分野では人工生成画像を用いて [Kataoka 22]，自然言語処理分野では文法などの基本を教示するために人工言語を用いて事前学習が行われている [Ri 22]．しかし，時系列予測においては，人工データを利用した研究は行われていない．

3. 人工データを利用した時系列予測

本研究では，時系列予測において図 2 のように，人工データを用いたデータ拡張を Neural Laplace の学習に組み込む．

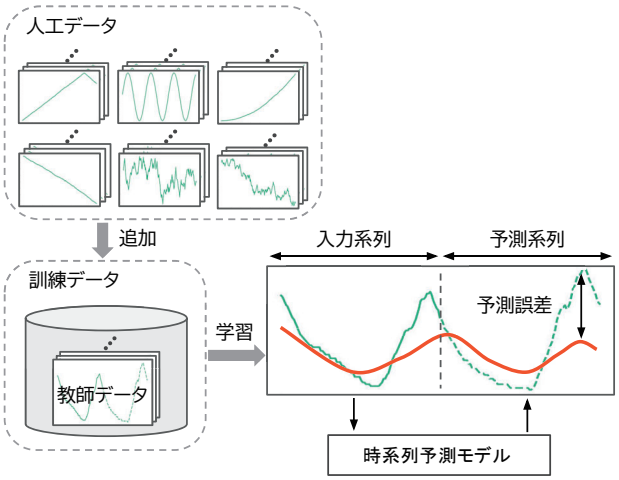

図 2: 提案手法の概要

データ拡張には Rotem らが作成した人工データを利用し，訓練データに追加して学習する．また，学習のための目的関数について，Neural Laplace で提案された目的関数の問題を指摘し，その改善案を示す．

Rotem ら [Rotem 22] は，多様な種類の関数から，周波数などのパラメタをランダムに変化させて人工データを自動生成した．例えば，傾向変動，季節変動，不規則変動のうち，傾向変動を表現するために単調増加（減少）関数が，季節変動を表現するために正弦波，三角波などの周期関数が利用されており，不規則変動を表現するためにそれぞれのデータに対して摂動が付与されている．

多様な種類の関数から，周波数などのパラメタをランダムに変化させて人工的に生成された合成波形から人工データセットを作成し，訓練データに追加する．そして，増強された訓練データでモデルを学習する．

目的関数については，Neural Laplace において提案された予測系列との二乗誤差を損失とした目的関数から，式 (2) のように，系列全体での二乗誤差を損失とするように変更する．Neural Laplace によりダイナミクスを関数 $F(s)$ でモデル化する際，予測系列のみを対象とした損失によりラプラス表現ネットワークを学習すると，入力系列 x_t $(0 \leq t < \tau)$ の時間領域での再構成に関する誤差が考慮されない．そこで，入力系列も目的関数に含めることで，系列全体の再構成を考慮することを目的としている．

$$J' = \sum_{0 \leq t \leq T} \|\hat{x}_t - x_t\|_2^2 \tag{2}$$

4. 実験

時系列予測において，人工データを追加した訓練データでモデルを学習することで，予測性能が向上するかを検証した．実験では，まず，提案した目的関数の変更について評価を行うため，目的関数の変更の有無による予測性能の違いを比較する実験を行い，その変更の妥当性を確かめた．その後，データ量を変更した場合の予測性能を比較した．

4.1 実験設定

本研究では，単変量時系列予測の標準ベンチマークの一つである ETT (Electricity Transformer Temperature) m2

表 1: 変更した目的関数の ETTm2 評価データにおける効果（誤差評価は予測系列のみを対象）

目的関数	MSE ↓	MAE ↓
予測系列	0.1159 ± 0.0042	0.2445 ± 0.0057
系列全体	0.1132 ± 0.0018	0.2393 ± 0.0035

(ETTm2) [Zhou 21] データセットを使用した．ETTm2 データセットは，電力使用量の予測を目的として，電力使用量予測の指標となる変圧器の「油温」の他，変圧器の電力負荷などの計 6 つの変量を 15 分間隔で 2 年間収集したデータである．

ETT データセットの単変量予測での標準的な設定に合わせ，入力と予測対象をともに「油温」とした．訓練データ，開発データ，評価データの割合も，同じく標準的な設定に合わせ，訓練：開発：評価＝ 6：2：2 とし，入出力の系列長はともに 96 とした．また，各データをスライド幅 1 で生成し，訓練データの平均値と標準偏差を用いて平均 0，分散 1 となるよう標準化した．

データ拡張には，Rotem ら [Rotem 22] が作成した既存の人工データセットからランダムに 32,000 個サンプリングし，系列長を 192 に揃えた人工データを使用した．

評価指標には，ETT データセットを対象とした系列予測において標準的に用いられている平均二乗誤差 (Mean Squared Error; MSE) と平均絶対誤差（Mean Absolute Error; MAE）を用いた．モデルの学習は最大 50 エポックとし，開発データの MSE が最小となるモデルのパラメタで評価データの評価を行った．評価は，乱数のシード値を変えて 5 回行った．

目的関数の変更の検証については，人工データを用いずに，元の訓練データのみを用いて，予測性能を比較した．また，人工データを用いたデータ拡張の有効性の検証においては，元々の教師である ETT データと人工データを増減させて，異なる割合の訓練データ量で予測性能を比較した．実際に比較した条件は，条件 (a) ETT 0% かつ人工 100%，条件 (b) ETT 100%かつ人工 0%，条件 (c) ETT 100%かつ人工 100%，条件 (d) ETT 50%かつ人工 0%，条件 (e) ETT 50%かつ人工 100%の 5 つとし，これらの条件で予測性能を比較した．

4.2 事前実験：目的関数の変更

変更した目的関数による MSE，MAE は表 1 のようになった．結果より，系列全体における二乗誤差を損失とすることで，MSE，MAE の低減が見られた．このため，データ拡張の実験においては変更後の目的関数 J' を採用した．

4.3 実験結果

評価データにおける MSE, MAE は表 2 のようになった．人工データのみの学習（条件 (a)）では，評価データである ETT データに適合できなかったことが分かる．しかし，人工データを訓練データに追加して学習する（条件 (b) と (c)，(d) と (e)）ことで，評価データの MSE を低減できた．さらに，データ数を半減した ETT データに人工データを加えた条件 (e) の場合のほうが，ETT データを全て使用した条件 (b) の場合よりも，評価データにおける MSE を低減できた．以上の結果から，ETTm2 データにおいては，人工データを用いたデータ拡張が有効であることを明らかにした．

5. 事例研究

人工データが予測に与える影響を確認するために，訓練時の人工データを利用した場合（人工あり）と利用しなかった場合（人工なし）について，ETTm2 開発データ 13,745 件の予測系列を可視化した．ここでは，人工データの追加で予測が改善した事例と改善しなかった事例について，いくつか特徴的なものを紹介する．

表 2: ETTm2 評価データの結果

訓練データ	MSE ↓	MAE ↓
(a) 人工	1.1392 ± 0.1693	0.8791 ± 0.0715
(b) ETT 100%	0.1132 ± 0.0018	0.2393 ± 0.0035
(c) ETT 100% + 人工	**0.1074 ± 0.0011**	**0.2323 ± 0.0030**
(d) ETT 50%	0.1127 ± 0.0027	0.2399 ± 0.0053
(e) ETT 50% + 人工	0.1104 ± 0.0023	0.2356 ± 0.0030

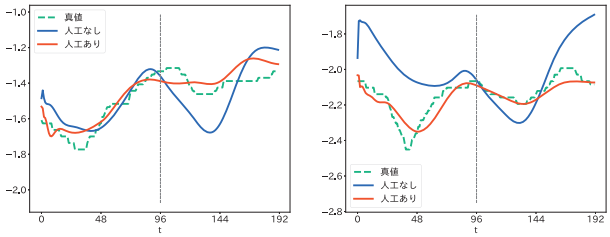

図 3: 予測が改善した例

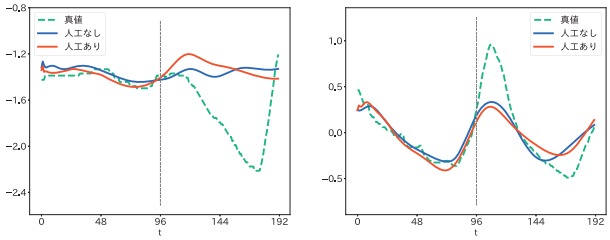

図 4: 予測が改善しなかった例

図 3 に予測が改善した事例を示した．提案する人工データの追加により，変化が緩やかな形状をモデル化できるようになった．これは，データ拡張により，ETTm2 訓練データへの過適合を回避できるようになったためであると考えられ，人工データの利用によりモデルの予測を改善できる可能性がある．

反対に予測が改善しなかった事例を図 4 に示した．入力系列の情報からその先の予測が困難であるような形状や，入力系列の変動と比較して，予測系列が大きく変動しているような形状に対しては，人工データの有無に関わらず，真値を予測できていない．対応策として，ETTm2 データセットの絶対的な時間情報（年月日，時刻など）をモデルに学習させることが考えられる．

6. おわりに

本研究では，時系列予測における人工データを用いたデータ拡張の有効性を検証するために，教師データと人工データを合わせて訓練データとするデータ拡張手法を提案した．ETTm2 データセットを使用した実験では，人工データの追加により，評価データの平均二乗誤差をわずかながら低減させ，人工データを用いたデータ拡張が有効であることを示した．

今後の課題として，予測が困難であった形状をモデル化するために，モデルの入力への絶対的な時間情報の追加や，モデルの学習を制御できるような人工データの生成手法の検討が挙げられる．また，サンプリング周期が一定でないような，より複雑な時系列データを対象として，人工データを用いたデータ拡張の有効性を検証する予定である．

参考文献

[Box 15] Box, G., Jenkins, G., Reinsel, G., and Ljung, G.: *Time Series Analysis: Forecasting and Control*, Wiley Series in Probability and Statistics, Wiley (2015)

[Cleveland 90] Cleveland, R. B., Cleveland, W. S., and Terpenning, I.: STL: A Seasonal-Trend Decomposition Procedure Based on Loess, *Journal of Official Statistics*, Vol. 6, No. 1, p. 3 (1990)

[Gemmeke 17] Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M.: Audio Set: An ontology and human-labeled dataset for audio events, in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780 (2017)

[Holt 22] Holt, S. I., Qian, Z., and Schaar, van der M.: Neural Laplace: Learning diverse classes of differential equations in the Laplace domain, in Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. eds., *Proceedings of the 39th International Conference on Machine Learning*, Vol. 162 of *Proceedings of Machine Learning Research*, pp. 8811–8832, PMLR (2022)

[Kataoka 22] Kataoka, H., Okayasu, K., Matsumoto, A., Yamagata, E., Yamada, R., Inoue, N., Nakamura, A., and Satoh, Y.: Pre-training without Natural Images, *International Journal of Computer Vision (IJCV)* (2022)

[Le Guennec 16] Le Guennec, A., Malinowski, S., and Tavenard, R.: Data Augmentation for Time Series Classification using Convolutional Neural Networks, in *ECML/PKDD Workshop on Advanced Analytics and Learning on Temporal Data*, Riva Del Garda, Italy (2016)

[Li 18] Li, Y., Yu, R., Shahabi, C., and Liu, Y.: Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting, in *International Conference on Learning Representations (ICLR '18)* (2018)

[Lim 21] Lim, B. and Zohren, S.: Time-series forecasting with deep learning: a survey, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol. 379, No. 2194, p. 20200209 (2021)

[Ri 22] Ri, R. and Tsuruoka, Y.: Pretraining with Artificial Language: Studying Transferable Knowledge in Language Models, in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7302–7315, Dublin, Ireland (2022), Association for Computational Linguistics

[Rotem 22] Rotem, Y., Shimoni, N., Rokach, L., and Shapira, B.: Transfer Learning for Time Series Classification Using Synthetic Data Generation, in Dolev, S., Katz, J., and Meisels, A. eds., *Cyber Security, Cryptology, and Machine Learning*, pp. 232–246, Cham (2022), Springer International Publishing

[Salinas 17] Salinas, D., Flunkert, V., and Gasthaus, J.: DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks (2017)

[Wen 21] Wen, Q., Sun, L., Yang, F., Song, X., Gao, J., Wang, X., and Xu, H.: Time Series Data Augmentation for Deep Learning: A Survey, in Zhou, Z.-H. ed., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 4653–4660, International Joint Conferences on Artificial Intelligence Organization (2021), Survey Track

[Zhou 21] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W.: Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, No. 12, pp. 11106–11115 (2021)