DILATEと後追い予測のための指標

Shape and Time Distortion Loss for Training Deep Time Series Forecasting Models

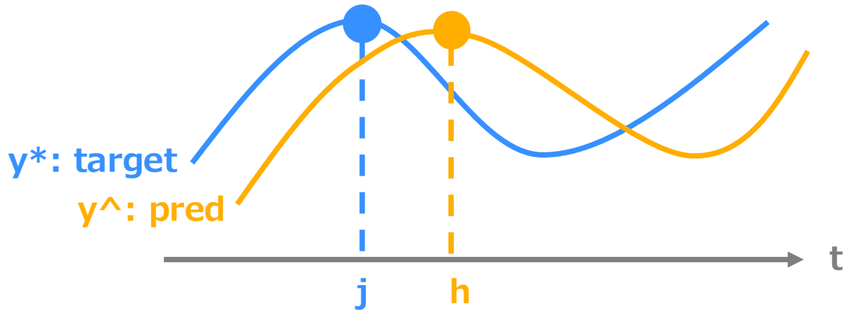
NTTドコモビジネス株式会社

石山隼

後追い予測

つなごう。驚きを。幸せを。
docomo Business

- 時系列予測において、モデルの予測値が実測値に比べて遅れて出力されていること
  - 先端AI数理PJでは後追い予測と呼んでいる



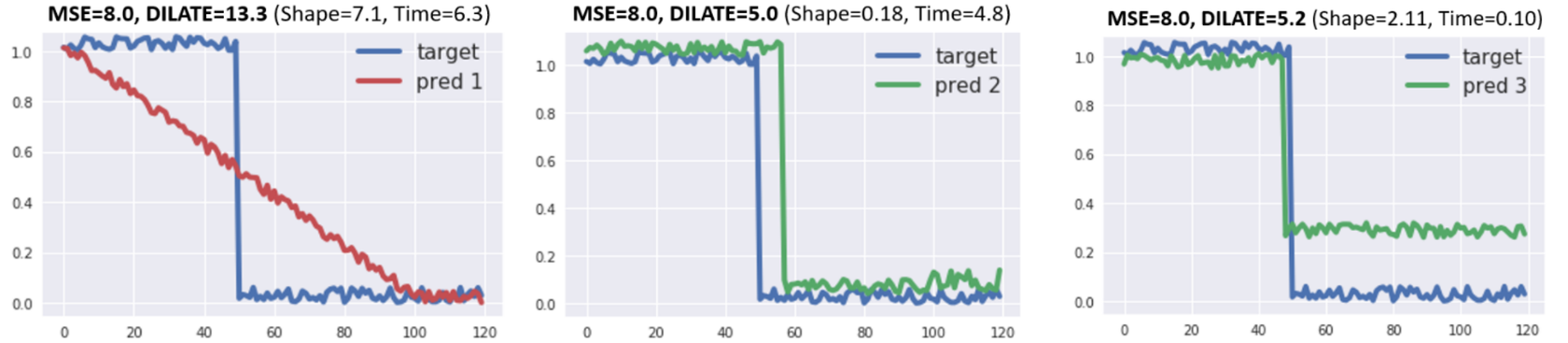
- Addressing Prediction Delays in Time Series Forecasting: A Continuous GRU Approach with Derivative Regularization (KDD2024) の論文では、prediction delayと呼ばれていたりする
  - もっと的確な用語があるかもしれない
- 見かけ上のMSE/MAEがよくても実務的には役に立たないことがある
  - 突発的な変化に対応できない
  - 実質的に1期前の実測値をそのまま出力するモデル(持続予測モデル)を学習してしまっている
  - streaming dataであれば、そもそも判断が早いモデルのほうが望ましい
- 後追い予測への対処法の1つとそもそもそれを測るための指標にはどのようなものがあり得るかを“Shape and Time Distortion Loss for Training Deep Time Series Forecasting Models” (NeurIPS 2019) の論文をもとに紹介

概要

- “Shape and Time Distortion Loss for Training Deep Time Series Forecasting Models” (NeurIPS 2019)
- DILATE (DIstortion Loss including shApe and TimE) と呼ばれる損失関数を提案
  - 微分可能なDTW(Soft DTW)とTDIを利用
- 課題：突発的変化を予測するために、損失関数としてのMSEは形状とタイミングが捉えられておらず不完全では？
- 提案手法：微分可能なDTW(Soft-DTW)とTDIをニューラルネットワークの損失関数とする
- 結果：
  - DILATEを利用したニューラルネットワークが、MSEの精度をほぼ失わずに形状・タイミングを改善
  - MSEのみの損失関数では、形状とタイミングが捉えられていない
  - Soft DTWのみの損失関数では、タイミングが捉えられず、精度(MSE)も低い
- 微分不可能だった形状・タイミングのメトリクスをSoft-DTW としてニューラルネットワークの損失関数に組み込んだ
  - また最大35倍に高速化した

課題

- 突発的変化を予測するとき、MSEは不十分
- Fig.1
  - 真の系列（青のステップ関数）に対する 3 つの予測（a, b, c）はいずれも類似したMSE
  - (a) はこれから生じる急激な低下を捉えられず制御目的には不適切
  - (b) は若干遅延しているが形状（shape）を捉えている
  - (c) は形状振幅がやや不正確だが、時間的位置（temporal localization）を捉えている



(a) Non informative prediction (b) Correct shape, time delay (c) Correct time, inaccurate shape

Figure 1: Limitation of the euclidean (MSE) loss: when predicting a sudden change (target blue step function), the 3 predictions (a), (b) and (c) have similar MSE but very different forecasting skills. In contrast, the DILATE loss proposed in this work, which disentangles shape and temporal decay terms, supports predictions (b) and (c) over prediction (a) that does not capture the sharp change of regime.

DILATE
(DIstortion Loss including shApe and TimE)

つなごう。驚きを。幸せを。
NTT docomo Business

- DILATEは、DTWとTDIの組み合わせからなる損失関数

$$\mathcal{L}_{DILATE}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) = \alpha\, \mathcal{L}_{shape}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) + (1-\alpha)\, \mathcal{L}_{temporal}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) \quad (1)$$

$$\mathcal{L}_{shape}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) = DTW_\gamma(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) := -\gamma \log \left(\sum_{\mathbf{A} \in \mathcal{A}_{k,k}} \exp\left(-\frac{\left\langle \mathbf{A}, \boldsymbol{\Delta}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) \right\rangle}{\gamma} \right) \right) \quad (2)$$

$$\mathcal{L}_{temporal}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) := \left\langle \mathbf{A}^*_\gamma, \boldsymbol{\Omega} \right\rangle = \frac{1}{Z} \sum_{\mathbf{A} \in \mathcal{A}_{k,k}} \langle \mathbf{A}, \boldsymbol{\Omega} \rangle \exp^{-\frac{\left\langle \mathbf{A}, \boldsymbol{\Delta}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) \right\rangle}{\gamma}} \quad (4)$$

- TDIはDTWをもとにした指標
- したがって、まずDTWを説明する

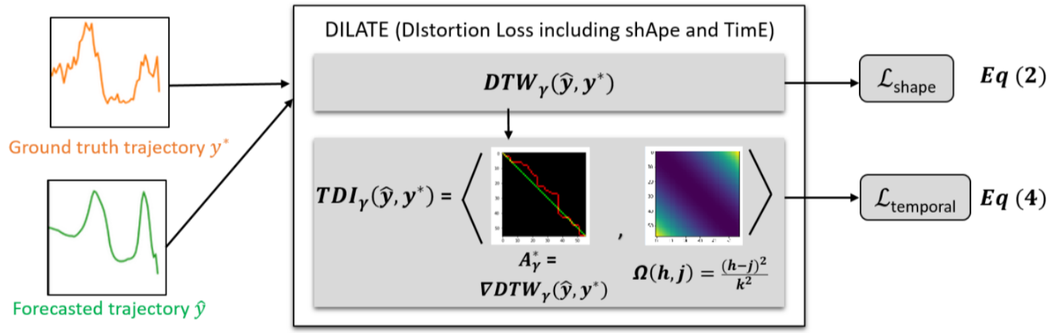


Figure 3: DILATE loss computation for separating the shape and temporal errors.

DTW(Dynamic Time Warping)

つなごう。驚きを。幸せを。
docomo Business

- 形状が異なる2つの時系列の類似度を測る手法。小さいほど似ている

$$DTW(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) = \min_{\mathbf{A} \in \mathcal{A}_{k,k}} \left\langle \mathbf{A}, \mathbf{\Delta}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) \right\rangle.$$

- A：最適な経路
- Δ(y^, y*)：コスト行列 $\mathbf{\Delta}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) := [\delta(\hat{\mathbf{y}}_i^h, \overset{*}{\mathbf{y}}_i^j)]_{h,j}$
  - δ：なんらかの距離関数

x ＼ y	2	4	8	10
1	1	3	7	9
2	0	2	**6**	8
5	3	1	3	5

アルゴリズムのイメージ

- コスト行列を作る
- 左上から→↘↓の移動のコストを計算
- 候補のうち最小となる累積コストを選択
- 終点の累積コストがDTW
- 右下から順に最小の累積コストを取った経路が最適経路

累積コスト

累積コスト→6+3で新たなコスト9

x ＼ y	2	4	8	10
1	1	4	11	20
2	1	**3**	6+3=**9**	16
5	4	2	5	10

最小

DTW

TDI(Time Distortion index)

- DTWにおいてマッチした時点間のインデックスの距離をタイミングのズレとして計算した指標

$$TDI(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) = \langle \mathbf{A}^*, \mathbf{\Omega} \rangle = \left\langle \underset{\mathbf{A} \in \mathcal{A}_{k,k}}{\arg\min} \left\langle \mathbf{A}, \mathbf{\Delta}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) \right\rangle, \mathbf{\Omega} \right\rangle \quad (3)$$

- A：最適な経路
- Δ(y^, y*)：コスト行列 $\mathbf{\Delta}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) := [\delta(\hat{\mathbf{y}}_i^h, \overset{*}{\mathbf{y}}_i{}^j)]_{h,j}$
- Ω：ペナルティの関数

- オリジナルは、target index × prediction indexの四角形に対する正確な面積(台形)の比として計算
  - この場合は0-1の範囲
- この論文では、Ω = (i-j)^2/k^2
- Ωの関数を変えれば、遅れの罰則を重くできる
  - 損失関数として利用する場合も問題ない

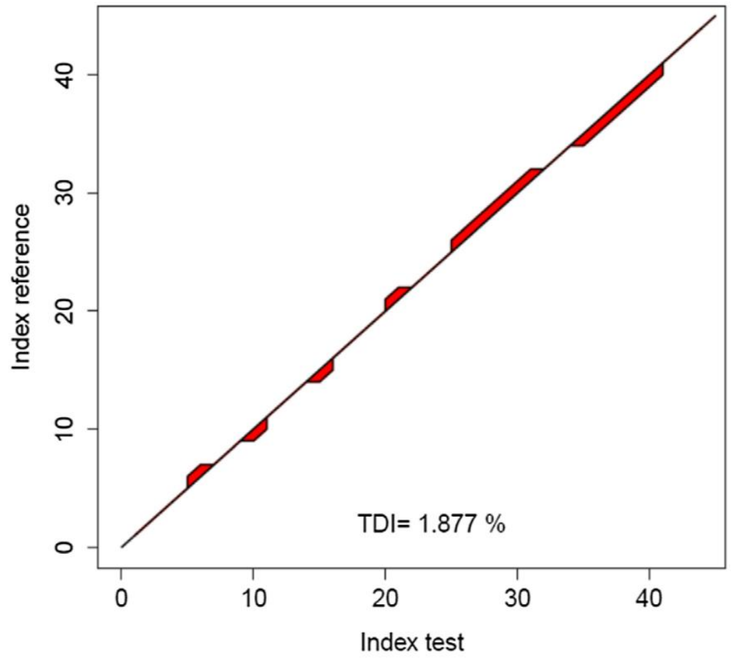


Fig. 21. Area bounded between the optimal path and the identity path.

Soft DTW

- Soft-DTW(DTW_γ)

$$\mathcal{L}_{shape}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) = DTW_\gamma(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) := -\gamma \log \left(\sum_{\mathbf{A} \in \mathcal{A}_{k,k}} \exp \left(-\frac{\left\langle \mathbf{A}, \mathbf{\Delta}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) \right\rangle}{\gamma} \right) \right) \quad (2)$$

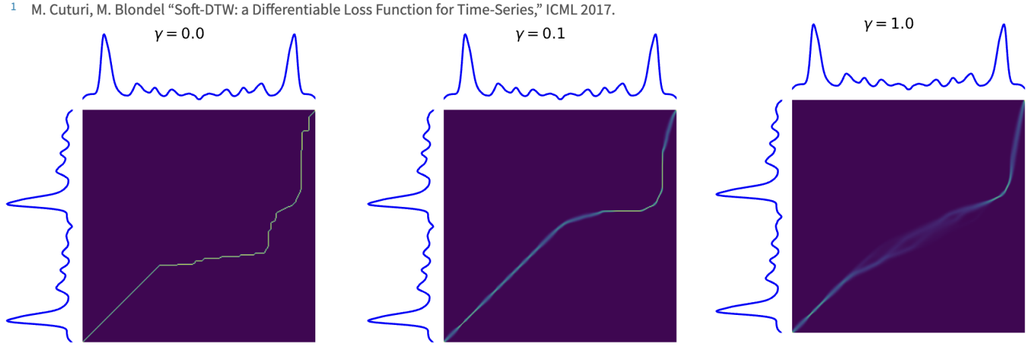
  - 微分可能

- TDI_γ

$$\mathcal{L}_{temporal}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) := \left\langle \mathbf{A}^*_\gamma, \mathbf{\Omega} \right\rangle = \frac{1}{Z} \sum_{\mathbf{A} \in \mathcal{A}_{k,k}} \langle \mathbf{A}, \mathbf{\Omega} \rangle \exp^{-\frac{\left\langle \mathbf{A}, \mathbf{\Delta}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) \right\rangle}{\gamma}} \quad (4)$$

- 損失関数として同時に最適化せず、
  DTW_γの計算→そのパスをもとにTDIを算出
  →損失を計算
- パスであるA*_γは確率的な値
  - 通常は最短経路に1がある行列

[1] M. Cuturi, M. Blondel “Soft-DTW: a Differentiable Loss Function for Time-Series,” ICML 2017.



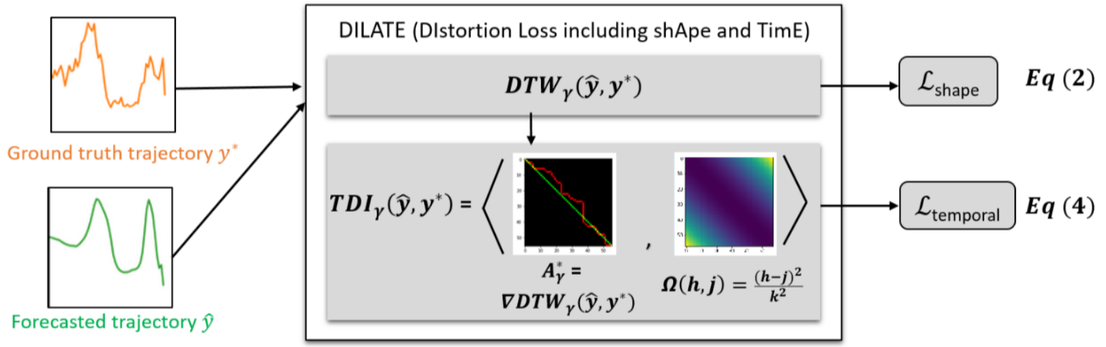


Figure 3: DILATE loss computation for separating the shape and temporal errors.

学習

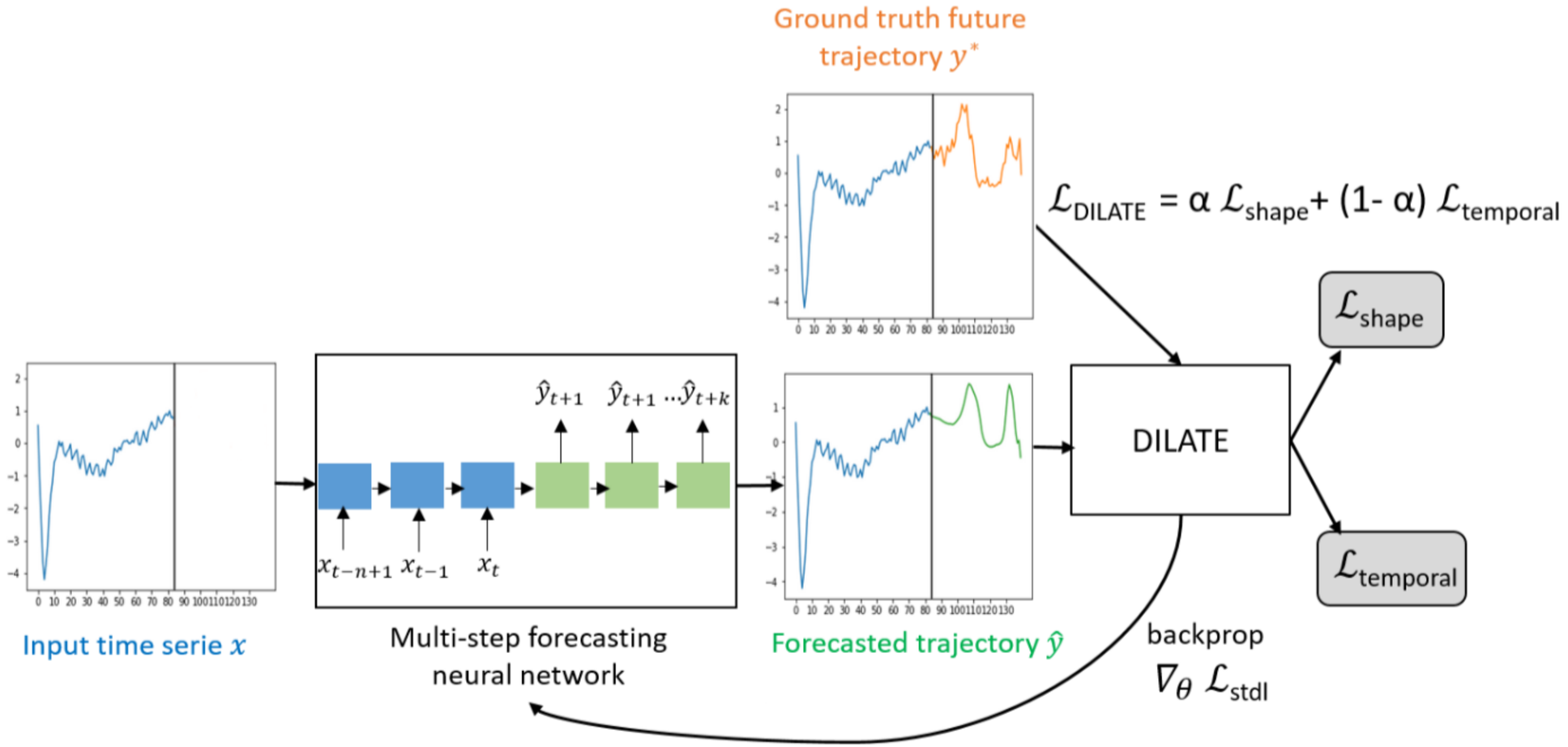


Figure 2: Our proposed framework for training deep forecasting models.

評価指標

- 後追い予測の検出として使える指標を考えたいので、少し長めに話します
- 3つの観点で5つの評価指標を使用している
- 予測精度　ユークリッド距離
    - MSE、MAE
- 形状
    - DTW(Dynamic Time Warping)
    - Ramp Score
- タイミング
    - TDI(Time Distortion Index)
    - Hausdorff距離

- 形状の方は立ち上がり？追従性能？などの評価指標に使えるかも
- タイミングの方は後追い予測の指標として使えるかも
    - 先行も同等に評価してしまうので、後追い評価の際には実装上の工夫が必要だが、それほど難しくはない

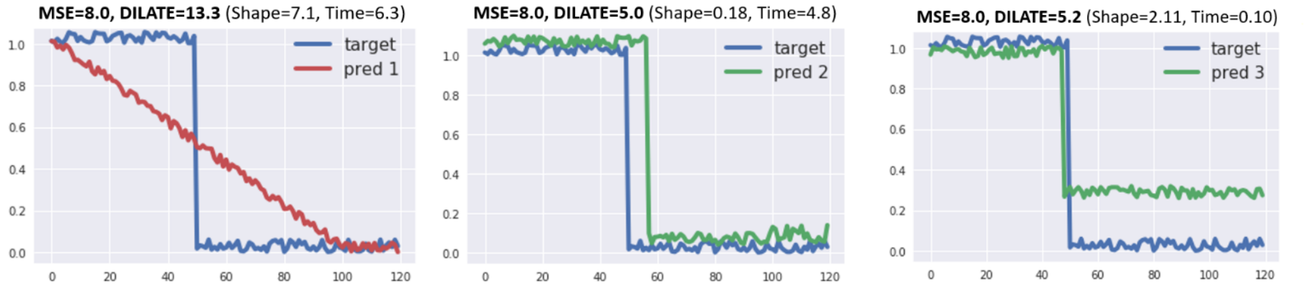
予測精度の評価指標(MSE、MAE)

つなごう。驚きを。幸せを。
NTT docomo Business

- ユークリッド距離にもとづく指標：小さい方が良い

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i - y_i^*\right)^2. \qquad \mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\,\hat{y}_i - y_i^*\,\right|.$$

- 問題点
  - 突発的変化を評価するときにMSEには問題がある



(a) Non informative prediction (b) Correct shape, time delay (c) Correct time, inaccurate shape

Figure 1: Limitation of the euclidean (MSE) loss: when predicting a sudden change (target blue step function), the 3 predictions (a), (b) and (c) have similar MSE but very different forecasting skills. In contrast, the DILATE loss proposed in this work, which disentangles shape and temporal decay terms, supports predictions (b) and (c) over prediction (a) that does not capture the sharp change of regime.

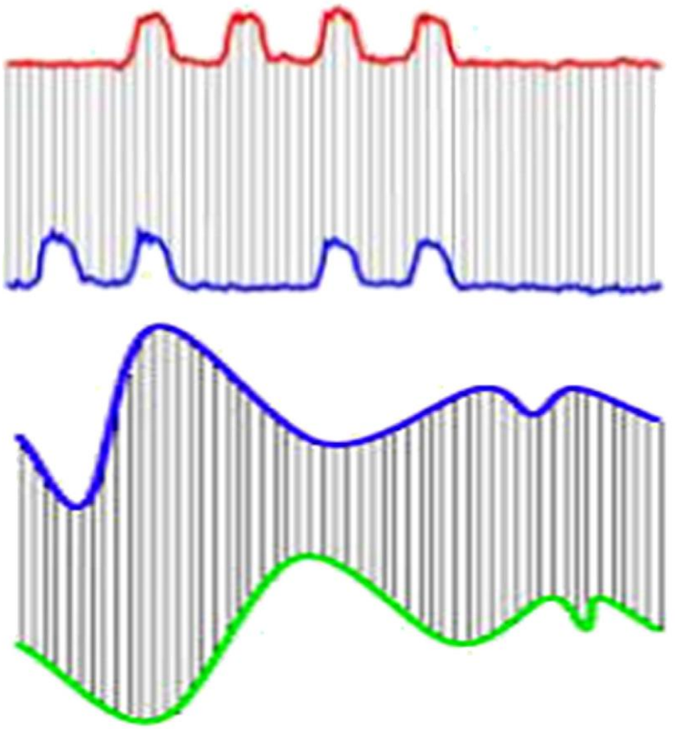


Fig. 2. Sequences are aligned one to one.

形状の評価指標(DTW)

- DTW：小さいほど類似している

$$DTW(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) = \min_{\mathbf{A} \in \mathcal{A}_{k,k}} \left\langle \mathbf{A}, \boldsymbol{\Delta}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) \right\rangle.$$

- 指標として使うなら、
  - 後追いの指標としては使えなさそう
  - パス長で割るなどして正規化距離（平均コスト）を算出すれば比較可能
  - プラントで使うなら、
    ステップ状の入力に頑健な手法のほうがいいかも？
- 系列長m,nに対して計算量O(mn)

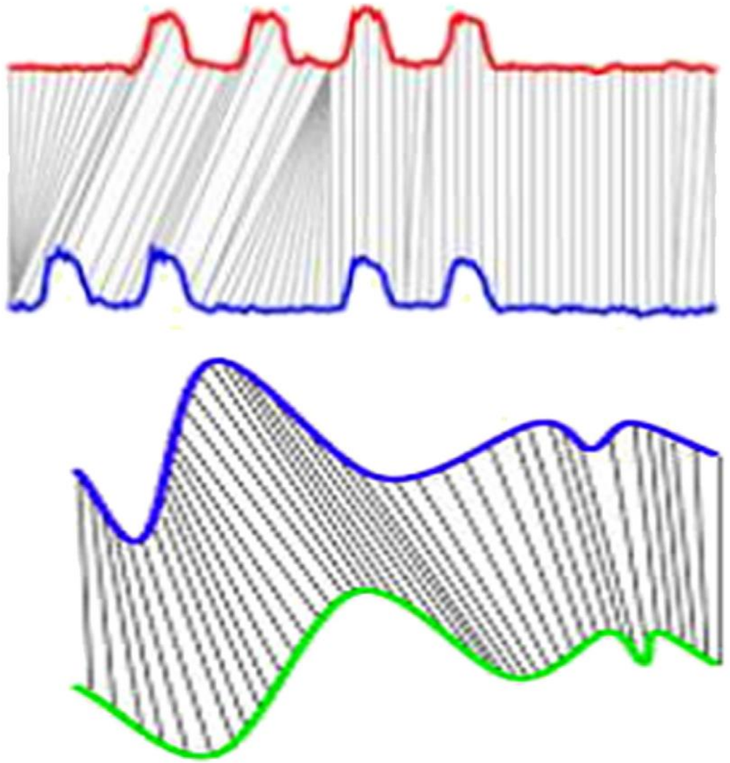


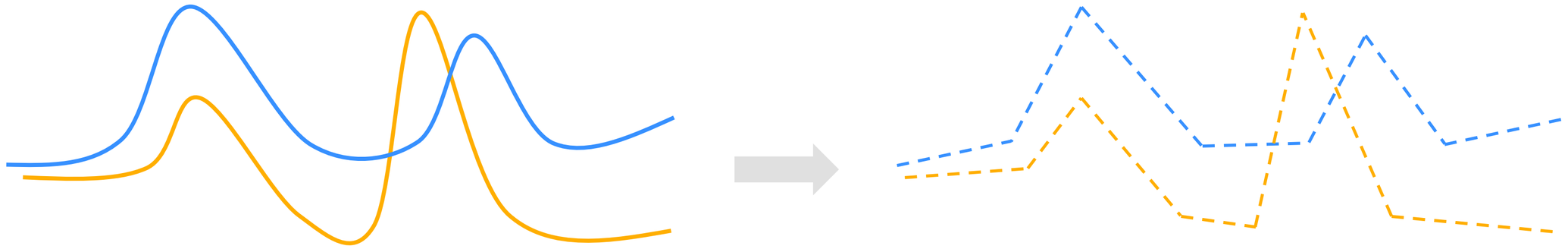
Fig. 3. Nonlinear alignments are possible.

形状の評価指標(Ramp Score)

- 電力予測における急激な変化を捉えるために開発された、形状の類似度を測る指標
  - 小さいほど形状が似ている

この論文でのアルゴリズム

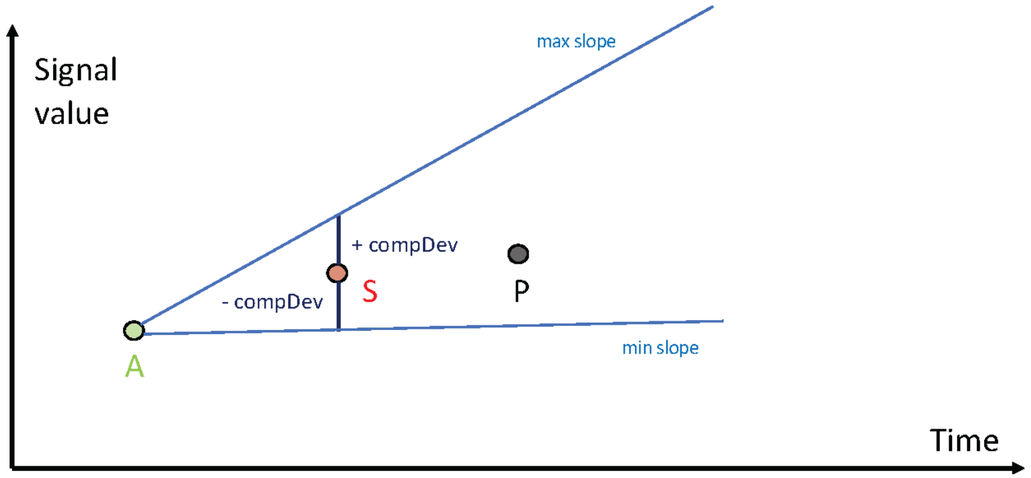
1. DTWでマッチングする(純粋に形状のみの評価指標とするため)
2. swing doorを利用して系列を圧縮+傾きを算出
3. DTWでマッチしたポイントごとに傾きの差を積分(通常はすべての点において計算？)



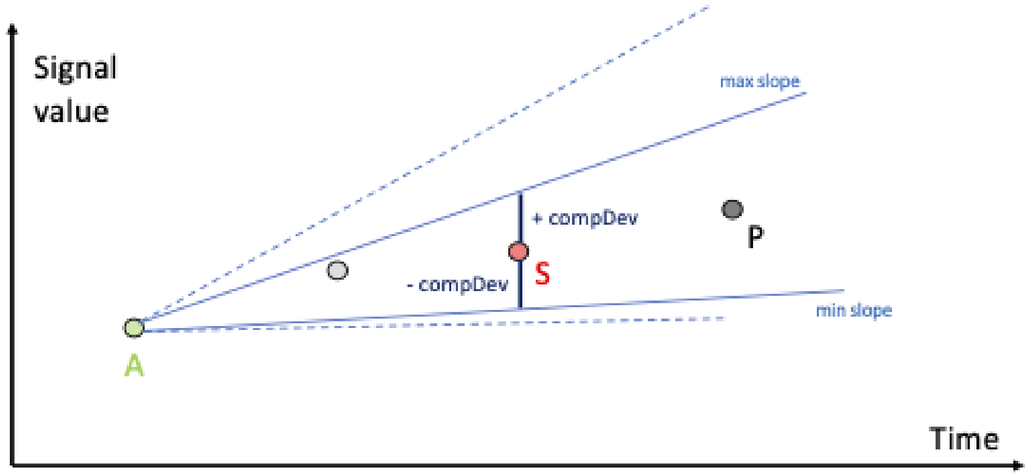
Swing Door

つなごう。驚きを。幸せを。
docomo Business

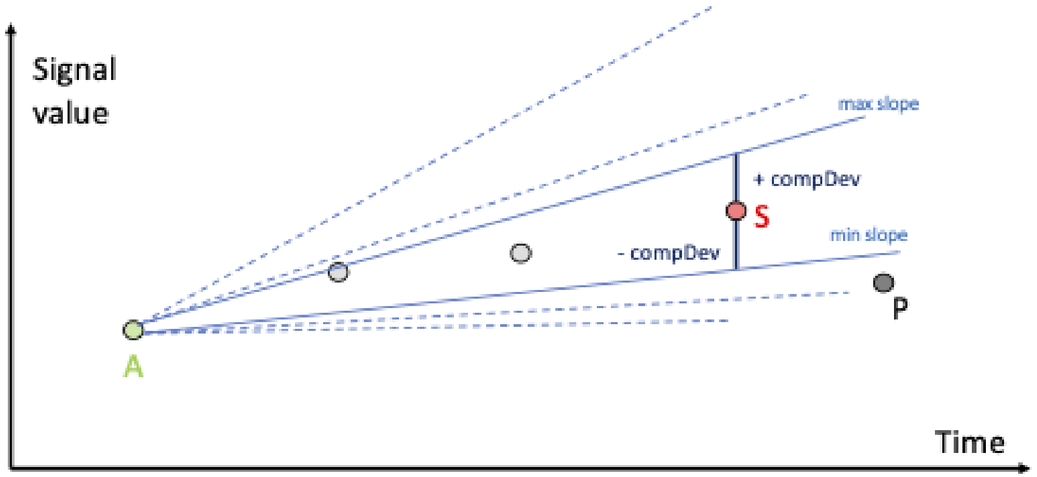
1 基準の点Aと現在の点S、幅を設定、幅の中に次の点があればSを更新



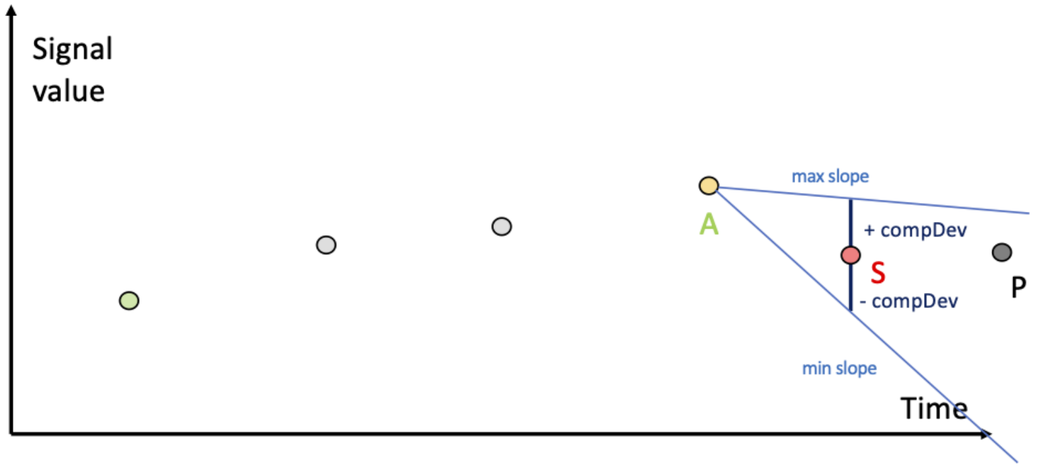
2 幅は常に小さくなる



3 次の点Pが幅から出たらAを更新する



4 1にもどる

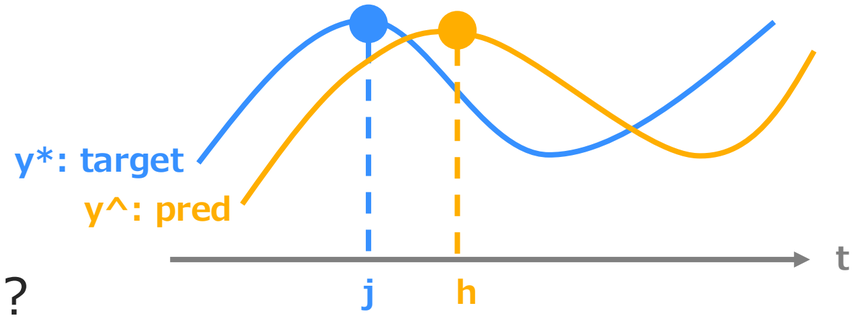


タイミングの評価指標(TDI)

- TDI：小さいほど時間的なズレがない
  - DTWで計算した最適な経路のindexの距離を指標にする

$$TDI(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) = \langle \mathbf{A}^*, \mathbf{\Omega} \rangle = \left\langle \underset{\mathbf{A} \in \mathcal{A}_{k,k}}{\arg\min} \left\langle \mathbf{A}, \mathbf{\Delta}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) \right\rangle, \mathbf{\Omega} \right\rangle \quad (3) \qquad \mathbf{\Delta}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) := [\delta(\hat{\mathbf{y}}_i^h, \overset{*}{\mathbf{y}}_i{}^j)]_{h,j}$$

- 後追いが発生したタイミングは分かりそう
- 純粋に後追いの指標として使うなら、Ωに重み付けをする
  - 例：if h > j then Ω = (h-j)^2/k^2 else Ω = 0
- あるいは、マッチした点やパスの割合を後追いの指標として出す？
  - 同じ著者のTDM(Time Distortion Metrix)が発想として近いかも
  - TDM：-1～1で先行と後追いの割合を評価



$$\mathrm{TDI} = \mathrm{TDI}_{adv} + \mathrm{TDI}_{late}, \qquad \mathrm{TDM} = 1 - 2\frac{\mathrm{TDI}_{adv}}{\mathrm{TDI}}.$$

- Laura Frías-Paredes, Fermín Mallor, Martín Gastón-Romeo, Teresa León, Assessing energy forecasting inaccuracy by simultaneously considering temporal and absolute errors, Energy Conversion and Management, Volume 142, 2017, Pages 533-546, ISSN 0196-8904, https://doi.org/10.1016/j.enconman.2017.03.056.

タイミングの評価指標(Hausdorff距離)

つなごう。驚きを。幸せを。
NTT docomo Business

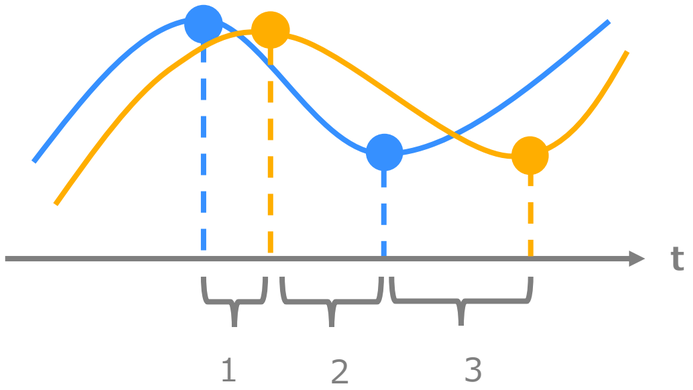
- 2つの系列がどれくらい離れているかを測るための指標：小さいほど似ている、時間的なズレがない

$$\mathrm{Hausdorff}(\mathcal{T}^*, \hat{\mathcal{T}}) := \max(\max_{\hat{t}\in\hat{\mathcal{T}}} \min_{t^*\in\mathcal{T}^*} |\hat{t} - t^*|, \max_{t^*\in\mathcal{T}^*} \min_{\hat{t}\in\hat{\mathcal{T}}} |\hat{t} - t^*|) \quad (1)$$

- 後追いが発生したタイミングは分からなさそう？変化点の検出次第

この論文でのアルゴリズム

- 変化点の特定
- 予測系列の変化点を基準に観測系列の変化点のうち、もっとも時間的に近い点をマッチング
- 予測と観測を入れ替えて再度マッチング
- マッチングした中でもっとも大きい時間の差を取るものを距離とする



Hausdorff距離 = 3

実験結果

- 一番右が提案手法

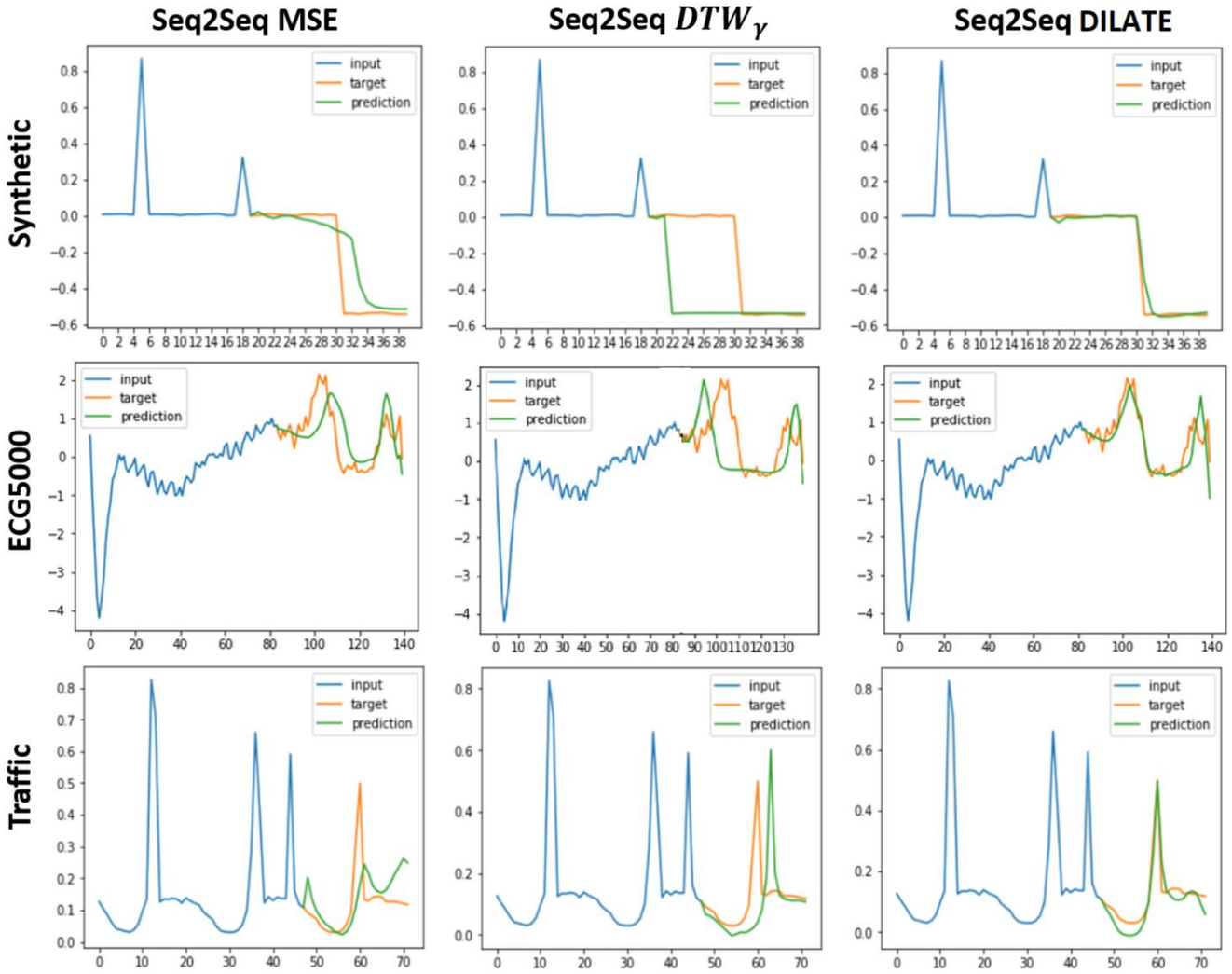


Figure 4: Qualitative forecasting results.

実験結果

つなごう。驚きを。幸せを。
docomo Business

- DILATEを利用したニューラルネットワークが、MSEの精度をほぼ失わずに形状・タイミングを改善
- MSEのみの損失関数では、形状とタイミングが捉えられていない
- Soft DTWのみの損失関数では、タイミングが捉えられず、精度(MSE)も低い

		Fully connected network (MLP)			Recurrent neural network (Seq2Seq)		
Dataset	Eval	MSE	DTW_γ [13]	DILATE (ours)	MSE	DTW_γ [13]	DILATE (ours)
Synth	MSE	**1.65 ± 0.14**	4.82 ± 0.40	**1.67± 0.184**	**1.10 ± 0.17**	2.31 ± 0.45	**1.21 ± 0.13**
	DTW	38.6 ± 1.28	**27.3 ± 1.37**	32.1 ± 5.33	**24.6 ± 1.20**	**22.7 ± 3.55**	**23.1 ± 2.44**
	TDI	15.3 ± 1.39	26.9 ± 4.16	**13.8 ± 0.712**	17.2 ± 1.22	20.0 ± 3.72	**14.8 ± 1.29**
ECG	MSE	**31.5 ± 1.39**	70.9 ± 37.2	37.2 ± 3.59	**21.2 ± 2.24**	75.1 ± 6.30	30.3 ± 4.10
	DTW	19.5 ± 0.159	18.4 ± 0.749	**17.7 ± 0.427**	17.8 ± 1.62	17.1 ± 0.650	**16.1 ± 0.156**
	TDI	**7.58 ± 0.192**	38.9 ± 8.76	**7.21 ± 0.886**	8.27 ± 1.03)	27.2 ± 11.1	**6.59 ± 0.786**
Traffic	MSE	**0.620 ± 0.010**	2.52 ± 0.230	1.93 ± 0.080	**0.890 ± 0.11**	2.22 ± 0.26	**1.00 ± 0.260**
	DTW	24.6 ± 0.180	**23.4 ± 5.40**	**23.1 ± 0.41**	24.6 ± 1.85	**22.6 ± 1.34**	**23.0 ± 1.62**
	TDI	**16.8 ± 0.799**	27.4 ± 5.01	**16.7 ± 0.508**	**15.4 ± 2.25**	22.3 ± 3.66	**14.4± 1.58**

Table 1: Forecasting results evaluated with MSE (×100), DTW (×100) and TDI (×10) metrics, averaged over 10 runs (mean ± standard deviation). For each experiment, best method(s) (Student t-test) in bold.

実験結果

- Ramp Score、Housdorff距離

		MSE	DTW_γ [13]	DILATE (ours)
Synthetic	Hausdorff	2.87 ± 0.127	3.45 ± 0.318	$\mathbf{2.70 \pm 0.166}$
	Ramp score ($\times 10$)	5.80 ± 0.104	$\mathbf{4.27 \pm 0.800}$	4.99 ± 0.460
ECG5000	Hausdorff	$\mathbf{4.32 \pm 0.505}$	6.16 ± 0.854	$\mathbf{4.23 \pm 0.414}$
	Ramp score	$\mathbf{4.84 \pm 0.240}$	$\mathbf{4.79 \pm 0.365}$	$\mathbf{4.80 \pm 0.249}$
Traffic	Hausdorff	$\mathbf{2.16 \pm 0.378}$	$\mathbf{2.29 \pm 0.329}$	$\mathbf{2.13 \pm 0.514}$
	Ramp score ($\times 10$)	6.29 ± 0.319	$\mathbf{5.78 \pm 0.404}$	$\mathbf{5.93 \pm 0.235}$

Table 2: Forecasting results of Seq2Seq evaluated with Hausdorff and Ramp Score, averaged over 10 runs (mean $\pm$ standard deviation). For each experiment, best method(s) (Student t-test) in bold.

実験結果(ほかの損失関数との比較)

- 提案手法が一番いい：soft-DTWによるパスの決定→TDIの計算
- ほか2つの手法は形状に関する行列とタイミングに関する行列を1つのsoft-minにまとめて最小化
- DILATE^t-W：時間的なずれの2乗でペナルティ
- DILATE^t-BC：時間的なずれが一定の範囲Tを超えるパスに無限大のペナルティを課す

Eval loss		DILATE (ours)	DILATE^t-W [28]	DILATE^t-BC [43]
Euclidian	MSE (×100)	**1.21 ± 0.130**	1.36 ± 0.107	1.83 ± 0.163
Shape	DTW (×100)	**23.1 ± 2.44**	**20.5 ± 2.49**	**21.6 ± 1.74**
	Ramp	**4.99 ± 0.460**	**5.56 ± 0.87**	**5.23 ±0.439**
Time	TDI (×10)	**14.8 ± 1.29**	17.8 ± 1.72	19.6 ± 1.72
	Hausdorff	**2.70 ± 0.166**	**2.85 ± 0.210**	3.30 ± 0.273

Table 3: Comparison to the tangled variants of DILATE for the Seq2Seq model on the Synthetic dataset, averaged over 10 runs (mean ± standard deviation).

DILATE^t-W

$$\mathcal{L}_{DILATE^t}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) := -\gamma \log \left(\sum_{\mathbf{A} \in \mathcal{A}_{k,k}} \exp \left(-\frac{\left\langle \mathbf{A}, \alpha \boldsymbol{\Delta}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) + (1-\alpha)\boldsymbol{\Omega} \right\rangle}{\gamma} \right) \right) \quad (5)$$

まとめ

- DILATE (DIstortion Loss including shApe and TimE) と呼ばれる損失関数を提案
  - 微分可能なDTW(Soft DTW)とTDI(Soft TDI)を利用
- 課題：突発的変化を予測するための損失関数として、
  MSEは形状とタイミングが捉えられておらず不完全
- 提案手法：微分可能なDTW(Soft-DTW)とTDI(Soft-TDI)をニューラルネットワークの損失関数とする
- 結果：
  - DILATEを利用したニューラルネットワークが、MSEの精度をほぼ失わずに形状・タイミングを改善
  - MSEのみの損失関数では、形状とタイミングが捉えられていない
  - Soft DTWのみの損失関数では、タイミングが捉えられず、精度(MSE)も低い

- 後追いの指標として使うなら、TDI、TDM、Housdorff距離などがありえそう