4P2-OS-38b-1in2

気象時系列データにおける変化点検知の基礎検討

A Fundamental Study on Change Point Detection for Weather Time-series Data

前原 宗太朗 *1　福井 健一 *2　冨田 智彦 *3　小野 智司 *1
Sotaro Maehara　Ken-ichi Fukui　Tomihiko Tomita　Satoshi Ono

*1鹿児島大学大学院理工学研究科情報生体システム工学専攻
Department of Information Science and Biomedical Engineering, Graduate School of Science and Engineering, Kagoshima University

*2大阪大学 産業科学研究所
The Institute of Scientific and Industrial Research, Osaka University

*3熊本大学 先端科学研究部
Faculy of Advanced Science and Technology, Kumamoto University

AMeDAS (Automated Meteorological Data Acquisition System) consists of about 1,300 observation points in Japan to observe weather data such as precipitation, temperature, etc. However, a small environmental change such as relocation of an observation station and contruction of buildings nearby it causes a change of obsevation data. In this paper, we attempt to detect such changes on the observation station environment from observed weather data using Recurrent Neural Network and training data synthesization.

1. はじめに

気象観測データの長期的な蓄積は，気候変動のメカニズム解明，将来予測や多様な気候モデルの解明のために不可欠である．降水量などの気象情報を観測する目的として，地域気象観測システム（アメダス）が全国約 1,300 か所に設けられている． 一方で，地域気象観測においては，観測地点の周囲の環境の変化（建造物の建立など）が生じたり，観測地点の変更が行われることがあり，この前後において観測結果にわずかな変化が含まれる可能性がある．上記のような変化によって観測値に何らかの傾向の変化が明確にみられる場合は，変更があった旨が公表されるものの，観測値に明確な変化がみられない場合はそのような情報が特に公開されないことがある．しかし，全国規模の地球温暖化現象と観測地点周辺の都市化の問題の区別，そして気候変動のメカニズムの正しい理解のために，上記のような変化の発生を把握することは重要である．

本研究では，上記の様な観測環境の変化を，再帰型ニューラルネットワーク（Recurrent Neural Network: RNN）[2] を用いて検出を試みる．変化点検知を行う際は一般的に，正常なモデルを学習し，予測値と実測値の差異に基づいて行う方式が多い．これに対して提案方式では，観測値の非線形的な変化も検出できるよう，変化の度合いを直接出力するネットワークを学習する．これは，気象モデルが複雑な非線形モデルからなるためである．

一方で，観測点の微小な変化の件数は限られており，十分な学習データを用意することは難しい．このため，本研究では，近傍にある 2 箇所の観測地点のデータを合成することで，学習データを人工的に生成する方式を提案する．これにより，教師情報を備えた十分な量の学習データを生成することが可能となる．実験により，本手法で観測環境変化の検知が可能であることを確認した．

連絡先: 前原宗太朗，鹿児島大学大学院理工学研究科情報生体システム工学専攻，sc113067@ibe.kagoshima-u.ac.jp

2. 関連研究

2.1 変化点検知の概要

変化点検知とは，時系列データ上の傾向や性質が変化した点を検出するタスクで，外れ値検出や異常部位検出などとは区別される．一方で，変化度を検出する一般的なアイデアとして，外れ値検出等と同様に確率分布が用いられる．教師ありデータでは，N 個の標本を含む訓練データ D として

$$D = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(N)}, y^{(N)})\} \quad (1)$$

のように入力 x と出力 y のペアが想定される．異常があった場合の $y^{(n)}$ を 1，正常な場合を 0 とした場合，異常度 $a(\boldsymbol{x}')$ は条件付き分布を用いて

$$a(\boldsymbol{x}') = ln \frac{p(\boldsymbol{x}'|y=1, D)}{p(\boldsymbol{x}'|y=0, D)} \quad (2)$$

と定義することができる．一方で，教師なしデータでは

$$D = \{(x^{(1)}), (x^{(2)}), \ldots, (x^{(N)})\} \quad (3)$$

のような観測データがあった場合，異常度が高ければ情報量が多いとして

$$a(\boldsymbol{x}') = -ln\, p(\boldsymbol{x}'|D) \quad (4)$$

と定義することができる [5].

教師なしの変化点検知の手法として特異スペクトル解析がある [1]．この手法では，時系列データとパラメータを与えると解析的に変化度を算出できる特徴をもつ．

一方，ニューラルネットワークを用いて時系列データを扱うモデルも提案されており，特に近年は再帰型ニューラルネットワーク（RNN）が注目されている [2]．RNN は，中間層に再帰的構造を取り入れることにより，時系列を考慮した入出力を可能とした特徴を持つ．予測問題や系列ラベリングに使われることが多く，変化点検知に適用した例は少ない．

表 1: 特異スペクトル解析におけるパラメータ

M	切り出す行列の列数
N	切り出す行列の行数
k	変化度スコア計算時に使う特異値ベクトルの数
L	履歴行列とテスト行列との相対位置

2.2 特異スペクトル解析

特異スペクトル解析による変化点検知では，行列の左特異ベクトルを用いて変化度を定義する．この手法は表 1 に示すパラメータを持ち，時系列データに対して，以下のアルゴリズムで変化度の計算を行う．

Step 1: 時刻 t における履歴行列とテスト行列の切り出し

時刻 t 周りの部分時系列を使って，過去側と現在側に 2 つの行列 $\boldsymbol{X}$ と $\boldsymbol{Z}$ を作成する．$\boldsymbol{x}^{(t)}$ は時系列データの時刻 t から長さ M の要素を切り出した部分時系列であり，$\boldsymbol{X}$ と $\boldsymbol{Z}$ は以下の式に基づいて切り出す．

$$\begin{aligned}\boldsymbol{X}^{(t)} &\equiv [\boldsymbol{x}^{(t-M-N+1)}, \ldots, \boldsymbol{x}^{(t-M-1)}, \boldsymbol{x}^{(t-M)}] \\ \boldsymbol{Z}^{(t)} &\equiv [\boldsymbol{x}^{(t+L-M-N+1)}, \ldots, \boldsymbol{x}^{(t+L-M-1)}, \boldsymbol{x}^{(t+L-M)}]\end{aligned} \tag{5}$$

$\boldsymbol{X}$ は t の直前までの情報を含むため履歴行列と呼ばれ，$\boldsymbol{Z}$ は現在の情報も含むためテスト行列と呼ぶ．L は履歴行列とテスト行列の相対位置を決めるパラメータで，ラグとも呼ばれる．

Step 2: 特異値分解

行列 $\boldsymbol{X}$ に対して以下のように特異値分解を行う．

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Gamma}^{1/2}\boldsymbol{V}^{\mathrm{T}} \tag{6}$$

$\boldsymbol{U} \equiv [\boldsymbol{u_1}, \ldots, \boldsymbol{u_M}]$ とするとき $\boldsymbol{u_i}$ は左特異ベクトルと呼ばれ，変化度推定の際はこのベクトルを用いる．履歴行列，テスト行列それぞれの左特異ベクトルを次のように定義する．

$$\boldsymbol{U}_k^{(t)} \equiv [\boldsymbol{u_1}, \ldots, \boldsymbol{u_k}], \boldsymbol{Q}_k^{(t)} \equiv [\boldsymbol{q_1}, \ldots, \boldsymbol{q_k}] \tag{7}$$

ここで k は左特異ベクトルの数を表すパラメータであり，特異値の数を $\boldsymbol{X}$ の階数よりも小さくとると，式 (6) は近似式とみなされる．$k < rank\boldsymbol{X}$ の際は，小さな特異値に属する特異ベクトルを無視することになり，ノイズ除去と同等の処理となる [5].

Step 3 : 変化度スコアの算出

変化度 $a(t)$ は次の式で定義される．

$$\begin{aligned}a(t) &= 1 - ||{U_k^{(t)}}^{\mathrm{T}} Q_k^{(t)}||_2 \\ &= 1 - ({U_k^{(t)}}^{\mathrm{T}} Q_k^{(t)} \text{の最大特異値})\end{aligned} \tag{8}$$

Step 4: Step 1 から 3 を，行列を切り出せるすべての t に対して繰り返し行う．

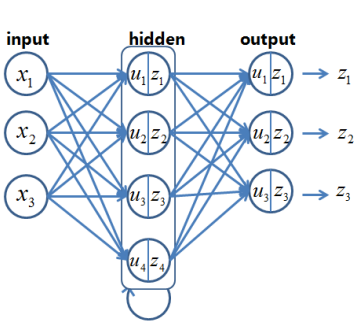


図 1: RNN の図

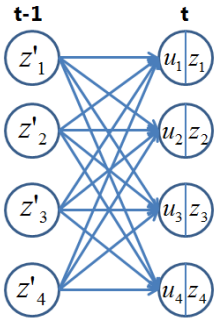


図 2: RNN における自己回帰

3. 手法

3.1 基本アイデア

気象時系列データの変化の要因には複数の要因が含まれる．大きく人工変動と自然変動に分類することができ，人工変動には観測地点の移転や建物の建立など，局所的な変動が含まれる．自然変動には自然気候変動や全地球規模の地球温暖化のいわゆる都市化といった比較的緩やかな変動が含まれる．

本研究では，上記のうち人工変動を検出することを試みる．人工変動による観測データへの影響は非線形的であることが予測されるため，ニューラルネットワークが好適であると考える． 特に，本研究では，時系列データを扱うことに適している RNN を用いて変化点検知を行う手法を提案する．

一方，アメダスのデータベースには，明確な観測データの変化がみられない場合は，観測地点の移転を行う場合であっても，その情報が掲載されないことがある． また，周囲の建造物の出現等による環境変化についても掲載されないことが多い．よって，アメダスデータに機械学習を適用する場合には，教師情報が付与されたデータの件数が限られてしまうという問題が生じる．

このため，本研究では，訓練データを人工的に生成する方式を提案する．これは，近傍（数十 km 程度）にある 2 点の観測地点における観測データを結合することで，観測地点を仮想的に移動させたことに相当する，教師情報付の訓練データを生成する方式である．生成に用いる観測地点の数を増やすことで，指数関数的に訓練データの量を増やすことが可能となる．

3.2 再帰型ニューラルネットワーク

再帰型ニューラルネットワークとは，中間層に再帰的構造を持つニューラルネットワークである．時刻 t において入力があると，$t-1$ の中間層の状態も利用して中間層の計算を行う．過去の状態を参照して中間層の計算を行うため，時系列データを扱うのに適している．簡易的な構造を図 1，2 に示す．

再帰型ではない一般のニューラルネットワークでは，第 1 層のユニットを $i = 1, \ldots, I$，第 2 層のユニットを $j = 1, \ldots, J$ で表すと，第 2 層の中間層は次の式で一般化される．

$$u_j = \sum_{i=1}^{I} w_{ij} x_i + b_i \tag{9}$$

$$z_j = f(u_j) \tag{10}$$

これに対して，RNN における時刻 t の中間層の出力は，上記

表 2: RNN におけるパラメータ

レイヤー数	5
ユニット数	96-24-24-24-1
活性化関数	expomemtial-relu, sigmoid
損失関数	平均 2 乗誤差
バッチサイズ	100
ドロップアウト係数	0.1
最適化	Adam

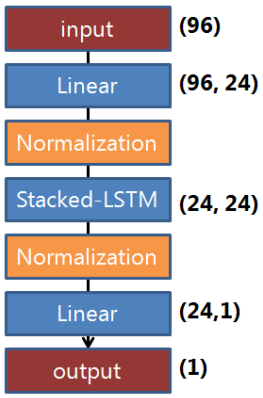


図 3: RNN の構造

を拡張した次の式で表される．

$$u_j^{(t)} = \sum_{i=1}^{I} w_{ij} x_i^{(t)} + \sum_{j'=1}^{J} w_{j'j} y_{j'}^{(t-1)} + b_j \qquad (11)$$

$$z_j^{(t)} = f(u_j^{(t)}) \qquad (12)$$

ここで w は結合の際の重みで，b はバイアス項である．f は活性化関数で，非線形関数が用いられる．

出力層のユニット数や損失関数はタスクによって選択する必要がある．代表的な例として，回帰問題の出力層の数は 1 つで損失関数に二乗誤差を用い，クラス分類ではユニットの数とクラス数が一致し，損失関数にはソフトマックスクロスエントロピーを用いる．変化点検知における出力は各時刻において変化があったかどうかを推定する 2 値分類となる．そのため出力層は 1 つのユニットを持ち，損失関数には二乗誤差を用いる．

本研究で用いる RNN のパラメータを表 2，構造を図 3 に示す．中間層には RNN を拡張した Long short-term memory（LSTM）[2] を使用している．LSTM は RNN に比べ長期依存の時系列問題を解くのに適しているためである．また，学習を安定させる目的でバッチ正則化（Batch Normalization）[3] を行う層を追加している．

3.3 人工データの作成

アメダスの観測データは，降水量や風速等の情報も含むものの，降水量や風速は分散が大きいため，本研究では気温のみを用いることとする．1 年のうち各季節（4 シーズン）の代表的な気温観測値（時間単位: 24 次元）に着目し，年ごとに 96 次元の特徴量を入力する．上記の代表的な気温観測値は各シーズンにおける平均値とする．

訓練データを人工的に生成する際は，近傍（数十 km 程度）にある 2 点の観測地点における観測データを結合することで，観測地点を仮想的に移動させたことに相当する変化を人工的に

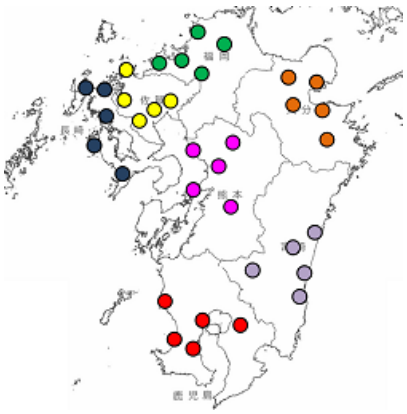


図 4: 観測所の分布

生成する．2 点の観測値の選択，および，2 点の観測データを結合する年の選択により，観測点が増えるにつれて指数関数的に訓練データを生成できる．これにより，実際の観測データにおける観測環境の変化に関する情報の不十分さを補い，RNN の学習を行うことが可能となる．

4. 評価実験

本研究では 2 つの実験を行った．実験 1 では，人工データによる RNN と特異スペクトル解析の比較を行った．この実験では，特異スペクトル解析のような解析的手法では気象時系列データの変化点を見つけるのは困難であり，かつ RNN ではその変化点を発見できたことを示す．

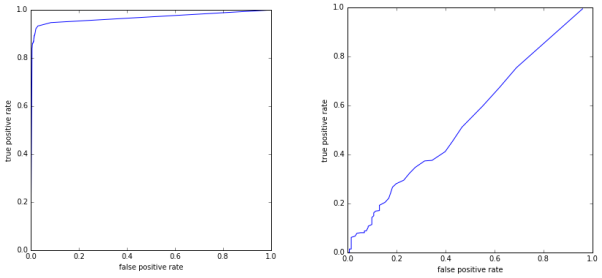
実験 2 では，人工データで学習を行った RNN を実データへと適用した．この実験では，人工データで学習した場合においても，実際の移転情報の検知ができたことを示す．このことより，適切な人工データの作成によって，建物の建立などその他の変化点も検知できる可能性が示唆された．

4.1 人工データによる比較実験

実験 1 では，人工データを用いて，RNN と特異スペクトル解析の性能の比較を行った．九州内の 7 県のうち鹿児島を除く 6 県それぞれにおいて，5 箇所の観測所で観測されたデータに対して，3 章で説明した人工データ生成方法を用いて，訓練用データの生成を行った (図 4 参照)．上記 30 箇所の観測所はそれぞれ 26 年分のデータを含んでおり，各県内 5 箇所のうち 2 箇所で観測されたデータをランダムに選択された時点で結合することで，合計 2,520 件の訓練データを生成し，RNN の学習に用いた．同様に，鹿児島県内 5 箇所の観測所で得られたデータから生成された 420 件のデータをテストデータとした．

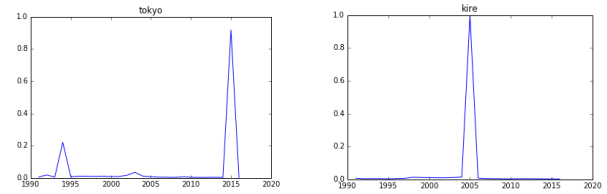
評価基準として，二値分類において広く使われている Receiver Operator Characteristics Curve（ROC）曲線を用いた． これは，閾値を変動させた際の偽陽性率（false positive rate: fpr）と真陽性率（true positive rate: tpr）の関係を示す．また，ROC 曲線の下部の面積は Area under ROC curve（AUR）と呼ばれ，この値が大きいほど検出器の性能が良いとされる．

RNN と特異スペクトル解析の ROC 曲線を図 5 に示す． それぞれの AUR の値は，RNN では 0.967，特異スペクトル解析では 0.506 となった．この結果より，特異スペクトル解析では数 km 単位の観測点の仮想的な移動を検知するが困難であったが，RNN では検出を行えることを確認した．



(a) RNN　　(b) 特異スペクトル解析

図 5: roc 曲線



(a) 東京　　(b) 喜入

図 6: RNN の推定例

4.2　実験 2: 実問題への適用

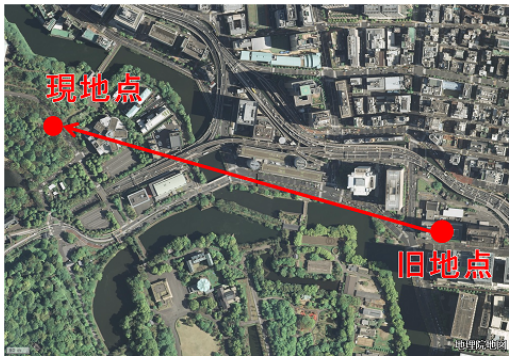
実験 2 では，実験 1 で生成した人工データを用いて学習を行った RNN を，実際の観測データへと適用し，何らかの変化を検出できるかどうかを試みた．なお，実験 2 を行うにあたり，著者らは対象となる観測所における観測環境の変化等について情報を事前に把握しておらず，RNN が変化があったと推定した場合に確認を行うこととした．

本実験では，RNN を用いることで，実際に観測地点が移動した事例を 2 件発見することができた．1 件目は東京都の事例である．実験 1 で学習を行った RNN を適用したとところ，図 6(a) に示すように，2015 年に何らかの変化が生じていた可能性を示唆した．このため，この観測地点における報道資料 [7] を確認したところ，図 7(a) に示すように，2014 年の 12 月におよそ 900m の移動があったことを確認した．2 つ目は鹿児島県喜入の事例である．上記と同様に RNN を適用したところ，図 6(b) に示すように 2005 年に何らかの変化が生じていた可能性を示した．このため，鹿児島地方気象台に確認を行ったところ，図 7(b) に示すように，2005 年の 3 月に観測地点がおよそ 200m 移動されたことが確認できた．

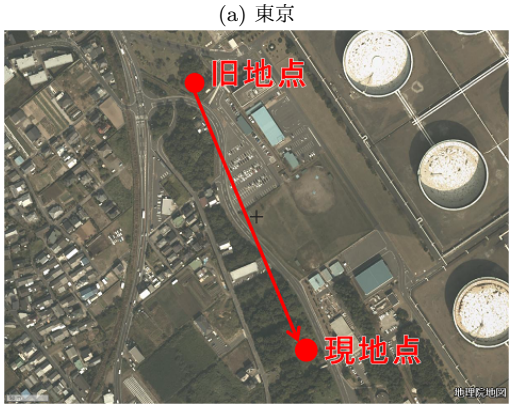
5.　おわりに

本研究では，気象時系列データにおける変化点検知手法を提案した．提案手法は，再帰型ニューラルネットワークを用いて直接変化点の予測を行う点，および，複数の観測データを結合することで人工的に変化点を生成し，訓練データとして用いる点に特徴がある．実験により，移転情報を検知するとができることを確認した．

今後，再帰型ニューラルネットワークが変化の推定を行った際に着目した特徴の明確化について検討を行う．



(a) 東京



(b) 喜入

図 7: 移転情報 (地理院地図 [8] を使用)

参考文献

[1] Idé, Tsuyoshi and Tsuda, Koji: *Change-point detection using krylov subspace learning*, Proceedings of the 2007 SIAM International Conference on Data Mining, pp.515–520, SIAM（2007）

[2] Hochreiter, Sepp and Schmidhuber, Jürgen: *Long short-term memory*, Neural computation, Vol. 9, No. 8, pp.1735–1780, MIT Press（1997）

[3] Ioffe, Sergey and Szegedy, Christian: *Batch normalization: Accelerating deep network training by reducing internal covariate shift*, arXiv preprint arXiv:1502.03167(2015)

[4] Radford, Alec and Metz, Luke and Chintala, Soumith *Unsupervised representation learning with deep convolutional generative adversarial networks* arXiv preprint arXiv:1511.06434(2015)

[5] 井出剛, 杉山将: 機械学習プロフェッショナルシリーズ 異常検知と変化点検知, 講談社（2015）

[6] 岡谷貴之: 機械学習プロフェッショナルシリーズ 深層学習, 講談社（2015）

[7] 国土交通省気象庁（2014）: 地上観測地点「東京」の移転について, http://www.jma.go.jp/jma/press/1410/03b/20141003_tokyo_rojo.pdf

[8] 地理院地図: http://mapps.gsi.go.jp/maplibSearch.do#1