F-010

k-medoids 法による農業環境データの分析
Analysis of Agricultural environment Data Based on k-medoids Clustering

鈴木 一矢 † 岩﨑 清斗 ‡ 大久保 誠也 † 斉藤 和巳 †
Kazuya Suzuki Kiyoto Iwasaki Seiya Okubo Kazumi Saito

1. はじめに

多様なトレンドなどを反映する SNS データ，企業株価などに象徴される経済データ，さらに，地球環境などをセンシングした測定データなどは，刻々と変化する時系列データの代表例である．過去の時系列データから将来の数値傾向を高い精度で予測することは，どの分野でも極めて高いニーズの基本タスクである．また，これら時系列データ生成の背後に内在する基本メカニズムに変化が起きたとき，その変化を効率良く高い精度で検出できれば，数値傾向の予測精度の向上が期待できる．それに加えて，その変化を起こした要因などを探求することで，データ生成に関する基本知識や法則の理解を深めることができる．我々は，このような変化検出をレジーム切換（regime switching）問題として定式化し，変化時刻などの検出技術を開発してきている [1]．

一方，農業では担い手の高齢化による労働力不足が深刻化しており，作業の合理化による生産性の向上や熟練農家が持つ技術の継承が課題となっている．しかし，これらは作業者の勘や経験に頼ることが大きいため，暗黙知として定量的に捉えることが困難である．

本研究では，レジーム切換問題として捉えることで，農業の課題解決に結びつくような基本知識や法則の獲得を目指す．そのための基本タスクとして，複数地点で観測された気圧，温度，湿度，日照時間といった気象情報を農場での作物の生育に結びつく環境情報として捉え，これらの基本関係を分析する．

2. 分析法

時刻 1 から T までの時系列データを T-次元ベクトル $\mathbf{x} = (x_1, \cdots, x_T)$ と表す．総数 N の時系列データ集合 $\mathcal{X}$ が与えられたとき，任意の時系列ペア $\mathbf{x}$ と $\mathbf{y}$ に対し，類似度関数 $s(\mathbf{x}, \mathbf{y})$ が定義されているとする．本稿では類似度として，機械学習やテキストマイニング分野 [**?**] で頻繁に採用される以下のコサイン類似度を用いる．

$$s(\mathbf{x}, \mathbf{y}) = \frac{\sum_{t=1}^{T} x_t y_t}{\sqrt{\sum_{t=1}^{T} x_t^2 \sum_{t=1}^{T} y_t^2}}. \quad (1)$$

k-medoids 問題とは，与えられた非負整数 k に対して，下記の目的関数 f を最大にする k 個のベクトルの集合 $\mathcal{B}_k \subset \mathcal{X}, |\mathcal{B}_k| = k$ を求める問題である．

$$f(\mathcal{B}_k) = \sum_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{z} \in \mathcal{B}_k} \{s(\mathbf{x}, \mathbf{z})\}. \quad (2)$$

ここで，$\mathcal{B}_h \subset \mathcal{B}_k$ とすれば，任意のベクトル $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ に対して，以下の関係が成立する．

$$\max_{\mathbf{z} \in \mathcal{B}_h \cup \{\mathbf{x}\}} \{s(\mathbf{y}, \mathbf{z})\} - \max_{\mathbf{z} \in \mathcal{B}_h} \{s(\mathbf{y}, \mathbf{z})\} \geq \max_{\mathbf{z} \in \mathcal{B}_k \cup \{\mathbf{x}\}} \{s(\mathbf{y}, \mathbf{z})\} - \max_{\mathbf{z} \in \mathcal{B}_k} \{s(\mathbf{y}, \mathbf{z})\}. \quad (3)$$

よって，目的関数 f がサブモジュラ関数となることは容易に確認できる．

サブモジュラ関数最大化の標準解法となる貪欲法の詳細を以下に示す．

1. 反復制御変数を $h = 0$ とし，結果を格納する集合を空 $\mathcal{B}_0 = \emptyset$ に初期化;
2. 集合 $\mathcal{B}_h$ を固定し，$\mathcal{X}$ から最良要素 $\mathbf{z}_{h+1} = \arg\max_{\mathbf{x} \in \mathcal{X}} \{f(\mathcal{B}_h \cup \{\mathbf{x}\}) - f(\mathcal{B}_h)\}$ を計算;
3. 最良要素 $\mathbf{z}_{h+1}$ を追加 $\mathcal{B}_{h+1} \leftarrow \mathcal{B}_h \cup \{\mathbf{z}_{h+1}\}$ し，$h = h + 1$ に設定;
4. $h = k$ ならば終了，さらもなければステップ 2. へ戻る．

貪欲法では厳密解は得られないものの，妥当な精度で最悪ケースの解品質を理論的に保証することができる．詳細には，厳密解を $\mathcal{B}^*$ とすると，貪欲法で求まる近似解 $\mathcal{B}$ の精度は，関係式 $f(\mathcal{B}) \geq (1 - 1/e) f(\mathcal{B}^*)$ で抑えられる．ここで，e は自然対数の底であり，貪欲解により，最悪でも厳密解の 63% 程度以上の性能が保証される [2]．

3. 実験による評価

本実験では，2006 年 1 月 1 日から 2016 年 12 月 31 日までの 11 年間の期間で，平均気圧，平均気温，平均湿度，平均秒速，及び，日照時間の 5 項目の日別データを用いた §．対象地点は，静岡県内の浜松市，三島市，及び，静岡市の 3 地点を選定した．

実験結果を図 2 に示す．図中では，浜松市，三島市，及び，静岡市を順番に H，M，S と略記し，平均気圧，平均気温，平均湿度，平均風速，及び，日照時間を順番に P，T，H，W，S と略記している．例えば，浜松市の平均気圧は HP となる．

図 2 (a) に，式 (1) のコサイン類似度に基づき，各地点それぞれの項目時系列ベクトルを Ward 法 [3] による階層的クラスタリングした結果を，デンドログラムとして示す．図 2 (b) から (f) には，$k = 5$ に設定した k-medoids クラスタリングで求めた代表ベクトル $\{\mathbf{z}_1, \cdots, \mathbf{z}_k\}$ のそれぞれの描画結果を示す．また，図 2 (a) のノードは，k-medoids クラスタリングの結

†静岡県立大学 経営情報学部
‡静岡県工業技術研究所 電子科

§http://www.data.jma.go.jp/obd/stats/etrn

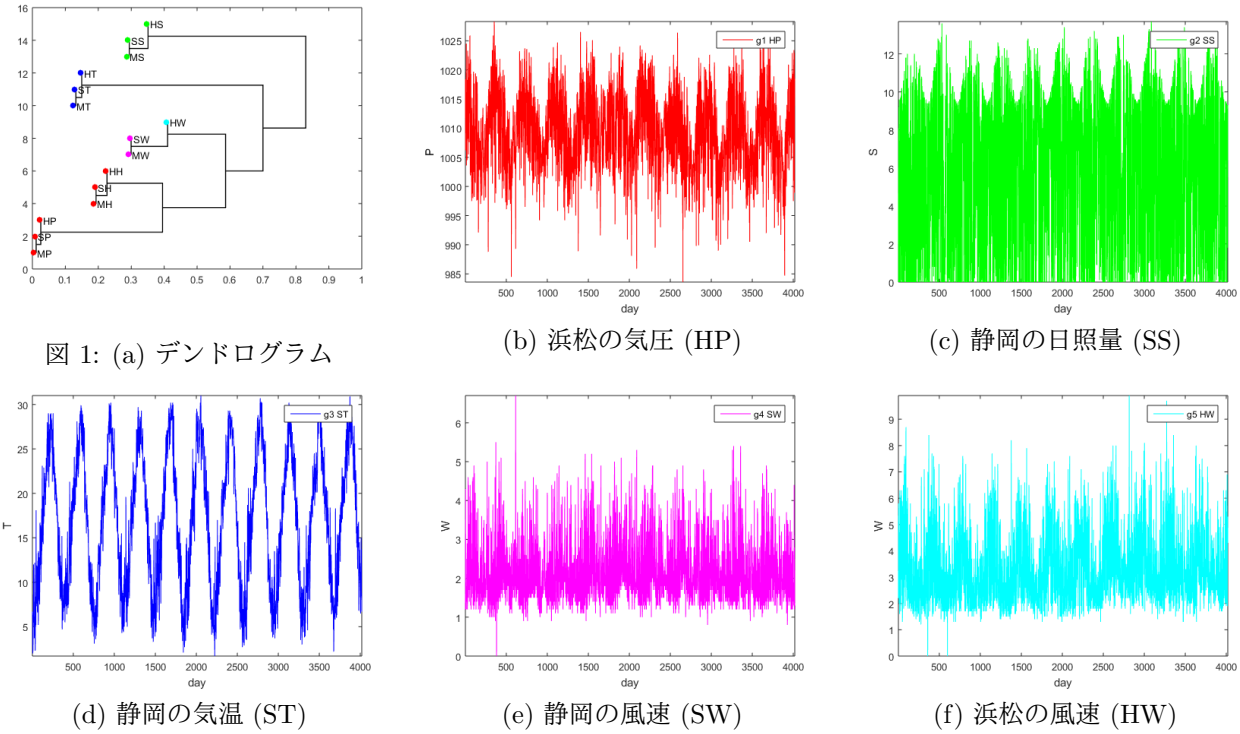

図 1: (a) デンドログラム

(b) 浜松の気圧 (HP)

(c) 静岡の日照量 (SS)

(d) 静岡の気温 (ST)

(e) 静岡の風速 (SW)

(f) 浜松の風速 (HW)

図 2: k-medoids クラスタリング結果のデンドログラムと代表ベクトル

果に基づき着色し，図 2 (b) から (f) の描画でも対応した着色にしている．

デンドログラムは 2 つの似ている要素を繋ぎ合わせ，早く繋がっているデータは，より類似したデータとなる．図 2 (a) より，気圧，気温，湿度，風速，日照量，どれをとっても静岡と三島の 2 つの都市が似ていると分かり，浜松だけこの 2 つの都市とは異なることが見られた．また，5 つの項目の内，気温と気圧については，3 都市共に類似しているものの風速については，浜松が他の 2 都市と大きく異なっていることが分かる．この実験結果は，浜松など静岡県西部では，「遠州のからっ風」と呼ばれ，冬に北西風が強まることに符合する．

項目毎のデータに着目してみると，どの項目も季節毎による周期的な変化をしていることがわかる．その中でも，日照量のデータは他のデータを比べて大きく異なっていることがわかる．これは，日照量の項目のみは，他のデータと異なり，天気の影響により季節によらず 0 となる場合が多いためと考えられる．また，気温は季節による，非常に安定した変化を見せており，気圧や風速，湿度のような揺れの多い変化をしていないことがわかる．これらのことは，0 が多い日照量，安定した変化の気温，それ以外のデータという形に分割されるかたちでデンドログラムにも表れている．以上の事から，日照量や気温データの個々のデータのみでは，他の項目との関係性を求めることが難しいことが分かる．一方で，他のデータと関係ないことから，予想を行う際には，個別に扱う必要がある可能性がある．

一方で，デンドログラムからも，k-medoids クラスタリングからも，気圧と湿度には，ある程度の関係がありそうなことが示された．

4. おわりに

本稿では，複数地点で観測された気圧，温度，湿度，日照時間といった農場環境情報の基本関係を分析した．具体的には，K-medoids 法による分析法を示し，その実験結果について報告した．今後は，さらに多様な環境情報や観測地点での評価実験を進める．

謝辞 本研究は，科学研究費補助金基盤研究 (C)(No.15K00429) の助成を受けた．

参考文献

[1] K. Saito, K. Ohara, M. Kimura and H. Motoda, ”Change Point Detection for Burst Analysis from an Observed Information Diffusion Sequence of Tweets,” Journal of Intelligent Information Systems (JIIS), 44:243-269, 2015.

[2] 室田一雄, ”離散凸解析の考えかた 最適化における離散と連続の数理.” 共立出版, 2007.

[3] J.H.Jr. Ward, ”Hierarchical grouping to optimize an objective function.” Journal of the American Statistical Association, 58:236-244, 1963.