クラスタリングされたヒートマップと折れ線グラフによる時系列データの可視化

Visualization of Time Series Data Using Clustered Heatmaps and Line Graphs

遠藤　麗香
Reika Endo
法政大学情報科学部ディジタルメディア学科
E-mail: reika.endo.3g@stu.hosei.ac.jp

Abstract

There are several methods to visualize time series data such as line graphs, stacked graphs, and heatmaps. Line graphs have an advantage that it is easy to read the values of the data and the increase or decrease in the values from the slopes of lines. However, line graphs lack visibility when the numbers of lines increase. By contrast, heatmaps do not lessen visibility by the increase of the data but require more visualizing space. Also, it is more difficult to read values from heatmaps than from line graphs. This paper proposes a method that combines heatmaps and line graphs to make it easier to recognize the features and the exact values of the data. The time series data are clustered by their similarity and each cluster will be drawn as a band shape rectangle. By clicking on the rectangle, the data belonging to the cluster will be drawn as a line graph next to the heatmap area. The number of drawn clusters can be changed by a slider. An experiment was conducted to compare the data readability of the proposed method with that of line graphs and that of heatmaps. The results show that the method provides users a higher readability of exact values than the line graphs and the heatmaps.

1. はじめに

時系列データの可視化には折れ線グラフや積み上げグラフ，ヒートマップがよく使われる．折れ線グラフはデータの正確な値を読み取りやすく，線の傾きから値の増減が分かりやすいが，線の増加によって線の重なりが増え，値の読み取りが難しくなるほか，値の範囲が広いと，折れ線と折れ線の間に無駄な空間ができることがある．ヒートマップは 2 次元データの値を色の濃淡で表したグラフであり，2 つの次元のうち，1 つの次元に時間を割り当てることで時系列データの可視化が可能となる．商業的にはウェブサイトのクリック箇所やよく見られている箇所を色の濃淡で表すことに利用されており，科学技術分野では，点が密集して見にくくなる散布図や，線の増加にともない折れ線同士の重なりが増える折れ線グラフの代わりによく使われている．複数の時系列データの可視化時には，各データを帯状のヒートマップで可視化し，帯を縦に並べることが多い．横軸が時刻となっており，その時刻での値を対応する色で帯を縦に塗りつぶす．データが増えるほど帯は増えるため，描画に必要な領域は増える．また，折れ線グラフに比べて値の読み取りが難しい．

本研究では，複数の時系列データの特徴や傾向，正確な値をより簡単に確認できるようにすることを目的として，データをクラスタリングし，ヒートマップによる特徴や傾向などの大まかな情報と折れ線グラフによる正確な値などの詳細な情報を組み合わせて可視化する手法を提案する．また，スライダで描画するクラスタ数を変えられるようにする．スライダによる対話的な可視化を行うことで，ユーザのニーズに合わせたクラスタ数での可視化を可能にし，データの読み取りをより容易にさせる．本手法のデータの読み取りやすさを確認するため，折れ線グラフのみの可視化，ヒートマップのみの可視化と比べる実験とアンケートを行った．実験の結果，本手法は他の 2 つの手法より高い読み取り精度を得た．また，アンケートより正確な値の読み取りやすさ，複数の時系列データの可視化に適していたかの 2 項目で最も高い評価を得た．データの探しやすさの項目では本手法が最も高い評価を得たが，他の 2 つの手法の評価と有意な差があると言えなかった．本手法はスライダの調節の時間がかかるため，データを探す時間が長くなることが分かった．

2. 関連研究

Kumatani ら [1]は時系列データをクラスタリングし，ヒートマップで可視化すると同時に，データ間の類似度と時刻間の類似度の位置関係を多次元尺度法で可視化する手法を提案した．クラスタリングはデータの相関係数から得られた距離を用いており，クラスタリングのためのしきい値はスライダで変えられるようにしている．他のデータと似ていないと判断されたデータはデータ数 1 のクラスタを作るため，重要でないと判断し，ヒートマップに描画しない．

Yagi ら [2]はタグ付けされた複数の時系列データを折れ線グラフで可視化し，表示する折れ線の数を対話的に調節する手法を提案した．データをタグや形状によってクラスタリングし，各クラスタの代表値のみを描画する．タグを指定することや折れ線をクリックすることでデータのフィルタリングが可能となっている．これら 2 つの提案手法はクラスタリングとフィルタリングによって重

Supervisor: Prof. Hiroshi Hosobe

要なデータのみを残して，表示するデータ数を抑えているため，広い描画領域を必要とせず，折れ線の重なりが減っている．しかし，ヒートマップではそもそも値の正確な読み取りが難しいという欠点が残る．

3. 準備

3.1. 階層型クラスタリング

クラスタリングには，階層型クラスタリングと非階層型クラスタリングの 2 種類が存在する．階層型クラスタリングは，似たデータ，つまり近い距離のデータを 1 つずつ階層的に連結させて木の葉の部分からボトムアップ式にクラスタを作成する手法である．階層型クラスタリングで得られるグラフは図 1 のようになり，このグラフをデンドログラムという．横軸が各データ，縦軸が距離を表している．このデンドログラムの階層を横に切る場所，しきい値によって構築されるクラスタの数が変わる．

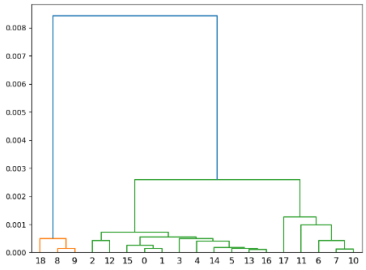


図 1 デンドログラム

4. 提案手法

本論文では，Kumatani ら [1]と同様，時系列データをクラスタリングし，ヒートマップで可視化する手法を提案する．さらに，正確な値の読み取りも可能にするために，ヒートマップは各クラスタのデータの平均値を，折れ線グラフはクラスタ内のデータの実際の値を可視化する．

入力データはm個あるとし，各データはn時刻を持つとする．データ群をA，i番目のデータのj番目の時刻における値をv_{ij}とすると，以下の族として表せる．

$$A = \{a_1, \ldots, a_m\}$$
$$a_i = \{v_{i1}, \ldots, v_{in}\}$$

i番目のデータa_iとk番目のデータa_kの相関係数を$c(i,k)$とすると，2 つのデータの距離d_{ik}は以下のように表せる．

$$d_{ik} = 1 - c(i,k)$$

距離d_{ik}は，類似するデータ同士では小さくなり，類似しないデータでは大きくなる．この距離d_{ik}を元に階層型クラスタリングを行う．

可視化結果は図 2 である．左側に各クラスタのヒートマップ，右側にクラスタをクリックで選択したときの，そのクラスタに属するデータの詳細を折れ線グラフで表示するスペースがある．ヒートマップの上部にはクラスタ数を変えるためのスライダを配置する．初期画面が図 2 (a)である．左側にヒートマップが表示されているが，右側に折れ線グラフは表示されていない．初期画面でのスライダによるしきい値は 0 であり，それぞれデータ数 1 のクラスタを作る．スライダを調節し，ヒートマップの帯を 1 つクリックした後の画面が図 2 (b)である．図 2 (a)と比べて，描画されているヒートマップの帯の数が変わっている．スライダを動かしたことでしきい値が変わり，そのしきい値でのクラスタ数が減ったことで描画される帯の数が変わっている．また，画面右側には折れ線グラフが表示され，折れ線グラフの下に地点名が表示されている．帯をクリックしたことで，クリックされたクラスタ内のデータが折れ線グラフとして表示されている．別の帯をクリックすると右側に表示される折れ線グラフは変わる．

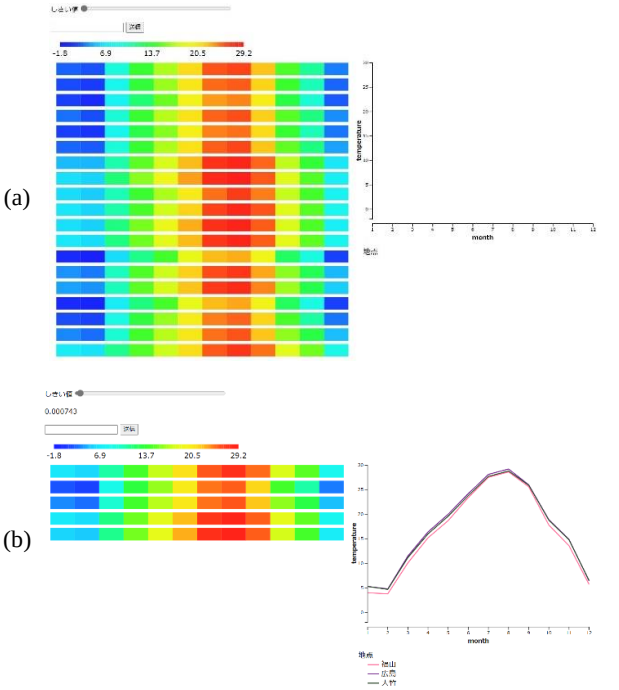


図 2 可視化結果：(a) 初期画面，(b) スライダを調節し，帯をクリックした後の画面

ヒートマップの可視化は，各データの各値を 1 つの長方形として，各長方形が横にn個並ぶようにする．また，それぞれデータの値から長方形の色を決定する．色の決定にはシグモイド関数を複数使用し，正規化した値から RGB 値を計算し，各長方形の色とする．αをゲインと呼ばれる任意の定数として，シグモイド関数は以下のように表せる．

$$\mathrm{sigmoid}(x) = \frac{\tanh(\alpha x/2) + 1}{2}$$

スライダを動かすたびに，切った場所つまりそのしきい値と階層構造が持つデータ間の距離を比較し，クラスタ数を計算して，描画されるヒートマップの帯の数が変わるようにする．再計算の結果クラスタ数が変わらなければ描画される帯の数は変わらない．各帯で描画される値は各クラスタ内のデータ値の平均値とする．あるクラスタCにs個のデータ$\{a_{i_1}, .., a_{i_s}\}$が含まれているとき，クラスタCの平均値a_{average}は値vを用いて以下の族として表せる．

$$a_{\mathrm{average}} = \left\{\frac{v_{i_1 1} + \cdots + v_{i_s 1}}{s}, \ldots, \frac{v_{i_1 n} + \cdots + v_{i_s n}}{s}\right\}$$

平均値a_{average}の任意の値vと，データ群A中のvの最大値$v_{\max}$，vの最小値$v_{\min}$の中点$(v_{\max}+v_{\min})/2$によって，平均値a_{average}のvを正規化し，シグモイド関数を使用する．これにより長方形の RGB 値を決定する．

ヒートマップの帯をクリックするたび，そのクラスタに含まれる全てのデータを右に折れ線グラフとして表示し，その各データ名を折れ線グラフの下に表示する．あるクラスタCにs個のデータ$\{a_{i_1},..,a_{i_s}\}$が含まれているとき，クラスタCの帯をクリックすると，s個のデータの折れ線グラフが表示される．スライダでしきい値を調節することで，クラスタ数が変わり，ヒートマップの帯の数，帯をクリックしたときに表示される折れ線グラフを変えられる．そのため，全てのデータをヒートマップで表示したときと比べて描画に必要な領域を減らすこと，折れ線グラフで表示したときと比べて，折れ線同士の重なりを減らすことが可能である．

5. 実装

可視化するデータは，気象庁の過去の気象データ・ダウンロードページから日本各地の月平均気温データを CSV ファイルで取得した．ファイルの読み込み，類似度計算，クラスタリングは Python 言語で実装し，ヒートマップと折れ線グラフでの可視化には JavaScript および JavaScript のライブラリ D3.js を使用する．ファイルを読み込み，必要な情報を抽出したら，各データ同士の相関係数を求める．相関係数は相関係数行列の形で出力されるため，行列の対角成分より上側の値のみ使用する．相関係数から距離を求めた後，階層型クラスタリングを行う．このとき得られるクラスタリング結果は階層構造になっていないため，階層の葉の要素である各データとデータ間の距離を階層構造にする必要がある．階層構造にしたクラスタリング結果，気温データ，日本各地の地点名は JSON ファイルに保存し，可視化で使用する JavaScript 側でファイルを読み込む．JavaScript 側ではファイルを読み込む処理，ヒートマップを描画する処理，ヒートマップの帯がクリックされたときに折れ線グラフを表示する処理，スライダを動かすたびに実行されるクラスタ数再計算処理を実装する．

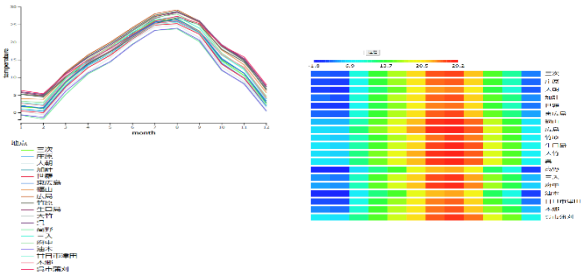
6. 実験

6.1. 実験方法

本手法の有用性を評価するために，本手法(図 2)，折れ線グラフのみ(図 3(a))，ヒートマップのみ(図 3(b))の 3 通りの手法で複数のデータを持つ時系列データを可視化し，どの可視化手法がデータの読み取りをしやすいか比べる実験を行った．折れ線グラフのみでの可視化では，全てのデータがそれぞれ折れ線として描画されており，各地点名が折れ線グラフの下に表示されている．ヒートマップのみでの可視化では，全てのデータがそれぞれ 1 つの帯として描画されており，各地点名が各帯の右に表示されている．使用したデータは茨城県の 14 地点，愛知県の 18 地点，広島県の 19 地点，東京都の 18 地点の月平均気温データである．

被験者は 6 名であり，15 歳から 49 歳までの男性 4 名，女性 2 名である．被験者には 3 通りの手法を被験者ごとに異なる順番で使用してもらい，指定したデータの該当する気温または地点を探してもらった．ただし，各被験者の 1 番目に使用する手法では愛知県のデータ，2 番目に使用する手法では広島県のデータ，3 番目に使用する手法では東京都のデータを可視化している．実験前の各手法を説明する際には茨城県のデータを使用した．探してもらったデータは各手法 2 つずつである．1 つ目は地点名と月を指定し，その時の小数第 1 位の桁までの気温(E1)，2 つ目は月と気温を指定し，条件に合う地点名である(E2)．各 E1，E2 の解は必ず 1 つずつしかないとする．E1 では平均二乗誤差と平均解答時間を，E2 では正答率と平均解答時間を算出した．

実験後に正確な値の読み取りやすさ，データの探しやすさ，複数の時系列データの可視化に適していたかの 3 項目で 5 段階の評価を問うアンケートを実施した．選択肢は 5 (とても良い)，4 (良い)，3 (どちらでもない)，2 (あまり良くない)，1 (全く良くない)の 5 つである．また，改善点や意見を書いてもらうための自由回答欄を設けた．



(a) (b)

図 3 (a)折れ線グラフのみ，(b)ヒートマップのみによる可視化

6.2. 実験結果

実験の結果をまとめたものが表 1 である．E1 の折れ線グラフのみとヒートマップのみにおいて，解答時間を正確に測定できていなかった回答は除いた．そのため，その 2 つの項目は 5 つの回答の平均となっている．地点と月から気温を答える E1 の平均二乗誤差は 3 つの内，本手法が 0.283 と最も小さくなり，ヒートマップのみの可視化が 1.60 と最も大きくなった．平均解答時間はヒートマップのみが最も短くなり，折れ線グラフのみが最も長くなった．月と気温から地点を答える E2 の正答率は本手法と折れ線グラフのみでは 83.3%であったが，ヒートマップのみでは 66.7%と最も小さくなった．平均解答時間は折れ線グラフが最も短くなり，本手法が最も長くなった．

表 1 実験結果

		E1		E2
本手法	平均二乗誤差	0.283	正答率	83.3%
	平均解答時間	74.3s	平均解答時間	107s
折れ線グラフ	平均二乗誤差	0.758	正答率	83.3%
	平均解答時間	92.3s	平均解答時間	58.4s
ヒートマップ	平均二乗誤差	1.60	正答率	50.0%
	平均解答時間	66.9s	平均解答時間	79.8s

6.3. アンケート結果

アンケートの正確な値の読み取りやすさ(Q1)，データの探しやすさ(Q2)，複数の時系列データの可視化に適していたか(Q3)の評価の平均が図 4 である．全ての設問において本手法が最も良い結果を得た．図 3(a)にある折れ線グラフのみでは折れ線同士の重なりが多く，Q2 のデータの探しやすさの評価が低い．ヒートマップのみの可視化は Q1，3 では最も評価が低いが，Q2 のデータの探しやすさは折れ線グラフのみより評価が高い．これらの設問の評価に対して分散分析を行ったところ，Q1 のp値は $p = 0.00273 < 0.05$，Q2 は$p = 0.0520 > 0.05$，Q3 は$p = 0.0471 < 0.05$となった．このことから，Q1 と Q3 では 3 つの手法の評価平均に有意な差があると言えるが，Q2 では有意な差があるとは言えないことが分かった．また，自由回答欄の回答として，「互いの良い悪い点をカバーできるのは良い．」「ヒートマップや折れ線の色設定が見難さを強くしていた．各手法とも見易くする工夫があればより使い易いと思う．」といった回答があった．

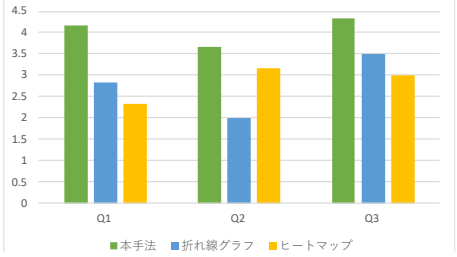


図 4 アンケート結果

7. 議論

7.1. 実験

実験の結果，地点と月から気温を答える E1 の平均二乗誤差は 3 つの内，本手法が最も小さくなった．折れ線グラフのみでは折れ線の重なりが多く，指定された値を見つけることが難しく，ヒートマップのみでは各長方形の色と基準のカラースケールから正確な値を読み取ることが難しかったのだと考える．本手法はスライダでクラスタ数を変えられるため，クラスタ数を変えて表示される折れ線の数を読み取りやすい数に調節したと考える．E1 の平均解答時間を見ると，ヒートマップのみが最も短く，折れ線グラフのみが最も長くなった．一方，月と気温から地点を答える E2 の平均解答時間は折れ線グラフのみが最も短く，本手法が最も長くなった．本手法はスライダによる調節に時間がかかるため，平均解答時間は折れ線グラフのみやヒートマップのみと比べて長くなると考えていたが，E1 では折れ線グラフのみより短くなっている．折れ線グラフのみの平均解答時間が本手法より長くなった理由として，折れ線グラフのみの可視化は折れ線の重なりが多く，該当する地点の折れ線を探す時間や値の読み取り時に繰り返し見比べることでかかる時間が長くなったためと考えた．しかし，E2 では平均解答時間が最も短いため，さらなる検証が必要である．E2 の正答率は本手法と折れ線グラフのみでは 83.3%であったが，ヒートマップのみでは 50.0%と最も小さくなった．ヒートマップのみでは E1 と同様に長方形の色とカラースケールから値を読み取ることが難しく，似た値を持つ地点を選んだが，本手法では折れ線グラフも用いて似た値を持つ地点から該当する地点を選んだのだと考える．

7.2. アンケート

Q1 の正確な値の読み取りやすさにおいて，本手法は 4 以上と最も高い評価を得ており，値の読み取りを容易に感じた被験者が多かった．データの探しやすさでは，本手法，ヒートマップのみ，折れ線グラフのみの順に評価が高かった．ヒートマップはデータ同士の重なりがなく，色の違いによってデータを探しやすいのだと考える．しかし，ヒートマップ単体では同じ色味のデータからさらに絞り込みをすることは難しい．そのため，本手法使用時にはヒートマップと表示させる折れ線の数を調節し，読み取りやすくした折れ線グラフを用いて条件に当てはまるデータを探したと推測する．分散分析を行った結果の Q2 のp値が 0.05 を超えた理由として，被験者の回答のばらつきが大きかったからだと考える．

8. おわりに

本論文では，クラスタリング，ヒートマップ，折れ線グラフを組み合わせて，複数のデータを持つ時系列データの値を読み取りやすく可視化する手法を提案した．ヒートマップのみ，折れ線グラフのみの可視化とデータの読み取りやすさを比べる実験では，より高い読み取り精度を得た．また，データの探しやすさに関してもヒートマップのみ，折れ線グラフのみより良い評価を得たが有意な差があるとは言えなかった．本手法はスライダによるクラスタ数の調節の時間がかかるため，データを探す時間が長くなることが分かった．

今後の課題として，データを探す時間を短縮すること，値の読み取りをさらに容易にすることがあげられる．ユーザがグラフを何度も見比べる必要がないよう，ヒートマップや折れ線グラフ部分の色設定や目盛りを見直し，読み取りやすさを向上させ，データを探す時間を短縮する必要がある．また，データを探しやすくするために，デンドログラムを組み合わせた可視化を行うことがあげられる．スライダのみではしきい値と階層構造の関係が分かりにくく，ユーザ自身で求めているクラスタ数に調節することが難しい．デンドログラムとしきい値を表す直線の可視化を導入し，クラスタの関係を見やすくすることでスライダの調節が容易になると考える．

文　献

[1] S. Kumatani, T. Itoh, Y. Motohashi, K. Umezu and M. Takatsuka, "Time-Varying Data Visualization Using Clustered Heatmap and Scatterplots," *Proc. 20th International Conference Information Visualisation*, pp. 63-68, 2016.

[2] S. Yagi, Y. Uchida and T. Itoh, "A Polyline-Based Visualization Technique for Tagged Time-Varying Data," *Proc. 16th International Conference on Information Visualisation*, pp. 106-111, 2012.