4ZF-07

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

遠藤 麗香† 細部 博史‡

法政大学情報科学部†‡

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．科学技術分野では，点が密集して見にくくなってしまう散布図や，線の増加にともない折れ線どうし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．

2. 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が，正確な値の読み取りも可能にするために，ヒートマップでは各クラスタのデータの平均値を，折れ線グラフではクラスタ内のデータの実際の値を可視化する手法を提案する．

実行後の初期画面が図 1 (a)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デンドログラムの切る場所を変えるためのスライダを配置する．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，各長方形が横に並ぶようにする．また，それぞれ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．ヒートマップの帯を 1 つクリックした後の画面が図 **1**(b)である．図 **1**(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，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．

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．

ヒートマップの帯をクリックするたび，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，各データ名を折れ線グラフの下に表示する．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ることができる．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．

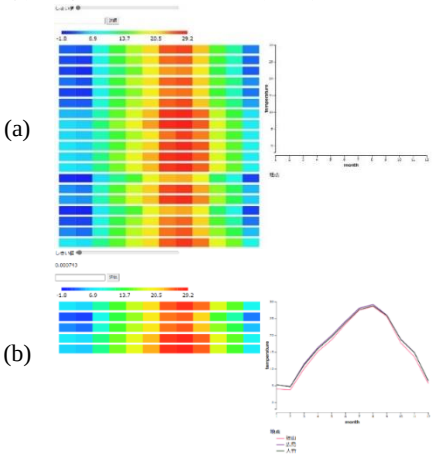

図 1 可視化結果：(a) 初期画面，(b) スライダを調節し，帯をクリックした後の画面

3. 実験

本手法の有用性を評価するために，本手法，折れ線グラフのみ，ヒートマップのみの 3 通りの手法で気象庁の日本各地の月平均気温データを可視化し，どの可視化手

Visualization of Time Series Data Using Clustered Heatmaps and Line Graphs

†Reika Endo, Faculty of Computer and Information Sciences, Hosei University

‡Hiroshi Hosobe, Faculty of Computer and Information Sciences, Hosei University

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．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同士で被りがないような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(Q1)，2つ目は月と気温を指定し，条件に合う地点名である(Q2)．各Q1，Q2の解は必ず1つずつしかないとする．Q1では平均二乗誤差と平均解答時間を，Q2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（問1），データの探しやすさ（問2），複数の時系列データの可視化に適していたか（問3）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）から1（全く良くない）の5段階である．

実験の結果をまとめたものが表1である．Q1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Q1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Q2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．アンケートの結果の平均が図2である．全ての問において本手法が最も良い結果を得た．

表1　実験結果

		Q1		Q2
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

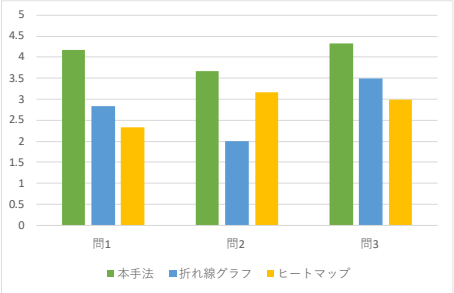

図2　アンケート結果

4.　議論

実験の結果，地点と月から気温を答えるQ1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．Q1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Q2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，Q1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，Q2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．Q2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Q1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．

問1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，自由回答欄の回答のように，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整し，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．複数の時系列データの可視化に適していたかの問では本手法が最も高い評価を得た．

5.　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．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．

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.