ヒートマップによる高次元可視化のためのクラスタリング手法の比較

熊谷 沙津希　(指導教員：伊藤 貴之)

1. 概要

高次元データの分析過程における重要な点に次元間の相関がある．高次元データの可視化においても，相関にもとづいて次元群にクラスタリングを適用し，クラスタを単位とした視覚表現を適用することで，相関のある次元どうしの類似性や差異に注目することが容易になる．本報告では 2 種類のクラスタリング手法を高次元データに適用し，ヒートマップを用いて可視化した結果について比較した結果を報告する．本研究ではクラスタリング手法として，最短距離法に基づく階層型クラスタリングと，非階層型クラスタリングの一種である k-medoids 法を用いる．本報告ではそれぞれのクラスタリング手法によって見られる特徴の違いを比較するとともに，ユーザ主観評価結果からも両者の違いを議論する．

2. 関連研究

画面空間の横軸に時刻を割り当てる形式で時系列高次元データを可視化する手法として，折れ線グラフとヒートマップが広く用いられている．しかし，折れ線グラフはデータを構成する個体数や次元数の増加とともに，折れ線同士の絡み合いも増加し，視認性低下の原因となる．これらの問題を解決するために，折れ線の表示数を対話的に調節する手法が提案されている[1]．また，数値の範囲が大きく，かつ数値分布が不均一である場合には，画面領域を浪費するような可視化結果になることが多い，という問題もある．

ヒートマップは値の大きさを色で表示する可視化手法である．ヒートマップを用いた高次元データの可視化には，データを構成する個体や次元を縦軸に沿って一列に並べ，横軸に何らかの次元（時系列データの場合には時刻）を割り当てて，両軸を分割して得られる各領域に色をつける．ヒートマップは次元数が高いデータにおいても，数値を表現する形状が画面上で重なり絡み合うことがないため，視認性の維持が容易である．さらに折れ線グラフと比較して，数値の範囲や分布により表示領域を浪費するような可視化結果を生じることもない．

ヒートマップを用いた可視化では，個体や次元を縦軸に沿って並べる順番によって，その効果が大きく変わる．この順番を決定するための一手段として時系列高次元データを効果的に可視化する手法が発表されている[2,3]．しかしこれらの手法では，k-means 法を単純に適用して個体や次元をクラスタリングしているため，類似性は低いが関係性の高い次元を視覚的に比較することが容易ではない．

3. 次元間距離にもとづくヒートマップ表示

前章で論じたヒートマップの問題点を解決するために，次元間の類似性に着目するのではなく，任意の次元間距離に基づいて生成された距離行列を用いて次元をクラスタリングすることを考える．

3.1 次元間距離

我々は正の相関だけでなく負の相関を有する次元間を観察したいと考え，相関係数にもとづいて次元間距離を算出してクラスタリングに用いることにした．現時点の我々の実装ではピアソン相関係数を適用して i 番目と j 番目の次元間の相関係数 $d(i,j)$ を算出する．この値を式(1)に適用することで，i 番目と j 番目の次元間の距離 $L_{i,j}$ を得る．

$$L_{i,j} = 1 - |d(i,j)| \quad (1)$$

以上の処理を全ての 2 次元ペアに対して適用することで，距離行列を得る．

続いてこの距離行列に，以下の 2 種類のクラスタリング手法のいずれかを用いて次元にクラスタリングを施し，ヒートマップでの可視化に用いる．

3.2 階層型クラスタリング

本研究では最短距離法を採用した階層型クラスタリングによって，ヒートマップで可視化した．可視化結果の例を図 1 に示す．同一クラスタに属する次元群は間隔を空けずにひと続きに並べて表示し，異なるクラスタに属する次元間には間隔をあけて表示している．またこの可視化結果では，1 個の次元だけで構成されるクラスタは表示しない，という制約を設けている．

3.3 非階層型クラスタリング

本研究では k-medoids 法を採用している．k-medoids 法は k-means 法から派生したクラスタリング手法であり，距離行列を参照して個体を分類できる点が特徴である．可視化結果の例を図 2 に示す．この可視化結果でも図 1 と同様に，同一クラスタに属する次元群は間隔を空けずにひと続きに並べて表示し，異なるクラスタに属する次元間には間隔をあけて表示している．

4. クラスタリング手法の違いによる可視化手法の操作性の比較

本章では，2 種類のクラスタリング手法における，クラスタ数決定までの過程，計算の実行回数，そして表示される次元数を比較した．

図 1,2 は全て売り上げに関する同一の高次元データの可視化結果例であり，横軸は日照時間である．色付けされ水平方向に伸びた一本の帯がひとつの次元を表し，縦軸には次元を表す帯が順に並べられている．青色に近いほど小さい値を表し，赤色に近いほど大きい値を表す．

非階層型クラスタリングではクラスタ数である整数値を操作するのみであるのに対して，階層型クラスタリングではユーザが実数値でクラスタ併合のための閾値を設定するため，同一のクラスタ数でも異なったクラスタリング結果を得ることが容易である．また現時点の我々の実装では，1 個で 1 クラスタとなっている次元を画面表示から割愛することで，いずれかの次元と相関を有する次元のみを可視化し，画面上の帯の数を適正化して視認性を維持している．この処理においても階層型クラスタリングのほうが有利である．

処理時間について考える．階層型クラスタリングの処理は一般的に「デンドログラム構築＋閾値設定に対応したクラスタ形成」の 2 ステップで構成される．現時点で我々が使用しているデータにおいて，処理時間の大半はデンドログラム構築が占めるが，これは前処理として 1 回だけ実施すればよい．結果として，閾値の対話操作設定にともなうクラスタリングの再実行に要する処理時間はとても小さい．一方で非階層型クラスタリングでは，クラスタ数の対話操作設定のたびに最初からクラスタリング処理をやり直さないといけないため，現状の実装では対話操作には不利である．

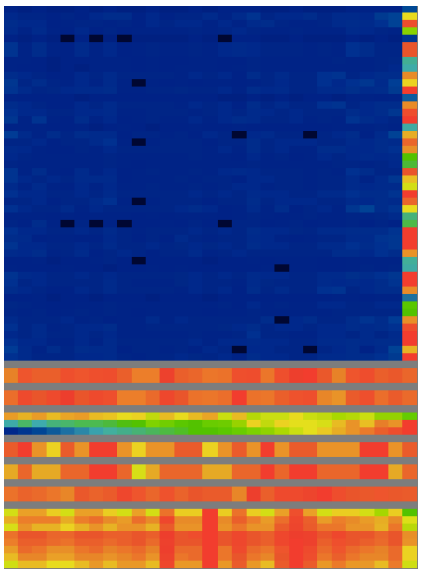

図 1　階層型クラスタリング実行結果例
(クラスタ数=8)

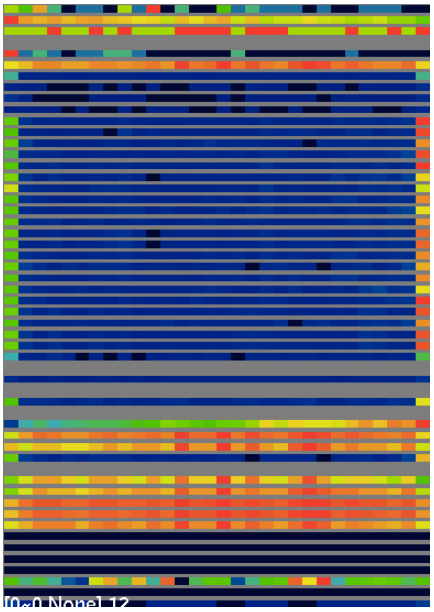

図 2　非階層型クラスタリング実行結果例
(クラスタ数=6)

5. まとめと今後の課題

本報告では，高次元データを構成する次元間の関係を観察するためにヒートマップを用いた可視化手法を紹介した．また最短距離法にもとづく階層型クラスタリングと，k-medoids 法にもとづく非階層型クラスタリングの 2 種類を実装し，その比較について論じた．

今後の課題として，クラスタリング手法の比較評価を網羅的に継続し，その結果をもとにクラスタリング手法の改善，およびヒートマップによる可視化手法の機能拡張をはかりたい．

6. 謝辞

本研究を進めるにあたり，ご指導とデータ提供を頂きました日本電気株式会社様に深く感謝致します．

参　考　文　献

[1] S. Yagi, Y. Uchida, T. Itoh, A Polyline-Based Visualization Technique for Tagged Time-Varying Data, 16th International Conference on Information Visualisation (IV2012), pp. 106-111, 2012.

[2] A. Hayashi, T. Itoh, S. Nakamura, A Visual Analytics Tool for System Logs Adopting Variable Recommendation and Feature-Based Filtering, 17th International Conference on Information Visualisation (IV2013), pp. 1-10, 2013.

[3] H. Suematsu, S. Yagi, T. Itoh, Y. Motohashi, K. Aoki, S. Morinaga, A Heatmap-Based Time-Varying Multi-Variate Data Visualization Unifying Numeric and Categorical Variables, 18th International Conference on Information Visualisation (IV2014), pp. 84-87, 2014.