散布図選択による多次元データ可視化における表示順の改良

理学専攻　情報科学コース　2240670　松尾　深景　（指導教員：伊藤貴之）

1. はじめに

データ分析は近年，企業や研究機関において非常に重要な役割を果たしている．多次元・多変数のデータを分析する際には，これらのデータを効果的に可視化することが重要である．

このような課題に対処するために，多次元データ分析のための効果的なデータ可視化手法が提案されている[1, 2, 3]．また，多次元データ分析のための有効な手法として，次元削減技術が活用されている[4, 5]．それに対して，散布図行列に代表されるように，任意の 2 変数を選択して生成される多様な散布図を表示し，任意の変数間の相関を網羅的に表現する可視化手法も知られている．この手法の応用として，本報告では任意の 2 変数から生成される散布図のうち特定の条件を満たす散布図を抽出して表示する可視化手法の改良を提案する．

多次元データ可視化のための散布図選択手法は多数発表されているが，その中でも伊藤らの先行研究[1]では，多様な指標を同時に参照することで有限個の多様な散布図を選ぶ手法を提案している．我々はこの手法を改良し，類似度の高い散布図をユーザが容易に比較できるような手法を実装している．伊藤らの先行研究では，各散布図に対して特徴ベクトルを生成し，コサイン類似度を用いて散布図間の類似度を算出する．そしてグラフ彩色問題を適用することで，多様な散布図を選出する．本研究ではこの手法を改良し，類似度の高い散布図をユーザが容易に比較できるような手法を実装した．具体的には，先行研究で定義された散布図間の類似度を参照し，散布図間の類似度から算出された距離行列を用いて階層クラスタリングを適用し，その結果を利用して散布図を配置する手法を採用している．階層クラスタリングには最短距離法，重心法，ウォード法の 3 つの計算手法を実装し，それらの結果を比較した．さらに，階層クラスタリング結果から算出された散布図のクラスタ群の画面配置手法として，行列形式に表示する手法と階層構造を視覚的に把握できる平安京ビュー[6]を採用する．また，その表示結果にもとづくユーザテストの評価結果を示す．

2. 散布図の表示順の改良

多次元データの任意の 2 変数を用いて生成される散布図には，それぞれの特徴がある．これを特徴量ベクトルとして数値化することで，抽象度の高い形で散布図の類似度を比較することができる．本報告では散布図の特徴量ベクトルを算出するために，先行研究[1]での散布図から相関・細さ・クラスの分離・クラスタと例外点の分離性の 4 つの数値的な特徴を抽出し，それをベクトルとして表現する手法を使用している．

特徴量ベクトルから散布図間の類似性を評価するために，提案手法ではコサイン類似度を算出する．コサイン類似度は，2 つのベクトルの間の角度の余弦を計算することで，類似度を測る指標である．この指標により，任意の 2 つの散布図について，その特徴の類似度を算出できる．

本手法では，類似度の高い散布図を近くに配置するために，散布図群の特徴量ベクトルに対して階層型クラスタリングを適用する．この処理の過程で，任意の散布図間の距離を計算し，その結果として得られる距離行列を参照することで，類似度の高い散布図が隣接するような樹形図（デンドログラム）を生成する．この樹形図における整列順に散布図を並べることで，類似度の高い散布図を近くに配置する．階層型クラスタリングの距離算出に最短距離法，重心法，Ward 法を採用した．以上の処理により，類似度の高い散布図が画面上で隣接するため，それらを効果的に比較することができる．そして，その類似した散布図を，行列形式で，または階層構造のデータを表現する平安京ビューを採用して表示した．

3. 実行例

3.1 使用したデータ

本研究では，伊藤ら [1] の先行研究でも用いた小売店の気象と売上の関係の データセットを実験に使用した．この事例では，2016 年 5 月 1 日から 2017 年 7 月 31 日 までの 457 日間のアパレルの小売店における各日の来客数や売上の 7 つの変数と，その各日の気象値 5 つの変数との関係のデータを題材にした．

3.2 実行結果

まず，階層クラスタリングの性能評価として，本研究では Python のライブラリである scikit-learn のシルエット係数を使用し，3 つの計算手法それぞれで 2 から 10 個のクラスタの場合を算出した(図 1)．その結果，重心法で 3 個のクラスタに分割した際にシルエット係数が最大となった．また，重心法はどのクラスタ数においても安定的に 0.5 以上の数値となっていることが確認できる．Ward 法においては，クラスタの分割数が増えるにつれて性能が向

上することが示唆されている．特に，10 個のクラスタで分割した場合にシルエット数が最も高くなった．一方で，最短距離法は分割数が増えても，1 つのクラスタに所属する要素の数が他の 2 手法に比べて多いため，シルエット係数は低い傾向が見られた．

また，実行結果として，重心法での行列形式での表示と重心法での 3 種類での平安京ビューでの表示結果を図 2，3 に示す．

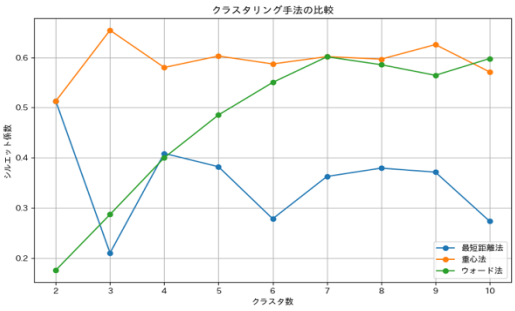

図 1 シルエット係数を用いたクラスタリング計算手法の性能比較

図 2　重心法で行列形式に表示した結果

4. 評価実験結果

ユーザテストは，情報科学を専攻する 20 代女性の被験者 53 名に対して行った．テストの目的は，各クラスタリング手法のユーザビリティと実用性を評価することであり，表示結果をもとにしたデータの理解しやすさに焦点を当てた．

可視化価値のある散布図を選択しやすい表示結果を確認するため，被験者にどの表示結果が好ましいか尋ねた．その中で，平安京ビューかつ重心法で表示した結果と平安京ビューかつ Ward 法で表示した結果を選択したのがそれぞれ 23 人だった．表示結果は似通っているため，これら 2 つの表示結果を選択する人で分かれた．これら 2 つの表示結果を選んだ理由として，「似ている散布図がグループで分かれていて，一目で比較しやすい」という意見が多く得られた．

それに対して，重心法かつ行列形式での行ごとに交互に方向を変える表示を選択した 2 人からは，「見やすく特定の枠がないことで固定概念を無くして考えられるから」という意見が得られた．5 段階評価においても，行列形式よりも平安京ビューの方が基本的に良い評価が得られたが，散布図の特徴を発見したいと考える時は，平安京ビューより行列形式で表示する方が良い場合があることが見られた．

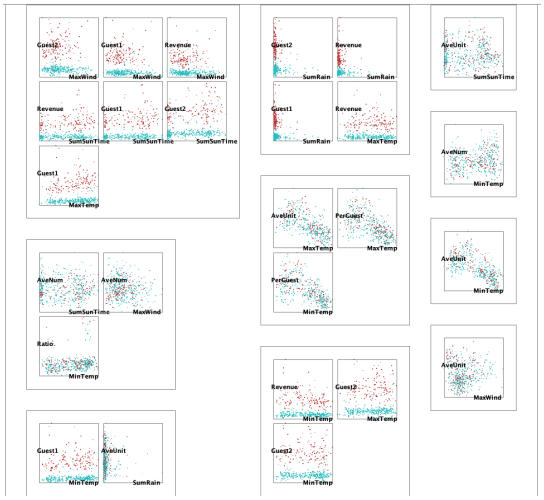

図 3　重心法で平安京ビューの形に表示した結果

5. まとめ

我々は，散布図同士の類似度を利用して，階層クラスタリングと表示順序を工夫し，表示結果の最適化を試みた．これにより，類似した散布図が近くに配置され，散布図同士の比較が容易になり，選択のしやすさが向上した．

しかし，現代のデータは多岐にわたり，異なるデータごとに異なる数値分布を示している．提案手法の柔軟性を確認するためには，本研究とは異なるデータセットやデータタイプにも適用できるかを検証する必要がある．

参考文献

[1] T. Itoh, A. Nakabayashi, M. Hagita, Multidimensional data visualization applying a variety-oriented scatterplot selection technique, Journal of Visualization, 26(1):199-201, 2023.

[2] Q. V. Nguyen, N. Miller, D. Arness, et al. Evaluation on interactive visualization data with scatterplots, Visual Informatics, 2020.

[3] G. Albuquerque, M. Eisemann, D. J. Lehmann, H. Theisel, W. Magnorm. Quality-based visualization matrices, Vision, Modeling and Visualization (VMV), 2009.

[4] J. H. Lee, K. T. McDonell, A. Zelenyuk. D. Imre, K. Muller. A structure-based distance metric for high-dimensional space exploration with multidimensional scaling. IEEE Transaction on Computer Graphics. 20(3):351-364. 2013.

[5] M. Ali, M. W. Jones, Dimension reduction applied to temporal data for visual analytics, The Visual Computer, 35:1013-1026, 2019.

[6] 伊藤，山口，小山田，長方形の入れ子構造による階層型データ可視化手法の計算時間および画面占有面積の改善，可視化情報学会論文集，26，6，51-61，2006.