

# 色付き Mosaic Matrix による高次元データの可視化

小林 弘明    三末 和男    田中 二郎\*

**概要.** 画面領域の pixel 数には限界があるため、各可視化手法で概観表示できる次元数には限界がある。よって次元の数が非常に多い高次元データの概観や、その特徴や傾向を把握することは難しい問題である。そこで限られた描画領域で高次元データの概観を閲覧するため、本研究では Mosaic Matrix をカテゴリ別に色付けする手法を提案する。また量的データのカテゴリ分割手法を工夫することで、狭い描画領域内で効果的にデータを表示することを目指す。本研究では提案手法と既存手法である Scatterplot Matrix とを比較評価することで、提案手法の有効性と表現の限界について考察する。

## 1 はじめに

多次元データの代表的な可視化手法としては、Parallel Coordinates Plot[1] や Scatterplot Matrix[2] などが挙げられる。しかし画面領域の pixel 数は限られているため、各可視化手法で概観表示できる次元数には限界がある。その限界を超えた次元数のデータは概観の表示ができず、Shneiderman が提唱した Visual Information-Seeking Mantra[3] に沿ったデータ分析ができない。よって次元の数が非常に多い高次元データの概観や、その特徴や傾向を把握することは難しい問題である。

本研究の目的は、30 次元以上の高次元データを対象に、その概観を一般的な大きさの画面領域内で閲覧可能にすることである。そのための視覚的表現手法として、Mosaic Matrix[4, Chapter4] を拡張し、カテゴリ別に色付けする手法を提案する。配色手法やカテゴリ分割手法によって可視化結果が変化するため、本研究では複数の配色や分割手法に対して評価を行う。さらに提案手法と既存手法である Scatterplot Matrix とを比較評価することで、提案手法の有効性について考察する。

## 2 色付き Mosaic Matrix

### 2.1 Mosaic Plot と Mosaic Matrix

Mosaic Plot は 100% 積み上げ棒グラフを拡張した、空間充填型のカテゴリデータ可視化手法である。Mosaic Plot の各矩形の幅と高さを、各軸に割り当てた次元のカテゴリ比率によって決定する。これによりデータ量が多いカテゴリに該当する矩形ほど、その面積は大きくなる。Mosaic Plot では複数のレコードを 1 個の矩形として表現することが可能であるため、狭い画面領域での可視化に向いている。

Copyright is held by the author(s).

\* Hiroaki Kobayashi, 筑波大学 情報学群 情報科学類, Kazuo Misue and Jiro Tanaka, 筑波大学 システム情報系

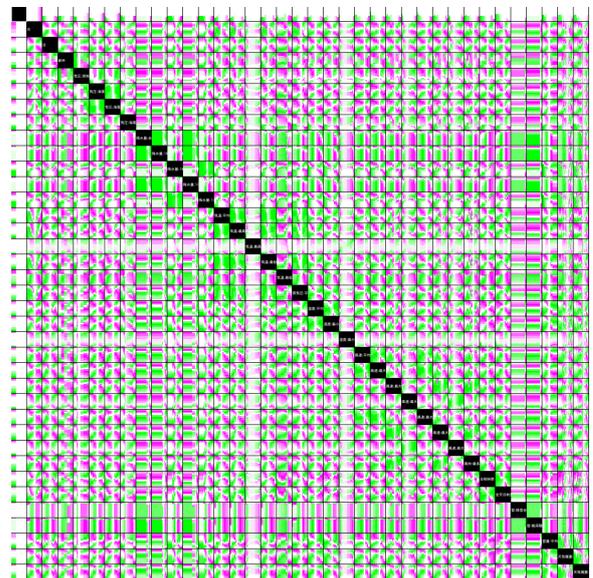


図 1. 色付き Mosaic Matrix の描画例

Mosaic Matrix は多次元データを可視化するための Mosaic Plot の拡張である。Mosaic Plot の X, Y 軸に 2 次元を割り当てて、全次元の組み合わせに対する Mosaic Plot を作成して行列状に配置する。

### 2.2 カテゴリ毎の色付け

Mosaic Matrix は原理的には何次元でも表現できるが、実際には描画領域に応じた限界がある。よって、数十次元のデータを行列状に配置すると可読性が低下してしまう。そこで高次元 Matrix 内での可読性向上のため、Mosaic Plot の色付けを工夫する。

提案手法では、Mosaic Plot の各軸のカテゴリ毎に色付け規則を設定し、それに従って矩形の色を決定する。このとき、データの偏りや次元間の相関の強さを色で把握できるような色付けを行う。例えば図 1, 2 は、左下と右上が緑、左上と右下が紫、その他の部分が白となるように色付け規則を設定した。

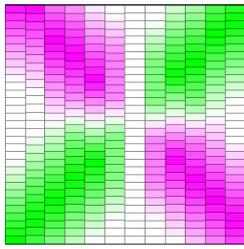


図 2. 各対角線上の矩形が共通の色となる色付き Mosaic Plot

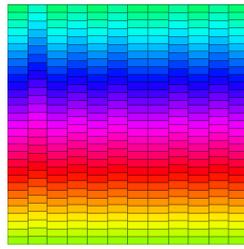


図 3. Y 軸カテゴリに色相を割り当てた色付き Mosaic Plot

## 2.3 カテゴリ分割手法

量的データを Mosaic Plot で可視化する場合、先にデータをカテゴリデータに変換する必要がある。その際のカテゴリ分割手法や分割数によっては、データの可読性などにも影響する。よって本研究ではカテゴリ分割手法として、最大値と最小値の間を等幅に分割する手法と、平均値をカテゴリの中心として標準偏差で分割幅を決定する手法を採用した。また各分割手法において、複数の分割数を採用した。

## 3 比較評価

気象庁ホームページ内 [5] から 5 都市の 1 年間の気象データを取得し、色付き Mosaic Matrix 及び Scatterplot Matrix を用いて可視化した。データの次元数は 37 次元で、レコード数は 1830 である。図 1 は提案手法による可視化例であり、これを元に比較評価を行った。

### 3.1 色付けに関する評価

図 2 のような色付けにより、相関の強さと正負を色の割合で知ることができた。例えば次元同士が強い正の相関を示す Mosaic Plot は緑、強い負の相関を示す Mosaic Plot は紫が多く見えた。また相関が弱い Mosaic Plot では、緑と紫の両色が同程度現れた。一方でデータが偏った次元があると、弱い相関の Mosaic Plot が強い相関の色に見える場合もあった。

図 3 のように Y 軸に色相を割り当てて可視化することで、Y 軸に割り当てられた次元のデータの偏りを把握することができた。またこの色付け規則は、文字データに対しても効果的であった。しかし小さな描画領域に大量の色相を割り当てた場合、偏りの少ない次元では色の区別が難しくなってしまった。

### 3.2 カテゴリ分割に関する評価

まず最大、最小値による分割の利点としては、値の偏りに関して誤解が少なく、直感的にデータを読み取れる点が挙げられる。しかし外れ値が分割の幅に影響してしまい、カテゴリが偏ってしまう次元も確認された。一方で平均と標準偏差による分割は、

外れ値の影響を受けずに分割を設定可能であった。

カテゴリ分割数については、5 分割と 10 分割の 2 種類について比較した。行列状に表示して両手法を比較したところ、狭い描画領域では分割数が少ない方が色を認識しやすいことが分かった。一方で分割数が減るとカテゴリの幅が広がるため、データの詳細な分布が見えなくなってしまった。

### 3.3 Scatterplot Matrix との比較

Scatterplot はデータ分布を点の位置で表現するため、相関やデータの偏りを直感的に理解しやすい。しかし 1 点が 1 レコードを表現するため、描画領域が狭くなると可読性が低下してしまう。

提案手法では色でデータの概観を表現するため、描画領域が狭くても可読性を維持することが可能である。また色付け規則を切り替えることで、相関やデータの偏りを見ることが出来る。一方で 3.1 節で述べたように、データの偏りによっては誤読を招く色表現になる場合もあった。しかしこれはカテゴリ分割手法を改善することで、ある程度解消できる問題であると考えられる。

## 4 まとめ

本研究では、データの偏りや次元間の相関といったデータの特徴を色で把握することが可能な、色付き Mosaic Matrix を提案した。本手法により、既存手法では難しかった高次元データの概観が閲覧できるようになった。一方で Scatterplot Matrix と比べて読解が難しい部分もあるため、今後は直感的により理解しやすい視覚的表現となるよう工夫する。

## 謝辞

本成果の一部は、株式会社富士通研究所からの受託研究によるものである。

## 参考文献

- [1] A. Inselberg and B. Dimsdale. Parallel Coordinates: A Tool for Visualizing Multi-dimensional Geometry. In *Proceedings of IEEE Visualization*, pp.361-378, 1990.
- [2] D. B. Carr, R. J. Littlefield, W. L. Nicholson and J. S. Littlefield. Scatterplot Matrix Techniques for Large N. In *Journal of the American Statistical Association*, Vol.82, No.398, pp.424-436, 1987.
- [3] B. Shneiderman. The eyes have it: A task by data-type taxonomy for information visualizations. In *Proceedings of the Symposium on Visual Languages*, pp.336-343, 1996.
- [4] M. Friendly. *Visualizing Categorical Data*. SAS publishing, 2001.
- [5] 気象庁 | 過去の気象データ検索. <http://www.data.jma.go.jp/obd/stats/etrn/index.php>.