テンソル SOM による関係データの可視化

Relational data visualization by Tensor SOM

◯[11] 岩崎　亘,　　[1] 古川　徹生

◯[1]Tohru Iwasaki,　[1]Tetsuo Furukawa

[1] 九州工業大学

[1]Kyushu Institute of Technology

Abstract: Tensor SOM organizes multiple topographic maps from relational dataset. By using Tensor SOM, it is easy to visualize and analyse higher-order relational data. In this presentation, the algorithm of Tensor SOM is described, and then some visualization techniques are introduced.

1　はじめに

SOM は高次元データを低次元のマップとして可視化するアルゴリズムである [1]．通常の SOM では 1 個のデータセットから 1 個のマップを生成するが，場合によっては複数のマップを同時に生成したい場合がある．典型的な例がマーケティング調査におけるユーザー・商品評価データの解析である．この場合はユーザーに関する情報と商品に関する情報の両方をマップとして可視化したい．さらに 2 つのマップの関係性まで可視化できれば，いっそう多くの情報をデータから引き出すことができる．本研究の目的は，複数のマップを生成することでデータの多面的解析を可能にする SOM の開発である．

ユーザー・商品評価データの場合，データはクロス集計テーブル，すなわち行列として表される [2]．このようなデータは関係データと呼ばれる．また解析対象であるユーザー，商品をそれぞれ「モード」と呼ぶ．一般にモード数が M の関係データは M 次の多次元配列，すなわちテンソルとして表現される．したがって本研究の目的は SOM をテンソルデータへ拡張することに他ならない．そこで以下，本拡張をテンソル SOM (Tensor SOM: TSOM) と呼ぶこととする．

TSOM の応用先はマーケティングデータ以外にも多く考えられる．たとえば電子メールのデータは「発信者×受信者×使用単語」という　3 次関係データとして表される [3]．同様に Twitter や Facebook など SNS のコミュニティ分析では，ユーザー同士のフォロー関係を関係データとして表現できる [4]．また顔画像データは「人物×表情×カメラアングル」のようにテンソルデータとして表せる [5]．このように関係データやテンソルデータはきわめて普遍的に存在するため，TSOM は SOM の応用範囲が大きく広げる．

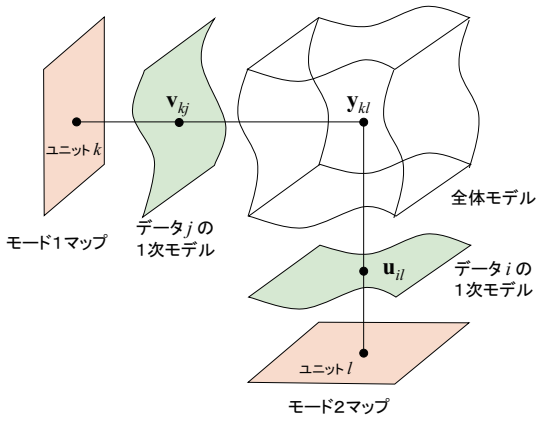


図 1: TSOM のアーキテクチャ．2 次の TSOM はマップを 2 組持つ．参照ベクトルはユニットの対に対して 1 個が割り当てられる．参照ベクトルの並びは積多様体となる（図では 3 次元に描かれているが，実際は 4 次元の積多様体になる）．

本稿では以下，TSOM のアーキテクチャとアルゴリズムを述べ，人工データと実データの分析結果を示すとともに，TSOM の多様な可視化法を紹介する．

2　TSOM のアーキテクチャとアルゴリズム

TSOM の説明に先立って，まず TSOM が扱うデータについて述べる．説明をわかりやすくするため，本稿ではユーザー・商品評価データを引き合いに説明する．この場合，ユーザーが第 1 モード，商品が第 2 モードの 2 次関係データとなる．なお TSOM 自体は任意の次数に拡張が可能である．

マーケティング調査において，各ユーザーは各商品に関して D 項目の評価を行ったとする．このときユー

ザー i の商品 j に対する評価は D 次元ベクトル $\mathbf{x}_{ij} \in \mathbb{R}^D$ となり，データセット全体は 3 次テンソル $\underline{\mathbf{X}} = (x_{ijd})$ で表される．通常の SOM ではデータの次元 D がマップの次元より大きくなければならない．しかし TSOM においてその制約はなく，評価値がスカラー x_{ij} であってもかまわない（その場合データセットは 2 次テンソルになる）．

2.1 TSOM のアーキテクチャ

2 次の TSOM はマップを 2 個持つ．すなわちマップ空間上にユニットを等間隔に配置したものを 2 組持つ．以下，第 1 モード（ユーザー）マップのユニット番号を $k \in \{1, \ldots, K\}$，第 2 モード（商品）マップのユニット番号を $l \in \{1, \ldots, L\}$ で表す．

通常の SOM では各ユニットに 1 つの参照ベクトルが割り振られる．すなわちユニット k に対して参照ベクトル $\mathbf{y}_k$ が対応する．一方 TSOM ではユニットのペア (k, l) に対して 1 つの参照ベクトル $\mathbf{y}_{kl}$ が割り当てられる（図 1）．したがって参照ベクトルは $K \times L$ 個存在し，モデル全体は 3 次テンソル $\underline{\mathbf{Y}} = (y_{kld})$ となる．

TSOM ではこれらのほかに「1 次モデル」と呼ばれる SOM の集合が存在する．今，特定のユーザー i についてのデータ $\{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iJ}\}$ をデータセットとみなしてモデル化した参照ベクトルを $\mathbf{U}_{i:} = (\mathbf{u}_{i1}, \ldots, \mathbf{u}_{iL})$ とする．この $\mathbf{U}_{i:}$ は特定のユーザー i の視点から見た商品モデルと見ることができる．これがモード 1 に関する 1 次モデルである．同様に特定の商品 j についてのデータ $\{\mathbf{x}_{1j} \ldots, \mathbf{x}_{Ij}\}$ をデータセットみなして得られた 1 次モデルを $\mathbf{V}_{:j} = (\mathbf{v}_{1j}, \ldots, \mathbf{v}_{Kj})$ とする． これは商品 j のみに限定したユーザーモデルとみることができる．これがモード 2 に関する 1 次モデルである．なお 1 次モデルに対して $\underline{\mathbf{Y}} = (\mathbf{y}_{kl})$ を「全体モデル」と呼ぶこととする．

2.2 TSOM のアルゴリズム

TSOM のアルゴリズムも SOM と同様に (1) 勝者決定 (2) 近傍計算 (3) モデルの更新の 3 ステップを繰り返し行う．

勝者決定

モード 1（ユーザー）に関する勝者決定は次のように行う．まず全体モデルを K 個の商品モデルの集合 $\{\mathbf{Y}_{k:}\}$ とみなす．ここで $\mathbf{Y}_{k:} = (\mathbf{y}_{k1}, \ldots, \mathbf{y}_{kL})$ は「（モード 1 に関する）第 k スライス」と呼ぶ．そしてユーザー i の 1 次モデル $\mathbf{U}_{i:}$ にもっとも近いスライス $\mathbf{Y}_{k:}$ をユーザー i に対する「勝者」とする．すなわち $k_i^* = \arg\min_k \|\mathbf{Y}_{k:} - \mathbf{U}_i\|^2$ である（図 2(a)）．同様に商品 j の 1 次モデル $\mathbf{V}_{:j}$ に

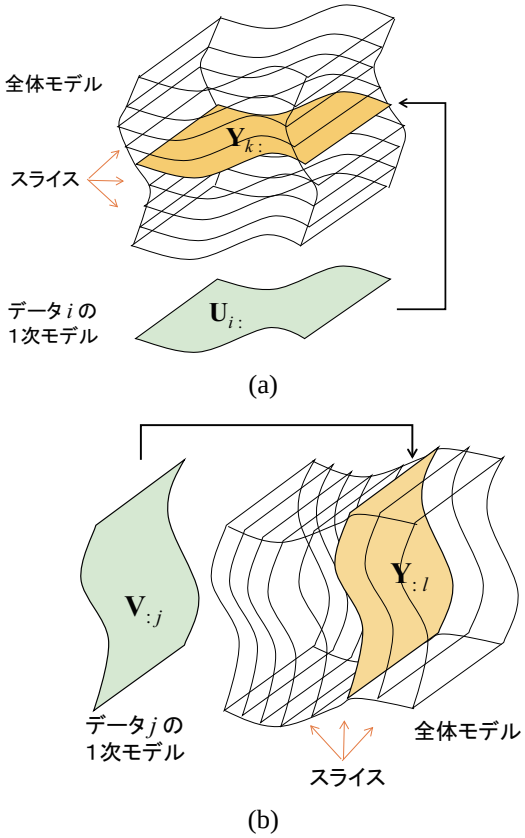


図 2: 勝者の決定法．各 1 次モデルに対して，全体モデルの中からもっとも適合するスライスを勝者とする．

もっとも合致する全体モデルのスライス $\mathbf{Y}_{:l}$ が商品 j に対する勝者 $l_j^* = \arg\min_l \left\|\mathbf{Y}_{:l} - \mathbf{V}_{:j}\right\|^2$ となる（図 2(b)）．

近傍計算

近傍については通常の SOM と同様に各モードごとのマップ上で定義される．第 1 モード（ユーザー）と第 2 モード（商品）に関する近傍は次のように与えられる．

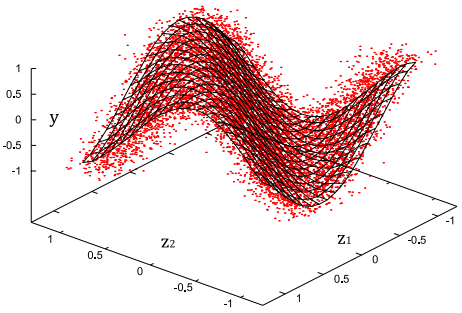
$$\alpha_{ki} = \exp\left[-\frac{1}{2\sigma^2} d^2(k_i^*, k)\right]$$
$$\beta_{lj} = \exp\left[-\frac{1}{2\sigma^2} d^2(l_j^*, l)\right]$$

とする．また各ユニットに割り当てられた近傍量の総和を $A_k = \sum_i \alpha_{ki}$, $B_l = \sum_j \beta_{lj}$ とする．

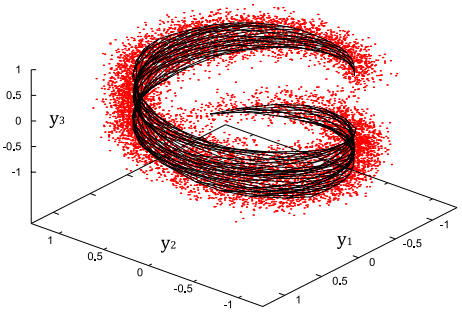
モデル更新

最後に全体モデルと各 1 次モデルを更新する．まず全体モデルは

$$\mathbf{y}_{kl} = \frac{1}{A_k B_l} \sum_{i=1}^{I} \sum_{j=1}^{J} \alpha_{ki} \beta_{lj} \mathbf{x}_{ij}$$



(a)



(b)

図 3: 人工データによる学習結果

として更新する．また各 1 次モデルについては

$$\mathbf{u}_{il} = \frac{1}{B_l} \sum_{j=1}^{J} \beta_{lj} \mathbf{x}_{ij}$$

$$\mathbf{v}_{kj} = \frac{1}{A_k} \sum_{i=1}^{I} \alpha_{ki} \mathbf{x}_{ij}$$

とする．以上を 1 ループとして，近傍半径を縮小しながら繰り返す．

3　実験結果

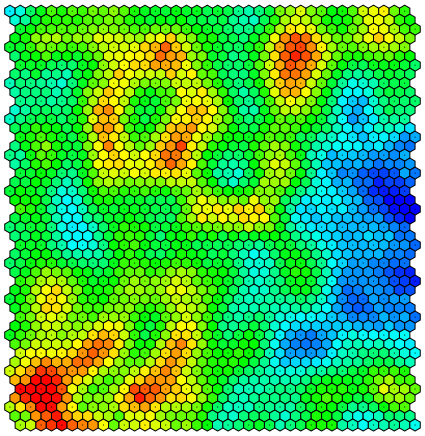
3.1　人工データによる学習結果

図 3 は人工的に生成したデータを用いて学習した結果である．2 つのモードのマップ空間はそれぞれ 1 次元とした．その結果，期待どおりの学習結果が得られた．

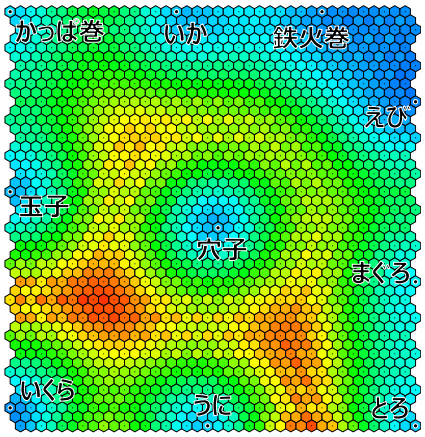
3.2　寿司データによる学習結果

実データとして寿司の嗜好調査データ [1] を用いた．このデータは 5,000 人の被験者に対し 10 種の寿司を好みの順位を回答してもらったものである [6]．ここでは嗜好の逆順を評価値として用いた．得られたマップを図 4 に示す．回答者マップ (a) は 5,000 人の回答者のマップである．後述する可視化法で調べたところ，類似した嗜好を持つ回答者がマップ上で近くに配置されており，期待通りの結果が得られた．一方，寿司マップ (b) ではトロとマグロ，軍艦巻きのイクラとウニが近くに配置されており，寿司の嗜好傾向を反映したマップが得られた．

[1] http://www.kamishima.net/sushi/



(a)



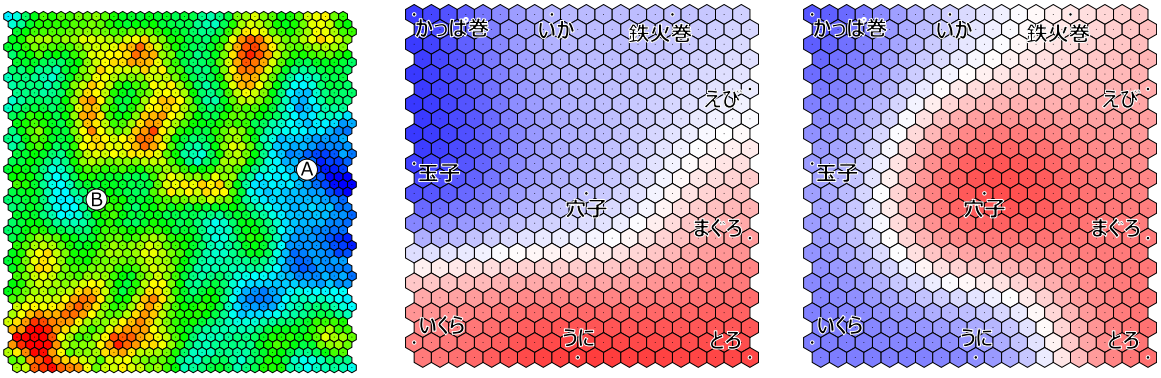
(b)

図 4: 寿司データによる学習結果．(a) ユーザーマップ (b) 回答者マップ．

4　多様な可視化法

SOM は単にマップを生成するだけでなく，多様な可視化法で多くの情報を引き出すことができる．代表的な可視化法のひとつは U-matrix 法であり，マップ上にクラスタ境界を明示することができる [7]．もうひとつの代表的な可視化法は component plane である [8]．これは参照ベクトルの特定の成分に着目し，その成分の大きさをグレースケール等で表現したものである．

TSOM でもこれらの手法を使うことができる．図 4 では U-matrix 法を用いており，軍艦巻き等のクラスタが見てとれる．さらに TSOM では conditional component plane という表示法を用いることで，2 つのマップ間の



(a) (b) (c)

図 5: Conditional component plane による寿司データの可視化．(a) 回答者マップ．回答者 A, B の位置を条件とした． (b) (c) 寿司マップ．カラースケールは回答者 A および B に対する評価値を表す．

関係を可視化することができる．たとえば第 1 モードのユニット k を条件として指定すると，第 2 モードのマップ上に y_{kld} の大きさをカラースケールで表示できる．これはユニット k という条件下での成分 d に関する component plane になる．図 5 は conditional component plane の例である．ここでは特定の回答者集団をマップ上で条件指定し，その回答者がどの寿司を好むかをカラースケールで表現したものである．条件指定を回答者マップの上で連続的に動かすことで，さまざまな回答者に対する寿司の嗜好を知ることができる．これとは逆に，寿司マップで条件指定すれば，特定の寿司に対する回答者の嗜好分布を可視化することができる．

5 おわりに

TSOM は単に高次元データを可視化するだけでなく，高次の構造を持つデータを複数マップとして可視化することができる．また多様な可視化法を駆使することにより，マップ間の関係性も見ることができる．昨今はビッグデータ解析などが注目されているが，データ規模の大きさではなくデータ構造の複雑さにも対応していくことが必要となる．そのような中で TSOM は強力な解析手法になるのではないかと考えている．

参考文献

[1] Kohonen, T.: *Self-Organizing Maps*, Springer-Verlag, Berlin Heidelberg (2001).

[2] Ricci, G., Gemmis, de M. and Semeraro, G.: Matrix and Tensor Factorization Techniques applied to Recommender Systems: a Survey, *International Journal of Computer and Information Technology*, Vol. 1, pp. 94–98 (2012).

[3] Berry, M. W. and Browne, M.: Email Surveillance Using Non-negative Matrix Factorization., *Computational & Mathematical Organization Theory*, Vol. 11, No. 3, pp. 249–264 (2005).

[4] Acar, E., Çamtepe, S. A., Krishnamoorthy, M. S. and Yener, B.: Modeling and Multiway Analysis of Chatroom Tensors, in *Proceedings of the 2005 IEEE International Conference on Intelligence and Security Informatics*, ISI'05, pp. 256–268, Berlin, Heidelberg (2005), Springer-Verlag.

[5] Vasilescu, M. A. O. and Terzopoulos, D.: Multilinear Image Analysis for Facial Recognition, in *ICPR (2)*, pp. 511–514 (2002).

[6] Kamishima, T., Kazawa, H. and Akaho, S.: A survey and empirical comparison of object ranking methods, in Fürnkranz, J. and Hüllermeier, E. eds., *Preference Learning*, pp. 181–201, Springer (2010).

[7] Ultsch, A. and Siemon, H. P.: Kohonen's self organizing feature maps for exploratory data analysis., in *Proc. INNC'90, Int. Neural Network Conf.*, pp. 305–308 (1990).

[8] Stefanovic, P. and Kurasova, O.: Visual analysis of self-organzing maps, *Nonlinear Analysis: Modelling and Control*, Vol. 16, No. 4, pp. 488–504 (2011).

連絡先
古川　徹生
E-mail: furukawa@brain.kyutech.ac.jp