クラスター構造の経時変化を可視化するための Time-Arrayed SOM の提案

石川 雅弘 †
つくば国際大学 産業情報学科 ‡

1 はじめに

種々の観測・統計データ，新聞などのテキスト，画像データ，またウェブドキュメントなど，様々なデータが大量に蓄積され続けている．大量のデータを俯瞰するためにはよくクラスター分析が行なわれる．しかし数値的な分析を行なっただけでは人間がそれを把握し実際に利用するのは難しい．そのため，分析結果を可視化しより人間に把握しやすい形で提示することが求められる．SOM(Self-Organizing Map)[1] は，そのような目的に利用可能なクラスター構造可視化手法の一つである．しかし，多くのデータには生成時刻がタイムスタンプとして付随しており，データセットを時間軸に沿って分割すれば，クラスター構造は時間とともに変化していると考えられる．例えばブログの各エントリを一つのデータと考えると，すべてのエントリからなるブログ空間はエントリの類似性によるクラスター構造を持つと考えられるが，それは時間と共に変化していることだろう．このような場合，ある期間におけるクラスター構造だけではなく，その経時変化も可視化する事が求められる．本稿ではこのような目的のための可視化手法として，SOM を利用した Time-Arrayed SOM を提案する．

2 Self-Organizing Map (SOM)

一般に SOM と呼ばれるのは Kohonen が Basic SOM と呼ぶ逐次学習型のものであるが，本研究ではよりクラスタリング手法的性格の強い一括学習型 SOM の Batch Map[1] を用いる．Batch Map は k-means クラスタリングに位相を持ち込んだものと見做せ，後述する近傍半径を 0 とした場合には k-means クラスタリングに一致する．以下では SOM を Batch Map を指す語として用いる．

データセットを $D = \{d_0, d_1, \cdots, d_{n-1}\}$ $(d_i \in \mathcal{R}^m)$，その上で定義された距離関数を $\delta : \mathcal{R}^m \times \mathcal{R}^m \to \mathcal{R}$ とする．またサイズが $X \times Y$ の二次元六角格子 (マップ) 状に並べられたセルを $c_{x,y}$ とする．各セルには参照ベクトル $v_{x,y} \in \mathcal{R}^m$ が関係付けられている．SOM は，k-means クラスタリングと同様，各 d_i を最も距離の近い参照ベクトルを持つセルに割り当てる．この時データ空間中で距離の近いデータ同士はマップ上でも近いセルに割り当てられるように徐々に学習 (参照ベクトルを更新) すること

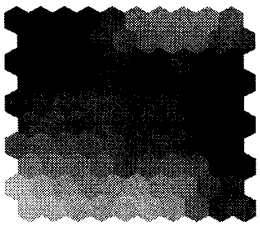

図 1 SOM による三次元データの二次元可視化の例

で，D のクラスター構造を二次元平面上に「再現」する．SOM の学習手順を示す:

1. 各セルの参照ベクトル $v_{x,y}$ を初期化する．
2. 各 d_i を $\delta(d_i, v_{x,y})$ が最小となるセル $c_{x,y}$ に割り当てる．
3. 各 $v_{x,y}$ を，$c_{x,y}$ と $c_{x,y}$ からのマップ上の距離が近傍半径 R 以内のセルに割り当てられたデータの平均値で更新する．
4. 収束するまで 2, 3 を繰り返す．ただし手順 3 の近傍半径 R は大きな値から徐々に減少させる．

学習後の各セルの参照ベクトルを何らかの方法で表示することで二次元可視化が実現される．図 1 は，$[0,1]^3$ 空間中にランダムに生成した 1000 個の点を学習させたサイズ 10×10 の SOM について，参照ベクトルの座標を R,G,B の強度と見做した RGB 色で表示したものである．類似色同士が固まって「クラスター」が可視化されている事が分かる．

3 Time-Arrayed SOM (TaSOM)

TaSOM では，さらに経時変化を可視化することが求められる．

3.1 満たすべき条件

データセット D をタイムスタンプに従って時間順に $D_0, D_1, \cdots, D_{T-1}$ に分割する．D_t から作成した SOM マップを S_t とすると，$\{S_t\}$ を時間順に並べることで経時変化を観察できる．しかし各マップを独立に作成した場合，複数の期間内に存在する同一のクラスターでもマップ上の位置は必ずしも同じにならないなどの問題がある．そのため，時間軸に沿った同一 (あるいは類似) クラスターを視認するのは難しい．時間軸方向での継続性や変化を可視化するためには，以下の条件を満足する必要がある．

条件 1: S_t と S_{t+1} 上に同一 (類似) クラスターが存在する場合，マップ上の位置もなるべく同じでなければならない．

条件 2: S_l と $S_m (l < m)$ 上に同一 (類似) クラスターが存在する場合，それが視認できなければならない．

Time-Arrayed SOM for visualizing cluster changes with time
†ISHIKAWA Masahiro (mi@tius.ac.jp)
‡Faculty of Industorial Informatics, Tsukuba International University

3.2 解決法

条件1を満たすために，$S_t(t>0)$の参照ベクトルを学習後のS_{t-1}の参照ベクトルで初期化する．これにより類似データはS_t上でもS_{t-1}上でも近い位置のセルに割り当てられることになる．

条件2を満たすためには，全マップの全セルを通して，類似した参照ベクトルを持つセルは類似色で彩色する．こうすることで，例え類似クラスターが時間的には離れたマップ上に存在する場合でもその類似性を視認できるようになる．

さらに，クラスターの移動，分裂，合併，消滅，出現，衰退，興隆などの様々な経時変化をできる限り自然に反映するため，またマップの縁にあるセルが斉一性を失なうボーダー効果を除去するために，マップは左右と上下がそれぞれ接続されているとみなしたトーラス型を採用する．

3.3 Coloring SOMによる彩色

類似性を反映した一貫した彩色を行なうため，学習後の全マップの全セルの参照ベクトルを一次元SOMでクラスタリングする．ここでも両端が接続されたリング型を用いる．この1次元環状SOMをColoring SOM(cSOM)と呼ぶ．cSOMのセルを$c_0, c_1, \cdots, c_{P-1}$，関係付けられた参照ベクトルを$v_0, v_1, \cdots, v_{p-1}$とする．学習後のcSOMの参照ベクトルから，下式に従い$R_i, G_i, B_i (0 \le i < p)$を求める．

$$\omega_i = \frac{2\pi \sum_{k=0}^{i-1} \delta(v_k, v_{k+1})}{\sum_{k=0}^{p-1} \delta(v_k, v_{(k+1) \bmod p})}$$

$$R_i = \frac{sin(\omega_i) + 1}{2}$$

$$G_i = \frac{sin(\omega_i + \frac{2\pi}{3}) + 1}{2}$$

$$B_i = \frac{sin(\omega_i + \frac{4\pi}{3}) + 1}{2}$$

各色の強度を0〜1としたRGB色$\langle R_i, G_i, B_i \rangle$をセル$c_i$に割り当てる．これはcSOMのリングを円と見なし，cSOMセル間の距離を考慮しつつ色相環から採色する事に対応する．従って，cSOMの隣接した二つのセルにはその距離に応じた類似色が割り当てられる事が保証される．

TaSOMの各マップ上の各セルは，それぞれの参照ベクトルが割り当てられたcSOMセルが持つ色で彩色する．

4 実装と実験

TaSOMの学習アルゴリズムとcSOMによる彩色アルゴリズム，およびその結果の3Dグラフィックスでの表示系をPython言語とVPython(Visual 5)モジュール[2]を使用して実装した．時間軸方向でのクラスターの存続や変化が可視化されることを確認するために，$[0,1]^3$空間中に生成したデータセットを用いた実験を行なった．図2はTaSOMによる三次元可視化の例である．ランダムに決定した点を中心に正規分布に従って生成した4つのクラスターからなる1000個のデータを一つのデータセットとし，各クラスターのデータ数を増減させながら作成した10世代分のデータセットを用いた．SOMマップのサイズは14×14であり，各セルは円柱で表示してある．cSOMのセル数は32，すなわち彩色には32色を用いた．図3は特定の色のセルのみを強調して表示した例である．時間と共にクラスターが成長していることが視認できる．マウス操作による回転・拡大なども可能である．

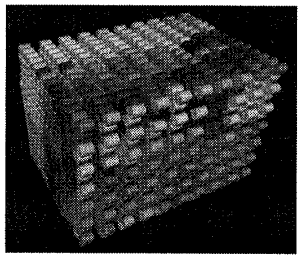

図2 Time-Arrayed SOMの3D表示例

図3 TaSOMによるクラスターの強調表示例

5 まとめと今後の展望

クラスター構造とその経時変化の可視化のためのTime-Arrayed SOMを提案した．また3Dグラフィックスの表示系を持つ実装を行ない，合成データによる実験でTaSOMの有効性を確認した．実験は限定的なデータセットで行なったため，今後はクラスター構造の様々な変化がどの程度可視化できるか，様々な合成データと実データによる確認を行ないたい．

また，各セルに割り当てられたデータ数や各色に属するデータ数とその経時変化など，可視化に利用できる情報は多い．それらの情報を取り込み，また補助的な表示系を追加し，さらに，インタラクティブな操作系を備えることでより探索的な利用ができるよう実装を進めている．

TaSOMは元々ブログや新聞データベースの可視化のために考案したものである．それらのデータへの適用も進めており，キーワード情報やリンク情報などと併わせた利用や検索支援への利用なども検討している．

参考文献

[1] Teuvo Kohonen. *Self-Organizing Maps, Third Edition*. Springer-Verlag, 2001.

[2] David Scherer and Bruce Sherwood. VPython: 3D programming for ordinary mortals. Website, 2010. `http://vpython.org/`.