クラスタリングに基づく情報の検索と視覚化

堀田, 政二

https://doi.org/10.15017/1398256

付 録D

自己組織化マップ

自己組織化マップ (Self-Organizing Map, SOM) は，高次元特徴ベクトルを利用してデータ分布を近似的に表現した地図（マップ）を作成する手法である [11]．SOM では低次元空間への写像の方法として教師なし競合学習を用いる．ここでは SOM のアルゴリズムを概説し，カラー画像を 2 次元平面上に配置した実験例を示す．

D.1 SOMのアルゴリズム

SOM では，一般に 1 次元あるいは 2 次元に配置したユニット群が用いられる．ここではユニットを 2 次元に並べた場合のアルゴリズムを示す．配置するデータは M 個あるとし，第 i データの n 次元特徴ベクトルを $\boldsymbol{x}_i = [x_1, ..., x_n]^T$ とする．準備として，$N \times N$ 個のユニットを図 D.1 のように格子状に並べる．k 行 l 列のユニットの持つ n 次元参照ベクトルを $\boldsymbol{m}_{kl} = [m_{kl1}, ..., m_{kln}]^T$ とする．

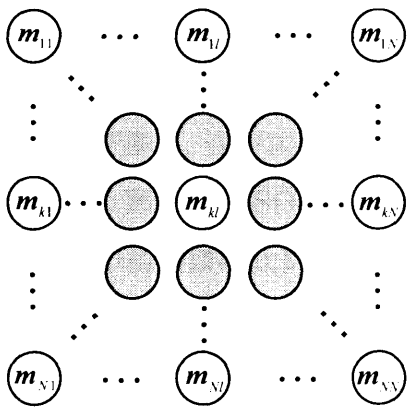

図 D.1: ユニット配列の概念図

アルゴリズム SOM:

Step 1: 各ユニットの参照ベクトル $\boldsymbol{m}_{kl}(k,l=1,...,N)$ を適当に設定する．学習係数の初期値 $0<\alpha<1$ を適当な値に設定する．

Step 2: 入力データ $\boldsymbol{x}_i(i=1,...,M)$ の順番をランダムに並び替えて，添字 i を付け直す．$\xi=1$ とする．

Step 3: 入力データ $\boldsymbol{x}_\xi$ について

$$\{k',l'\}=\arg\min_{k,l}=\|\boldsymbol{x}_\xi-\boldsymbol{m}_{kl}^{(\xi)}\| \tag{D.1}$$

を求める．ここで $\|\boldsymbol{x}_\xi-\boldsymbol{m}_{kl}^{(\xi)}\|$ は，$\boldsymbol{x}_\xi$ と $\boldsymbol{m}_{kl}^{(\xi)}$ とのユークリッドノルムである．

Step 4: ユニットの参照ベクトルを以下のように更新する:

$$\boldsymbol{m}_{kl}^{(\xi+1)}=\begin{cases}\boldsymbol{m}_{kl}^{(\xi)}+\alpha^{(\xi)}(\boldsymbol{x}_\xi-\boldsymbol{m}_{kl}^{(\xi)}) & (k,l\in S)\\ \boldsymbol{m}_{kl}^{(\xi)} & (k,l\notin S)\end{cases} \tag{D.2}$$

ここで S は k' 行 l' 列のユニットと，その 8 近傍（図 D.1 の灰色部分）のユニットからなる集合を表す．

Step 4: $\xi=M$ ならば Step 5 へ．そうでなければ $\xi=\xi+1$ とし，学習係数を $\alpha^{(\xi+1)}=\alpha^{(\xi)}(1-\xi/10000)$ と更新して Step 3 へ．

Step 5: すべてのユニットの参照ベクトルの変化が小さくなったら終了．そうでなければ Step 2 へ．

なお，上記のアルゴリズムの Step 3 の近傍の選択方法および Step 4 の学習係数の更新の方法はこの限りではない．詳しくは文献 [11. 48. 49] 等を参照されたい．

D.2 実験

B.2.2 節の画像データを SOM により 2 次元平面上に配置する実験を行った．すべての画像を 512 色のヒストグラムで表し，式 (B.5) により 2 次形式距離で標準化したものを画像の特徴ベクトルとした．図 D.1 に SOM による画像配置を示す．ユニットは 25×25 個用いた．学習係数 α の初期値は 0.1 とした．収束にかかった時間は約 5 時間であった．

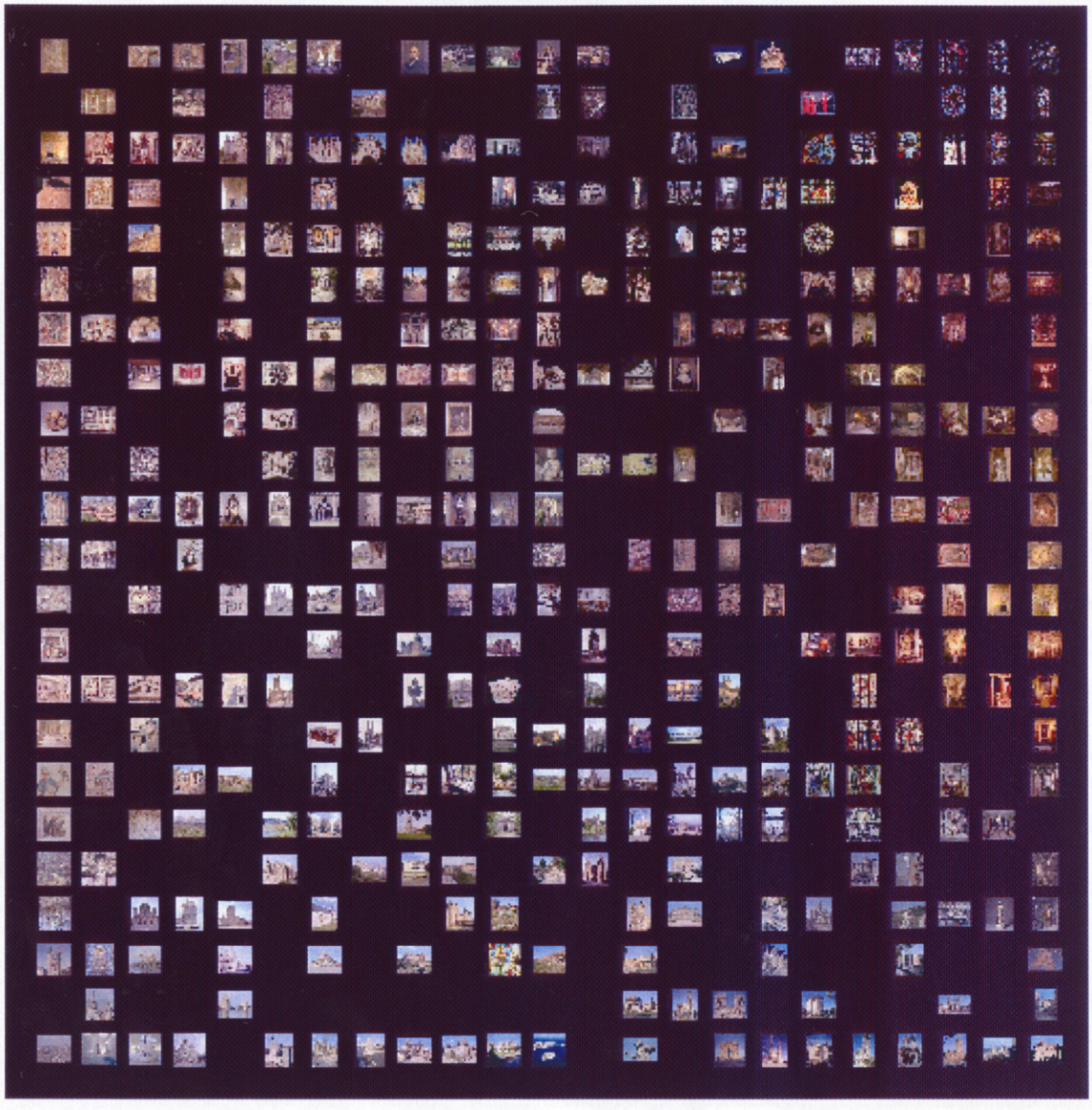

図 D.2: SOM による画像の配置

D.3 まとめ

SOM はデータが均等に配置されるので見やすいが，1) 予め用意する最適なユニット数が不明であること，2) データ数に対して十分多くのユニットを用意しないと，マップ上でのデータ分布のひずみが大きくなること，3) データ数，特徴ベクトルの次元が増加すると計算時間と使用するメモリの量が膨大になること，などの問題点がある．