パーセプトロンの変形たち：競合学習と自己組織化

視覚の特徴抽出系，記憶のシステム，脳のマップ構造
自己組織化：特別な指示なしに構造が発生すること．

教師あり学習と教師なし学習

パーセプトロンのような教師あり学習(分類すべき正解が与えられている)に対して、ただデータのみが与えられていて、それを学習していると自ずとその分類が見えてくる学習方式がある。このような場合は教師なし学習と呼ばれ、自己組織化といわれる自律的な構造形成過程の一つのバリエーションである。ただし、自己組織化といってもそこにはまったく仕掛けが無いわけではなく、自己組織化の学習機構には非明示的に分類のメカニズムが埋め込まれており、そのメカニズムとデータの分布の相互作用により分類が決定されているだけのことである。

競合学習とは

複数のパーセプトロンが同一のデータに対して反応し、その中で最も入力に強く反応したものがそのデータに対して学習するという、競争による学習の基本メカニズムである。英語では competitive learning, winner take all rule などと呼ばれている。

例　例えば、n 次元空間中に分布したデータを二つに分割することを考えよう。その場合，パーセプトロン類似の細胞を二つ用意し，入力データがその二つのどちらに「近い」かということを判定し，近いほうに所属するとして分類する．これが二つではなくもっと多い場合でも同様で,「最も近い(最近)」「最もそれらしい（最尤）」という判断基準で分類できる．

距離の定義
「最短」などのパターン間の類似性の判定基準としては、Euclid 距離、ベクトルの Cosine などが考えられる。

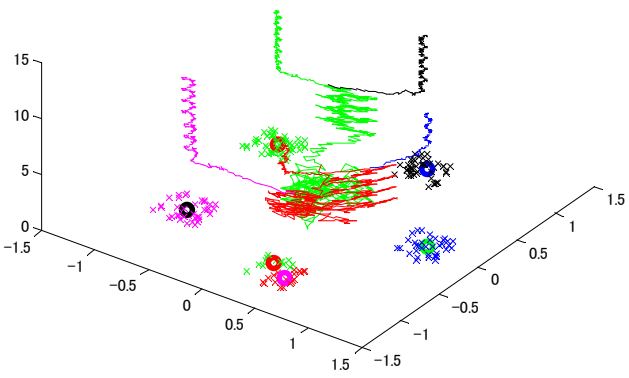
対策　学習細胞の初期値によって競合学習はそのままではうまくいかない場合もある。それに対しては、細胞のデータの分担を自立的に制御するメカニズムが必要となる。

1. 疲労学習：分割するクラスの数が最初から決まっている場合にはその数で制御
2. 細胞分割：学習が収束した時点で、細胞の分担を評価して、まずければ分割する

疲労学習アルゴリズム／細胞分割アルゴリズム

課題 2　配布データの自己組織的分類をやってみよう

細胞の初期状態は 1 個とする。各細胞によって分担されたデータの分散を求め、その分散がある程度小さくなるまで細胞の分割を繰り返す。サンプルのデータでは、クラスタが 5 個になったとたんに分散は小さくなる。



明治大学 理工学研究科 集中 2022　6-7 月集中　（大森隆司）

課題 2 に関して

プログラムの流れの例

1. 変数群係数 W[i][2]；学習データ X[n][2]; 学習係数 α；
2. 入力データセット X[n][2] (n=1…N) を用意する（ファイルから読み込み）
3. 細胞を 0 個(index I=0)用意し，それに係数 W[i][2] を用意する．
4. repeat　　% 細胞数を増やして
   (ア) 細胞を一つ増やす　I++
   (イ) 誤差の蓄積量 D[i]=0 for all i
   (ウ) repeat
      ① for j=1…N　　% すべての学習データに対して
         1. C = argmin_i (// X[j][*] – W[i][*] //)　　入力に最も近い細胞を発見し，
         2. d = // X[j][*] – W[C][*] //　　まずその近さを記録し，
         3. ΔW[C][*] = α (X[j][*] – W[C][*])　　その細胞が入力に近寄っていく
         4. d を変数 D に蓄積していく
            誤差の蓄積量 D の作り方はいろいろある．例えば，D を一定比率で減衰させながら d を足していくと，ある種の移動平均ができる．
            D[C] = 0.9 D[C] + 0.1 d
         5. W[*][*] , D[*] をログとしての残していく
      ② next j
      ③ for 文(1 ～ 5) をもう一回反復する
   (エ) until (全ての W の変化が収束)
5. until （D が十分小さい）
6. W の経過の二次元プロット
7. D の経過の学習回数に対するプロット

示してほしい図

A. 学習回数（時間）とともに W がどう変化していったか．
　　細胞の分割はグラフの上でも線を分岐させる
B. W の軌跡の二次元プロット W の初期値をいろいろ変えて
　　想定：初期値で分岐パターンは変わるが（本当か？）決まった場所に収束する

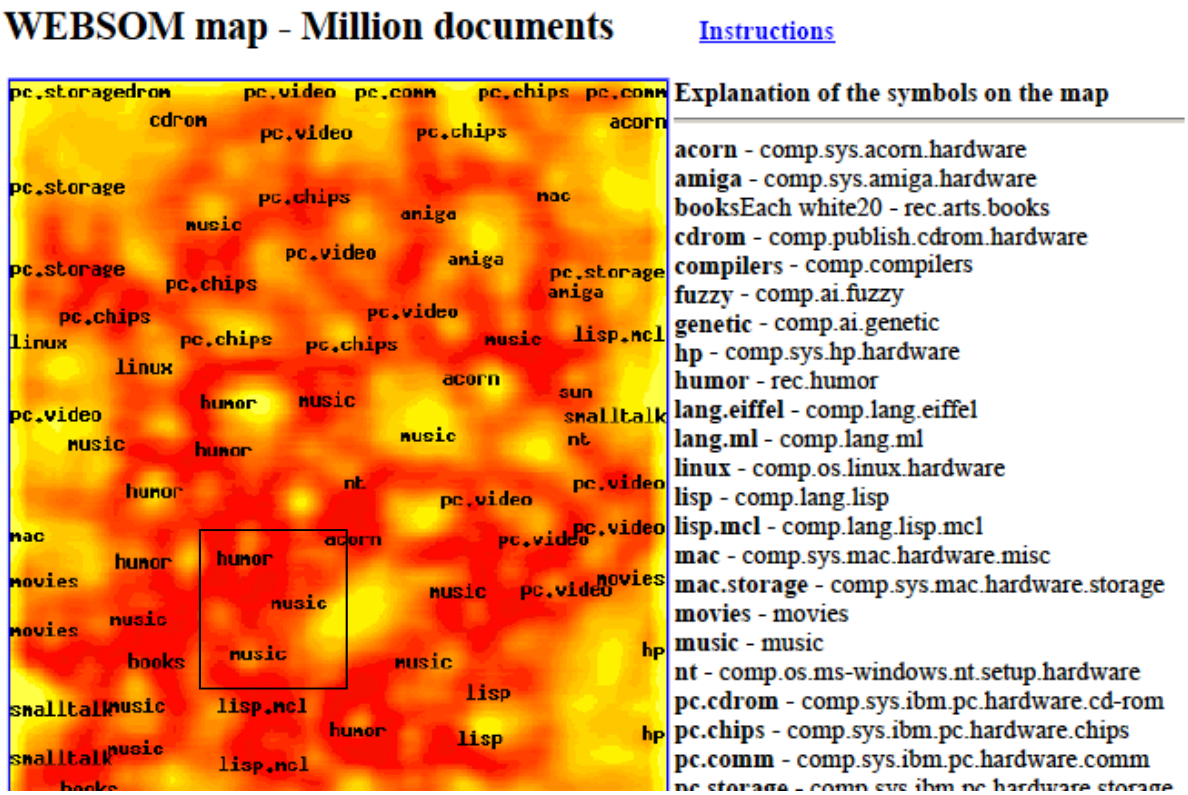
知識の構造　：　特徴マップ

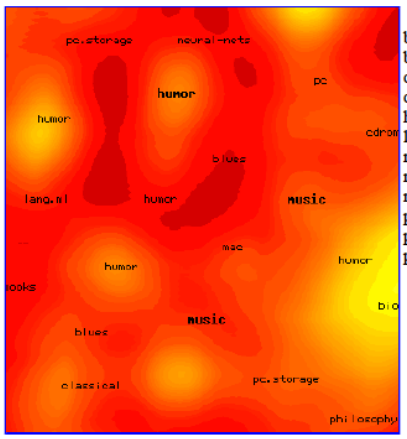
脳には多数の領野が存在し、それらが個々に異なる機能を持っているといわれている。しかし、具体的にそれらの領野がどのような関係で機能分担を行っているかという点については判らないことが多い。　→　領野，記憶，

細胞間の関係　→　自己組織化マップ

競合学習の問題は、学習して認識ができるようになったデータ群の相互関係が判らないことである。そこで、細胞間の相互関係が目に見える形で学習を進行させ、その結果としてデータ群のトポロジーを理解しようとする方法がある。基本的な方法は T.Kohonen によって提案された多次元のマップを形成するものであるため、Kohonen マップ、自己組織化マップ、などと呼ばれる。

応用例　WebSOM：　/http://websom.hut.fi/websom





自己組織化マップの構造と学習則

入力：N次元ベクトル $\mathbf{x}=\{0,1\}^N$。細胞：m次元のメッシュ上に配置された細胞 $C_{\mathbf{m}}$。
結合：各細胞は入力 $\mathbf{x}$ との間に結合 $\mathbf{w_m}$ をもつ。距離：$\|\mathbf{x}-\mathbf{w_m}\|$ が最小の $\mathbf{m}$ とその近傍の細胞 $\mathbf{n}$ が学習する。

$$\mathbf{W_n}=(1-\alpha)\mathbf{W_n}+\alpha\mathbf{x}$$

ここで m の近傍 n は例えば 1 次元なら 2 近傍，2 次元なら 4 近傍，8 近傍さらにはより遠くまでがありえる．

二次元マップの実例

三次元（一般には N 次元）のデータに二次元のマップを当てはめる。
マップの隣り合わせの細胞には、それぞれの細胞の特徴空間での位置が情報として保持されているため、その差分として各細胞の近辺でのマップの基底ベクトルを求めることができる。それから、新たなデータがどのくらいそのマップに属しているか、逆にどのくらいそのデータがマップの分布から外れているかを推定可能となる。その量から、マップ構造とデータ分布のマッチングのよさを議論することができる。

WebSOM の作り方

1. 対象とする Web テキストデータ群を集めてくる
2. Web テキストからキーワードを抽出する
   (ア) 全ての単語を切り出し，格変化などを取り去り，無意味な語を排除する．
   (イ) 各単語の発生頻度をカウントし，語彙発生ベクトルを作る
   (ウ) 必要に応じて，特徴を追加する　→　結果として各 Web を表す特徴ベクトル
3. 対象とする SOM の空間次元を決定する（普通は二次元）
4. SOM を構築する
5. SOM の各部分を表現する代表的なラベルを作る

生物の実体としての脳マップ

大脳皮質には、多数の領野（機能の異なる情報処理を行っていると考えられる部位）が存在し、その中には明確にマップ構造が見られるものが存在する。例えば、視覚系の初期の領野には、網膜上のトポロジーをそのまま表現するマップが存在する。

因果関係の部分性

実世界で観測されるデータ群の間には関係がある．特徴マップはそれを検出する関係抽出アルゴリズムである．関係の多くは因果関係によって結ばれている．因果関係には時間差のあるものが普通であるが，同一の原因によって発生したデータには同時性の因果関係が観測される（これは因果とはいわないかもしれないが,,,）．

問題は，観測されるデータ群（またはデータ源）のすべての間に因果関係があるとは限らないことである．さらに，同時に観測される因果関係は一つとは限らない．よく整理された学習課題ではこの問題は人間が解決して因果関係のある変数のみを抽出したデータを用意して学習問題とするが，現実世界の問題ではそのような因果関係はおかまいなしにデータが観測されてくる．その中から因果関係のある変数群を見つけなければならない．これは現在も未解決の課題である．

課題 3　特徴マップの学習

二次元ランダム分布データ $x=(x1,x2)$：$0<x1,x2<1$ に対して　4×4　程度のサイズの特徴マップの学習を進行させ，その学習過程でのマップの広がり方を観測する．
なお，学習回数を増やしてもマップはデータの分布に対して端まできれいには拡がらない．その理由を理論的に考察せよ．

マップの次元の推定

多くの事例では，マップの次元は人間が事前に与える．しかし現実にはデータの次元が事前にわかることは少なく，試行錯誤でマップの次元や細胞数を整える必要がある．そのような状況に対応した方式も存在する．

例えば，

・マップの各軸の曲がり具合を推定する（ex. 二次元のデータに 1 次元のマップ）

・マップの各細胞の受け持つ領域の広がりを推定し，そこから最適な細胞数を決める

・マップの次元で表現しきれない次元を推定し，そこから新たな次元を決める

等が考えられる．

マップの誤差の推定方法，

あるデータに特徴マップをフィットさせたとき,そのフィットの良さ（悪さ）を評価することで,そのマップの設定のよさを知ることができる．良さの指標としては，たとえば誤差があり，個々のデータの最近傍の細胞からの距離の総和，あるいはデータがマップから外れた方向への距離（シュワルツの直交化法で計算可能）などを考えればよいであろう．

テキスト　T.コホネン著　自己組織化マップ，シュプリンガフェアラーク東京 1996

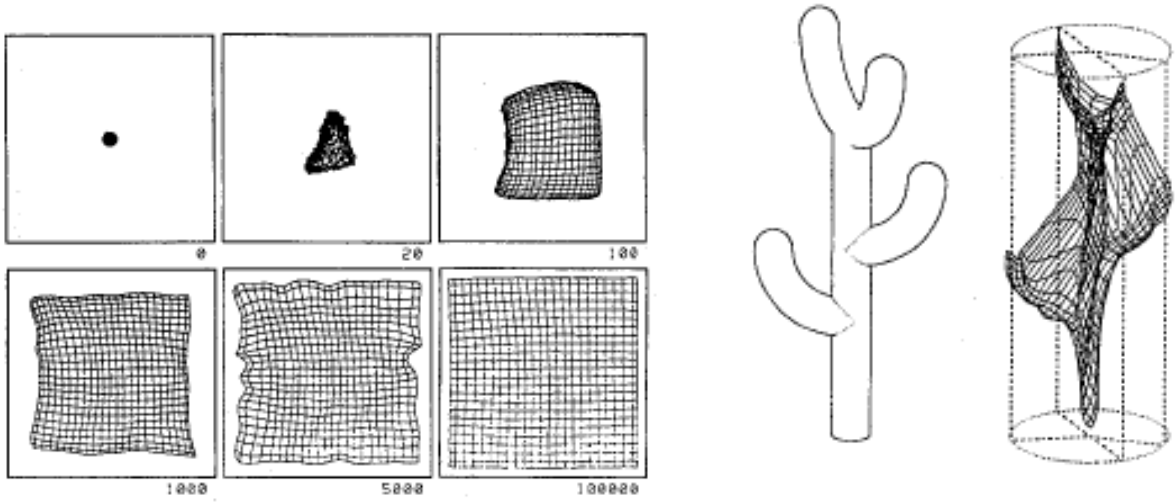


図 3.4　順序づけ過程中の参照ベクトル（正方形配列）．右下の隅に書かれている数字は学習回数を示す．

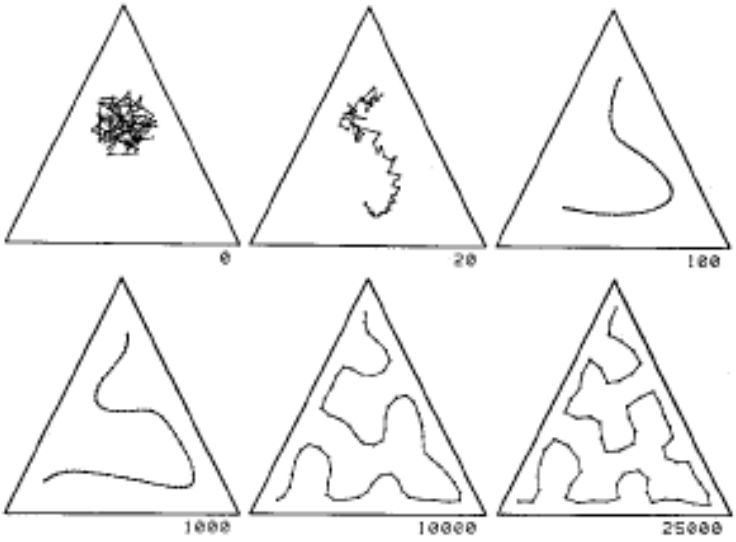


図 3.5　順序づけ過程中の参照ベクトル（線形配列）．

課題 3　特徴マップの学習

1. 二次元データ n_k=(Xk,Yk) k=1…N の生成　N～10000 くらい
   Xk = rand() [0..1], Yk = rand() [0..1] の値を与える
2. 二次元に配置された細胞 Wij = (w_{ij}^x,w_{ij}^y) の生成．1<= i,j <= 4 ．初期値は乱数[0..1]
3. 学習係数 α=0.05
4. for　Loop = 1 to 集束するまで
   (ア) if　Loop 1000 回に一回　then
      ①　Wij = (w_{ij}^x,w_{ij}^y)を二次元面上にプロットし，隣り合わせの i,j を線でつなぐ．
      理想的には 4*4 のグリッドがプロットされるはず．
   (イ) end if
   (ウ) for k = 1…N
      ①　I,J = $\mathrm{argmin}_{i,j}$ ∥ Wij − n_k ∥
      ②　$W_{IJ} = (1-\alpha) W_{IJ} + \alpha\ n_k$；
      ③　for I,J の 8 近傍
         1.　$W_{IJ} = (1-\alpha/2) W_{IJ} + \alpha/2\ n_k$；
      ④　end for
   (エ) end k
5. end Loop

α の値は調整が必要

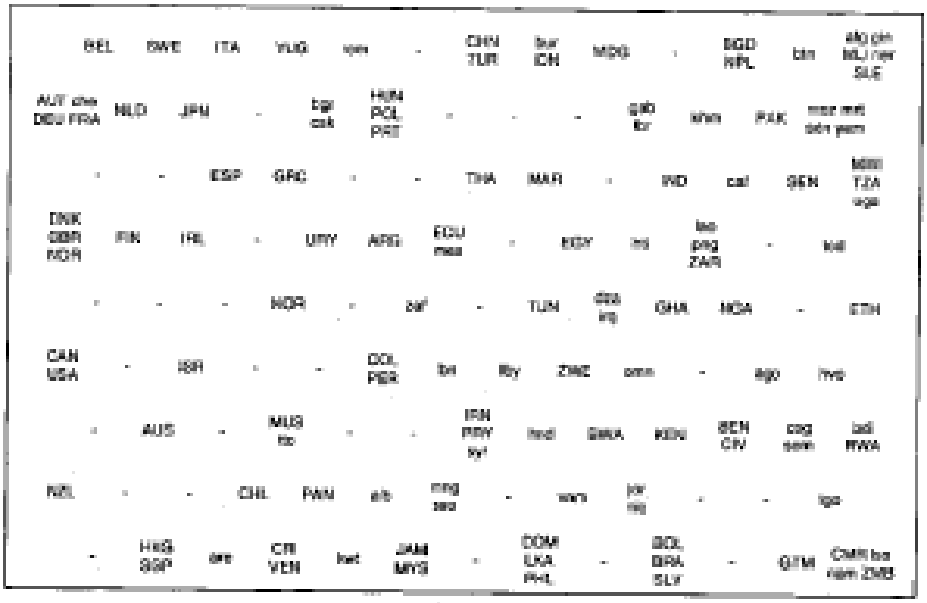


図 3.23　全世界 126 国の "貧困マップ"．大文字で書かれた記号はマップを形成するのに使用した国名と対応する；残りの記号（小文字）は 11 以上の属性値が欠けている国を意味する．そしてこれらの記号は学習段階後にこの SOM 上に位置づけされた．