論　文

モジュール競合学習を用いた適応的クラスタリング

鈴木　　敏† 　　　上田　修功†

Adaptive Clustering Method using Modular Learning Architecture

Satoshi SUZUKI† and Naonori UEDA†

あらまし
クラス内のデータが連続的に分布しクラス間で重なりがない場合において，クラス数未知の状態から逐次処理によりデータをクラス分けする手法を提案する．提案する手法はモジュール競合ネットワークに統計力学のアナロジーを用いる手法である．即ち，入力データを粒子と見なし，粒子の分布に適したエネルギー状態を各モジュールが推定することにより学習を行なう手法である．この学習の結果，連続的に分布する粒子は同一のモジュールに振り分けられ，同時に，粒子の存在しないモジュールも現れる．従って，予めモジュール数を多めに用意し，学習後，粒子の集まったモジュールを１つのクラスと見なすことで，入力データのクラス分けとクラス数推定が可能となる．本論文では提案手法の詳細と，２次元人工データ及び顔画像を用いた実験結果を示す．

キーワード
クラスタリング，教師無し学習，競合学習，逐次処理，自由エネルギー，統計力学

1.　ま え が き

人間は視点に依存して多様に変化する３次元物体を２次元網膜像を通して認識・類別している．この処理は，教師信号無しの逐次処理学習であり，随時新しい入力が与えられる追加学習でもある．この機能の一部を計算機上で実現するため，筆者の一人は，モジュール競合学習法 [1] [2] を応用して，様々な視点方向の２次元射影像から教師無しで３次元物体を認識・類別するモデルを提案した [3] [4]．この手法は，連続的に広がる分布を１つのクラスと見なす教師無しのクラス分けを可能にしたが，物体数 (クラス数) が既知という前提条件を必要とし，また，追加学習には対処できなかった．なぜなら，従来のモジュール学習法 [1] [2] が入力空間を全モジュールを使って表現する手法だからである．そこで，本論文では，追加学習に対処できるクラス数未知のクラスタリング手法について検討する．

従来のクラスタリング手法の代表としては決定木による手法や k-means 法， isodata アルゴリズム等が挙げられる [5]．しかし， k-means 法に代表される，クラス数を既知として入力空間を切り分けるクラスタリング手法は，連続的に分布する１つのクラスに対しても，分布の形状によっては複数のクラスとしてクラスタリングを行うこともあるので，上記の問題には適用できない．一方，決定木や isodata アルゴリズムを逐次処理へ適用するには，データ間距離を計算するために予め全データを記憶しなければならない．更に，クラス分けのためのしきい値が必要になるが，その有効な設定方法は一般に知られていない．以上の理由より，従来の古典的なクラスタリング手法は本論文の対象とする追加学習可能な逐次型クラスタリングには適用できない．

最近では，寺島らが自己組織化特徴マップ (SOM) [6] を用いたクラス数未知でのクラスタリング手法を提案している [7]．この手法では，データ密度の高い空間に代表ベクトルが集まるという SOM の性質を利用して，クラスタリングを行っている．この手法はオンラインのクラスタリング課題に対する有効な解決法であるが，例えば，学習途中で新しいクラスのデータが追加される場合，新しいクラスを表現するには，他のクラスを表現していた代表ベクトルを利用しなければならないため，データから代表ベクトルへの写像が大きく変化する可能性がある．即ち，追加学習には対処していない．

本論文では先のモデル [3] [4] に対する学習法の工夫により，上記目的の実現を目指す．ところで，近年，スピングラスの理論を中心に，統計力学の手法がニューラル

† NTT コミュニケーション科学基礎研究所，京都府
NTT Communication Science Laboratories, Kyoto, 619-0237 Japan

ネットワークの研究に応用され，数多くの重要な知見が得られている [8]．従って，本論文でも学習アルゴリズムへの統計力学的アナロジーの適用を試みる．

特定の条件下にある粒子系空間では自由エネルギーを低くしていくと，全ての粒子がエネルギーの低い状態に集まるか，或は空間内に粒子が殆んど存在しなくなるかの 2 種類の状態に収束するという性質がある(注 1)．この粒子の性質を利用し，複数あるモジュールが特定の粒子の集合に対応するモジュールと，対応する粒子の無いモジュールとに自律的に分かれるように学習を行なう．即ち，入力データを粒子と見なし，各モジュールを粒子系空間と見なすことにより，連続的に分布する粒子が同一のモジュールに安定して存在するように学習させることができる．これにより，追加学習にも対処できるクラス数未知のクラスタリングの実現が期待でき，先の学習法 [3] [4] における問題点が解決できる．

2. モジュール競合ネットワーク

本論文で用いるネットワークは，一つのモジュールが複数のガウス関数から成る 2 階層の競合ネットワークである．各モジュールはガウス関数の平均ベクトルを組み合わせて入力データの再構成を試み，次いで，モジュール間でその精度を競合する．この競合結果により，入力は各モジュールに確率的にクラス分けされる．以下にこのネットワークの詳細を述べる．

入力 $\boldsymbol{x}$ が与えられた時，第 i モジュールで再構成された情報 $\boldsymbol{y}_i$ は

$$\boldsymbol{y}_i = \sum_{j=1}^{U_i} h_{ij}\boldsymbol{c}_{ij} \tag{1}$$

で与えられる．但し，U_i は第 i モジュール内のガウス関数の総数であり，$\boldsymbol{c}_{ij}$ は第 i モジュール，第 j ガウス関数の平均ベクトル (未知パラメタ) を表す．また，h_{ij} はモジュール内の競合における寄与率を表し，

$$h_{ij} = \frac{t_{ij}}{\sum_{k=1}^{U_i} t_{ik}} \tag{2}$$

で与えられる．但し，t_{ij} は σ^2 を定数として，

$$t_{ij} = \exp\left(-\frac{\|\boldsymbol{c}_{ij}-\boldsymbol{x}\|^2}{2\sigma^2}\right) \tag{3}$$

で表される．即ち，入力に近いガウス関数ほど寄与度が大きくなるようにし，$h_{ij} > 0, \sum_{j=1}^{U_i} h_{ij} = 1$ となるよう正規化している．

(注 1)：式の上での性質である (3.参照)

入力 $\boldsymbol{x}$ の帰属クラスは，モジュール間競合により決定される．モジュール間競合は，各モジュールで再構成された情報 $\boldsymbol{y}_i$ と入力 $\boldsymbol{x}$ とのユークリッド距離 $\|\boldsymbol{y}_i-\boldsymbol{x}\|$ に基づいて行われ，この距離が最小となるモジュールのクラス指標 i を入力 $\boldsymbol{x}$ の属するクラスとする．但し，確率モデルとするため，モジュール間の競合は次式で定義される g_i を用いて行う．

$$g_i = \frac{s_i}{\sum_{k=1}^{M} s_k} \tag{4}$$

ここで，M は全モジュール数を表し，s_i は ρ^2 を定数として，

$$s_i = \exp\left(-\frac{\|\boldsymbol{y}_i-\boldsymbol{x}\|^2}{2\rho^2}\right) \tag{5}$$

とする．入力と再構成された情報の近さは数学的な取り扱い易さから式 (5) に示す指数関数とした．

学習にあたって，モジュール内のガウス関数を適応的に追加する機能を与えることにより，各モジュールが適応的に学習できるようになる．ガウス関数の追加を含む本ネットワークの学習法については 4.で詳しく述べる．

3. 粒子系空間のモデル

上記のネットワークをモジュール間の偏りを許容して学習させるために，本論文では，統計力学的手法を導入し，逐次的に与えられる入力データを粒子の観測位置と見なす手法を用いる(注 2)．

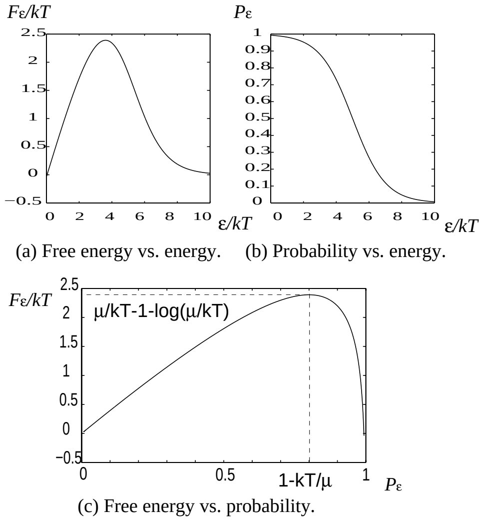
以下，利用する粒子系空間の性質を説明し，次に，その性質をモジュール競合ネットワークに適用する方法について述べる．

3.1 粒子系空間の性質

例えば電子のような粒子の粒子源を含む低温の空間(注 3)では，多くの粒子が自由に動ける状態が空間の自由エネルギーが高い状態である．逆に，自由エネルギーが低い状態は空間内の粒子がエネルギーの低い状態に集まっている場合か，或は空間内に粒子が殆んど存在しない場合である．このことは，以下に記す粒子のエネルギーと存在確率及び自由エネルギーの関係式により示される．

粒子源を持つ粒子系ではエネルギー ϵ を持つ粒子が存在する確率は，μ, k, T を定数として，

(注 2)：この時，入力データ自体を粒子の位置を表すパラメタと見なすため，粒子自体はお互いに区別のない同種粒子であることが必要となる．但し，ボース粒子は自由エネルギー最小化により全粒子が基底状態に収束してしまうため不可．
(注 3)：フェルミ粒子が低温のグランドカノニカル状態にある場合を考える．この時，空間内の粒子数は空間の状態に従って増減する．



(a) Free energy vs. energy. (b) Probability vs. energy.

(c) Free energy vs. probability.

図 1 粒子の性質
Fig. 1 Characteristics of the particles.

$$P_\epsilon = \frac{1}{\exp\left(\frac{\epsilon-\mu}{kT}\right)+1} \tag{6}$$

で与えられる [9]．また，粒子系の自由エネルギー F は

$$\begin{aligned} F &= \sum_\epsilon \left(\mu P_\epsilon + kT\log\left(1-P_\epsilon\right)\right) \\ &\equiv \sum_\epsilon F_\epsilon \end{aligned} \tag{7}$$

で表される [9]．但し，$\sum_\epsilon$ は可能な全てのエネルギー状態の和を表す．

$T < \mu/k$ の低温では，個々の粒子のエネルギー ϵ と自由エネルギー F_ϵ との間には図 1(a) の関係がある（図では $\mu/kT = 5$ の場合を示している)．この図から，自由エネルギーを低くするためには各粒子のエネルギー ϵ が非常に大きいか 0 に近いかのどちらかになることが必要であることが分かる．ところが，粒子のエネルギーと存在確率の間には図 1(b) の関係があり，粒子のエネルギーが高いと存在確率は非常に低くなる．故に，自由エネルギーが低い状態では，エネルギーの低い粒子が集まっているか空間内に殆んど粒子が存在していないかのどちらかとなる (図 1(c))．

従って，このような粒子系を複数組み合わせた空間を考え，自由エネルギーを下げることで，ある部分空間では粒子が低エネルギー状態に集まり，ある部分空間では粒子が存在しなくなることが期待できる．次節で，この粒子系空間の性質をアナロジーとして用いた，モジュール競合ネットワークの詳細について述べる．

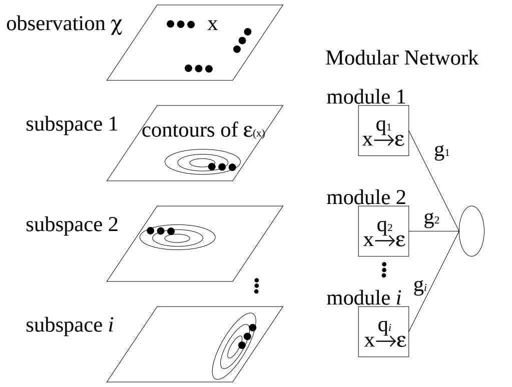


図 2 提案モデルのイメージ
Fig. 2 An image of the proposed model.

3. 2 ネットワークへの適用

複数の部分空間から成る空間が，低温で安定した (自由エネルギーの低い) 状態にあるとし，各部分空間は独立した座標系を持ち，それぞれが異なるエネルギー状態にあるとする．今，空間内の粒子 N 個を観測し，その観測結果は $\chi = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$ で与えられたとする．$\boldsymbol{x}_j$ は j 番目の粒子が観測された位置を表す．但し，各粒子がどの部分空間で観測されたかは与えられない．2.のネットワークを用いて，この観測結果から各空間のエネルギー状態を推定するモデルを考える．

図 2に示すように，まず粒子系の部分空間の数をネットワークのモジュール数に 1 対 1 に対応させる．これにより，各モジュールに座標 $\boldsymbol{x}$ からエネルギー ϵ_i への変換 $\boldsymbol{x} \overset{q_i}{\longmapsto} \epsilon_i$ を学習させる．この関数 q_i により部分空間 i のエネルギー状態が与えられ，従って，部分空間内の粒子の存在が確率的に与えられる．また，モジュール間の競合 $g_i(\boldsymbol{x}_j)$ を $\boldsymbol{x}_j$ の粒子が部分空間 i で観測された確率と見做す．

以上の対応関係により，以下に示す粒子の存在確率を高め，且つ，自由エネルギーを小さくするようにネットワークに最適なエネルギー状態を学習させる．この結果，各モジュールは粒子の存在するモジュールと存在しないモジュールとに分けられ，目的とするクラスタリングが実現される．

[粒子の存在確率]

観測された粒子 $\boldsymbol{x}$ がネットワークで作られる空間に存在する確率 $P_{\boldsymbol{x}}$ は，粒子 $\boldsymbol{x}$ が部分空間 i で観測された確率 P_i とモジュール i の中で粒子 $\boldsymbol{x}$ が存在する確率 $P^i_{\boldsymbol{x}}$

を用いて，

$$P_{\boldsymbol{x}} = \sum_{i=1}^{M} P_i P_{\boldsymbol{x}}^i \tag{8}$$

で表される．

P_i は前述の仮定より，$g_i(\boldsymbol{x})$ で表され，$P_{\boldsymbol{x}}^i$ はモジュール i の作る部分空間での粒子の存在確率なので，式 (6) から

$$P_{\boldsymbol{x}}^i = \frac{1}{\exp\left(\frac{\epsilon_i(\boldsymbol{x})-\mu}{kT}\right)+1} \tag{9}$$

で与えられる．

ここで，各モジュールでの座標 $\boldsymbol{x}$ からエネルギー ϵ_i への変換 $\boldsymbol{x} \xmapsto{q_i} \epsilon_i$ を表す関係式を

$$\begin{aligned}\epsilon_i(\boldsymbol{x}) &= \frac{kT}{2\rho^2}\|\boldsymbol{y}_i - \boldsymbol{x}\|^2 \\ &= -kT\log(s_i)\end{aligned} \tag{10}$$

で定義すると，$P_{\boldsymbol{x}}$ は次式で表される．

$$\begin{aligned}P_{\boldsymbol{x}} &= \sum_{i=1}^{M}\frac{g_i}{\exp\left(\frac{\epsilon_i(\boldsymbol{x})-\mu}{kT}\right)+1} \\ &= \sum_{i=1}^{M}\frac{s_i}{\sum_{k=1}^{M}s_k}\frac{s_i}{\exp\left(-\frac{\mu}{kT}\right)+s_i}\end{aligned} \tag{11}$$

$P_{\boldsymbol{x}}$ の最大化は，観測された粒子がネットワークの作り出す空間に存在する可能性を高めることに相当する．

[自由エネルギー]

一方，余分な粒子が空間内に入り込まないために，空間全体の自由エネルギーを最小化しなければならない．但し，全自由エネルギー F は，部分空間の自由エネルギー F_i の和で表されるから，$\mu' = \exp(-\mu/kT)$ として，式 (7) より

$$\begin{aligned}F &= \sum_{i=1}^{M}F_i \\ &= \sum_{i=1}^{M}\sum_{\epsilon}F_{\epsilon}^{(i)} \\ &= \sum_{i=1}^{M}\sum_{\boldsymbol{x}\in\chi}\frac{\mu s_i}{\mu'+s_i} + kT\log\left(\frac{\mu'}{\mu'+s_i}\right) \\ &= -kT\sum_{\boldsymbol{x}\in\chi}\sum_{i=1}^{M}\frac{\mu}{kT}\frac{\mu'}{\mu'+s_i} + \log\left(\mu'+s_i\right) \\ &\equiv -kT\sum_{\boldsymbol{x}\in\chi}f(\boldsymbol{x})\end{aligned} \tag{12}$$

で表される(注4)．

4. 学習アルゴリズム

本節では前節で説明したモデルに基づいて，学習アルゴリズムを導出する．本モデルにおける学習パラメタはガウス関数の平均ベクトル $\boldsymbol{c}$ であり，学習は平均ベクトルの更新とガウス関数の適応的追加により行われる．但し，ガウス関数の連続性を保つために，ガウス関数の追加は $f(\boldsymbol{x})$ の最大化に適用し，この時追加されるガウス関数の平均ベクトルは入力 $\boldsymbol{x}$ と一致させる．

$P_{\boldsymbol{x}}$ の最大化と $f(\boldsymbol{x})$ の最大化ではパラメタが必ずしも一致しないので，学習はそれぞれ独立して行う．但し，$P_{\boldsymbol{x}}$ の最大化は特定粒子に関する自由エネルギー最小化でもあり，2種類の学習は最終的には同じ収束方向へ向かうことになる．

4.1 観測粒子の存在確率最大化

まず，$P_{\boldsymbol{x}}$ の最大化について示す．第 i モジュール第 j ガウス関数の平均ベクトル $\boldsymbol{c}_{ij}$ の更新は，α を学習係数として，

$$\begin{aligned}\Delta\boldsymbol{c}_{ij}^{p} &= \alpha\frac{\partial P_{\boldsymbol{x}}}{\partial \boldsymbol{c}_{ij}} \\ &= \frac{\alpha g_i}{\rho^2}\left(p_i^2 - 2p_i + \sum_{k=1}^{M}g_k p_i\right)\boldsymbol{G}_i\end{aligned} \tag{13}$$

をその変化量とすることで表される．但し，

$$\begin{aligned}p_i &= P_{\boldsymbol{x}}^i \\ &= \frac{s_i}{\mu'+s_i}\end{aligned} \tag{14}$$

および，

$$\begin{aligned}\boldsymbol{G}_i &= \sum_{l}(\boldsymbol{y}_{il}-\boldsymbol{x}_l)\frac{\partial y_{il}}{\partial c_{ijl}} \\ &= h_{ij}\left(a_{ij}(\boldsymbol{c}_{ij}-\boldsymbol{x}) + (\boldsymbol{y}_i - \boldsymbol{x})\right)\end{aligned} \tag{15}$$

であり，$\sum_l$ は入力の各次元についての和を表す．また，a_{ij} は，

$$a_{ij} = \frac{(\boldsymbol{y}_i-\boldsymbol{x})^t(\boldsymbol{y}_i-\boldsymbol{c}_{ij})}{\sigma^2} \tag{16}$$

を表すものとする．

(注4)：式 (7) では自由エネルギーはエネルギー状態の和として表されているが，χ が粒子の存在可能位置を表していることと，一つのエネルギー状態には一つの粒子しか存在できないというフェルミ粒子系の性質を利用して，エネルギー状態の和を観測される粒子に関する和に置き換えている．即ち，全ての可能なエネルギー状態は $q_i(\chi)$ で表されるとしている．

4. 2 自由エネルギーの最小化

一方，自由エネルギーの最小化は $f(\boldsymbol{x})$ の最大化で置き換えられ，ガウス関数の追加と平均ベクトルの更新の２種類の手法を用いて行う．

[ガウス関数の追加]

まず，ガウス関数追加のためのコストについて検討する．ガウス関数の追加は，与えられた入力 $\boldsymbol{x}$ に対して，以下に示す追加条件を満たすモジュールに，$\boldsymbol{c} = \boldsymbol{x}$ として新しいガウス関数を追加する．

ガウス関数の追加条件は，

$$\Delta f(\boldsymbol{x}) > \delta > 0$$

を用いる．但し，Δf はガウス関数追加前後での $f(\boldsymbol{x})$ の増加量を表す．

$$\Delta f(\boldsymbol{x}) = (\text{追加後の} f(\boldsymbol{x})) - (\text{追加前の} f(\boldsymbol{x}))$$

式 (12) より，

$$f(\boldsymbol{x}) = \sum_{i=1}^{M} \left(-F_{\epsilon}^{(i)}/kT\right)$$

であり，また，ガウス関数に変化のない部分空間では自由エネルギーは変化しないことから，モジュール i にガウス関数が追加されたとき，$\Delta f(\boldsymbol{x})$ は部分空間 i の自由エネルギーの変化量により，

$$\Delta f(\boldsymbol{x}) = (\text{追加前の} F_{\epsilon}^{(i)}/kT) - (\text{追加後の} F_{\epsilon}^{(i)}/kT)$$

と表される．

それぞれの部分空間においてエネルギー ϵ を持つ粒子が存在する確率 P_{ϵ} と自由エネルギー F_{ϵ} との関係は図 1(c) に示されている．この図は $T = \frac{1}{5}\mu/k$ の低温における P_{ϵ} と F_{ϵ}/kT との関係を表している (式 (6),(7) 参照)．

図 1(c) よりガウス関数追加後は $P_{\epsilon} \simeq 1$ 及び $F_{\epsilon}^{(i)}/kT \simeq 0$ となるため，

$$\Delta f(\boldsymbol{x}) \simeq (\text{追加前の} F_{\epsilon}^{(i)}/kT)$$

と近似でき，ガウス関数追加の可否判定は追加前の $F_{\epsilon}^{(i)}/kT$ の値により決められる．図 1(c) より適当な $\delta(F_{\epsilon}/kT$ の値; $0 \ll \delta < \mu/kT - 1 - \log(\mu/kT))$ を定めることで，ガウス関数追加の条件を

$$F_{\epsilon}^{(i)}/kT > \delta$$

とできる．この条件により，確率 P_{ϵ} が特定の範囲内にある時にガウス関数の追加が可能となる．逆に，ガウス関数追加が可能となるときの P_{ϵ} の値は決められた範囲内にあるため，ガウス関数が追加されるときにはその近傍に粒子が存在していることになり，従って，部分空間内での粒子の連続性が保たれていることになる．

[平均ベクトルの更新]

もしも，条件 $F_{\epsilon}^{(i)}/kT > \delta$ を満たす i が存在しない場合には，β を定数として，

$$\begin{aligned}\Delta \boldsymbol{c}_{ij}^{f} &= \beta \frac{\partial f_i(\boldsymbol{x})}{\partial \boldsymbol{c}_{ij}} \\ &= \frac{\beta}{\rho^2}\left(\frac{\mu}{kT} - \frac{1}{1-p_i}\right)\left(p_i^2 - p_i\right) G_i \quad (17)\end{aligned}$$

により平均ベクトル $\boldsymbol{c}_{ij}$ の更新を行い，$f(\boldsymbol{x})$ を最大化する．

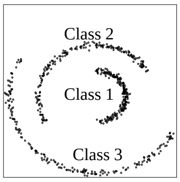
4. 3 アルゴリズムのまとめ

以上から，学習アルゴリズムは次のようになる．

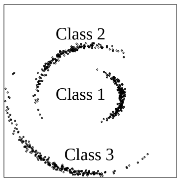
1. 入力サンプルの平均値に微少ノイズを加え，平均ベクトル $\boldsymbol{c}$ の初期値とする．
2. 逐次与えられる観測データ $\boldsymbol{x}$ に関し，順次，全てのモジュールに同じ入力 $\boldsymbol{x}$ を与える．
3. 式 (13) を用いて平均ベクトルの更新を行う．
4. モジュール i^* ($i^* = \arg\max_i g_i$) に於いて，$F_{\epsilon}^{(i^*)}/kT > \delta$ であれば，平均ベクトルを $\boldsymbol{x}$ とするガウス関数を追加，それ以外は式 (17) を用いて平均ベクトルの更新を行う．
5. 既定の学習回数に達すれば終了，さもなければ 2 〜 4 を繰り返す．

5. 計算機実験

提案したモデルの有効性と特性を調べるため，複雑な分布を持つ２次元人工データを用いて，様々な条件下で



(a) Uniform distribution



(b) Gaussian distribution

図 3 入力分布
Fig. 3 Input data distributions.

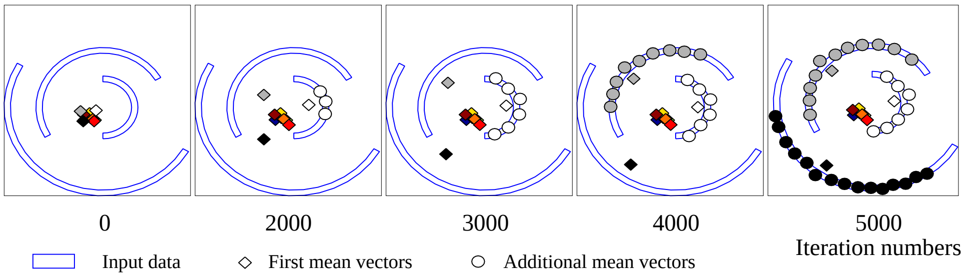


図 4　学習過程の例
Fig. 4　An example of clustering process.

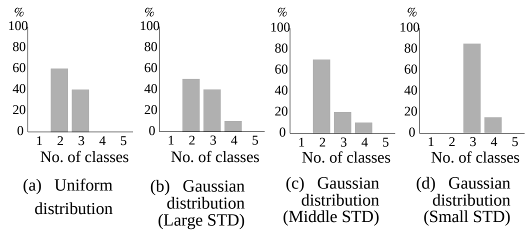


図 5　クラス数推定
Fig. 5　Estimation results of the number of classes.

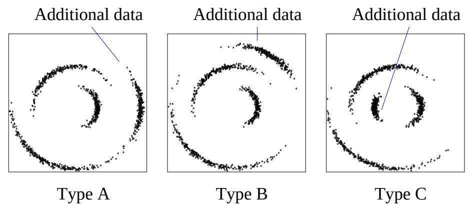


図 6　追加学習での入力分布
Fig. 6　Input data distributions used by additional learning.

表 1　類 別 精 度
Table 1　Classification accuracy.

Distribution	Uniform	Gaussian		
STD		large	middle	small
%	66.7%	74.9%	70.4%	98.0%

学習を試みた．更に，現実問題への応用として，様々な方向から見た複数人の顔画像をクラス分けする実験を行った．

5. 1　2 次元人工データによる評価

図 3に実験で用いた 2 次元人工データを示す．最も近接する他クラスまでの距離が一定になるように，データは半径の異なる 3 種類の半円弧上に分布しており，図 3(a) では一様分布，図 3(b) では正規分布の広がりを持っている．正規分布のデータは標準偏差 (STD) を 3 種類 (「大」，「中」，「小」) 用意し(注 5)，一様分布のデータと合わせて 4 種類の入力データセットについてそれぞれ学習を行なった．

ネットワークの初期状態ではモジュール数は 10，各モジュール内のガウス関数は 1 個とした．以下，このガウス関数の平均ベクトルを便宜上第 1 平均ベクトルと呼ぶ．モジュール間の競合を明確にするため，実験では第 1 平均ベクトルの初期値は，任意に選ばれた入力 200 点の平均ベクトルに若干のノイズ (サンプルの分散の 5%) を加えて設定した．各パラメタは $\mu/kT = 5$, $\alpha = 0.1$, $\beta = 0.01, \delta = 2, \sigma = \rho = 1$ とした (これらのパラメタ設定については付録を参照)．学習は，ランダムに選ばれた入力データをモデルに繰り返し与えることで行う．学習回数は 10000 回とした．ここで，1 回の学習とは，1 入力データに対してモデルの更新を 1 度だけ行うことを意味する．

図 4に図 3(b)(標準偏差「小」) のデータに対する学習過程の例を示す．図 4では入力データの分布範囲を曲線で囲い (input data)，各第 1 平均ベクトルを '◇'，学習により追加されたガウス関数の平均ベクトルを '○' の記号で示した．各々の記号が属するモジュールは，濃淡によって区別した．

学習初期は，第 1 平均ベクトルが競合により入力データに近づいていく．第 1 平均ベクトルと入力データの距離がある程度以上近づくと，新たなガウス関数を追加して自由エネルギーを大きく減少させる．学習が進むにつれて，ガウス関数の数は増え，1 つのモジュールが特定のクラスの入力空間を覆うようになる．この結果，モ

(注 5)：原点から見た入力範囲の中心と入力座標とが成す角度が正規 (ガウス) 分布に従うように生成された点の集合であり，標準偏差の「大」，「中」，「小」は各々 180 度，120 度，60 度とした分布で，幅が 5% の半円弧内にある点の集合である．

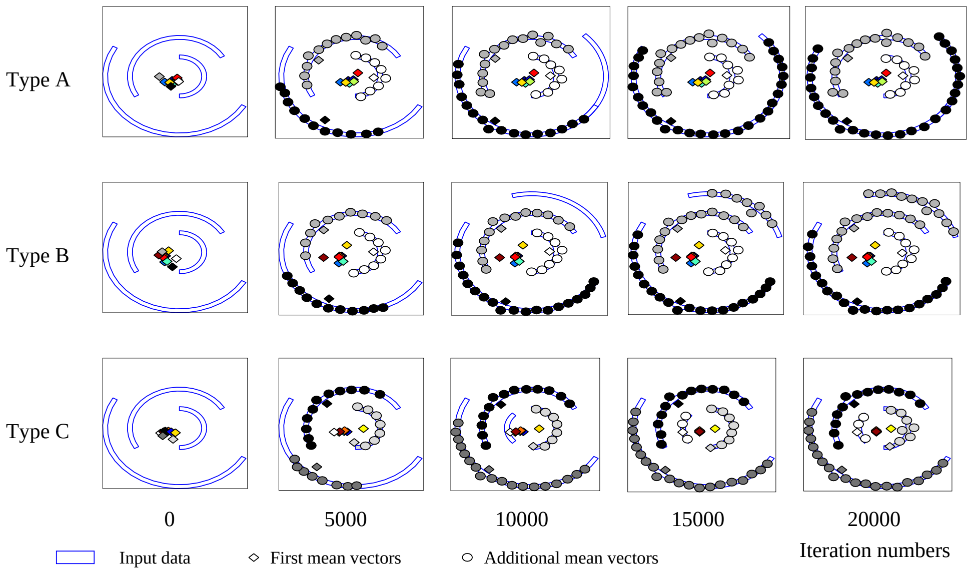


図 7 追加学習での学習過程
Fig. 7 Examples of clustering process with additional learning.

ジュールとクラスの対応関係が定まるようになる．

学習により得られた結果は，クラス数の推定および類別精度により評価される．クラス数は，学習によるガウス関数追加の有無により推定される．たとえ１つでもガウス関数が追加されれば，当該モジュールは１クラスに対応するものと判断される．一方，類別精度は各データが対応するモジュールに正しく分類される割合を示す．本人工データは，各データの生成時に所属クラスが既知であるため，クラスとモジュールの最適な１対１の対応関係を学習結果から推定できる．この対応関係に従うクラス分けを正しいクラス分けと見なし，テストデータが正しく分類される割合を計算する(注 6)．以下，第１平均ベクトルの初期値をランダムに変えて 20 回行った実験の結果を示す．

図 5にモデルにより推定されたクラス数の 20 回の試行における統計を示す．正しく３クラスと推定された割合は，標準偏差「小」の場合に約 85% であった．これに対し，これより標準偏差が大きい例では，１つ少なく２クラスと推定された割合が多い．異なるクラスのデータが近接する部分で連続したデータと見なされ，その結果，クラス数が過小に推定されている．

また，表 1は，それぞれの場合の類別精度を表している．類別精度もクラス数推定と同様に標準偏差「小」の場合に極めて高かった．図 5(d) の結果と合わせると，クラス数を多く推定した場合でも，大部分の入力データは対応するモジュールに分類されていることがわかる(注 7)．

以上から，本手法はクラス間の分散に比べ，クラス内の分散が小さいときに有効な手法であることがわかる．

5.2 追加学習

次に，追加学習の実験を行った．前節の実験の標準偏差「小」の場合について，実験終了後に新しいデータを加えて，さらに 10000 回学習を行った．追加したデータは図 6に示すように，既存のクラスの分布範囲を広げる場合 (Type A) と，新しいクラスを追加する場合 (Type B, Type C) の３種類を用意した．

図 7に学習過程の例を示す． 10000 回の学習を終えた時点で新しいデータを加え，更に学習を続けた． Type A,C では推定されたクラス数が各々３，４となり，正しい結果を示したが， Type B の場合には１つのモジュールが２クラスの入力空間を覆う結果となり，前節の結果と同様に連続したデータとして学習された．いずれの学習に於ても，追加学習によるそれ以前の学習結果への影響はほとんど無いことが，ガウス関数の位置の変化がほ

(注 6)：テストデータは各クラスの分布空間から等間隔で 181 点を取り出したデータであり，学習データとは異なる．

(注 7)：図 5(a)(b)(c) のように，クラス数が少なく推定された場合には，対応するモジュールが存在しないデータクラスが存在するため，類別精度は明らかに低くなる．

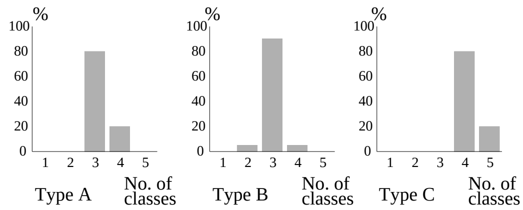


図 8　追加学習でのクラス数推定
Fig. 8　Estimation results of the number of classes with additional learning.

とんど無いことからわかる．クラス分け結果への影響はこの実験では全く見られなかった．

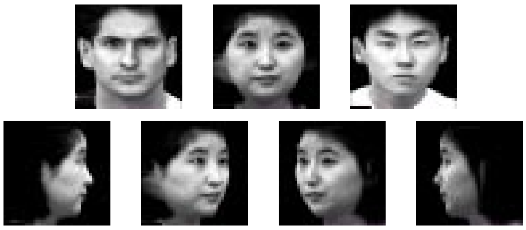
20 回の学習から得られたクラス数推定結果の統計を図 8に示す． Type A,C では共に 80% 以上の割合でクラス数を 3， 4 と正しく推定している．新しいデータの追加前後で比較すると， 20 回の実験全てにおいて，追加されたクラス数は Type A では 0， Type C では 1 であった．この時の類別精度は Type A が 93.7%， Type C が 90.5% であった．一方， Type B では約 90% の割合でクラス数を 3 と，実際よりも少なく推定しており，類別精度は 73.4% であった．

Type B で学習を失敗した主な原因は，ガウス関数の初期値周辺が他の分布により囲まれているため，第 1 平均ベクトルが目的の分布の近傍まで移動できなかったことにある．本手法ではこのような場合，正確なクラス分けができない．しかし，このように経路を塞がれる状況は高次元入力では比較的起りにくいことが期待できる．

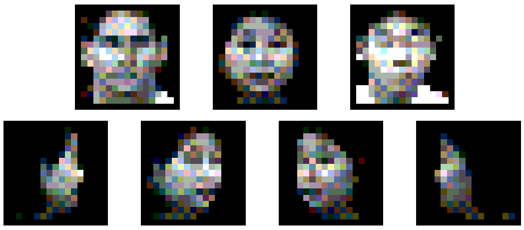
5. 3　顔画像への応用

提案するクラスタリング手法の応用として，顔画像を入力とした顔の認識実験を行う．入力として用いる画像は，正面顔を中心として左右に 1 度ずつ 90 度まで回転し，合計 181 個の画像を 1 クラスとしたものである [10]．同様の画像を 3 人分用意し (図 9(a))，各画像を 15 × 15 次元に圧縮して入力に用いた (図 9(b))．入力画像の生起確率は正面顔を中心とした正規分布 (標準偏差「小」) に従うとし，選ばれた入力画像には若干のノイズ (各画素の濃淡に 5% の分散) を加えた(注 8)．ネットワークのモジュール数は 5 および 10 の 2 種類を用意し，各第 1 平均ベクトルの初期値は，ランダムに選ばれた 200 個の入力画像の平均に若干のノイズを加えて設定した．

学習回数を 50,000 回 (注 9) としたときの実験結果の例を図 10に示す．各パラメタは $\mu/kT = 5$, $\alpha = 0.1$, $\beta = 0.01$, $\delta = 2.3$, $\sigma = \rho = 10$ とした．図にはモジュール数を 5 とした場合の学習前および学習後の第 1 平均ベクトルと追加された平均ベクトルの例が獲得された順に示されている．各モジュールにはさまざまな方向からの顔のイメージが獲得されており，モジュールとクラスの対応がとれていることがわかる．



(a) Original images.



(b) Input images.

図 9　入力として用いた顔画像の例
Fig. 9　Examples of facial image.

モジュール数が 5 の場合と 10 の場合各々において， 20 回の試行におけるクラス数推定の結果，正しくクラス数 3 を推定した割合は，モジュール数 5 の場合で 65%，モジュール数 10 の場合で 70% であった．一方，全入力画像に対する類別精度はモジュール数 5 の場合に 90.3%，モジュール数 10 の場合に 91.9% であった．クラス数を間違って推定した場合でも，殆んどの画像は正しいクラスにクラス分けされていることが分かる．

6.　む　す　び

クラス数未知のデータを教師信号無しで逐次処理によりクラス分けする手法を提案し， 2 次元人工データおよび顔画像を用いた実験によりその有効性について検討した．その結果，第 1 平均ベクトルの移動経路が塞がれず，クラス間の分散に比べクラス内の分散が小さい場合には，分布の形状に依存することなく，連続した分布を

(注 8)：このノイズは学習には本質的な影響を与えていない．
(注 9)：計算時間は人工データによる 20000 回の学習で約 20 分，画像データによる 50000 回の学習で約 10 時間 (Ultra sparc 200Mhz, Matlab)．

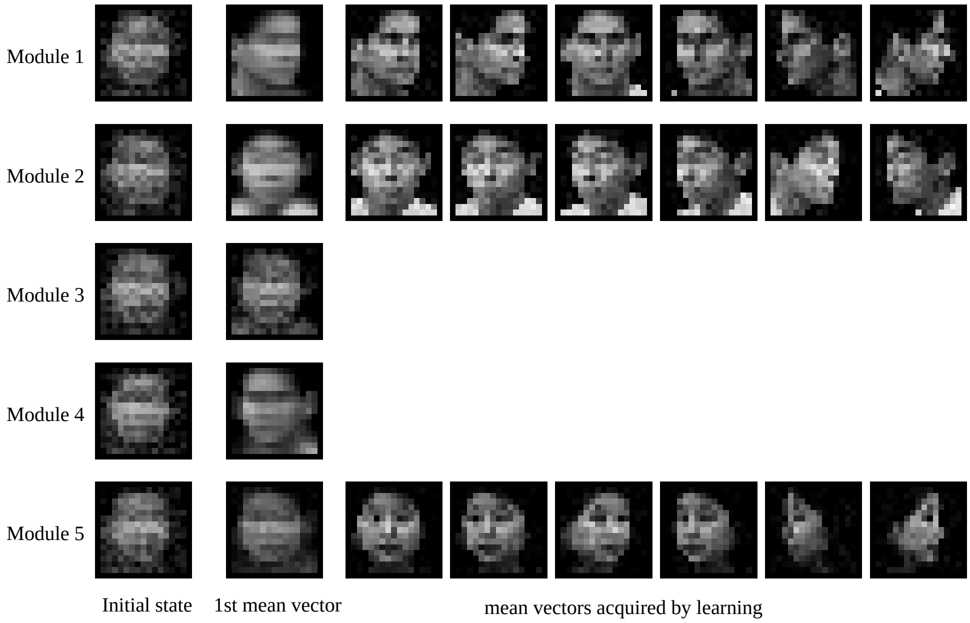


図 10　顔画像のクラスタリング結果
Fig. 10　Clustering results of facial images.

1つのクラスとして分類可能であることを確認した.

謝辞　貴重な画像データを提供していただいた ATR 人間情報通信研究所の赤松茂室長と磯野勝宣研究員に感謝します.

文　　献

[1] M.I.Jordan and R.A.Jacobs, “Hierarchies of adaptive experts,” In Advances in Neural Information Processing Systems 4, San Mateo, CA, Morgan Kaufmann Publishers, pp.985-992, 1992.

[2] R.A.Jacobs, M.I.Jordan, S.J.Nowlan and G.E.Hinton, “Adaptive mixture of local experts,” Neural Computation, vol.3, pp.337-345, 1993.

[3] 鈴木　敏, 安藤広志, “2次元射影像からの3次元物体の認識と類別 — モジュール構造を用いた教師なし学習モデル —,” 信学論, vol.J79-DII(7), pp.1291-1300, 1996.

[4] H.Ando, S.Suzuki and T.Fujita, “Unsupervised visual learning of theree-dimensional objects using a modular network architecture,” Neural Networks, vol.12, pp.1037-1053, 1999.

[5] R.Duda, and P.E.Hart, “Pattern Classification and Scene Analysis,” John Wiley & Sons, NY, 1973.

[6] T.Kohonen, “Self-organization and associative memory,” Third Edition, Springer-Verlag, Berlin, 1989.

[7] 寺島幹彦, 白谷文行, 山本公明, “自己組織化特徴マップ上のデータ密度ヒストグラムを用いた教師無しクラスタ分類法,” 信学論, vol.J79-D-II, no.7, pp.1280-1290, 1996.

[8] 西森秀稔, “ニューラルネットワークの統計力学,” パリティ物理学コース, 丸善, 1995.

[9] 長岡洋介, “統計力学,” 岩波基礎物理シリーズ7, 岩波書店, 1994.

[10] 磯野 勝宣, 赤松 茂, “PCA による顔画像の生成のための3次元顔データの対応表現,” 信学技法, HIP96-17, 1996.

（平成 x 年 xx 月 xx 日受付）

鈴木　　敏　（正員）

平2東大・教養・基礎科学科卒. 同年 NTT 入社. 平4 ATR 人間情報通信研究所出向. 平9より NTT 復帰. 現在, コミュニケーション科学基礎研究所所属. 物体認識に関わる計算モデルの研究に従事. 平6日本神経回路学会研究賞受賞. 日本神経回路学会会員.

上田　修功　（正員）

昭57阪大・工・通信卒. 昭59同大大学院修士課程了. 同年 NTT 電気通信研究所入所. 以来, 画像処理, パターン認識・学習, ニューラルネットワーク, 統計的学習理論の研究に従事. 現在, NTT コミュニケーション科学基礎研究所知能情報研究部 学習理論研究グループ 主幹研究員 (特別研究員), 奈良先端大客員助教授. 平5〜6米国 Purdue 大学客員研究員. 1992年日本神経回路学会研究奨励賞, 1997年電気

通信普及財団賞 (テレコムシステム技術賞) 受賞，工博．日本神経回路学会，日本統計学会，IEEE 各会員．

付　　録

パラメタ設定について

学習時におけるパラメタ μ/kT, δ は，図 1(c) からおおよその値が決められる．提案学習手法では自由エネルギーを下げることによって粒子の存在確率を 0 か 1 に振り分けることを行うので，グラフの頂点位置が基準となり，存在確率の収束方向が決められる．従って，グラフの傾きは大きく，頂点の X 座標は偏りが少ない方が望ましい．図 1(c) では μ/kT が大きい程，グラフの傾きは大きく，頂点の X 座標も大きくなり， $4 \leqq \mu/kT \leqq 6$ 程度が妥当であると考えられる．

また，δ の値はガウス関数追加の基準を定めるもので，基準は 0 に近いほど緩く，逆に $\mu/kT - 1 - \log(\mu/kT)$ に近いほど厳しくなる． $\mu/kT = 5$ の場合には， $1.5 \leqq \delta \leqq 2.3$ 程度が妥当である．

$\mu/kT = 5$, $\delta = 2$ の場合には，図 1(c) からおよそ $0.57 < P_\epsilon < 0.93$ が得られ，式 (6),(10) を使って計算すると，およそ $2.20 < \|\boldsymbol{y}_i - \boldsymbol{x}\|/\rho < 3.07$ となる．これはガウス関数が追加される時の，入力データと既存のガウス関数間の距離 $\|\boldsymbol{y}_i - \boldsymbol{x}\|$ が標準偏差 ρ の 3 倍前後になるときにガウス関数が追加されることを意味しており，入力データの次元や大きさの見当がつけば， ρ を決定する目安となる．同様に $\mu/kT = 5, \delta = 2.3$ の場合には，およそ $2.48 < \|\boldsymbol{y}_i - \boldsymbol{x}\|/\rho < 2.88$ となる．

また， α, β は $1 > \alpha > \beta > 0$ となるように，設定する．これは， $P\boldsymbol{x}$ の最大化による粒子の存在位置の決定を， $f(\boldsymbol{x})$ の最大化による余剰粒子の排除に優先させるためである．