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

第2章

類似度行列に基づくファジークラスタリング

2.1 まえがき

本章では類似度行列からファジークラスタを逐次に抽出する方法を提案する [20]. データ間の互いの類似度を要素とした行列を類似度行列と呼ぶ. この類似度行列からファジークラスタを逐次に抽出する方法として, 最適解が解析的に求まる固有値問題に帰着させる方法 [21, 22] を使う. これはグラフスペクトル法 [6] と総称される方法の一つであり, ノイズデータに頑健なクラスタが抽出される. 各データのクラスタへのメンバシップ値は類似度行列の第1固有ベクトルによって与えられる. 各データの残存率を類似度行列の要素に掛けていくことにより, 抽出済みのクラスタを取り除きながら逐次にクラスタを抽出していく. 抽出処理はクラスタの凝集度の変化に基づき, ある程度の大きさのクラスタが抽出されたら終了する. 津田らの方法 [21] もグラフスペクトル法の一種であるが, クラスタを抽出するときにペナルティ項を次々に付加するためペナルティ係数の設定が必要である. 本章で提案する方法ではペナルティ項がないため, 津田らの方法 [21] のような係数の設定不良による性能劣化の心配がない. 提案手法の有効性を, 簡単な2次元データに対する実験とグレイスケール画像のセグメンテーション, およびビデオのセグメンテーションの実験例で示す.

2.2 類似度行列に基づくファジークラスタ抽出

データが n 個あるとする. データ i とデータ j の類似度を s_{ij} とする. s_{ij} が大きい, すなわち類似したデータをグループ化する. まず最初に, 最も主要なクラスタを1つ抽出することを考える. 各データ i がこのクラスタに所属する度合いを x_i とし, クラスタの凝集度を $\sum_{i=1}^{n}\sum_{j=1}^{n} x_i s_{ij} x_j = \boldsymbol{x}^T S \boldsymbol{x}$ で評価する. ここで $S = [s_{ij}]$ は類似度

行列であり，$\boldsymbol{x} = [x_1, \ldots, x_n]^T$ である．この凝集度が最大となる $\boldsymbol{x}$ を求める．ただし，$\|\boldsymbol{x}\|^2 = \boldsymbol{x}^T\boldsymbol{x} = 1$ と制約する．凝集度が最も高いクラスタの $\boldsymbol{x}$ は

$$\begin{aligned} \max_{\boldsymbol{x}} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} x_i s_{ij} x_j \\ \text{subj.to} \quad & \sum_{i=1}^{n} x_i^2 = 1 \end{aligned} \tag{2.1}$$

の解である．この最適化問題の解法は Lagrange 乗数法により固有値問題に帰着することができる [23]．式 (2.1) を行列 S とベクトル $\boldsymbol{x}$ で表せば

$$\begin{aligned} \max_{\boldsymbol{x}} \quad & \boldsymbol{x}^T S \boldsymbol{x} \\ \text{subj.to} \quad & \boldsymbol{x}^T\boldsymbol{x} = 1 \end{aligned} \tag{2.2}$$

となり，Lagrange 関数は

$$L = \boldsymbol{x}^T S \boldsymbol{x} - \lambda(\boldsymbol{x}^T\boldsymbol{x} - 1) \tag{2.3}$$

となる．λ は Lagrange 乗数である．式 (2.1) の解は

$$\frac{1}{2}\frac{\partial L}{\partial \boldsymbol{x}} = S\boldsymbol{x} - \lambda\boldsymbol{x} = 0 \tag{2.4}$$

を満たす．式 (2.4) から

$$S\boldsymbol{x} = \lambda\boldsymbol{x} \tag{2.5}$$

となり式 (2.1) の最適化問題の解法は固有値問題に帰着できる．第 1 クラスタへの所属の度合い $\boldsymbol{x}$ は行列 S の最大固有値の固有ベクトルである．行列 S は非負行列であるので固有ベクトル $\boldsymbol{x}$ の要素 x_i はすべて非負である．凝集度の値は固有値で与えられる．Sarkar ら [22] はこの $\boldsymbol{x}$ のしきい値処理でクラスタを抽出したが，ファジークラスタリングとして考えた場合，最大の x_i でも 1 よりかなり小さいので $\boldsymbol{x}$ をそのまま所属度とするのは不自然である．そこで x_i が最大のデータ i^* がクラスタの代表データであるとし，代表データの所属度が 1 になるように $m_i = x_i/x_{i^*}$ と規格化する．本論文では，この規格化した所属度をメンバシップと呼ぶ．

2.3 重み付きデータのクラスタ抽出

津田ら [21] や Sarkar ら [22] は重み付きデータは扱っていないが，ここでは各データが重みを持つ場合を考える．前節では重みはすべて 1 であるが，例えばあるデータ

が他のデータに重なった場合，それらを一つのデータにまとめて重み v_i を 2 とすれば同じことになる．このとき式 (2.1) は

$$\begin{aligned} \max_{\boldsymbol{x}} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j s_{ij} x_i x_j \\ \text{subj.to} \quad & \sum_{i=1}^{n} v_i x_i^2 = 1 \end{aligned} \tag{2.6}$$

となる．ここではこれを一般化して重み v_i は正の実数値をとれるとする．特に以下では $0 \le v_i \le 1$ である．$y_i = \sqrt{v_i} x_i$ と変数変換すると式 (2.6) は

$$\begin{aligned} \max_{\boldsymbol{y}} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} \sqrt{v_i}\sqrt{v_j} s_{ij} y_i y_j \\ \text{subj.to} \quad & \sum_{i=1}^{n} y_i^2 = 1 \end{aligned} \tag{2.7}$$

となる．この最適化問題の解法も前節と同様に Lagrange 乗数法により固有値問題に帰着できる．式 (2.7) の最適解 $\boldsymbol{y} = [y_1, \ldots, y_n]$ は行列 $\tilde{S} = [\tilde{s}_{ij}]$: $\tilde{s}_{ij} = \sqrt{v_i}\sqrt{v_j} s_{ij}$ の固有ベクトルであり，式 (2.6) の解 $\boldsymbol{x}$ は $x_i = y_i / \sqrt{v_i}$ で与えられる．$\sqrt{v_i}\sqrt{v_j} s_{ij} y_i y_j = v_i v_j s_{ij} x_i x_j$ であるから $\tilde{S}$ の固有値が凝集度を与える．x_i が最大の i を i^* とすると $m_i = x_i / x_{i^*}$ をデータ i のメンバシップ値とする．なお $v_i = 0$ のときは，このデータを取り除いて上のことをすればよいが，$v_i = 0$ だと $y_i = 0$ となるので $x_i = 0$ とすることにすれば $v_i = 0$ も含めてよい．

2.4 ファジークラスタの逐次抽出

以上ではデータからクラスタを 1 個だけ抽出した．そこでこれを拡張して，抽出したクラスタに含まれるデータを取り除きながら逐次にクラスタを抽出する方法を考える．津田ら [21] はペナルティ関数を付け加えていく方法を提案しているが，ここでは重みをかけることによってデータを取り除いていく．

まず 1 個めのクラスタを前節の方法で取り出す．次に 2 番めのクラスタ抽出ではデータから 1 番めのクラスタを取り除いて同じことをすればよい．各データ i が 1 番めのクラスタに所属するメンバシップを m_{1i} とするとデータの残存率は $1 - m_{1i}$ となり，これが 2 回めでのデータの重み v_i となる．そこで $v_i = 1 - m_{1i}$ として前節の結果を適用すると，行列 $S_2 = [s_{2ij}]$: $s_{2ij} = \sqrt{1 - m_{1i}}\sqrt{1 - m_{1j}} s_{ij}$ の第 1 固有ベクトルを $\boldsymbol{y}_2 = [y_{21}, \ldots, y_{2n}]$ とすると $x_{2i} = y_{2i} / \sqrt{1 - m_{1i}}$ となり (ただし第 1 クラスタの代表データ i_1^* では $v_{i_1^*} = 1 - m_{1i_1^*} = 0$ となるので前記のように $x_{2i_1^*} = 0$ とする)，この x_{2i} が最大となるデータ i_2^* が第 2 クラスタの代表データであり，$m_{2i} = x_{2i} / x_{2i_2^*}$ が各

データiの第2クラスタへのメンバシップ値となる．以下同様にして第kクラスタへのメンバシップは行列$S_k=[s_{kij}]$: $s_{kij}=\prod_{l=1}^{k-1}\sqrt{1-m_{li}}\sqrt{1-m_{lj}}s_{ij}$の固有ベクトル$\boldsymbol{y}_k$を求めれば，$x_{ki}=y_{ki}/\sqrt{\prod_{l=1}^{k-1}(1-m_{li})}$により$m_{ki}=x_{ki}/x_{ki_i^*}$と与えられる．

このようにして次々にクラスタを抽出していくと，抽出したクラスタの凝集度は単調に減少していく．なぜなら行列$S_k=[s_{kij}]$は非負行列であり，$s_{kij}=\prod_{l=1}^{k-1}\sqrt{1-m_{li}}\sqrt{1-m_{lj}}s_{ij}$からわかるように，各要素は抽出の度に単調に減少していく．すなわち，$0\leq s_{kij}\leq s_{k-1,ij}.\forall i,j$である．したがって，非負行列の性質により，固有値も単調に減少する．すなわち，S_kの固有値をλ_kとすると$0\leq\lambda_k\leq\lambda_{k-1}$である．$\lambda_k$が第$k$クラスタの凝集度であるので凝集度も単調に減少する．この凝集度の変化に基づいてクラスタ抽出の打ち切りを決めることができる．

2.5 実験例

2.5.1 2次元データを使った実験

図2.1のような2次元データに本方法を適用してみた．データiの2次元座標値を(x_i,y_i)とする．データiとjの類似度を$s_{ij}=e^{-\alpha[(x_i-x_j)^2+(y_i-y_j)^2]}$により計算した．$\alpha=100$とした．まず2.4節の逐次抽出を行ったときの固有値，すなわちクラスタの凝集度の変化を図2.2に示す．図から6番め以降のクラスタから凝集度の変化が緩やかになっているので，主要なクラスタ数は5個であると判断できる．なお，クラスタの数の決定は自動的に行うのが望ましく，例えば凝集度がある閾値よりも小さくなった時点で終了するという方法が考えられるが，この閾値の設定は一般には難しいので，本論文では上記のように主観的に決定した．図2.3に第1クラスタと第5クラスタのメンバシップ値を示す．他のクラスタのメンバシップ値も同様の形状である．このメンバシップ値に基づいて，しきい値を0.5としてクリスプなクラスタを抽出した結果を図2.4に示す．図中の番号は抽出の順番を表している．このように本手法はノイズにロバストである．

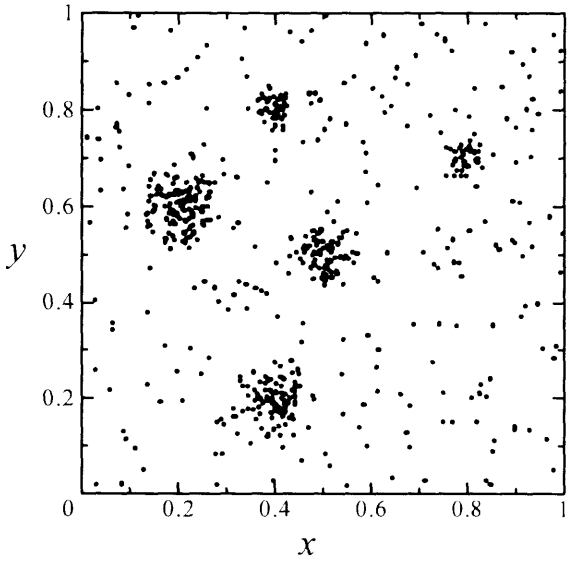

図 2.1: 2次元データの例

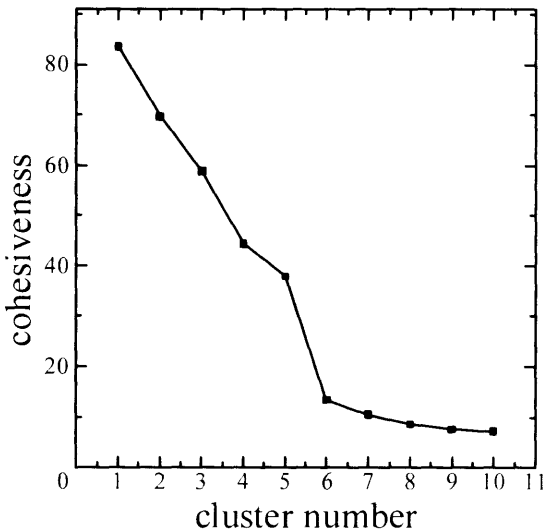

図 2.2: 凝集度の変化

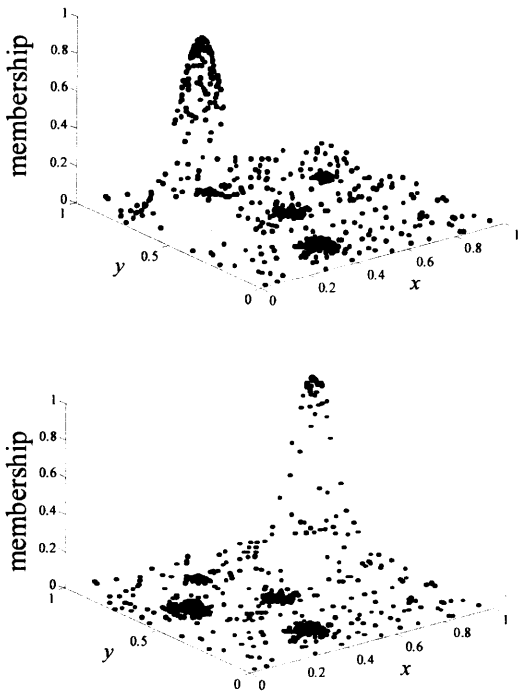

図 2.3: 第 1 クラスタ (上) と第 5 クラスタ (下) のメンバシップ値

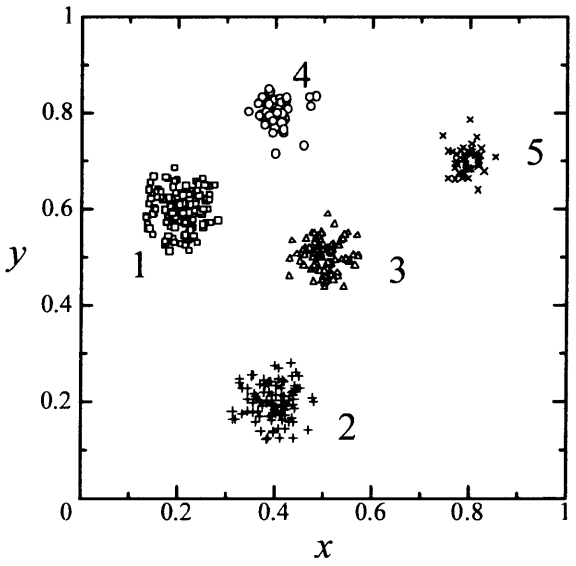

図 2.4: 抽出されたクラスタ

2.5.2 ビデオのセグメンテーション

実データの例としてビデオのフレーム画像をクラスタリングした．第 i フレームをデータ i とし，フレーム画像間の類似度 s_{ij} を次のようにして求めた．フレーム画像 i を n 色のカラーヒストグラム $\boldsymbol{h}_i = [h_{i1}, \ldots, h_{in}]^T$ で表し，色 p と q の類似度を $t_{pq} = \exp(-\beta \|\boldsymbol{L}_q - \boldsymbol{L}_p\|)$ として，フレーム i と j の類似度を $s_{ij} = \exp(-\alpha[\sum_p^n \sum_q^n (h_{ip} - h_{jq}) t_{pq} (h_{ip} - h_{jq})])$ とした．ここで $\boldsymbol{L}_p$ は色 p の Lab カラーベクトル $\boldsymbol{L}_p = [L_p^*, a_p^*, b_p^*]^T$ であり，β は定数である．実験には元のフレーム画像サイズが 160×120 ピクセル，フレーム数 437 のウィンドサーフィンのビデオを用いた．フレーム画像間の類似度を測るときにはサイズを 20×15 ピクセルに縮小し，色数 $n = 64$ に減色した．$\beta = 0.001$, $\alpha = 100$ とした．図 2.5 に凝集度の変化を示す．6 番目のクラスタからは凝集度の減少が比較的緩やかになっており，それ以後はあまりまとまったクラスタは抽出されていない．そこで，ある程度まとまったクラスタは 5 個であると判断した．この 5 個のクラスタへの各フレームのメンバシップ値を図 2.6 に示す．クラスタを色で区別し，抽出された順番に番号をつけている．図を見てわかるように，比較的メンバシップ値が大きな山状の部分 (ショット数) は 8 個あるが，提案手法では 5 個のクラスタを抽出するだけで 8 個のショットを得ることができる．ショットの境界でのメンバシップの変化を見ると，クラスタ 1 とクラスタ 4 の境界 (図中実線矢印) と，クラスタ 4 とクラスタ 5 の境界 (図中点線矢印) ではメンバシップは不連続に入れ替わっており，カットであることがわかる．その他のショットはすべて連続的である．各クラスタでメンバシップが最大となるフレーム画像を図 2.7 に示す．画像の下の数字はクラスタの番号である．このようにメンバシップによってショット切替りの種別やショットの繰返し構成を把握することができる．

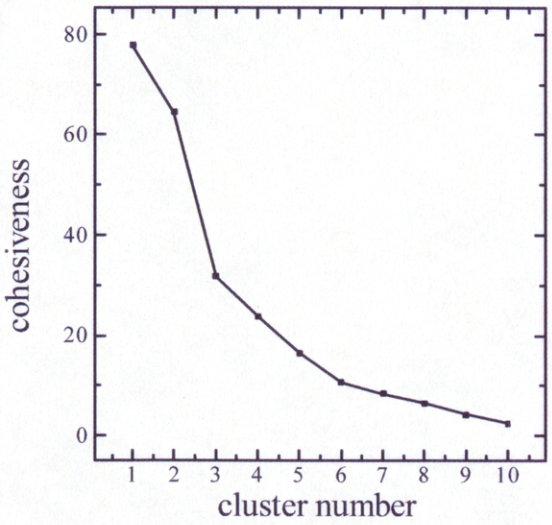

図 2.5: 凝集度の変化

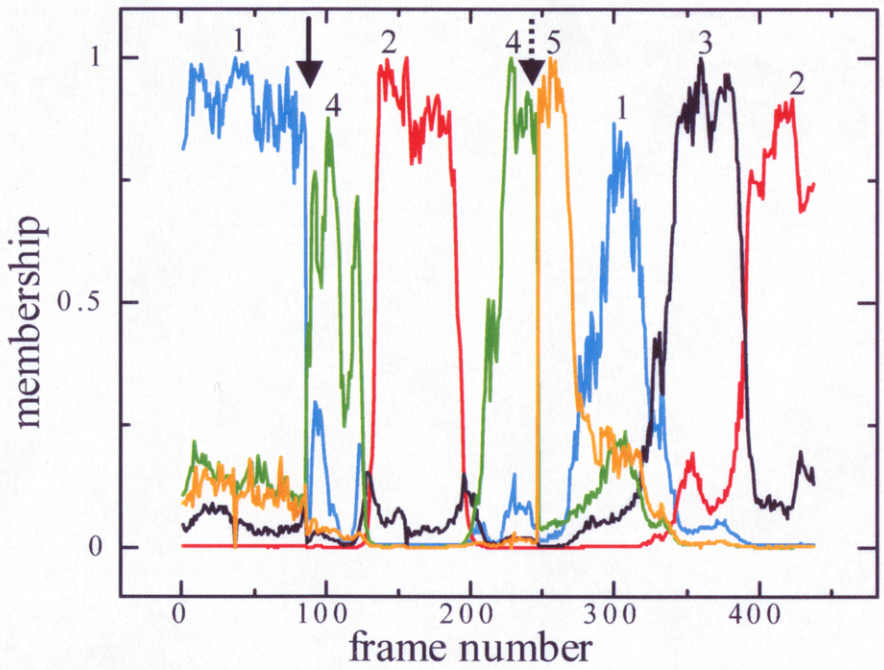

図 2.6: メンバシップ値

図 2.7: クラスタの代表フレーム

2.5.3 画像のセグメンテーション

重み付きデータの例として2.4節の逐次抽出を静止画のセグメンテーションに応用した．画像をヒストグラムで表し，ヒストグラムの各ビンをデータとする．ここでは256レベルのモノクロ画像を扱い，第iグレイレベルの画像中の頻度h_iを求めヒストグラムを構成した．これは1次元重み付きデータとなり，h_iがデータiの重みv_iとなる．ビンiのグレイレベルを256で割ったものをg_iとする．ビンiとjの類似度を$s_{ij} = e^{-\alpha(g_i - g_j)^2}$として計算した．$\alpha = 50$とした．図2.8にサイズ$202 \times 253$ピクセルの元画像とヒストグラムを示す．凝集度の変化は図2.9のようになり，4番目以降のクラスタからは凝集度はほとんど変化していないのでクラスタの個数を3個とした．図2.10にセグメンテーションされた画像とクラスタ抽出後のヒストグラムを示す．各領域のグレイレベルは等間隔に付けており，元の画像のグレイレベルとは直接関係は無い．

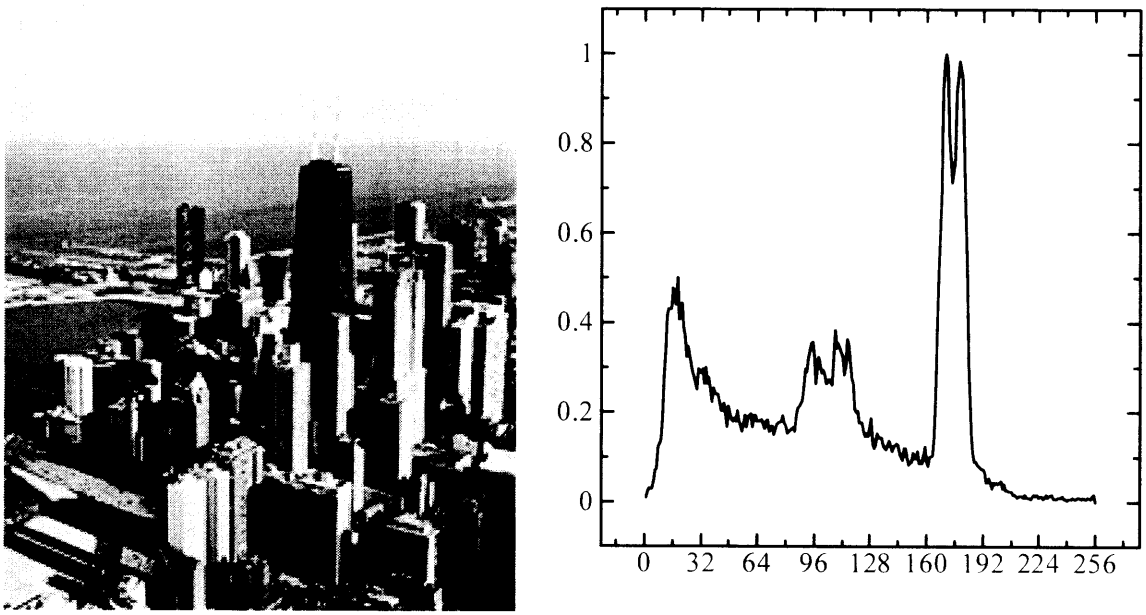

図 2.8: 元画像 (左) とヒストグラム (右)

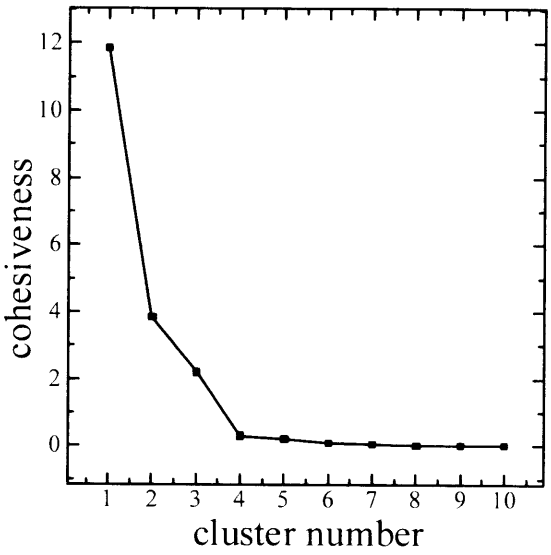

図 2.9: 凝集度の変化

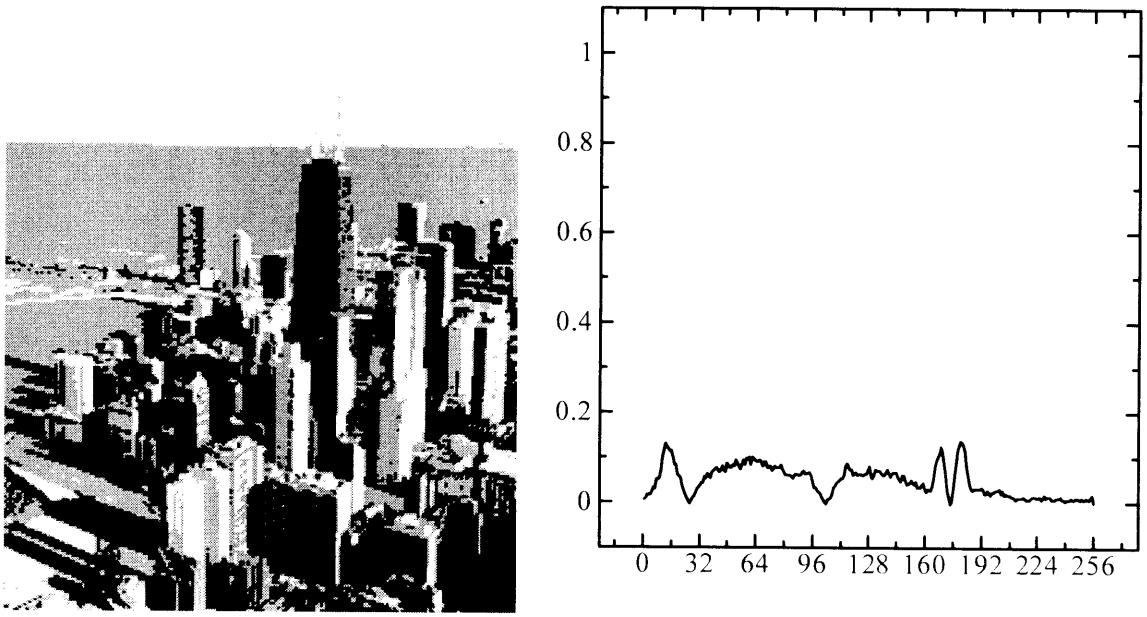

図 2.10: セグメンテーション結果 (左) とクラスタ抽出後のヒストグラム (右)

2.6 むすび

データの類似度行列に基づくファジークラスタリング法を提案して実験で性能を検証した．本方法の特長は，1) データが特徴ベクトルで与えられる必要がなく類似度でよいこと，2) クラスタ数を凝集度の変化に基づいて推定できること，3) ファジーであるためノイズにロバストであること，4) 逐次抽出のとき主要なクラスタから順に抽出されるのでクラスタの主要順位がわかること，などである．