

パターン認識による臨床検査データの判読

奈良県立医科大学中央臨床検査部

岡本 康幸

INTERPRETATION OF CLINICAL LABORATORY DATA BY PATTERN RECOGNITION

YASUYUKI OKAMOTO

Central Clinical Laboratory, Nara Medical University

Received October 5, 2009

Abstract : 臨床検査データの解釈にパターン認識の手法を用いることによって、コンピュータによる自動判読や、パターンの中に隠された新たな病態情報を抽出することが期待される。この目的のために筆者がこれまで行ってきた試みについて紹介する。内容は、数値データセットについてのパターン類似性を用いたマッチング検索、ニューラルネットによる学習を用いた診断支援、波形データについてのゆらぎ解析による機能評価、画像データについての複雑性を指標とした質的評価の定量化、である。

Key words : pattern recognition, neural network, detrended fluctuation analysis, image analysis, fractal

1 はじめに

パターンとして提示される臨床検査データは少なくない。心電図や脳波のような時系列波形データ、そして細胞の検鏡所見などの画像データは言うまでもなく、数値データも複数項目のデータセットによって生体の生理的または病的状態に対応するパターンを表現する。検査データのパターンは、様々な特徴の集まりに再構成され、それらの特徴から臨床的に意味のある情報を取り出すことが可能であると考えられる。通常、この作業プロセスは、知識・経験の学習結果として判読者に記憶されている。もしこのプロセスを客観的に記述することができれば、すなわち、パターンから特徴を量的パラメーターとして抽出し、適切なルールにあてはめることによって意味別に分類することができれば、コンピュータを用いて自動でデータ判読することが可能である。さらに、判読者にも認識できないような新たな特徴から隠れた情報を検出できれば、より優れた診断能力を発揮することも期待される。

本稿では、パターン認識に基づく検査データ判読の試みについて、筆者がこれまで検討を行ってきた成績を中

心に紹介する。その前に、パターン認識の手法についての概説を加える。

2 パターン認識の手法¹⁾

パターン認識のシステムを図1に示す。アナログデータは量子化・標本化(サンプリング)によりデジタル変換を行う。得られた一群のデジタル情報(数値データのセット、時系列波形・画像の生データ)が表すパターンから抽出した特徴量・特徴ベクトルを分類規則や関数などに入力し、出力されたクラスにより評価を行う。評価には定性的な識別と定量的な回帰分析がある。特徴量としては、前処理・正規化されただけのデータやデータ間の関係を示す計算値(差や比など)のような単純なものから多変量解析的な主成分や因子、周波数領域でのパワー分布や複雑性の指標など様々なものが利用される。いずれにしても、評価に最適な特徴量を選択することが重要である。

評価方法は、“教師あり”(評価基準に対応するようにクラスを出力する)と“教師なし”(出力されたクラスに評価基準を付与する)の2法に分かれるが、さらにそれぞれ様々な手法が知られている。“教師あり”の方法では、テ

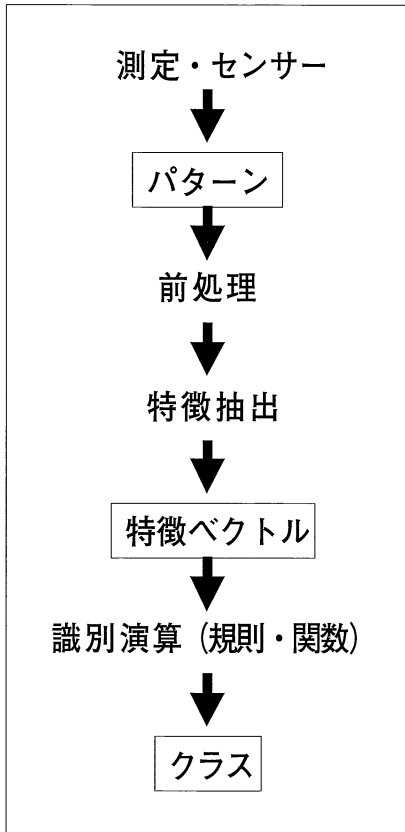


図 1. パターン認識のシステム

ンプレートやプロトタイプとの類似性による識別法や学習により得られた識別関数を用いる方法があり、識別関数として線形識別関数、ニューラルネット、サポートベクターマシンなどがある。“教師なし”の方法ではクラスター分析や競合学習などがあるが、評価基準について発見的であることが特長である。

3 数値データセットのパターン認識

1) パターンマッチングによる症例検索

事例ベース推論(Case-based reasoning, CBR)²⁾の1手法として、検査データセットが表現するパターンを用いて、過去の症例データベースから得られるデータセットとの類似性を評価し、診断支援や精度管理に利用することを試みた。この際、検査項目によってデータの値域・分布が異なるため、対数正規分布型を示す項目(血清酵素、トリグリセリドなど)は値を対数変換したうえで、基準範囲がすべて 1.0 ± 0.2 となるように標準偏差(SD)を用いて正規化した。これはSDを0.1とし、0.8～1.2を基準

範囲とする1種のSDI(Standard deviation index)³⁾表現となる。

類似性の評価は、正規化データセットを特徴ベクトルとみなし、ベクトル間の相関係数または距離(ユークリッド距離など)を評点とするパターンマッチングにより行った。しかし、検査項目には異常値が出やすいもの(血糖、トリグリセリドなど)と異常値となることが稀なもの(抱合型ビリルビン、Naなど)とがあり、それらを等価には扱えないため、項目ごとのデータへの重み付けが必要となる。そこで、データベースでの異常値の出現頻度を算出して項目ごとのデータへの係数を設定した。すなわち、データを異常値出現頻度で割ることによって、異常値の出やすい項目の評価への寄与を少なくし、稀な異常値が重視されるようにした。

表1は15,368件のデータベースの中から検索症例(肝不全を伴う肝硬変)と高い相関係数を示した上位5例を抽出した成績であり、類似症例が選択されていることが分かる。肝硬変は頻度の高い疾患であるが、検索症例が稀なものであれば、データベースから抽出できる確率が低下し、類似症例のパターンが元々データベースに含まれていない可能性も高くなることが予測され、これが本法の限界といえる。ただし、多くの施設で構築したデータベースをネットワーク化することによって、稀な症例にも対応できるようになるかもしれない。また、データベースのパターンをいくつかのプロトタイプに集約できれば、より効率的な検索が可能になると考えられる。

本法では、すべてのデータが基準範囲内にあっても別人同士での相関は低く、一方、異常値があっても同一人では高い相関が認められており、データパターンの比較による個人の識別が可能である。したがって、時系列での相関の変動を観察することにより、病態の急激な変化やあるいは検体取り違いの検出にも応用できると考えられる。

2) ニューラルネットによる診断支援

代表的な人工ニューラルネットとして知られている誤差逆伝播学習アルゴリズム(Back propagation algorithm, BP)⁴⁾は、データベースと評価基準との関係を教師あり学習により獲得するためのツールとして、広く用いられている。学習されたネットワークは、一種の非線形識別関数として、新たなデータセットに解釈を与えることが可能となる。ネットワークは、複数のニューロンを持つ入力層・中間層・出力層からなる多層パーセプトロンモデルによる階層型で、後方層(中間層または出力層)の各ニューロンへの入力値となる前方層(入力層また

表 1. パターンマッチングによる類似症例検索

症例	病名	TB	ZTT	ALP	AST	ALT	LD	γGTP	TP	Alb	BUN	UA	CRE	CRP
検索症例		1.2	13.9	238	100	98	365	172	6.8	4	11	5.2	0.6	0.1
1位抽出例	肝硬変	1.1	15.8	253	116	109	393	58	7.4	3.9	12	4.1	0.5	0.1
2位抽出例	肝硬変、肝癌	1.1	13	308	69	91	345	98	6.4	3.6	15	6	0.7	0.1
3位抽出例	慢性肝炎	0.8	14.4	226	82	86	400	40	6.8	3.8	13	4.6	0.5	0.1
4位抽出例	肝硬変、肝不全	1.2	18.3	248	86	96	317	195	7.1	3.5	11	5	0.7	0.1
5位抽出例	慢性肝炎	0.6	13.4	193	100	137	305	148	7.2	4	11	4	0.6	0

症例	病名	TC	TG	Na	K	CL	Ca	GLU	RBC	Ht	Hb	WBC	PLT
検索症例		160	95	142	4	103	8.7	87	434	40.9	14.4	49	8.4
1位抽出例	肝硬変	157	179	141	4.1	104	9.6	100	421	38.4	12.1	58	11.8
2位抽出例	肝硬変、肝癌	156	109	141	3.9	106	8.7	83	425	42.4	14.4	37	5.4
3位抽出例	慢性肝炎	153	97	141	3.9	107	9	91	416	38.1	12.8	38	5.9
4位抽出例	肝硬変、肝不全	148	63	140	3.5	105	9	92	418	40.1	14	56	11.2
5位抽出例	慢性肝炎	191	115	141	3.8	104	8.7	90	456	41.8	14.5	24	9.7

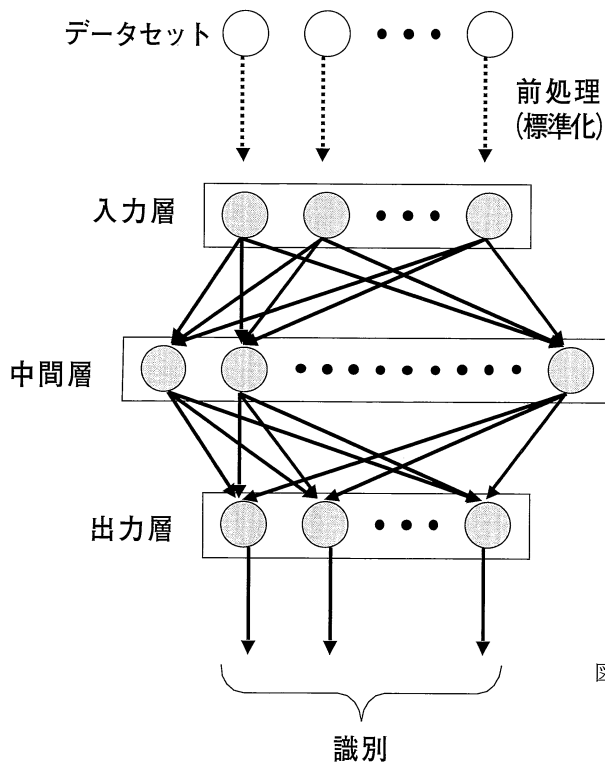


図 2. ニューラルネットワーク
 入力層・中間層・出力層からなる多層パーセプトロンモデルによる階層型ニューラルネットワークの構造

は中間層)のニューロン間結合に重み付けパラメーターが設定されている(図2)。通常、ニューロン間の結合にはロジスティック関数(下式)が用いられる。

$$\frac{1}{1+e^{-\left(\sum_{i=1}^N W_i X_i - \theta\right)}}$$

ここで N はニューロンの数, X_i はニューロンの出力値, W_i は重み付けパラメーター値, θ は閾値である。

BP は, 出力層ニューロンへの出力値と評価基準との誤差が最小となるようにすべての重み付けパラメーターを

修正(学習)する方法で, 誤差をパラメーターで微分した値を修正項に含んだパラメーター更新式でパラメーター値を収束させていくようにコンピュータを用いて反復計算する。

本法を用いて, 種々の慢性肝疾患の検査データセットと肝臓の病理組織診断との関係を学習し, 新たな症例のデータセットによる組織診断の推定を行った成績を示す⁹⁾。ネットワークの入力層には基準範囲で標準化した14項目の検査データセット(性別, 年齢, TB, ZTT, ALP, AST, ALT, ChE, γ GTP, 総胆汁酸, TP, Alb, A/G

表 2. ニューラルネットによる慢性肝疾患の病理組織診断推定

病理組織診断	例数	ニューラルネットによる診断					正診率
		肝硬変	慢性肝炎活動型	慢性肝炎非活動型	脂肪肝	判定不能	
肝硬変	3	3					100%
慢性肝炎活動型	15	2	9	2	1	1	60%
慢性肝炎非活動型	4			3	1		75%
脂肪肝	2		1		1		50%

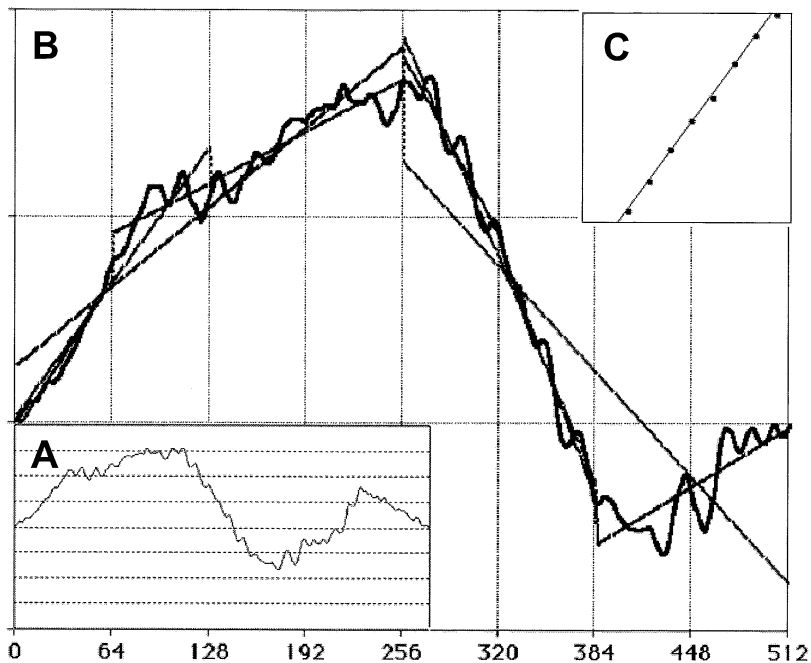


図 3.トレンド除去ゆらぎ解析(DFA)

A: 積分時系列データ, B: 区間サイズごとのトレンド, C: 区間サイズとトレンド除去後変動の標準偏差との関係(両対数プロット)

比, TC)を入力し, 中間層と出力層のニューロン数はそれぞれ9および4とし, 4種の慢性肝疾患(肝硬変17例, 慢性肝炎活動型40例および非活動型20例, 脂肪肝11例)の組織診断を評価基準として出力値との誤差を最小とするように学習を行った. 試験症例として, 肝硬変3例, 慢性肝炎活動型15例および非活動型4例, 脂肪肝2例を用いて組織診断との一致率を検討したところ(表2), 全体で66.7%の正診率が得られた. なお, 同じ試験症例のデータに対する肝臓病の専門医20名の正診率は20.8~62.5%であった.

ニューラルネットは汎用性の高いシステムとして既に多くの実績があるが, 本法が有効であるためには, 識別可能性のあるテーマを設定することと識別に十分な症例を含んだデータベースを用いることが重要である. また, ネットワークの構成・計算のパラメーターなどにも工夫が要求される. しかし, 本法はほとんどの識別問題に適用が可能であることとその正診率の高さによって利便性が高く, 今後とも幅広く利用されていくことは間違いのない.

4 波形データのパターン認識

— 脳波のゆらぎ解析による精神作業の検出 —

時系列波形データの変動はトレンド成分, 周期的成分およびゆらぎ成分(ノイズ)を含む. 生体信号におけるこれらの成分の解析は生理機能の評価に応用されており, 例えば, 心拍変動(Heart rate variability, HRV)⁶⁾での特定の周波数成分およびゆらぎ成分の解析は自律神経機能・ストレスの評価に用いられており, 脳波のゆらぎ成分の解析によって感性や心理的側面の評価を行うことも試みられている⁷⁾. 多くの生体信号の時系列変動でのゆらぎは, 時系列での長期相関性を有しており, 自己アフィ

ン性フラクタルの性質(自己相似性)を示す. このフラクタル性の指標がストレスや感性・心理的側面を表す特徴量として意味を持つと考えられる.

トレンド除去ゆらぎ解析(Detrended fluctuation analysis, DFA)⁸⁾は, フラクショナル・ブラウン運動(自己アフィンの性質を有する過程)におけるゆらぎの長期相関性を解析する方法として知られている. 時系列データでトレンド(非定常な変動成分)を除いたゆらぎのフラクタル性指標を求め, 心拍変動や脳波などの生理機能評価に応用されてきているが^{9),10)}, 心理的側面の評価にも応用できる可能性がある. そこで, 自発的精神作業時に記録した脳波を用いて, DFAによる心理的变化の検出を試みた¹¹⁾.

方法としては, 脳波記録中, 約1分間の間隔を挟んで, 1分間の暗算(2から始まって数値二倍の連続計算)と1分間の仮想描画(風景の絵を描く)を実施させ, それぞれの作業中とその前後の計5エポックで記録された約5秒間(1024ポイント)の波形生データをDFAに供した. DFAは, 以下の手順で実施した. ①データ X_k を平均値との差により正規化し, 各ポイントまでの積分時系列データ Y に変換(下式, 図3A).

$$Y(i) = \sum_{k=1}^i (X_k - \bar{X}) \quad (i=1 \dots N)$$

②全時点数 N をサイズ n の区間に分割し, 区間ごとのトレンド Y_n を最小二乗法により算出後, トレンドを除いたデータ変動の標準偏差を求めて $F(n)$ とし, 変数 n について $F(n)$ 配列を得る(下式, 図3B).

$$F(n) = \sqrt{\frac{1}{N} \sum_{i=1}^N [Y(i) - Y_n(i)]^2} \quad (0 < n < N)$$

③ n と $F(n)$ を両対数グラフにプロットした傾きをスケーリング指数 α として算出(図3C). 得られたスケーリング

表3. 脳波の部位別スケーリング指数 α

Left side		Right side	
	Fp ₁ : 1.08±0.13	Fp ₂ : 1.08±0.15	
F ₇ : 1.11±0.19	F ₃ : 0.99±0.09	F ₄ : 0.98±0.11	F ₈ : 1.03±0.16
T ₃ : 0.99±0.18	C ₃ : 0.97±0.09	C ₄ : 1.00±0.12	T ₄ : 1.00±0.16
T ₅ : 0.95±0.10	P ₃ : 1.00±0.08	P ₄ : 0.96±0.09	T ₆ : 0.97±0.08
	O ₁ : 0.94±0.09	O ₂ : 0.92±0.09	

平均±SD (n=10)

電極は国際式10/20法に従って, frontal pole (Fp), frontal (F), central (C), parietal (P), occipital (O) に配置

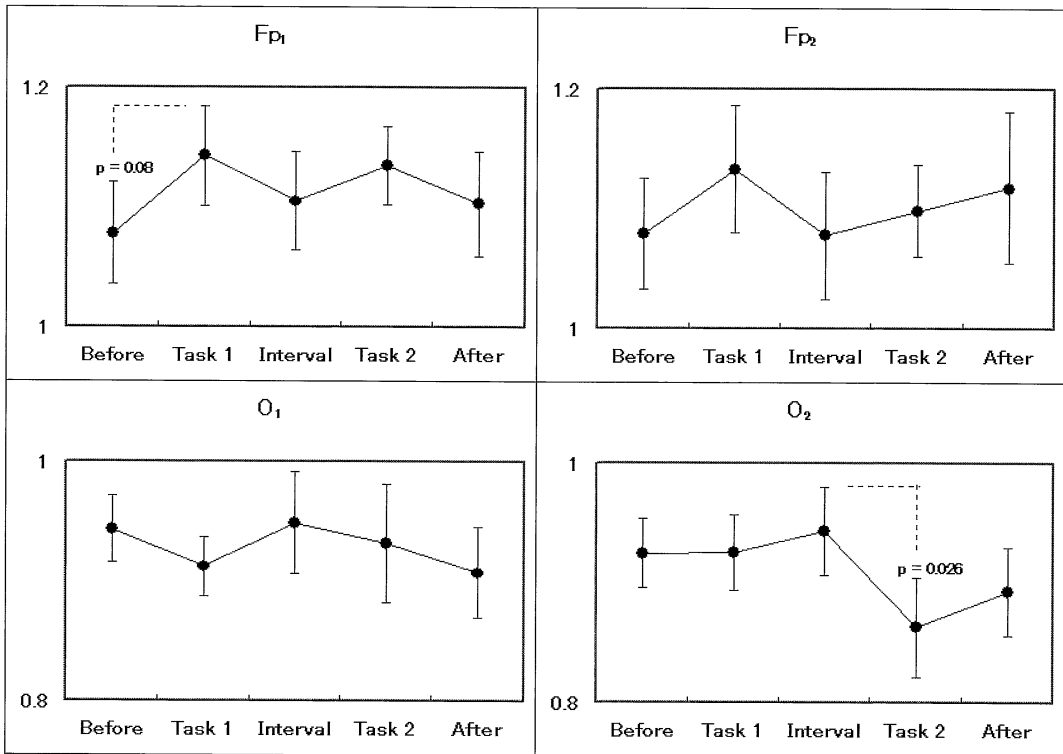


図 4. 精神作業に伴う DFA スケーリング指数 α の変化
 平均 \pm SD (n=10)
 Task 1: 暗算、Task 2: 仮想描画

指数を特徴量として、その変化を検討した。

その結果、作業前脳波のいずれの電極でも α はほぼ 1 前後の値を示していたが、前頭極では高めの値、後頭では低めの値となり、部位による機能の質的な差が表現されていると考えられた(表 3)。また、暗算中は F_{P1} の α 値が上昇する傾向を示し、一方、仮想描画中では O_2 の α 値が有意に低下していたことから、作業内容の違いによって、部位特異的に変化の方向に違いのあることが示唆された(図 4)。これらの意義は不明であるが、暗算は左脳の前頭部での、仮想描画は右脳の後頭部での、それぞれ異なる方向への機能変化として捉えることが可能かもしれない。

時系列波形データのゆらぎから得られる特徴量としては、フラクタル次元 D_d やパワースペクトル密度から得られるスペクトル指数 β などがあり、 $\beta = 2\alpha - 1 = 5 - 2D_d$ の関係が成立し得る。 $\beta = 1$ ($\alpha = 1$) の時がいわゆる f^1 ゆらぎに相当する。脳波の各部位からこのような特徴量を抽出することによって、脳の機能的異常の診断や¹²⁾、快・不快などの感性の評価への応用が検討されている¹³⁾。

5 画像データのパターン認識

— フラクタル次元による画像解析 —

前項でも触れたが、フラクタル次元は、自己相似性のパラメーターで、パターンの複雑さを数値的に表現するものとして用いることができる。例えば、平面上の滑らかな線(一次元)から全体を埋め尽くす面(二次元)までの範囲内で、曲折のある線が面を覆う程度を非整数の次元として表現できる。とくに生物学的な複雑さにはフラクタルの性質を持つものが多いことが知られており、フラクタル次元によりその複雑さの背景にある生物学的特徴を定量的に表現できる可能性がある¹⁴⁾。ここでは、塗沫染色されたリンパ球のデジタル画像を用いて、その細胞核構造のフラクタル次元値を測定し、細胞鑑別への応用を検討した成績を紹介する¹⁵⁾。

方法は、まず血液塗沫標本のデジタル画像を 256 レベルの grayscale 画像(図 5A)に変換し、grayscale のレベルごとに 2 値化および細線化処理を行った後、Box-counting 法¹⁶⁾により順次フラクタル次元値を求めていっ

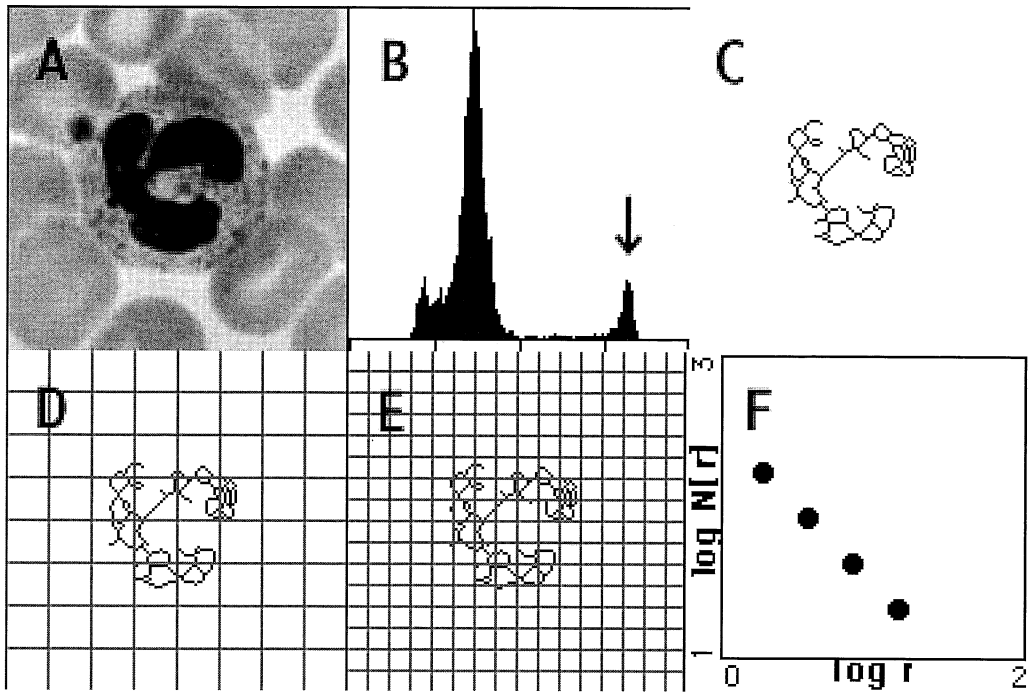


図5. Box-counting 法によるフラクタル次元の測定

A: サンプル画像(分葉核好中球)、B: grayscale レベルのヒストグラムと核部分のピーク(矢印)、C: 核部分の細線化処理、D および E: 異なる Box サイズの背景、F: Box サイズと画像含有 Box 総数との関係(両対数プロット)

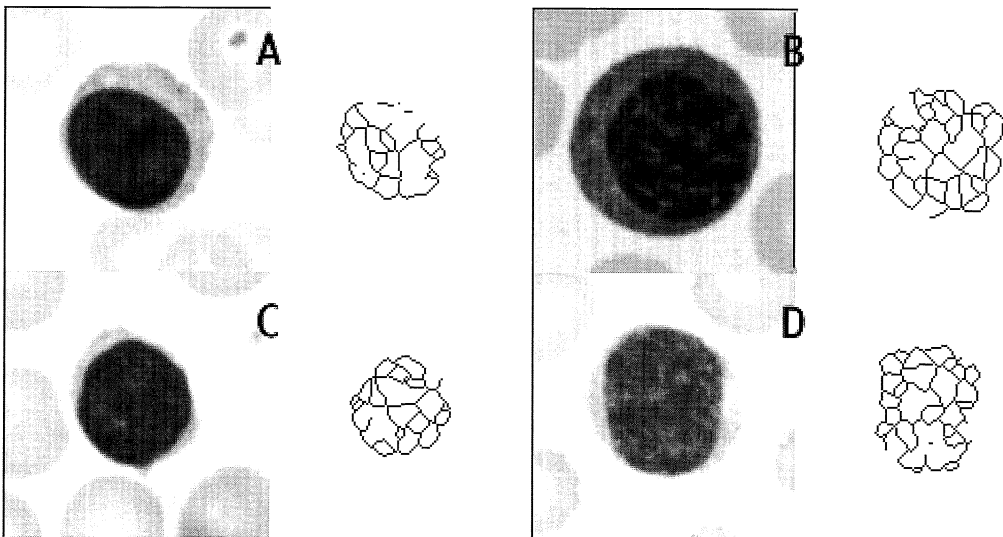


図6. リンパ球像(左)およびその核部分の細線化処理像(右)

A: 正常リンパ球、B: 異型性リンパ球、C および D: 慢性リンパ性白血病における腫瘍性リンパ球

た。ここで、濃度ヒストグラム(図 5B の矢印)が示すように、細胞核に対応する部分画像は濃度の高いレベルでの分画として分離可能であるため、その分画の中で得られた最も高い次元値を代表値として採用した。サンプル像(分葉核好中球)を 2 値化および細線化処理した後、最も高い次元値を示した時の図を図 5C に示す。

Box-counting 法による次元値(以下、Db 値)は、画像含有の有無を判定する枠の幅(grid の pixel 数=Box の一辺= r)と r において計測される画像含有 Box 総数(画像を捕える Box の数= $N[r]$)との関係 $\log N[r]/\log r$ を異なるサイズの r について求め、plot した時の傾きの絶対値で表現される。図 5D, E に、それぞれ 2 種類の r (16, 8pixel) に対応する Box 像を示し、図 5F に 4 段階で($\log r, \log N[r]$) を plot した図を示す。

以上に述べた方法で、正常リンパ球、異型リンパ球および慢性リンパ性白血病(CLL)での腫瘍性リンパ球のそれぞれについて Db 値を測定した。それぞれの代表的な細胞像と核の細線化像をペアにして図 6 に示す。それぞれの Db 値は、正常リンパ球が 1.18(図 6A)、異型リンパ球が 1.29(図 6B)、そして CLL のリンパ球が 1.28(図 6C)と 1.31(図 6D)であった。さらに多数例を用いた成績での Db 値の平均は、正常リンパ球では 1.16 ~ 1.18、異型リンパ球では 1.29 ~ 1.31、CLL のリンパ球では 1.23 ~ 1.31 であり、異型リンパ球および CLL のリンパ球で高い Db 値が認められている。

図 6C の細胞のように、CLL のリンパ球は比較的小型で核構造にも明らかな異型性が認められないことがあり、この場合形態的には正常リンパ球との判別も困難となる。しかし、Db 値を測定することによって、構造的差異の存在を示唆する有力な情報が得られる可能性があると考えられる。フラクタル次元は一つの特徴量であるが、他の複数の特徴量を組み合わせることによってさらに詳細な細胞鑑別システムを構築することが期待される。

6 おわりに

本稿では、臨床検査データ判読のために有用と考えられるパターン認識の手法と、それによって得られる特徴量を用いて、隠された病態情報を読み取る試みについて述べた。本法により得られた成績を臨床検査データの報告に付加することによって、より詳細な診断の支援に寄与することが期待される。

文 献

- 1) 麻生英樹：パターン認識と学習。in 統計科学のフロンティア 6. パターン認識と学習の統計学。(甘利俊一, 竹内啓, 竹村彰通, 伊庭幸人, 編). 岩波書店, 東京, pp.1-95, 2003.
- 2) Dussart, C., Pommier, P., Siranyan, V., Grelaud, G. and Dussart, S. : Optimizing clinical practice with case-based reasoning approach. *J. Eval. Clin. Pract.* 14 : 718-720, 2008.
- 3) Foulis, P.R., Norbut, A.M., Losos, F.J.3rd., Kessler and G.F.Jr. : Quality-control statistical interpretation by microcomputer. *Am. J. Clin. Pathol.* 74 : 636-644, 1980.
- 4) Rumelhart, D.E., Hinton, G.E. and Williams, R.J. : Learning representations by back-propagation errors. *Nature* 323 : 533-536, 1986.
- 5) 岡本康幸, 中野博, 吉川正英, 松岡弘樹, 阪本たけみ, 辻井正：人工ニューラルネットワークを用いた臨床検査診断支援システムに関する研究 —肝機能検査による肝疾患組織診断推測の試み—。臨床病理 42 : 195-199, 1994.
- 6) Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology: Heart rate variability. Standards of measurement, physiological interpretation, and clinical use. *Eur. Heart J.* 17 : 354-381, 1996.
- 7) Natarajan, K., Acharya, U.R., Alias, F., Tiboleng, T. and Puthusserypady, S.K. : Nonlinear analysis of EEG signals at different mental states. *Biomed. Eng. Online.* 16 : 7, 2004.
- 8) Peng, C-K., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Sciortino, F., Simons, M. and Stanley, H.E. : Long-range correlations in nucleotide sequences. *Nature* 356 : 168-170, 1992.
- 9) Peng, C-K., Havlin, S., Stanley, H.E. and Goldberger, A.L. : Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. *Chaos* 5 : 82-87, 1995.
- 10) Lee, J-S., Yang, B-H., Lee, J-H., Choi, J-H., Choi, I-G. and Kim, S-B. : Detrended fluctuation analysis of resting EEG in depressed outpatients and healthy controls. *Clin. Neurophysiol.* 118 : 2489-2496, 2007.
- 11) 今井竜子, 岡本康幸：脳波のトレンド除去ゆらぎ解析(DFA)による精神作業の検出。臨床病理 56:383-386, 2008.

- 1) 麻生英樹：パターン認識と学習。in 統計科学のフロンティア 6. パターン認識と学習の統計学。(甘利俊

- 12) **Stam, C.J., Montez, T., Jones, B.F., Rombouts, S.A.R.B., van der Made, Y., Pijnenburg, Y.A.L. and Scheltens, Ph.** : Disturbed fluctuations of resting state EEG synchronization in Alzheimer's disease. *Clin. Neurophysiol.* **116** : 708-715, 2005.
- 13) **飯塚拓也, 中川匡弘** : 脳波のフラクタル次元解析と感性処理への応用. 電子情報通信学会技術研究報告 . **104** : 53-58, 2005.
- 14) **Cross, S.S.** : Fractals in pathology. *J. Pathol.* **182** : 1-8, 1997.
- 15) **丹羽欣正, 吉田崇, 森分和也, 岡本貴史, 倉本智津子, 山崎英子, 松田多恵子, 増谷喬之, 岡本康幸** : デジタル画像を用いたリンパ球核構造のフラクタル次元の測定. *臨床病理* **50** : 702-705, 2002.
- 16) **Falconer, K.** : *Fractal Geometry. Mathematical Foundations and Applications.* 2nd ed., Wiley, NewYork, pp.41-47, 2003.