

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

統計科学のフロンティア 6

甘利俊一 竹内啓 竹村彰通 伊庭幸人 編

パターン認識と  
学習の統計学

新しい概念と手法

麻生英樹 津田宏治 村田昇

岩波書店

## 編集にあたって

# パターン認識と学習——統計学の手法の新展開

与えられた対象をデータをもとに分類することはよく見られる作業である。データは視覚や聴覚的なパターンで与えられることが多いから、この問題はパターン認識と総称され、人間には比較的容易にできるのにコンピュータに行わせるのはたいへん難しい問題として、注目されてきた。

データやパターンは統計的な搖らぎを伴うから、パターン認識は統計学の問題でもある。統計学では古くから判別分析などの手法が開発されてきたが、多変量解析というガウス分布を暗黙のうちに想定する枠の中であったため、文字認識、音声認識などの現実の問題にそのまま適用できなかった。

文字認識や音声認識は、コンピュータの発展とともに、パターンとしての独自の構造に着目した手法が開発されてきたが、一方ではより一般的な認識の手法への要望も高まってきた。これに一つのきっかけを与えたのがいわゆるニューラルネットワークの手法である。汎用の人工ニューラルネットワークをもとに、例題を多数与えて学習させ、パターン認識問題を解こうという発想である。この手法は、例題は利用できるもののどういう仕組みで識別のアルゴリズムを構築したらよいかわからない多くの問題に、光明をもたらすものであった。

パターンやデータには搖らぎが付きものであるから、ニューラルネットワークの手法も当然統計学の対象である。これは、統計学にガウス性を想定しない非線形の手段と学習という逐次的な手法をもたらし、新しい流れを作った。さらに、非線形性と学習を利用する流れは統計学だけで閉じたものではなくて、人工知能、情報理論、脳のモデル、制御理論、統計物理学などを巻き込んだ新しい分野をなしている。この中で統計科学の果たす役割は大きい。

本巻では、第Ⅰ部に麻生英樹氏によるパターン認識の多様な手法の紹介を置く。ここではこうした手法が統計科学における確率モデルとどのように結びつき、またその理論的な基礎がどう確立されているのか、その根拠

をわかりやすくかつ統一的な視点から述べている。これにより、パターン認識や学習といった情報にかかる現代技術の多様な広がりと、その中の統計科学の果たす中心的な役割が理解いただけるものと思う。

パターン認識は、構造が複雑であるから線形の手法でこなすわけにはいかない。しかしがニューラルネットワークなどの非線形の手法は局所解に落ち込み、最適な解が必ずしも得られないなど、問題点も多い。そこで登場したのがサポートベクトル機械(SVM)である。これはパターンを多次元の空間に非線形に埋め込むことによって、その後は線形の手法で識別を行うものである。カーネル法はこの埋め込みを自動的に行う巧妙な仕掛けであって、これによって SVM が脚光を浴びた。第Ⅱ部は津田宏治氏が SVM とカーネル法、さらにその適用範囲の拡大を説明している。

パターン認識は仕組みが複雑で一筋縄でゆくものではない。どの手法がよいかは問題による。それでは、いい加減な識別装置を学習によって多数作っておいて、その知恵を集めて理想的な識別装置が作れないものだろうか。通常の統計的推論では、衆愚を集めるというこのような手法は、考え抜いた最適な手法にかなうはずがなく、問題にならない。しかし、ニューラルネットワークをはじめ非線形を用いる手法では、局所解が多数あってそこに落ち込むという問題点がある。これを解決しようというのがブースティングという、多数の弱解を学習によって作り出し、それを統合してよい識別方法を生み出すという手法である。村田昇氏は、その原理と実際を情報幾何にまで根拠を拡大しながら解説している。これが第Ⅲ部である。

本巻によって、情報の諸分野と協調融合しながら発展していく新しい統計科学の姿を見ていただければ幸いである。

(甘利俊一)

# 目 次

編集にあたって

第Ⅰ部 パターン認識と学習 統計科学からの展望	麻生英樹	1
第Ⅱ部 カーネル法の理論と実際	津田宏治	97
第Ⅲ部 推定量を組み合わせる バギングとブースティング	村田昇	139
索引	223	

裝丁 蟻名優子

# I

---

## パターン認識と学習 統計科学からの展望

麻生英樹

## 目 次

1	パターン認識と統計科学	4
1.1	パターン情報処理	4
1.2	パターン認識システム	6
1.3	統計的パターン認識	10
2	いろいろなパターン識別手法	12
2.1	テンプレートマッチング法	12
2.2	$k$ -最近傍識別法	14
2.3	部分空間法	16
2.4	識別関数の最適化による方法	17
2.5	決定木による方法	20
2.6	ニューラルネットワークによる方法	23
2.7	識別関数の評価	31
3	統計的意思決定としてのパターン識別	33
3.1	パターン生成過程のモデル	33
3.2	損失関数	35
3.3	事後確率最大化識別	36
3.4	多次元正規分布による推定	38
3.5	判別分析	40
3.6	ノンパラメトリックなクラス分布の推定	41
3.7	確率分布モデルとしての決定木やニューラルネットワーク	44
3.8	グラフィカルモデルとナイーブベイズ識別	45
3.9	パターン認識と統計的モデル選択	54
4	クラスタリングとベクトル量子化	57
4.1	ボトムアップとトップダウンのクラスタリング法	58
4.2	$K$ -平均法	60
4.3	競合学習による方法	61
4.4	混合分布による方法	61
4.5	クラスタリング結果の評価	64
5	時系列パターン情報の認識	66
5.1	時系列パターン情報のモデル	66
5.2	音声認識	70
6	学習と統計科学	72
6.1	機械学習	73
6.2	統計的学習理論	74
6.3	経験損失と期待損失	77
6.4	経験損失最小化	80
6.5	構造的損失最小化	83
6.6	真の分布とヒューリスティクス	84
6.7	強化学習の理論	86



通信のための情報理論以来、統計科学と情報処理技術、情報通信技術とは不可分の関係にあるが、その中でも、統計科学との関係が深い分野の1つとして、パターン認識と機械学習があげられる。

近年、情報ネットワークの普及によって、テキストや音声、画像をはじめとして、多種多様で大量のデータが電子的にネットワーク上に蓄積されるようになっている。こうした中で、パターン認識技術の中心は、音声、画像、といった1つのモダリティでの認識率や検索精度を競う技術から、複数のモダリティを組み合わせた、構造をもつ、さらに複雑な情報を扱う技術へとシフトしてきている。また、機械学習技術は、大量のデータからの潜在的な関係の抽出を目指すデータマイニングや、ネットワーク上に分散しているエージェントによる学習、などへと広がってきている。

このような社会的ニーズを受けて、パターン情報処理技術や機械学習技術は、グラフ構造や隠れ変数をもつ確率分布のような複雑な確率分布モデル、オンライン学習のようなデータ利用、あるいは、複数の学習者によるインタラクティブな学習、などの新規な技術シーズの実験場となり、統計科学のダイナミックな展開の原動力の1つとなっている。

1章から5章では、パターン認識技術を統計科学の観点から展望する。とくに、パターン認識の問題を統計的決定理論によって定式化することを通じて、研究の歴史の中で提案してきたさまざまなパターン識別手法が、複雑な形状のデータの分布を捉えるための工夫であることを示す。さらに、6章では、機械学習技術の理論基盤の1つである統計的学習の理論を紹介し、期待損失の最小化という一般的な問題設定によって、パターン認識や確率分布推定といった多くの問題をより広い観点から統一的に扱うことができる事を示す。

読者が現在も活発に発展しつつある領域にアプローチする際に、本稿が1つの指針となれば幸いである。

# 1 パターン認識と統計科学

「パターン認識」という言葉を聞いたことはなくても、郵便番号の自動読み取りが実用化されていることを知っている人は多いだろう。また、個人情報端末の手書き文字入力や、パソコン・コンピュータの音声認識を使ったことがある人も少なくないと思う。この章では、まず、パターン情報とは何か、パターン情報の認識技術とはどのような情報処理かについて述べ、その中で統計的手法がどのように利用されているかを説明してゆく。

## 1.1 パターン情報処理

人間は、体中に配された莫大な数の感覚器(センサ)からの大量の情報の流れを、日常的に処理しながら生きている。いわゆる五感である、視覚、聴覚、触覚、嗅覚、味覚の他にも、温度や痛み、あるいは、筋肉の伸張、頭の回転や傾き、などの、多様な種類のセンサがあり、それらから時々刻々と脳神経系へと流れ込む情報をリアルタイムに並行処理し、それに基づいて体中の莫大な数のアクチュエータを制御している情報処理の複雑さと柔軟さは、驚嘆すべきものである。このように、人間の情報処理系が扱っている情報は「時間的・空間的に広がりをもって分布する莫大な数の変数値の組み合わせ」である。こうした情報は「パターン情報」と呼ばれる。すなわち、人間は、パターン情報を入力として受け取り、パターン情報を出力する「パターン情報処理」を行っている。

センサから得られるパターン情報は、莫大なバリエーションをもつていて、時々刻々とうつろいゆく。われわれが、一生の間に、完全に同じパターンのセンサ入力を二度受け取ることはないとと思われる。パターン情報の膨大なバリエーションに対処するために、われわれは、センサから得られるパターン情報を分節(segmentation)・分類(classification)・認識(recognition)・理

解(understanding)する。すなわち、うつろいややすく、莫大なバリエーションをもつ表層的なパターン情報から、直接は観測できないが、生活や生存のために本質的で安定的な情報を推測し、それらを統合することによって、予測可能性、制御可能性の高い世界像を構築している。

例として、視覚情報の処理を考えてみよう。ヒトの単眼の網膜にはおよそ1億個の視細胞が埋め込まれているといわれている。その網膜が捉えている情景には、通常、多くの対象が含まれている。神経系は、進化の過程で獲得した、さまざまな暗黙の知識を使ってそれらを適切なまとまりに分節し、それぞれのまとまりのもつ性質、および、まとまりの間の関係を推測する。そのようにして、目の前で動いているものが人間であるのかどうか、人間であれば誰であるのか、その人が何を意図して、今何をしているのか、を推測し、最終的には、たとえば、「お年寄りが切符を買おうとして、どうしてよいかわからずに困っている」というように、言語によって情景を記述することができるようになる。さらに、理解の結果である意味情報とすでに持っている知識を利用して推論を行い、切符を買うのを助けるために声をかけること、ができる。

情報処理技術の研究を、人間の行っている情報処理を代替することをめざすものと、人間の情報処理を支援・強化することをめざすものとに大別した場合、パターン認識(pattern recognition)の研究は、第一義的には前者であり、人間の行っているパターン情報処理の中でも最も基本的な機能の1つである「パターン情報の中の分節された1つのまとまりを、その属する範疇(category)・クラス(class)に分類する処理」を、コンピュータを使って工学的に実現することをめざすものである。

人間は、文字を読み、人の顔や物を見分け、音声を聴き取る、といったように、さまざまな種類のパターン認識をいとも簡単に日常的に行っている。こうした機能をコンピュータによって代替できれば、いろいろな応用が可能になる。そこで、コンピュータによる情報処理の黎明期から、コンピュータにパターン認識を模擬させることをめざした研究が行われてきた。その結果明らかになったことは、人間のようなパターン認識をコンピュータに行わせることは、予想よりもずっとむずかしい問題だ、ということで

あった。

むずかしさの本質は、パターン認識をするためには、時間的、空間的に広がって分布する非常にたくさんの情報の間の関係を総合的に考慮する必要がある、という点にある。そして、われわれの神経系がどのようにそれを行っているかはわれわれの意識にのぼることではなく、言葉によってその手続きを明示的に説明することはとてもむずかしい。

それにもかかわらず、多くのパターン認識研究者の長年にわたる研究、コンピュータの処理能力・記憶能力の飛躍的な増大、研究者によって収集され電子的に蓄積されたパターン情報データの増大、などの結果として、近年になって、印刷文字認識、ナンバープレート読み取り、手書き文字認識、音声認識、顔認識、指紋認識などのようなパターン認識技術が、次々と実用化・商品化のレベルに到達しつつある。

それらの成功の背後では、さまざまな確率分布モデルに基づいた多種多様な統計的な手法が大きな役割を果たしてきた。すなわち、パターン認識に代表されるパターン情報処理技術において、統計科学が非常に有効であることは歴史的に証明されてきた。現在では、パターン情報処理は、医療・遺伝・心理・教育・農業などと並んで、統計科学の応用の主要な1分野となっている。

## 1.2 パターン認識システム

音声認識のように必然的に時間的変化を扱うパターン認識技術もあるが、文字認識や顔認識など、時間的変化をあまり考慮せずに扱えるパターン認識技術もある。まず、簡単のために、そのような時間的変化を考慮しないパターン認識について考える。この場合、パターン認識の問題は、ある観測されたパターン情報  $o$  を、その属するクラス  $c$  に対応づける写像  $\Psi: o \rightarrow c$  を求める問題になる。

入力となるパターン情報とクラスの関係について、顔認識システムを例としてより具体的に考察してみよう。オフィスの入口に CCD のカメラを置いて、その前に立った人の顔を認識することを考える。この場合、クラ

ス  $c$  にあたるのは「カメラの前に立った人物の名前」であり、パターン  $o$  は「カメラが撮影した画像」である。

クラスとパターン情報との関係について考えるとき、多くの場合、因果関係に沿って考えるのが自然である。この例の場合には、ある人がカメラの前に立ち、画像が撮影される、つまり、 $c$  から  $o$  が生成されるという方向である。入口に人が来るたびに撮影することを繰り返すと、同じ人が、条件をいろいろ変えて何度も撮影されることになる。当然ながら、カメラの前に立つ人が同じ人物であっても、得られる画像がまったく同じものになることはない。

画像の生成過程に即して考えると、どのような画像が得られるかにかかる要因、すなわち、画像の変動の要因は、被写体が誰であるか  $c$ 、に加えて、明るさと光線の状態  $l$ 、服装(めがねも含む)や化粧  $d$ 、髪型  $h$ 、体調と表情  $e$ 、立ち位置と向きと姿勢  $p$ 、背景の状態  $b$  などが考えられる。これらの要因による変動に比べればカメラに使われている CCD の雑音などによる変動は十分小さいと考えられるから、 $c$  と  $o$  の関係は  $o = f(c, l, d, h, e, p, b)$  のような関数関係と考えてよいだろう。

このことを「通信の情報理論」の観点から見ると、カメラで撮影される画像は、「誰であるか」という情報  $c$  を画像  $o$  という非常に冗長な形に符号化(encoding)したもの、と見ることもできる。符号化の過程で、光線の具合などのさまざまな情報が一緒に混ぜ込まれてしまっている。それらの情報は、誰であるかを知りたい、という立場から見れば無視すべき雑音である。したがって、画像を認識・理解することは、雑音の加わった冗長な符号の誤り訂正復号化(decoding)、あるいは暗号の解読と似ている。

逆に、 $o$  から  $c$  を求める過程を、画像から、必要な意味情報だけを抽出する情報圧縮過程と見れば、それを非常に効率の良い符号化過程と考えることもできる。実際、「昨日 A さんが来たよ」と伝えるだけで、そう言わされた人の頭の中に、A さんのイメージが復元されるのであるから、言葉を交わすことは、大変効率のよい通信方法である。

画像の解像度(画素数)が  $1024 \times 1024$  で、1 画素あたりの色表現が R(赤の輝度), G(緑の輝度), B(青の輝度)それぞれ 256 階調とすると、この表

現系の情報容量は数 MByte である。実際に撮影される画像の分布は偏っているため、その情報量はこれよりはだいぶ小さいが、それでもかなり大きい値になるだろう。

一方、クラス  $c$  に関する情報量は、たとえば、オフィスに出入りする人が  $K$  人で、その出現確率が均一とすると、 $\log_2 K$  bit である。これは、 $K = 1024$  でもわずか 10 bit に過ぎない。すなわち、 $o$  から  $c$  を得る写像は、莫大な情報を捨てる情報圧縮写像であり、顔の認識に必要な情報だけを残して、いかに上手に不要な情報を捨てるか、が問題になる。

パターン情報の生成過程で複雑に混じり合ってしまった情報を逆に分解し、 $o$  から  $c$  を推定することは一見絶望的に思われるが、顔認識システムの研究では、他の多くのパターン認識システムの研究と同様に、この問題を、図 1 のようないくつかの段階に分けて解いている。

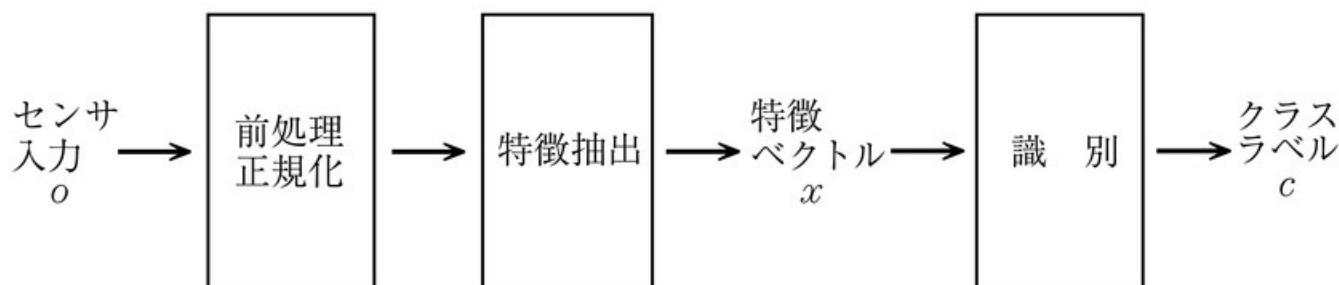


図 1 パターン認識システムのブロック図

まず、 $c$  以外の要因による  $o$  の変動のうち、 $c$  による変動との相互作用が少なく、比較的簡単に除去可能なものをできるだけ除去する。具体的には、画像全体の中から、顔の認識に重要な部分(たとえば、目、鼻、口を含む領域)だけを切り出して一定の大きさに正規化したり、画面全体の明るさの分布を正規化したりする。これによって、明るさ、服装、髪型、立ち位置、背景、などによる変動を、完全にではないが、ある程度排除することができる。このような処理は前処理(preprocessing)とか、正規化(normalization)と呼ばれる。さらに、目や鼻などの位置関係から、顔の 3 次元モデルを使って顔の向きを推定して、向きを補正したりすることも試みられている。

こうして、必要な情報をできるだけ損なわずに不要な情報を除去して得られる画像を  $o'$  とすれば、 $o'$  と  $c$  との関係はだいぶ簡素化されている。しかし、 $o'$  には、まだ、表情や体調、化粧、めがね、など、顔の認識にとって

は雑音となる情報が含まれている。これらの要因による画像の変動は、 $c$  の変化による変動と混じり合っていて、簡単な処理で除去することはできない。また、 $o'$  を各画素の値を要素とするベクトルとして処理する場合、次元がまだまだ高く、その後の処理が大変である。そこで、必要のない情報を捨てるのとは逆に、顔の認識にとって重要なと思われる情報を抽出することが行われる。

たとえば、目、口などの顔の部品の形やそれらの間の距離は顔の認識にとって重要そうに思われる。実際、たくさんの画像を使って正規化された顔画像の各画素の明るさと、クラス  $c$  との相互情報量を計算すると、相互情報量が大きいのは、目や鼻の周辺の画素であることが知られている。しかし、もっと他にも、重要な情報があるかもしれない。人間は暗黙のうちに顔認識をしてしまうため、何が重要な情報かはなかなか明示的にはわからない。そこで、主成分分析や判別分析といった統計的な情報圧縮手法も用いられている。正規化された画像を画素数次元のベクトルと考えて、多数の顔画像のデータを主成分分析したり、判別分析したりすれば、顔画像の変動の主軸となる座標軸や、顔画像をクラス分けするときに重要な情報をもつ座標軸が抽出され、個々の顔画像は、その次元圧縮された座標系での座標値によって表される。

このような情報抽出過程は特徴抽出(feature extraction)と呼ばれ、特徴抽出の結果は特徴量(feature)と呼ばれる。特徴量は、数次元から数百次元程度の実数値のベクトルとなることが多いため、特徴ベクトル(feature vector)とも呼ばれる。特徴量の空間は特徴空間(feature space)と呼ばれる。

こうして、入力顔画像は、正規化されて特徴抽出された結果、数次元から数十次元の特徴ベクトルへと変換される。しかし、この段階になっても、あるクラスに属するデータから得られる特徴ベクトルが、1つの値に完全に集約されるということはない。排除できないゆらぎや、正規化、特徴抽出の影響が残るのである。この、残された変動を除去して、特徴ベクトルを、最

終的にその属するクラスへと写像する処理が、識別(discrimination)<sup>\*1</sup>である。

### 1.3 統計的パターン認識

正規化や特徴抽出では、主に、個々のパターン認識問題に固有の知識を用いて、認識のために重要な情報を抽出し、重要ではない情報を捨てた。どのような特徴量を抽出するかは、パターン認識が成功するかどうかを左右する重要なポイントであるが、その解決法は問題固有性が高く、一般的に論じることがむずかしい。主成分分析や判別分析のような汎用的な情報圧縮写像を特徴抽出に使う方法、あるいは特徴抽出の問題を「入力パターンの変換群に対して不变性をもつ特徴量を求める」という形でできるだけ一般的に扱おうとする試みもあるが、一般的には、良い正規化や良い特徴量を得るためにには、認識対象とするパターン情報の性質を深く理解する必要がある。すなわちパターン認識研究の主な目的の1つは、音声や画像など、認識対象とするデータが固有にもっている性質を発見することである。

これに対して、識別過程では、すでに問題固有の知識を使い尽くした後であるため、問題固有の知識によらない、汎用的な手法が必要とされる。ここが、統計的手法の主たる活躍の場となる。

前節で述べたように、パターン識別とは、特徴ベクトル  $x$  をクラス  $c$  へと対応づける処理である。そのためには、 $x$  の空間をいくつかの部分に分割して、それぞれの部分に、対応する  $c$  の値(クラスラベル)を振ればよい。ここでの問題は、

- (1) 空間の分割をどのように表現するか？
- (2) 適切な分割をどのように求めるか？
- (3) 新しい入力  $x$  がどの部分に入っているかをどのように判定するか？

\*1 研究分野によって、判別、弁別と訳されることもある。また、classification という言葉があり、こちらは通常は分類、あるいは類別と訳されるが、これを識別と訳している場合もある。さらに、decision という言葉もあり、通常は(意思)決定と訳される。これらの言葉の使われ方は、それらが使われてきた分野での慣用によっており、使用されている文脈や分野を考慮して意味を汲み取る必要がある。

である。

良い識別方式を得るためにには、各クラスに属するパターンから計算される特徴ベクトル  $x$  が、特徴空間の中でどのように分布しているかを的確に捉えて、その偏りを利用する必要がある。逆にいえば、各クラスのデータの特徴ベクトルが特徴空間の中でランダムに分布しているようでは、良い識別方式を得ることは不可能である。すなわち、特徴量の取り方が悪かったということになる。

$x$  の分布のしかたを、パターンの生成過程や特徴ベクトルの計算過程から明示的に導出することはほとんど不可能であるため、事例データに基づく方法、すなわち、正解のわかっているパターンをたくさん用意して、そこから得られる特徴ベクトルの分布の様子に従って識別方式を構成する方法が必要とされる。顔認識システムの例でいえば、あらかじめ、誰が写っているかがわかっている画像をたくさん集めて、そこから計算される特徴ベクトルの分布のしかたを利用することになる。このためのデータ、すなわち、特徴ベクトル  $x$  と正解クラス  $c$  のペアの集合  $D = \{(x_1, c_1), (x_2, c_2), \dots, (x_N, c_N)\}$  は学習データ (learning examples)、学習用データ、あるいは、訓練データ (training samples) などと呼ばれる<sup>\*2</sup>

十分な数の学習データを集めれば、そこから、特徴量の分布のしかたについての情報が得られる。たとえば、あるクラスに属する特徴ベクトルは、空間の狭い領域に集中しているかもしれない。あるいは、特徴空間の特定の部分空間に集中しているかもしれない。このように、あらかじめ収集されたデータに基づき、統計的手法を用いて識別を行う方法の研究は、統計的パターン認識 (statistical pattern recognition) と呼ばれ、パターン識別研究の主要な流れの 1 つとなってきた。

いくつかの特徴量によって表される対象を、それが属するクラスへと分類・識別する問題は、病気の診断、生物の種の分類などに共通するもので、古くから統計学の主要な応用対象の 1 つである。そこでは、データに基づいて最適な識別・分類を行うための手法として、判別分析、数量化 2 類、ベ

---

\*2 なぜ、「学習」、「訓練」といった言葉が用いられるのか、不思議に感じられるかもしれない。それについてのひとつの説明は 2.7 節にある。

イズ識別、といった手法が提案されてきた。統計的パターン認識の分野でも、こうした手法は用いられているが、それに加えて、特徴量の分布のしかたを捉えるためのいろいろな工夫も発明されてきている。次章ではそれを具体的に見てゆく。

## 2 いろいろなパターン識別手法

特徴空間内の学習データの分布のしかたをどのように捉え、どのように利用するか、という問題に答えるために、さまざまな観点から多くの方法が考えられてきた。この章では、代表的なアイデアとして、

- (1) 代表ベクトル(テンプレート)による方法
- (2) 近傍パターンの投票による方法
- (3) 部分空間による方法
- (4) 識別関数による方法
- (5) 決定木による方法
- (6) 階層型のニューラルネットワークによる方法

の6つを簡単に紹介する。いずれも、異なった履歴をもち、歴史的に生き残り、長く研究され、多くの問題で有効性が示されてきたものである。各手法にはさまざまな改良や拡張が行われてきており、詳しく述べれば、それぞれについて1巻の書物を必要とする。技術的な内容の詳細については、すでに多くの良書があるため、ここでは、それぞれの手法の核となっている基本的なアイデアをできるだけわかりやすく記述することを心がけた。

### 2.1 テンプレートマッチング法

良い正規化と特徴抽出が行われて、不要ない情報が十分に捨てられた場合には、各クラスのデータの分布は、かなりよくまとまった局所的なものになることが予想される。そのような場合には、クラスごとのまとまり

を 1 つの代表ベクトル(テンプレート(template)あるいは、プロトタイプ(prototype)と呼ばれる)で表現することが考えられる。識別を行うときには、入力ベクトル  $x$  と各クラスの代表ベクトルとの間の距離(distance)，あるいは類似度(similarity)を，何らかの尺度で評価して，最も近いクラスに識別すればよい。このような手法は，テンプレートマッチング(template matching)法と呼ばれる。

代表ベクトルの求め方，および，代表ベクトルとの距離の測り方によって，多くのバリエーションが考えられる。代表ベクトルの求め方としては，たとえば，クラス  $c$  に属する学習データの分布から，平均  $\mu_c$  を求めるのが最も自然かつ簡単である。新しい入力  $x$  の識別方式としては，各クラスの代表ベクトル  $\mu_c$  との間のユークリッド距離を計測して，距離が一番短いもの，つまり，一番近いクラスへと識別することが考えられる。

この場合， $(x, y)$  をベクトル  $x$  と  $y$  の内積として，

$$\|x - \mu_c\|^2 = \|x\|^2 - 2(x, \mu_c) + \|\mu_c\|^2$$

なので，クラス平均までの距離(の 2 乗)を最小にする  $c$  を選ぶことは，クラス  $c$  ごとに定義される

$$g_c(x) = 2(x, \mu_c) - \|\mu_c\|^2$$

という  $x$  について 1 次(線形)の関数を最大にする  $c$  を選ぶことと同値である。また，クラス  $c$  の領域と  $c'$  の領域との間の境界は  $g_c(x) = g_{c'}(x)$  から

$$2(x, (\mu_c - \mu_{c'})) - \|\mu_c\|^2 + \|\mu_{c'}\|^2 = 0$$

という超平面(hyper plane)となる。

この方法は直観的に大変わかりやすく簡単だが，いつもうまくゆくわけではない。たとえば，特徴空間内で，あまりよくまとまっていないクラスと，とてもよくまとまっているクラスとがあった場合に，それぞれの平均とのユークリッド距離によって近さの評価をすると，広がっているクラスのデータはコンパクトなクラスに間違えられやすくなる。そこで，それぞれのクラスの分布のひろがりを考慮した距離として，

$$(x - \mu_c)^T \Sigma_c^{-1} (x - \mu_c)$$

のように，各クラスの分散共分散行列  $\Sigma_c$  の逆行列で重みづけた距離(の 2 乗)を使うことが考えられた。ここで  $x^T$  は，列ベクトル  $x$  の転置である。こ

の距離はマハラノビス距離(Maharanobis distanceあるいはMaharanobis' distance)と呼ばれている<sup>\*3</sup>. この距離に基づいて識別を行うと, クラス間の識別境界は2次曲線になる.

さらに特徴抽出が不十分で, 1つのクラスのデータが長く伸びたり, 複雑な形に分布をしていたり, いくつかのまとまりに分裂してしまっていたりすることも考えられる. このような場合の対策としては, 代表ベクトルを複数用意することが考えられる. この方法は, マルチテンプレート(multiple template)法と呼ばれる.  $x$  の識別を行う場合には,  $x$  と, すべての代表ベクトルとの間の距離を評価して, 最も近い代表ベクトルの属するクラスへと識別すればよい.

しかし, 複数の代表ベクトルの決め方は自明ではない. 最も単純には, 何らかのサンプリングルールに従ってランダムに選択する, という方法が考えられる. また, 各クラスに属するデータをクラスタリングして, いくつかのクラスタに分け, まとまりごとに平均を求めることが行われる. この場合, どのようなクラスタリング手法を用いるかや, クラスタの数をどう決めるか, などが問題となる.

## 2.2 $k$ -最近傍識別法

古典的なパターン識別方式の中で, 最もよく知られているものの1つが,  **$k$ -最近傍識別法**( $k$ -nearest-neighbours classification rule,  $k$ -NN rule)である. テンプレートマッチングが, 学習データから1つ, あるいは少数の代表ベクトルを求めて利用するのに対して, この方法では, 学習データをすべてそのまま記憶しておく. すなわち, 学習データ全部で学習データの分布を表現する. このように, 多くのデータをほとんどそのまま記憶して利用する方法は, 学習データの分布の様子を少数の代表ベクトルなどによって圧縮し

---

\*3 この呼び名は, Maharanobis が 1930 年頃の論文において 2 つの正規分布  $N(\mu_1, \Sigma)$  と  $N(\mu_2, \Sigma)$  の間の距離を  $(\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)$  によって計量することを提案したことによる. 分布間の距離を指す場合には, マハラノビス汎距離(Maharanobis' generalized distance)とも呼ばれる.