怪奇!! 次元の呪い

ー識別問題，パターン認識，データマイニングの初心者のためにー

（後編）

（株）NTTデータ　坂野　鋭　sakano@rd.nttdata.co.jp

NEC　山田 敬嗣　yamada@ccm.cl.nec.co.jp

■理想と現実ー識別器の選びかた（承前）ー■

前編では識別問題の検討に入る際に，やってはいけないことを中心にお話した．後編では，ではどうしたらいいのか？ということを中心にお話した上で，評価の仕方を説明し，実例を通して検討の仕方を解説しよう．

理想の識別器とはどのようなものだろうか？　一言でいうと，対象となるカテゴリの分布，もしくはカテゴリ間の識別面を正確に表現できる識別器ということになる．

では，いくらでも複雑な識別面を表現できるニューラルネットやSVM（Support Vector Machines）はとてもGoodな識別器ではないか？ということになるのだが，現実はそう簡単ではない．

というのは有限のサンプルから真の分布や識別面を推定することが困難な問題だからである．

普通のパターン認識の教科書を見ると，学習時にはカテゴリごとの確率密度関数を推定して，認識時には確率の高い方，つまり確率密度関数の値の大きな方を認識結果として出力するなどと書いてある．しかし，この確率密度関数の推定が結構難しい問題なのである．例として，人工データによる散布図を示そう．

図-1は2つのカテゴリがある場合の2次元の識別問題について，ある確率密度関数を仮定して発生させた人工データである．この散布図から，我々がどのような確率密度関数を仮定したかが分かるであろうか？　あるいは，最適な識別面がどのようなものか推定できるであろうか？　ちょっとここで本稿を読むのを休んで，このデータから推定できる最適な識別面を想像して欲しい．

休　憩

想像できたであろうか？　**図-2**はまったくの素人の方にお願いして引いてもらった識別面である．多くの方が想像した識別面はこれとあまり変わらないのではないだろうか？　実は我々が仮定したカテゴリごとの確率密度関数は**図-3**に示すような2変数正規分布である．したがって，真の識別面は図-3で見える2次曲線になる．読者の想像はどの程度あたったであろうか？

筆者らの想像では正解を当てられた方はほとんどないかと思う．計算機で自動的にこれらの問題を解くことがどのくらい難しい問題であるかの片鱗は理解してもらえたであろうか？

この散布図の問題は，複雑な識別面を作ることができるニューラルネットなどの識別器がもたらす問題を具現化している．図-1の問題では，2次の識別面が正解なわけだが，いくらでも複雑な識別面を作れるアルゴリズムでは，図-2のように過度に学習データに依存した識別面を作ってしまい，テストデータではかえって識別精度の低下を引き起こす．識別精度を高くするためには，識別面が複雑であればよいというものではなく，分布の性質を正確に反映した識別面を作ることが望ましい．つまり，複雑な識別面を作るということと，適切な識別面を作ることができるということは違った能力なのである．

では，結局どのような識別器を選ぶのが正しい識別器の選択法なのであろうか？　実のところ，この問題に対する正しい解はない．我々のお勧めする方法は，

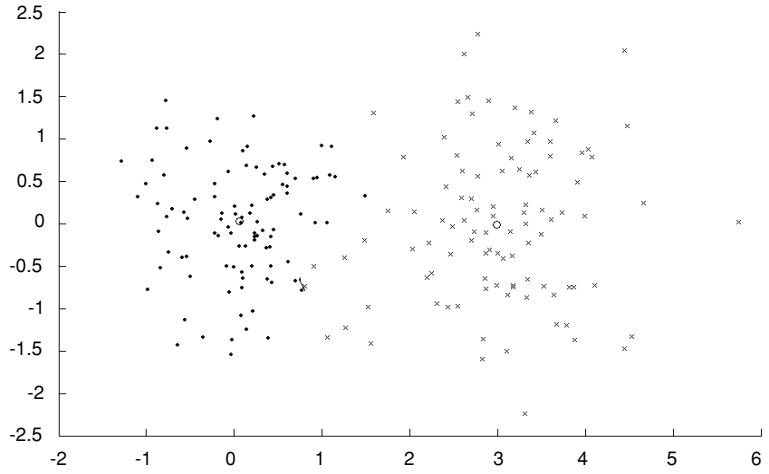


図-1　人工データによる散布図の例

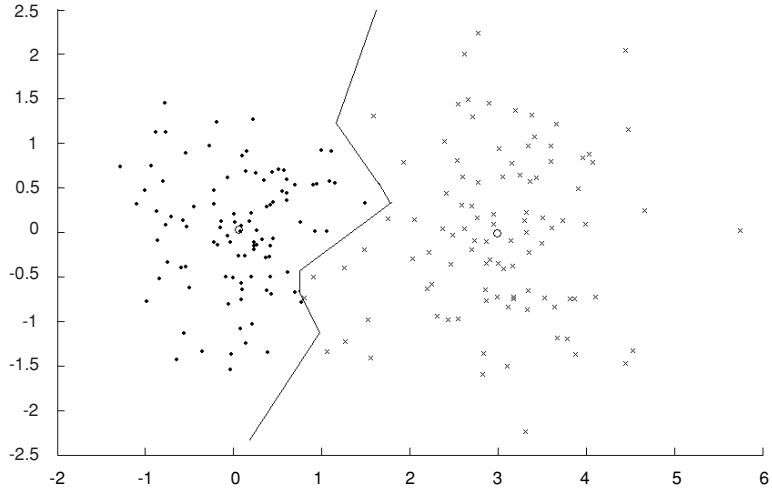


図-2　図-1の散布図に識別面を書き込んだもの

なるべく簡単な，つまりパラメータの少ない識別器を採用することである．次章で述べる識別器の評価方法に基づいて評価したとしても，有限のデータから正確に識別器の良し悪しをいうことは難しい．しかし，はっきりいえることは，サンプルが有限である以上，パラメータが増えれば，推定精度が下がり，未知データに対する頑健性が失われるということである．特に多層パーセプトロンや学習ベクトル量子化のようにパラメータ数を調整可能な場合には，識別精度が最大のものより，多少識別精度が低くても，パラメータが少ないものを選ぶのが経験的には正しい．

つまり，S君は多少遠回りでも，中間層のユニット数が少ない方から試してみるべきであった．また，現在では過学習という現象は，単純にサンプル数が少ないために過度に複雑な識別面を作ってしまう現象として理解されている．つまり，これもまた，次元の呪いに

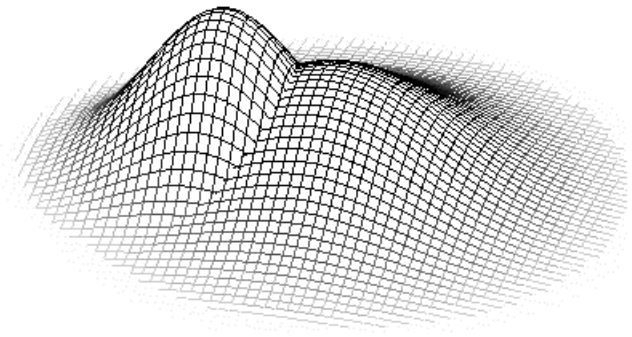


図-3　図-1，図-2の乱数を発生させた正規分布

他ならないのである．現実的には学習を途中でやめることには意味があるが，サンプルを増やしたり，中間層のユニット数を削減したりする方がはるかに効果がある．また，この分野の進歩は速い．速いだけではなくて，直感が通用しない分野であるだけに，誤解が通説として流布される状況も珍しくはない．古い文献を無批判に信用することもS君の失敗といえるかもしれない（その意味では，本稿も本質的な誤りを含んでいる可能性を否定できない）．

■評価の仕方■

さて，実験系が大体組めたら，いよいよ評価である．識別問題に関する評価が理論的に行えることはまずなく，ほとんどが実験的な評価となるわけだが，実験を開始する前に，まず識別精度の目標を定めることと，どのくらいの精度が有為であるかを考えること，の2つを行っておくことをお勧めする．

無論，100%の精度が得られれば，それに越したことはないが，現実の問題ではそのようなことはまずない．問題の性質，応用の場面を想定して，識別精度の目標を設定することは実験の前に必ず行う必要がある．たとえば，現在手動でやっている（すごく大変な）仕事を半自動化するようなテーマであれば，識別精度が50%でも役に立たないとは言いきれない[1)]．

また，実現した実験系は現実世界の小さな小さなサブセットでしかない．ここでのわずかな差が現実世界でどのくらいの意味を持つかを考えておかないと，説得力のあるデータを出すことはできない．たとえばS君の実験系では，テストデータが1,000個（＝100個×10字種）しかないのだから，1%の違いはわずかに10個のサンプルが正しく識別できたということでしかない．この場合でいえば，2～3%くらいの差がないと有為な差とは言い難い（したがって，このくらいの差がない場合には，パラメータの少ない識別器を採用すべきである）．

さて，これらのことを考慮した上で実験的な評価にかかるわけだが，必ず注意しなくてはならないのは，

学習データとテストデータを分ける

ということである．何度も言うが，学習に用いるサンプルは有限であり，世界のごく一部しか反映していない．学習に用いていないデータによる評価は絶対に必要である．なんだ，そんなことと言うなかれ，著者らは年に数回はそのような論文に出会っている（無論，そのような論文はリジェクトとなるので一般読者の目に触れることはあまりない）．これは，どのような識別器を使うか，どのようなデータを相手にするかに依存しない．

当然のことだが，次元圧縮行列の生成や特徴選択アルゴリズムの構成のときに，テストデータのカンニングは禁物である．ついつい，テストデータでの識別結果での誤りを見て，特徴選択や特徴抽出の再設計をしてしまうかもしれないが，これは許されないカンニングになる．

次の問題は学習データとテストデータをどのように分けるのかということである．一番簡単に思いつくのはS君がやったように半分ずつに分けることである．この方法は交差検定法（Cross varidation，以下CV）と呼ばれ，最も広く用いられている．

しかし，データを分けるとデータ数が減ってしまうじゃないか！と思われた方もあるだろう．文字データのように比較的多数のデータが利用できる場合には，多少減っても問題はないが，医療データのようにデータ収集コストがきわめて高価である場合には深刻な問題になる（なんといってもデータの1レコードごとに人間1人の生死が絡んでいるわけで，これ以上高価なデータはちょっと思いつかない）．このような場合に推奨されるのは「1つ取って置き法」（Leave One Out，以下，LOO）と呼ばれる方法である．

この方法では，n個のデータがあった場合に，1つをのぞいて$n-1$個を用いて学習を行い，残りの1つをテストデータとして用いるという操作をn回繰り返す．つまり，n回の実験を繰り返すことで，すべてのデータを，テストデータとして用いることを可能にするのである．S君の検討の場合でいえば，LOOを用い，実験を100回繰り返すことでテストデータの個数を1,000個から2,000個に増やすことができたことになる．

しかし，一方で，100回も実験するのは大変だという意見もあるだろう．実のところ著者らも面倒なのでLOOよりCVですませてしまうことが多い．

中間的な方法として推奨されるのがn fold CVと呼ばれる方法である．この方法は，データをn個の集合に分け，$n-1$個の集合を学習に，残りをテストに用いるという操作をn回繰り返す．分野によってはローテーション法などとも呼ばれているが，データ数が少なく，検討を急ぐ場合に推奨される方法である．この方法の問題はデータの分割数の決め方である．我々の推薦する方法は，学習データ数が次元数を超える，最小の分割数を採用することである．無論，データ数が十分に大きいと考えられる場合にはCVでも構わない．また，何らかの理由で，学習データ数が実質的な次元数を上回っていると考えられる場合もCVで構わない．

いずれにせよ最も重要なのは

学習データとテストデータを分ける

ということであり，実験系のすべての段階でテストデータの情報が学習系に入り込まないように努力しないと，信頼できる実験結果は得られない．

■その他の重要な検討項目■

以上で，通常の識別器の設計に関する議論は終了しているが，これ以外にも重要な検討が存在する．ここでは，通常役に立つかどうかは分からないが，時として役に立つかもしれない方法を並べてみたい．

高次元のデータを扱っていてとてもフラストレーションがたまるのが，データの分布を目で見ることができないということである．教科書には1～3次元のいかにもありそうな図が書かれているが，たとえば100次元空間での文字データが本当に正規分布をしているかなどというのは，ものすごく気になる話題であるにもかかわらず，どうすればそれを検証できるかは誰も教えてくれない．

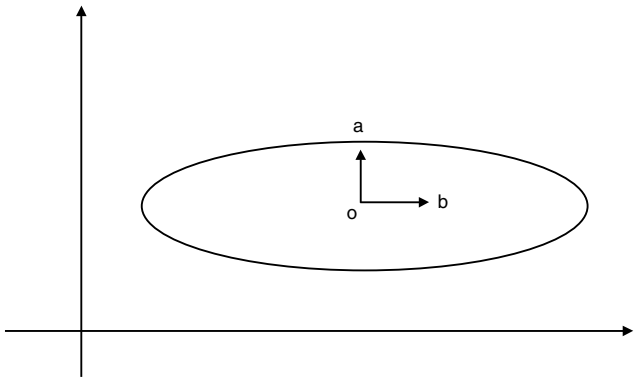
このため，高次元データを2～3次元に写像する方法に関する研究は数多く行われている．代表的なのは主成分分析や多次元尺度構成法[2]のような方法であるが，ニューラルネットの1つである自己組織化特徴写像[3]のような方法も広く用いられている．最近では，クラスタ判別法[4]のような簡単だが実用的な方法も提案されている．

もう1つ，我々の経験では特徴の妥当性は常に大きな問題である．S君が扱っている文字認識の問題では文字データを（少なくとも素人には）わけの分からない方法で高次元のベクトルに変換し，わけの分からない空間で学習するわけであるが，S君がやったように学習データから抽出した特徴ベクトルを平均した場合に，平均が果たしてどのようなものになっているかは知りたくてたまらない情報ではないだろうか？　このための研究は数少ないがたとえば文献5）のような例も存在するので，参考になることもあるかもしれない．

また，一方で多変量解析の分野では，この問題は早くから認識され，顔グラフなど，さまざまな多次元データの表現法が工夫されている[6]．

■人工データの例■

ここでは，人工データを用いて識別器がどのように振る舞うかをみてみよう．実験に使ったデータは，図-3の正規分布から発生させた乱数である．したがって，正解は確率密度関数の値が等しくなる面で，これが2次関数で与えられることは前章までに説明した．



図中，oaとobはユークリッド距離の意味では等距離だがマハラノビス距離の意味ではoaの方が大きくなる．ユークリッド距離の等距離面は平面となるが，マハラノビス距離の等距離面は2次曲面になる．

図-4　マハラノビス距離は，分散の大きさを考慮した距離

このようなデータが与えられたときに，識別器がどのように振る舞うかをみてみよう．ここで比較してみるのは，平均値からのユークリッド距離を用いる1次識別器と，マハラノビス（Mahalanobis）距離を用いる2次識別器（**図-4**参照），*k*-NN識別器で$k=1$とした場合の3種類である．直感的には，マハラノビス距離を用いた識別器が最高となるように思えるし，人によっては全学習データを記憶している*k*-NN識別器が高性能を出すと考えるかもしれない．

論より証拠である．**図-5**にそれぞれの方法の識別精度の学習データ数に関する依存性を示す．この図でテストデータは同じ分布から発生させた10,000個のデータを用いた．

一目見て分かる通り，この場合には1次識別関数と2次識別関数では大きな差がついていないばかりか，1次識別器の方が性能がよい場合さえある．*k*-NNに至っては，かなり低い精度しか出ていない．

なぜこのようなことになるかというと，2次識別器では，平均値のほかに分散共分散行列を計算しなくてはならない．そのため，1次識別器より数多くのデータがないと，正確な識別面を推定することができない．このため，1次と2次ではかなりのサンプル数がないと明確な差が出ない．また，*k*-NNでは，さらに複雑な識別面を作るために，さらに多数のデータがないと性能が出ないことになる．

図の中では，2次元で100個という，現実の識別問題からするとずいぶん条件のよい場合の結果を示しているのであるが，それでも十分とはいえないのである．

このような状況は現実の識別器でもよくみられ，一

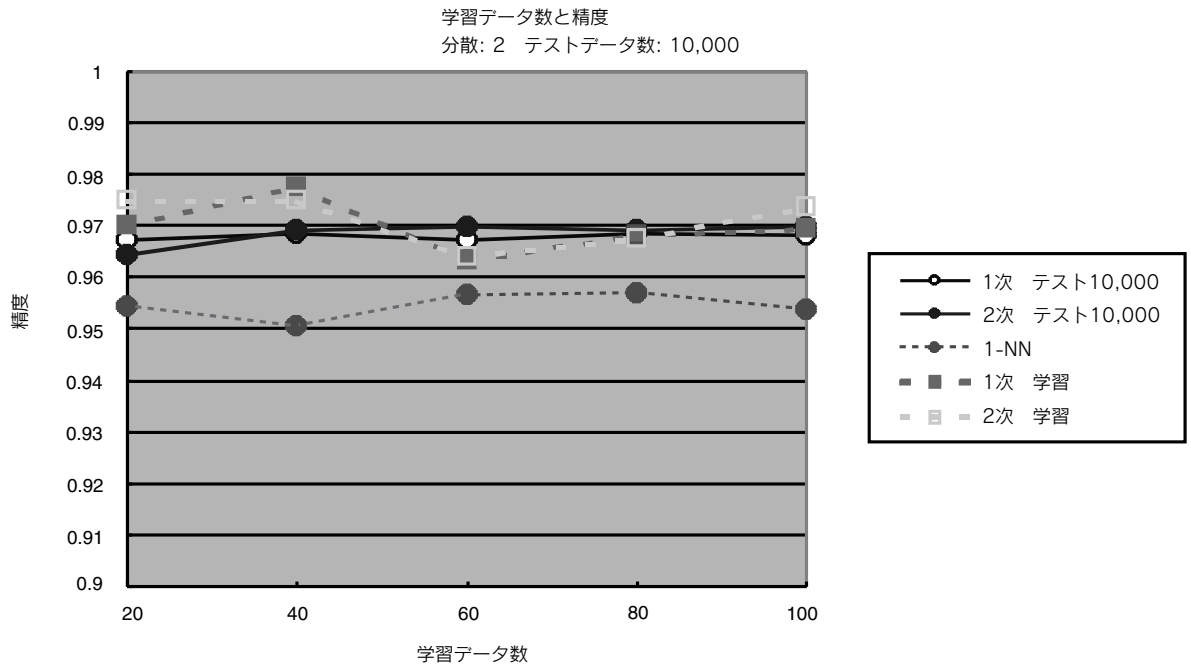


図-5　**図-2**のデータに対する識別精度の学習データ数依存性

時期は盛んに研究された[7), 8)].

■具体的な識別問題－顔認識の問題■

本章では，具体的な識別問題の例として，顔画像による個人認証の問題を扱ってみたい．顔による個人識別は，画像による監視など，ここ半年くらいはビジネスも巻き込んで大きな話題になっている技術であるが，ここではあまり難しい技術は使わず，ここ10年の顔認識研究の活性化のきっかけになった固有顔の方法を試してみる．

固有顔の方法は，1991年にマサチューセッツ工科大学のTurkらによって提案され，条件さえ整えれば比較的簡単な方法で顔による個人識別が可能であることを示した歴史的な方法である．

方法自体はきわめて簡単であり，顔画像を画素を要素としたベクトルと考え，これを主成分分析した特徴を用いて，平均値からのユークリッド距離を用いて認識を行う．

実験のためには（株）NTTデータで収集した顔画像データを用いた．前処理としては，このデータから手動で顔部分を切り出し，16×12画素のサイズに正規化を行った（**図-6**）．

このデータは79人分の顔画像を1人当たり20枚収録している．画素を要素としたベクトルとみなした場合，192次元となるため，少々少なすぎるデータであるということができる．そのため，画像を左右反転して1人あたりの画像を40枚に水増しし，評価方法としてはCVを用いた．

この手の認識問題に関して，よくある誤解は，やはり情報量の多い，高次元のデータ，つまり解像度の高い画像の方が認識率が高いということであるが，現実には必ずしもそうではない，高解像度が次元の呪いを呼び込むために，適度に粗い画像の方が認識性能は上がるのである．

実験結果を**表-1**に示す．比較のために，次元圧縮を行わない単純パターンマッチングの精度も掲載した．当然，1次元の低い次元では，識別性能は惨澹たるものである．しかし，10次元あたりでパターンマッチングを超える認識性能を出し始めて30次元でピークに至っている．このように，次元数の削減は「次元の呪い」に対して有効に働くのである．

■終章－もっと勉強したい人のために■

以上，不気味な罠である「次元の呪い」を中心に，識別問題に挑戦する際に注意すべき事柄をあげてきた．この解説では，基本的に教科書にあまりかかれていないことを中心に記述したために，基礎知識の側面ではかなりの不足がある．

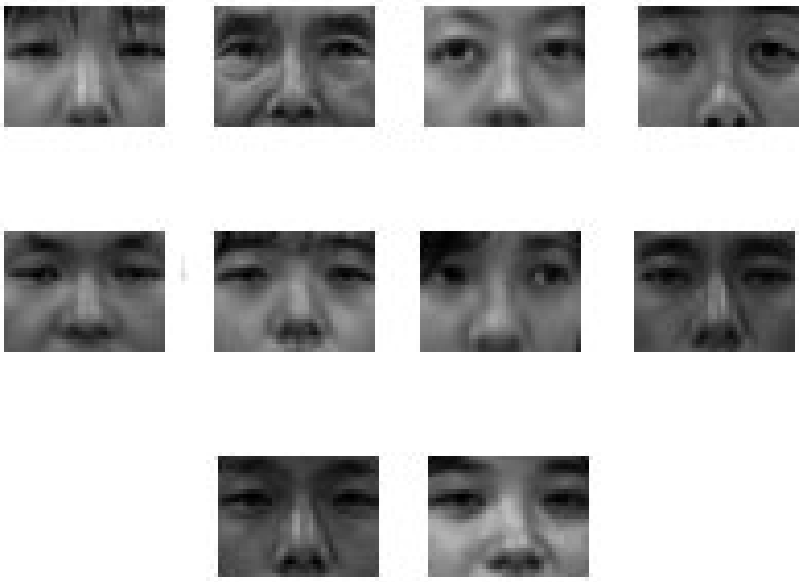


図-6　実験で用いた顔画像データの例

認識法	画像マッチング	固有顔	固有顔	固有顔	固有顔	固有顔
次元数	192	1	5	10	20	30
認識率	92.23%	18.1%	84.3%	94.8%	96.7%	97.1%

表-1　固有顔法，画像マッチング法における認識率

これを補う最も手っ取り早い教科書が文献9）である．基礎的な事柄が分かりやすく書かれているだけではなく，著者らの豊富な経験に基づく示唆に富んでおり，初心者から上級者にまで推薦できる．最近になって名著である文献10）の第2版の邦訳が出版された．第1版に比較して情報が増えすぎている感があり，初心者には薦めにくい分量となっているが，豊富な情報は研究を進める上で役に立つ．

また，最近の発展については文献11）が優れた解説である．実験例や文献も豊富で初心者にもプロの研究者にも役に立つ．最近の識別器であるニューラルネットやSVMに関しては文献12），13）などがよい解説である．

これらの文献は，パターン認識サイドに偏っている．理由はいうまでもなく著者らが2人ともパターン認識コミュニティの住人であるからだが，データマイニング，バイオインフォマティクス，機械学習，テキスト処理などさまざまな分野で類似の研究が行われており，今後はこれらの分野同士の交流が盛んになることを期待しつつ筆を置きたい．

謝辞　本稿執筆にあたり，実験などの手助けをしてくれた（株）NTTデータの末永高志氏，佐藤新氏．大学の研究室の様子についてコメントをいただいた早稲田大学大学院理工学研究科の三枝亮君，吉澤大樹君．文献の収集などに助力をいただいた産業技術総合研究所の坂野貴子博士に感謝します．

参考文献

1）東　陽子, 他: 投稿準備中.
2）林知己夫, 飽戸　弘: 多次元尺度解析法－その有効性と問題点－, サイエンス社（1976）.
3）Kohonen, T.: Self-organized Formation of Topologically Correct Feature Maps, Biological Cybernetics, Vol.43, pp.59-69（1982）.
4）末永高志, 佐藤　新, 坂野　鋭: 分布の構造に着目した特徴空間の可視化－クラスタ判別法－, 信学技報, PRMU2001-44, pp.39-44（2001）．あるいは, クラスタ構造に着目した特徴空間の可視化－クラスタ判別法－, 信学論DII, Vol.J85-D-II, No.5, pp.785-795（2001）.
5）坂野　鋭, 木田博巳, 武川直樹: 遺伝的アルゴリズムによる文字識別系の解析, 信学論D-II, Vol.J78-D-II, No.7, pp.1687-1694（1997）.
6）竹内啓編集: 統計学辞典, 326p, 東洋経済（1989）.
7）杉浦, 木村, 他: サンプル数と識別関数の識別率の関係, 信学技報PRU88-24（1988）.
8）山田, 上, 他: ニューラルネットを用いた文字認識, 信学技報PRU88-58（1988）.
9）石井健一郎, 上田修功, 前田英作, 村瀬　洋: わかりやすいパターン認識, オーム社（1998）.
10）Duda, R. O., Hart, P. E. and Stork, D. G.: Pattern Recognition, John Wily & Suns, 2nd ed.（2000）.
11）Jain, A. K., Duin, R. and Mao, J.: Statistical Pattern Recognition: A Review, IEEE, Trans., PAMI, Vol.22, No.1, pp.4-37（2000）.
12）山田敬嗣, 佐藤　敦: ニューラルネットによるパターン認識, 電子情報通信学会誌, Vol.82, pp.852-859, pp.977-984, pp.1046-1053, pp.1248-1255, Vol.83, pp.50-56（1999-2000）.
13）前田英作: 痛快！サポートベクトルマシン, 情報処理, Vol.42, No.7（July 2001）.

（平成14年4月18日受付）