6. パターン認識適用時の留意事項及び使い分け法

パターン認識や統計等の解析手法を実行して得られた結果は充分に信頼性の高いものでなければならない。 パターン認識や統計の実行結果の信頼性が低ければ、実行後に行われる解析結果の評価や解読作業自体が無意味なものとなってしまう。

ここではパターン認識による解析時に、信頼性の高い結果を得る為に必要となるいくつかの留意事項についてのべる。 これらの留意事項は解析を限定するものではなく、条件さえ満たせば解析結果は信頼性の高いものが得られるという指針である。 ここで示される留意事項を満たさないで得られた結果は信頼度が低く、従って結果について様々な議論を行う事は危険であるし、学会等での発表も出来ないものとなる。

ここでのべる留意事項の多くは、判別分析に関するものであり、その内容はパターン（サンプル）数及び数値データ（記述子／次元数）数に関するものである。

計算機が身近に存在する現在では、数値データさえあればパターン認識を実行して解析結果を出す事は容易である。 この制限事項について充分理解しておくことが信頼性の高い結果をえる為の基本であり、信頼性の低い解析結果にふりまわされてしまう事を防ぐ基本でもある。

6．1． パターン認識と偶然性との関係（線型2クラス分類機の古典的問題）

パターン認識の線形／非線型手法を用いた分類／予測問題で常に考慮すべき留意点として、「偶然性（ＣＨＡＮＣＥ ＣＯＲＲＥＬＡＴＩＯＮ）の回避」がある。 この偶然性の回避が保証されなければ、解析結果の信頼性を失い、パターン認識による解析そのものが無意味となる。

□ 偶然性の問題とは？

線形分類機において、特に意味がないのに分類出来ることを「偶然による分類」という。

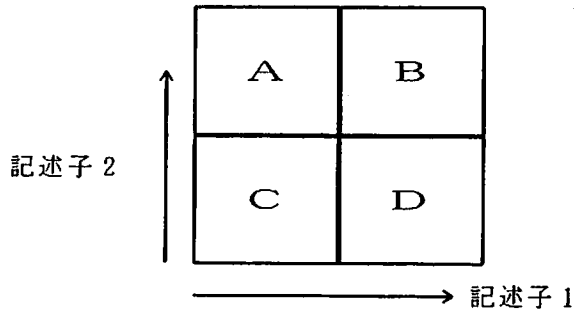
この問題は2クラス分類において、解析に用いる記述子がＮ個の時、2^N 種類の分類パターンが存在し、従ってパターン数が少ない時は意味の無い分類が実行される可能性が高くなるという、単なる確立的な問題に由来するものである。 この時、用いた記述子とパターンのクラス情報間には特に重要な相関は存在せず、単に分類が出来たという事実だけに止まる。

通常の解析業務では単に与えられたデータに関し分類するだけでなく、その結果の解読を行い新たな展開に結びつける事、クラス未知パターンの分類予測を行う事にある。 この目的で解析を行う時、データとクラス情報間に何の相関もない単なる偶然で得られた結果について解読することは無意味であるし、正確な分類予測も困難である。

以下ではこの「偶然による分類」について、記述子数とサンプル数を基本として検討をくわえる。

□ 記述子数とサンプル数との関係について （基本）

例）以下には2クラス分類の時を例に取って述べる。



パターンA，B，C，Dとが、記述子1及び2とで4分割されている状態を示す

Ｎ＝2で分類可能な場合の数は 2^2 で4通り存在する。

図1． 2次元（記述子）によるパターンA，B，C，Dの分類

図1では2次元データを用いて4個のパターンA、B、C及びDとを分類している状態を示す。 分類パターンは1次元（記述子）増加する毎に2倍となる。 従って、2次元の時は $2^2 = 4$ 種類の分類が可能である。 従って、4個以下のパターンを分類する時、この4領域内に各パターンが納まるため必ず分類可能となる。

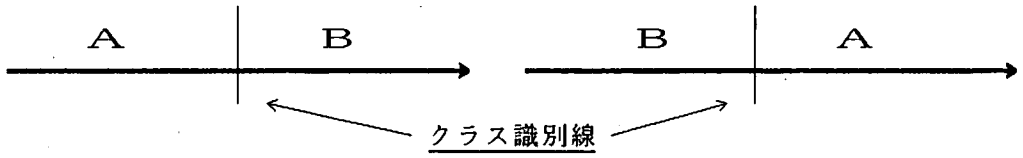


図２．１次元によるパターンＡ，Ｂの分類

図２ではパターンＡ，Ｂが１次元上に存在している時、クラス識別線で２分割される事を示しているが、この時の分割可能性のケースは図２に示されるようにパターンＡ，Ｂとが互いに反対に入れ代わった２ケースだけである。　従って、１次元だけでクラス分類できるケースはこの２ケースであり、分類対象とするパターンが２個しかない時には偶然に分類する可能性が高い（この場合５０％）。

しかしながら、この１次元でパターンが３個以上のものを２分割する時、例えばサンプルが１００個のものを１次元で２分割出来た時、この分類が全く偶然に出来たと考える事は不可能である。　つまり、この解析結果は必然的に分類出来たものであり、分類に用いたデータとサンプルのクラスとの間には何らかの必然的な相関関係があるといえる。パターン認識における解析ではこの必然的な解析状態を意識的につくり出すことで、解析結果の信頼性を高める事が可能である。

この２ケース（１次元で２個のサンプル、１次元で１０個のサンプル）について偶然により分類される可能性について求めてみると以下のようになる。

①　１次元上で２個のサンプルを２クラス分類する時

先ず、図２から１次元で２クラス分類される場合の数Ｒは２ケースしかない。
次に、２個のサンプルが２クラスに分類される場合の数Ｃは対象性を考慮すると、やはり２ケースしかない。　従って、分類可能となる可能性は

$$P = \frac{\text{パターン数による２分類可能な場合の数（C）}}{\text{次元数による２分類可能な場合の数（R）}} = \frac{2}{2} = 1$$

により、この場合はＰ＝１であり、。　つまり、サンプルがどんな分布をしていても必ず２分割出来る事が保証される事になる。　この事から、次元数を増やすことで分類パターンが急激に増大し、必ず２分割出来るようになる。　これは簡単に証明出来る。　たとえこのようにして分類を強引に実行しても、その得られた結果は偶然により分類された事になり、結果の信頼性は全く無い事になる。

②　１次元上で１０個のサンプルを２クラス分類する時

先ず、図２から１次元で２クラス分類される場合の数Ｒは２ケースしかない。
次に、１０個のサンプルを２グループに分割する場合の数Ｃは以下の式によりもとめられる。

$$C = \sum_{d=1}^{5} \frac{10!}{d \times (10-d)!} = 7603$$

従って、この場合分類可能性Ｐ（％）の値は

$$P(\%) = \frac{2}{7603} \times 100 = 0.026\%$$

という極めて低い値となる。

＊パターン数が２０の時、２クラス分類の為の場合の数は７４５１２１６２５３６通りとなる。　パターン数が増えるにつれて急激に場合の数が増大する事がわかる。

即ち、このように低い分類可能性にもかかわらず分類が成功したならば、この分類は偶然に分類されたのでなく、必然的に分類されたのであるという結論になる。　この結果、①のケースは全くの偶然によって分類される事が多いが（結果の信頼性が無い）、②のケースでは逆に偶然による分類の可能性は低く、必然による分類結果（信頼性の高い結果）となる。

Pの値がどの程度の値になれば信頼の出来る結果となるかについては様々な議論があるだろう。　一般的に統計解析の分野では５％及び１％信頼度という値がよく用いられている事を考えれば、１％以下であれば充分な信頼度が得られているものとして考えてもおかしくないであろう。

□　「偶然性」問題における次元数とサンプル数との関係（一般化）

次元（記述子）が一つ増える毎に分類可能な場合の数は２倍ずつ増加する。
従って、次元数ｄにより定まる分割可能な場合の数Ｒは以下の式で示される。

$$R = 2^{d} \qquad (1)$$

この結果、次元数がＮで、サンプル数が 2^{N} 以下の時には必ず分類出来、この分類結果は偶然により支配されている事は明白である。
一方、サンプル数がｎの時、このサンプルを２クラスに分類出来る場合の数Ｃは単なる組み合わせ問題であり、以下の式で示される。

$$C = \frac{1}{2} \sum_{k=1}^{n} \frac{n!}{k \times (n-k)!}$$

これらの項目を考慮し、与えられた記述子（次元数）ｄでサンプルｎを２分割出来る可能性Ｐは

$$P = \frac{\text{サンプルnに対する2分割の場合の数}}{\text{記述子dによる2分割の場合の数}} = \frac{C}{R}$$

で示される。　つまりＰが１の時、与えられたサンプルは理論上必ず２分割出来、Ｐが小さい値になればなるほど２分割できる可能性が小さくなる。　このＰが充分に小さい時、与えられたサンプルが２分割できたならば、この分割は「偶然による２分割」の可能性の少ない、信頼性の高い結果を意味する。
いま、サンプル数ｎを次元数ｄで割った値をαとする。　この値をＸ軸に、ＰをＹ軸に取りサンプル数の値毎にプロットしたものが図３に示されている。

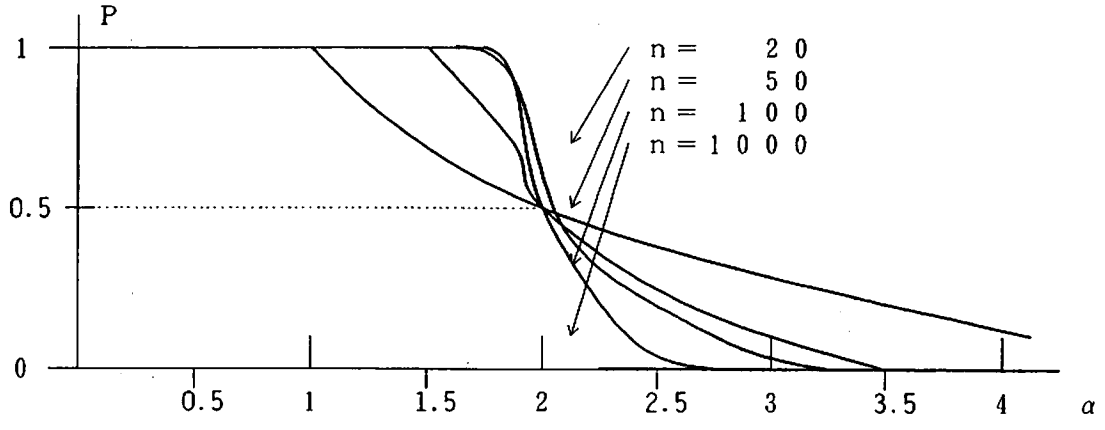


図４．　２分割の可能性に対するα（サンプル数ｎ／次元数ｄ）とＰの関係

この図より、分類可能性Ｐを小さくする要因として２つ存在する事がわかる。　一つはサンプル数ｎであり、もう一つはα＝ｎ／ｄである。　ともに値が大きくなる程Ｐの値は小さくなり、解析の信頼性が高まることを示している。　信頼度の高い解析はＰが０．０５～０．１であるという事を考慮した時、サンプル数が少ない時を含めて考え、αが４以上であれば充分である事がわかる。　従って、線型２クラス分類機を用いる時、αは４以上にすることで解析結果に「偶然性」の入る可能性を充分に小さくする事が可能である。

６．２．　線型／非線型分類問題における適用制限事項（間違った適用をしない為に）

ここでは主としてパターン認識による解析時に前説で述べた偶然性の回避を行なう為にはどのような点に留意する事が必要になるかについてのべる。　この考慮がなされなければ、解析結果は偶然性に支配される可能性が高く、折角の苦労が水泡と帰するので注意が必要である。

以下には、現在提唱されている偶然性回避の為の基準を示す。　これらの基準は解析に用いる記述子の数とサンプル（パターン）の数との関係で示されている。

□　使用サンプル数と記述子数との関係

基準：　解析に用いる全サンプル数と記述子数との間には以下に示す関係を満たす事が必要である。

$$4 \leqq \frac{\text{サンプル数}}{\text{記述子数}}$$

＊この基準は最低の必要条件である。　従って、より信頼性の高い分類／予測率を望むのであるならば、この値は１０以上が望ましいとされている。

＊この基準はパターン認識による解析のみならず、重回帰手法による解析（例えばＨＡＮＳＣＨ／ＦＵＪＩＴＡ法等）時にも有効な基準である。　従って、重回帰手法を用いた解析時にはサンプル数と説明変数（記述子数）との比は４以上にして解析する事が必要である。

□　クラスサンプル数と記述子数との関係

基準：　クラスサンプル数と記述子との間には以下のような関係を満たす事が必要であり、すべてのクラスについてこの要求を満たさなければならない。

記述子数　≦　クラスサンプル数

先の全サンプル数と記述子数との関係を満たしたとしても、クラス間の構成比率に大きな差異がある時、解析の信頼性は大きく落ち込む事になる。

例えば、２クラス分類において全１００個のパターンを用いた時、クラス１が９９パターンでクラス２が１パターンしかない時の解析と、クラス１、２がともに５０パターンずつある時とでは結果の信頼性は大きく異なる。

ここで示された制限事項はこのような問題点をクリアするものであり、解析に用いる記述子数は最少ポピュレーションを持つクラスのパターン数以下である事が必要である。

□　最小使用サンプル数と分類率との関係

基準：　最小使用サンプル数が小さくなればなるほど、そのサンプルを用いた解析で得られる分類率は高い値が要求される。

サンプル数	分　類　率	サンプル数	分　類　率
小 ⟷ 大		２０パターン	１００％
大 ⟷ 小		４０パターン	７０％以上

図　．サンプル数と分類率との一般的関係

前記２制限事項を満たしたとしても充分とは言えない。　残る問題として、全パターン数に関する制限事項が存在するであろう事は容易に想像出来る。　例えば、前記２条件を満たして１００％の分類率を得たとしても、解析に用いたパターン数が１０個の時と１００、１０００個の時とでは当然信頼性は異なっている。

この最少サンプル数の厳密な値をきめる事は困難であるが、著者の計算によれば、２クラス分類を行う時、解析に支障の無い程度の信頼性を達成する為には、約２０サンプルの時１００％の分類率が必要であり、４０パターン用いる時は７０％以上の分類率が必要という事が求められている。　但し、これは解析を行う為の最低限度である。　解析に用いるサンプル数は、常にこれよりも大きな分類率を達成している事が必要である。

一般的傾向として、同程度の信頼性を得る為にはサンプル数が少ない時程高い分類率を達成する事が要求される。

□　適用制限事項のまとめ

前記制限事項を以下に簡単にまとめる。

2クラス分類
（判別分析等）

①	化合物数（パターン数）／記述子数　≧　3～4
②	クラスを構成するパターン数　≧　使用記述子数
③	使用パターン数が40個の時 分類率／予測率　≧　70％

＊　基本的な線型判別分析法はこれらの制限事項に束縛される。

＊　最近隣法は上記の制限事項に束縛されない手法である。

メモ1：　2ʳ の数について

r＝10；　1024

r＝20；　1048576

r＝30；　1073741824

r＝40；　1099511627776

r＝50；　1125899906842624

r＝60；　1152921504606846976

メモ2：　N！について

N＝10；　3628800

N＝20；　2432902008176640000

6．3．　パターン認識適用時の留意事項

パターン認識や多変量解析手法は、ある程度のデータを与えれば相応の結果を出す。しかしこの結果が常に正しいものとは限らない。　パターン認識や多変量解析が恐ろしいのは、特殊な場合を除いて出力結果をながめただけでは信頼正の高いデータなのか、信頼性のないデータであるかの判定が出来ない点にある。

従って真に正しい解析を行う為には、解析を行う前や解析途中で、常にこの解析データの信頼性という点をチェックしながら解析することが大事である。　このチェック項目としては前節で述べた適用制限事項の他にも幾つかの留意事項があり、これらの留意事項について必要事項を満たしながら解析を行わなければならない。

□　手法間の特徴／制限による適用限界及び使い分け

パターン認識手法及び多変量解析手法は多数存在するが、個々の手法の特徴を生かした形で使用する事が必要となる。　作業目的や使用データの特性／種類の差異による手法の使い分けは大事である。　間違った適用をすれば、得られた結果に対する信頼性の保証は無くなる。

クラス分類機の適用限界（２／多クラス分類）

以下にはクラス分類を行う時に用いられる手法について、２クラス分類と多クラス分類とにわけて、それぞれの場合に多用される手法をまとめてある。

表　．分類手法と適用範囲（１）

	２クラス分類	多クラス分類
線型学習機械法（パーセプトロン）	○	
最小二乗アルゴリズムによる判別分析	○	
シンプレックスアルゴリズムによる判別分析	○	
ＢＡＹＥＳ線型判別分析	○	
ＢＡＹＥＳ非線型判別分析	○	
ＡＬＳ法	○	○
最近隣法（Ｋ－ＮＮ法）	○	○
ＳＩＭＣＡ法	○	○
ニューラルネットワーク	○	○

多クラス分類可能な手法は当然２クラス分類も可能である。　また、２クラス分類にしか用いられない手法でも、多クラス問題を２クラス問題の積み重ねと見なし複数回実行する事により解析可能である。　しかし、この場合は運用上でのカバーであり、本質的な機能／特徴としての多クラス分類ではない。

□　データのパターン分布特性に依存する使い分け

様々な手法を分類問題に適用する時、解析対象となるデータセット（母集団）のデータ構造の差異を考慮しながら解析する事が必要となる。

以下には解析対象となるデータセットが２分割可能時と不可能時とにわけ、そのおのおのに用いられる解析手法をまとめてある。　この、２分割可能時とは線型に分割する時必ずインライヤーが存在する為１００％の分類は不可能な時である。　ＢＡＹＥＳ非線型、最近隣法、ＳＩＭＣＡ法、ニューラルネットワークを除いた手法は線型分類機であるのでこの場合１００％の分類は不可能である。　しかし、線型学習機械法を除いたアプローチでは分類時に引かれる分類面が大きく偏ることはないので、２分割不可能時でも比較的信

表　．分類手法と適用範囲（２）

	線型２分割可能	線型２分割不可能
線型学習機械法（パーセプトロン）	●	
最小二乗アルゴリズムによる判別分析	●	○
シンプレックスアルゴリズムによる判別分析	●	○
ＢＡＹＥＳ線型判別分析	●	○
ＢＡＹＥＳ非線型判別分析	●	○
ＡＬＳ法	●	○
最近隣法（Ｋ－ＮＮ法）	●	●
ＳＩＭＣＡ法	●	●

頼性の高い結果が得られる。　表　では白抜きの○で示している。
この分類は必ずしも絶対的な分類ではない。　総ての手法は線型２分割の可能／不可能にかかわらず適用可能である。　しかし、最終的に得られる判別関数等に手法の特徴に依存する特殊な癖が出てくる事がある為、その癖を回避し信頼性の高い結果を得る為にはこのような使い分けが必要になるという事である。
予めデータパターンが非線型の分類をしている事がわかっているならば、非線型の分類機を用いて解析することが１００％分類を得る近道である。
解析のアプローチや目的によってはこの考えは必ずしも正しくはなくなるので注意が必要である。　例えば、単に分類する事が問題でなく、寧ろパターンの分布状態をチェックすることが重要な時には、線型学習機械法を用いる事でインライヤーのパターンを取り出す事が可能であり、時としてインライヤーの情報が重要となる時もある。

□　数値データの種類の差による適用手法の差
数値データは連続変数と不連続変数とに大きく２分類可能である。　連続変数は文字通り総ての取りうる連続な値をとるものであり、計算機的には実数として表現される。また不連続変数は０、１、２といったように間隔をおいた値であり、計算機的には整数として表現されるものである。　この不連続変数のうち、０と１からなるデータをバイナリーデータと呼ぶ。　また、このバイナリーデータは構造－活性・物性相関分野のＨＡＮＳＣＨ－ＦＵＪＩＴＡ及びＦＲＥＥ－ＷＩＬＳＯＮ法においてよく利用されるが、これはダミー変数と呼ばれている。

(1)　連続変数の時
(2)　不連続変数の時

□　情報の重複（相関係数）
(1)　クラスデータの分布状態　（第４章参照）
(2)　クラス間の重なり状態（第４章参照）

□　数値データ（記述子）間のスケールの差
用いる数値データ間に大きなスケールの差がある時には注意が必要である。　事実上、スケールの差があっても解析の実行に支障は無いが、このスケールの差は得られた判別関数／回帰式等の係数の値に大きな影響を与える。　一般的に単位の大きな記述子の係数は小さくなり、単位の小さな記述子の係数は大きくなる。
単に分類／予測だけを行うのであるならばこのような係数の大小の差は問題にならないが、構造－活性相関におけるHansch-Fujita 法における回帰式の係数の解読時には問題となる。　このようなスケールの統一されないデータから得られた判別関数／回帰式の係数についてその軽重を論じる事は困難なことであり、無意味でもある。　一般的にはスケールを合わせたデータを用いる事が必要である。
Hansch-Fujita 法のように回帰式の係数を直接比較して解析するという事が無い限り、解析に用いる記述子に大きなスケールの差があるならばオートスケーリング等によりスケールを統一しておく方が安全である。　但し、このオートスケーリングをしたデータを用いて解析を行う時は用いた記述子の係数を比較する事はできない。　出来るのは、係数の正負の符号の比較程度である。　また、オートスケーリングを行った時予測をおこなうことが困難となるので留意が必要である。

□　解析結果の解読
解析結果の解読は極めて重要な過程である。　解読は解析手法に応じて行い、個々の手法の限界を考慮しつつ行う事が必要である。
例えば判別分析から得られる判別関数と重回帰から得られる回帰式とで、その形式上の差異は存在しない。　この為、回帰式と判別関数とで解読を同じ精度で行うと判別関数の方は過剰評価に陥ることになる。

６．４．　欠損値（ミッシングデータ）の取扱について

□　欠損値について
パターン認識による解析では、一部のデータが存在しないものを入力データとして解析する事はできない。　記述子中にデータの存在しない部分がある時、その記述子を取り除くか、またはその欠損値を有するパターンを解析母集団から取り除くかの手続きが必要で

ある。
　表　で＊マークのデータが欠損している時、考えられる対応策としては、欠損値の多い記述子２とサンプル５とを取り除く事である。　しかし、これだけでは不十分であり、さらにサンプル２、３を除くか、記述子４、８を除くかの処置が必要となる。　最終的にはサンプル数が５、記述子数が６のデータとなってしまう。
　通常の解析では、サンプルを集める事とその関連データとを収集する事が最も時間と人手のいる作業となる。　このような状況下でただ一個の欠落データの為に、関連するサンプルや記述子を解析母集団から取り除くのは大変能率の悪い事である。

表　　ミッシングデータを持つデータマトリクス（６ｘ９）
　　＊印はデータ欠損を示す

サンプル＼使用記述子	1	2	3	4	5	6	7	8	9
サンプル１									
サンプル２		＊						＊	
サンプル３				＊					
サンプル４		＊							
サンプル５		＊		＊	＊			＊	
サンプル６									

□　欠損値を補足する為のアプローチ
　①　平均値を代入する
　②　重回帰手法等を用いて内挿／外挿する
　③　相関係数の高い記述子を選び出し、この選びだされた記述子が持つ値に比例してスライドさせた値を用いる。

　これらの手段をとる事で欠損値になんらかの数値データを補足する事は可能であるが、この様な手続きをへて得られた記述子はあくまでも解析の為の便宜であると認識している事が必要である。　解析過程の特徴抽出でノイズデータとしてとり除かれれば良いが（この場合でも、特徴抽出自体がこのような欠損値により実際とは異なった結果をもたらす可能性があるという事を認識しているべきである）、最重要記述子として残った時は、改めて欠損値を求め直す等の対策を施す事が必要である。　このような欠損値が存在した時に、記述子やサンプルを取り除く為の基準や、欠損値補足を行うか否かの判断基準等について定まったきまりは無い。　ただ、特殊な場合を除き、欠損値を混在させたまま解析出来る手法は殆ど存在しないのが現状である。

６．５．　データ数が少ない時のアプローチ

　パターン認識による解析において、解析パターン数が小さいことは様々な点で問題をもたらす。　解析信頼度が低下すること、解析自体が実行出来なくなること、解析手法が限定されてくる事などである。
　パターン数が少ない時大きな影響があるのは分類問題である。　しかし、最近隣法やＳＩＭＣＡ法は分類手法のなかでもパターン数の大小に左右されることのない数少ない手法である。
　パターン数が少ない時は分類以外の問題でパターン認識を利用する事が安全である。例えば、主成分分析その他のマッピング手法やクラスタリング手法を用いる事は可能である。　データの分布状態がわかれば、その分布状態を基準にしてクラス未知化合物の分類を人間が行うことができる。
　パターン認識による解析は多数のサンプルを扱う事が基本であるが、少ないサンプルしかない時は解析手法を限定する事でそれなりの結果を得る事は可能である。　しかし、多数のサンプルを扱う時に比して解析信頼度は小さい事は事実であり、この点での留意が必要である。

6．6．　内挿／外挿の問題

解析の最終目的はクラス未知化合物のクラス予測等である事が多い。　このクラス予測の時、常に問題になるものとして、内挿及び外挿の問題がある。

一般的傾向として、クラス未知パターンの外挿は線型解析システムの法が良い結果を与える事が多い。　非線型解析システムは、識別平面が複雑な形状を呈しているために、内挿の点では強力であるが、外挿には不向きであるという問題がある。

この問題を回避する目的で、線型性をました形での非線型問題の解決手法へと手法自体を改造する事も考えられる。　しかし、一般的には解析母集団と掛け離れたパターンの予測は行わないことが原則である。　どうしても解析する時、他のアプローチや先見的知識と合わせて解析する事が望ましい。

6．7．　その他のパターン認識適用に当たっての留意／制限事項

■　化学的問題に起因する問題点

□　解析目的とは化学的に何の関係の無い数値データの利用

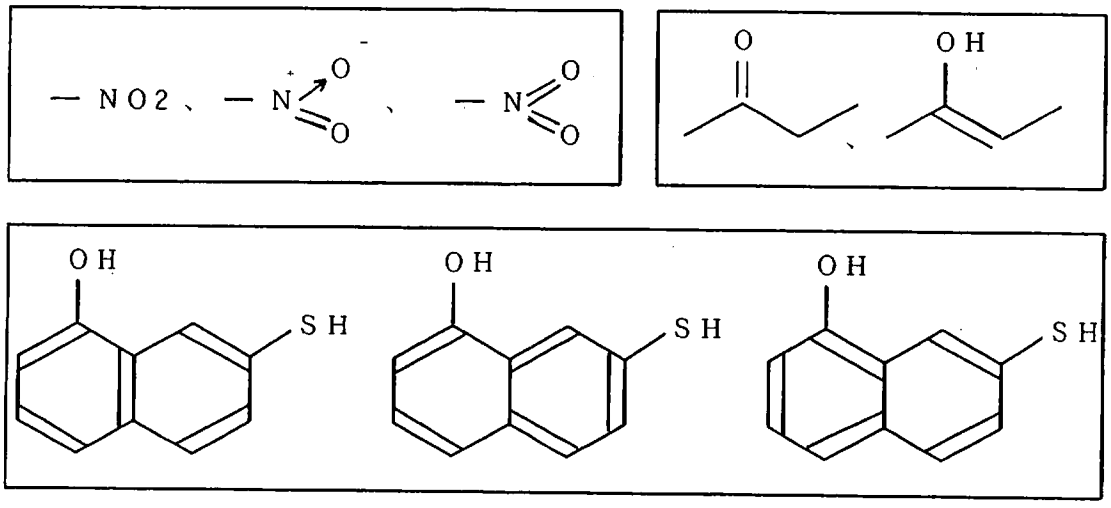
化学分野で入手可能な数値データは多数存在する。　解析目的に沿った数値データを利用する事は当然であるが、無意識に解析とはなんの関係も無い数値データを用いる事もあるので注意が必要である。　これらの数値データは特徴抽出過程において取り除かれるので問題ないが、解析を困難にする一因となるので出来る限り取り除く事が望ましい。

例；　分子量予測にＩＲスペクトルを用いる

例；　構造―活性相関解析で、化合物名に関する数値データを用いる

□　化学的に矛盾を含むデータの利用

①　情報的に統一されない化合物構造式の利用



ここに示された構造式は化学的には全く同一物を表しているものである。　しかし、これらの化合物はシステムに入力されたとき異なる化合物として認識されてしまう。　これらの化合物を同一化合物として処理しているか否かは解析システムに依存する。　従って、構造式を用いた解析を行う時にはこれらの問題がどのようにシステム内で処理されているのかについて予め確認しておく事が必要である。

扱いが異なる時は化合物のシステムへの入力時に、システムで記載法を統一する等の処置が必要である。

□　化学的に極めて重要であっても、解析には邪魔となるデータ

化合物の立体情報は化学では極めて重要な情報である。　しかし、パターン認識や多変量解析手法による解析時には必ずしも常に重要になるとは限らない。　寧ろ場合によっては、邪魔な情報になる事もある。

例えば、構造－活性・物性相関の分野ではこの立体情報はしばしば解析不能に陥いる要因となる。　このような例としては化合物の薬理活性はＤ体とＬ体とで活性が大きく異な

る場合が該当する。　この時、化合物の薬理活性の強度に従った分類はＤ体かＬ体かという１次元の問題となってしまい、多変量を用いる理由が無くなる。

□　化学的に極めて重要であっても、その情報を数値データに変換出来ない為に無駄なデータとなるもの

化合物の立体情報を計算機に入力する事は困難である。　現時点では化合物の３次元構造を情報（３次元座標データ／α、β等の特殊表記／その他）として入力する事は出来ても、そのデータを用いて化合物の３次元情報を的確に反映する（取り出す）数値変換技術は存在しないといった方が良い。

■　化学以外の問題に起因する問題

□　プログラムから生じる制限事項

パターン認識による解析を手で行う事は出来ない。　必ず計算機を使う事が必要となる。この為、解析システムに由来する使用制限等が必ず存在する。　以下に化学システムに存在する一般的な適用制限事項について簡単にのべる。

・使用化合物の原子数制限

水素を含む時と含まない時とで使用可能化合物には大きな差が生じる。

一般的に、パターン認識や多変量解析による解析プログラムで化合物を扱う時は水素原子を省略する事が多い。　これは水素原子の情報は解析時に特別な意味を持たない事が多い為である。　しかし他の目的、例えば分子軌道計算では水素原子は重要なものであり省略不可能であるし、また１Ｈ－ＮＭＲ等の解析では必須である。

・使用化合物の原子種制限

特に重金属や貴金属を含む化合物の時は注意が必要となる。

・使用化合物の種類の限定

例えば立体情報を用いるか、用いないかで適用化合物の範囲や仕事の範囲が大きく変化してくる。

・使用化合物の数の限定

システムの設計により解析に使用出来る化合物の数は制限をうける。

□　解析データ自体に潜む問題

①　データ等にノイズを含む時

種々スペクトルデータの時

・測定誤差（初心者／エキスパート等）
・機器設定ミス
・測定化合物（純度／安定性等）の問題
・測定条件（温度／湿度／溶媒等）の変化

②　単純ミス

・データ入力ミス