② 線型重回帰によるＬＯＧＰ値予測

守口らは１２個の構造パラメータを用いて線型重回帰手法により様々なタイプの化合物についてＬＯＧＰ値の予測を行っている。 この回帰式を求める為に用いたサンプル数は１２３０化合物で、相関係数は０．９４７、標準偏差は０．４３２である。 母集団として構造式の異なる化合物を多数用いているにも関わらず高い相関係数を得ている事は驚異的な事である。

この手法の利点は先のフラグメント付加方式と異なり、フラグメントパラメータの存在に関係なく計算が出来るという点である。 従って、殆どの化合物について計算がかのうである。 現時点で対象化合物として適用不可能な化合物としては錯体等がある。

以下にはこの推定で利用されるパラメータについて示す。

表 ． ＬＯＧＰ値推定に用いられる構造パラメータ

パラメータ内容	係 数 値
(1) 炭素／ハロゲン数*1	０．３４３
(2) 窒素原子及び酸素原子の総数	－０．８４５
(3) オルト（－ＯＨ，－ＣＯＲ），（－ＯＨ，－ＮＨ２），（－ＮＨ２，－ＣＯ２Ｈ），その他	０．５４８
(4) 不飽和結合の総数	－０．０９８
(5) Ａｒ－極性置換基結合の数	０．２８１
(6) Ｎ，Ｏの近接効果*2	０．４１７
(7) ４級窒素及びＮ－ｏｘｉｄｅ	－３．６９４
(8) 両性アミノ酸	－２．１９７
(9) 飽和炭化水素（ダミー変数）（但し１個の不飽和結合を含むものもカウントする）	１．０４５
(10) 環の存在（ダミー変数）（ベンゼン及びベンゼンに直接縮合した環を除く）	－０．３７４
(11) ニトロ基の数	０．５２８
(12) Isothiocyanato (-N=C=S) : 1.0 Thiocyanato (-S-CN) : 0.5	１．５３６
＊ ＣＯＮＳＴ．	０．２５３

＊１， Ｃ＝１．０、Ｆ＝０．５、ＣＬ＝１．０、Ｉ＝２．０

＊２， $X^{2}\cdots Y$、$X^{1}\cdots A^{1}\cdots Y$（Ｘ，Ｙ：Ｎ／Ｏ、 Ａ：Ｃ，Ｓ，Ｐ）であり、修正項として、－ＣＯＮ＜、－ＳＯ２Ｎ＜ について－１を与える。

3．3． その他のパラメータ

その他のパラメータとしては様々な物が考えられる。 これらは解析の目的、分野、解析手法等によりケースバイケースで使いわける事が必要である。 以下にはいくつかの典型的な例について述べる。

□ 画像パラメータ

□ 音声パラメータ

3．4．　スペクトルデータ（NMR，IR，UV，Mass，他）概論

スペクトルデータはパターン認識の化学分野への応用で最初に利用されたパラメータであり、現在までに利用されたスペクトルデータは様々な種類に及んでいる。　パターン認識に利用されるデータとしては、当初は計算機に入力し易い形式を持つマススペクトルや汎用性の高いガスクロマトグラフィーのデータ等が多用されていた。　現在では計算機向きのデータ構造（スペクトル構造が単純な形をしている）を持ち、且つ構造解析ツールとして重要なスペクトルであるC－13NMRスペクトルの利用が多くなっている。　また、スペクトル形状が複雑であっても、得られる情報の価値が高いのでプロトンNMRやIR等のスペクトルデータも利用されている。

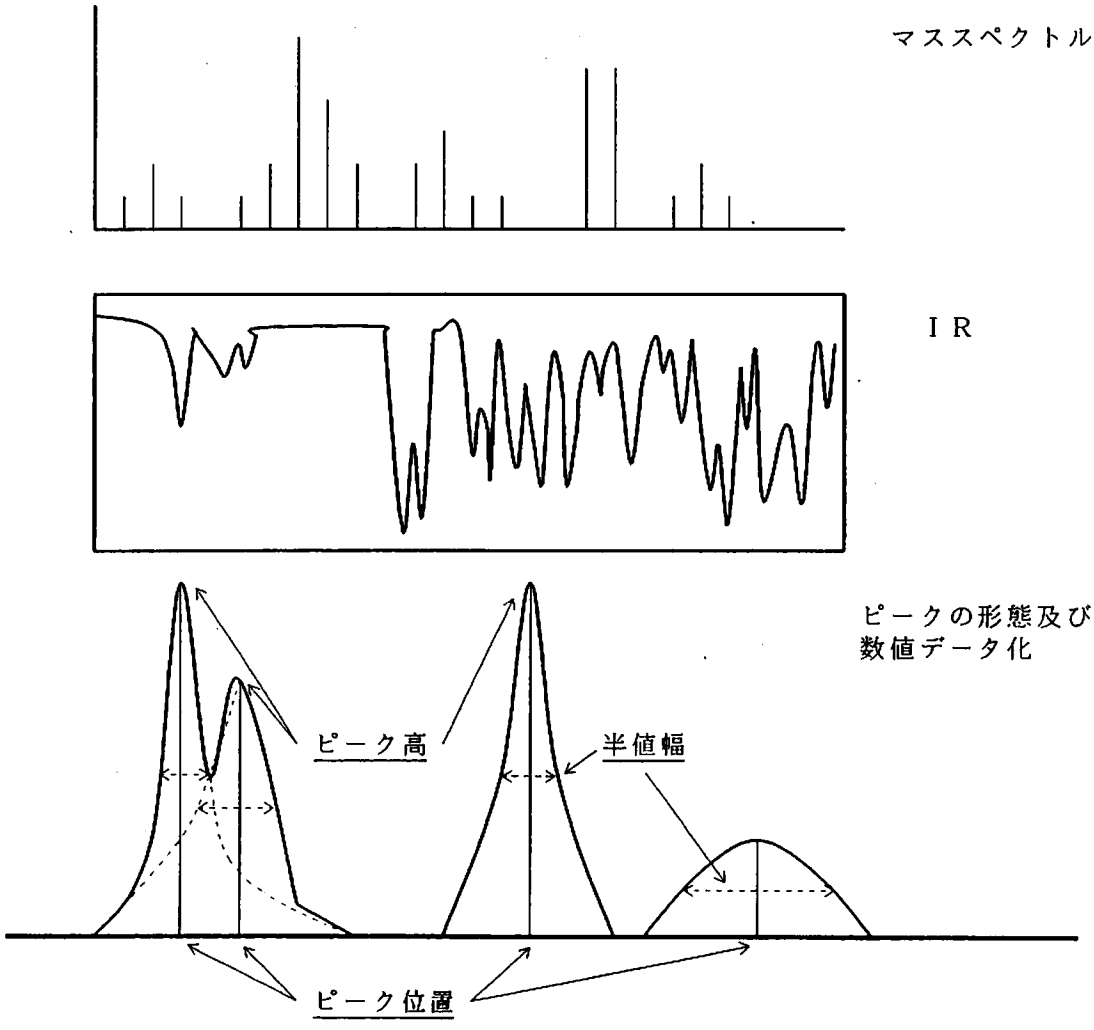


図1．　代表的スペクトルチャートとピークの数値データ化

図1にはスペクトル構造が単純でデータが求めやすい（ピーク位置とピーク強度）マススペクトルと、スペクトル構造が複雑でスペクトルを正確に数値データに変換する事が困難なIRスペクトルを示す。　最下段の図はスペクトル形状を数値データへ変換する典型的な方法について示す。

□　スペクトルチャートから数値データへの変換

一般にスペクトルチャートから得られる重要な情報として、ピーク位置／ピーク強度／ピーク形状／面積強度／ピーク分裂／カップリング／他　といった様々なものがある。これらの情報をスペクトルから取り出す事で、化合物構造／その他に関する情報を能率良く取り出す事が可能となる。

一般的に行われるスペクトルチャートを数値データに変換するアプローチとして、以下に示されるような2種類存在する。

(a)チャート中に潜む重要な情報を取り出し、数値データに変換する。

(b)単に機械的に数値データへと変換する。
(a)のアプローチは先に述べたスペクトル中の様々な情報を効率よく数値データに変換する事を意味している。　マススペクトルではピーク形状が単純であり、得られる情報としてはピーク位置とピーク強度程度である。　従って、(a)のアプローチをマススペクトルに適用する事は容易である。　しかし、ＩＲやプロトンＮＭＲ等のスペクトルチャートになるとピークの形状は極めて複雑であり、スペクトルからピーク位置／強度／その他の意味ある情報を取り出す事は困難となる。
計算機等を用いた解析に利用するデータは、(b)のように単純なアプローチを取って収集されることが多い。　例えば、スペクトルチャートを横軸にそって一定間隔毎に切断し、その切断位置のピーク強度を入力データとするようなアプローチが行われている。　このアプローチでは解析に必要となる重要な情報をとりこぼす可能性もあるが、簡単に再現性良く実行出来る事が最大の魅力である。

□　スペクトルチャートから数値データへの変換における問題点
スペクトルデータを数値データに変換する手法として(a)、(b)の２種類ある事を示した。ピーク位置／ピーク強度／ピーク形状／面積強度等を忠実に数値データにする(a)の手続きは、スペクトルチャート中に潜む情報を効率良く取り出す観点で理想的なものである。(a)のアプローチは、人間がスペクトル解析を行う時のアプローチそのものである。　従って、このアプローチではスペクトルチャートから数値データへと変換する過程が人手による複雑な過程となる。　この過程でデータ再現性の問題が重要となる。
一方、機械的に行う(b)の変換方法は、データ変換過程での再現性等の問題は少ない。しかし、この種のアプローチをとるとデータ量（次元数）が大きくなり、ノイズデータが増えて必要な情報をとりこぼしやすい。　この問題点を解決する為に、測定幅を大きくし測定点を減少させる事が行われる。　これは解析に必要な情報量の減少を意味する。従って、解析精度を保ちつつ観測点（データ量）をへらすことがパターン認識でスペクトルデータを扱う為の重要なポイントとなる。

□　スペクトルデータを扱う時の留意点
パターン認識に利用される数値データが満たすべき条件として幾つか有る。　なかでも最も大切な条件は　①最少のデータ量で充分な解析が出来ること、②数値データの再現性が高いことの２点である。

①の条件はデータの情報密度が高く、良質のデータである事と同意である。　①の条件を満たすとき、ノイズデータが含まれる可能性は少なくなり、解析精度も向上する。　さらに、データ量が少ないと、計算時間／記憶容量等様々な点で有利となる。
②の条件は測定条件の差異により、たとえ同一化合物であっても異なったチャートになることを意味する。　従って、このようなスペクトルデータを用いて解析を行っても、精度の高い解析を行うことは困難である。

□　スペクトル解析におけるヌルデータの重要性
スペクトル解析においてはピークが有るという情報と同程度に、ピークが無いという情報も重要である。　この考えかたはスペクトルのデータベース検索や構造決定等にも利用されている。
パターン認識等で扱う時は０データとなるので、手法的な制限等いくつか留意すべき点がある。　これら留意点を考慮する事で、より効果的な解析につながる事がある。

□　スペクトルチャートそのものに潜んでいる問題点
スペクトルチャートデータを扱う時、チャートその物に潜む問題点をクリアしなければならない。　この問題は、スペクトル計測時点でのサンプルの純度、実験条件、オペレータの個人差、機械の設定／コンディションの差等に由来するものである。　これらの変動要因はスペクトルを数値データに変換する前に解決しておく事が必要である。

3．5．　部分構造記述子

部分構造記述子は、ある特定の部分構造が化合物中に含まれているか否かについての情報を数値データ化したものである。　このパラメータは部分構造の取扱方により様々なバリエーションが生じる。

① 化合物中に部分構造が含まれているか否かの情報
② 化合物中に組み込まれている部分構造の数に関する情報
③ 化合物中に存在する部分構造を抜き出し、その部分構造廻りの環境を考慮して数値データへと変換する。　変換方法は、トポロジカル、トポグラフィカル、及び物理化学的パラメータ等で行なう変換手法が適用される。

部分構造記述子には他の記述子にはない利点があり、構造－活性・物性相関等では強力な記述子として利用される事が多い。　以下に部分構造記述子の利点についてまとめる。

① アルゴリズムが明確である。
② 構造式を基本として数値データへ容易に変換出来る。
③ 記述子の持つ情報の内容が理解し易い。
④ 構造として持っている知識を、そのまま数値データに変換出来る。

部分構造記述子は単に化合物を分類する為に利用されるだけでなく、より高度な目的を持って利用される事が多い。

① 先験的知識／情報を確認する。
② 部分構造情報を用いて、より高度な化合物へと導く。

①の目的で利用する事が多い。　解析の結果、最終記述子として残った部分構造記述子の部分構造は解析目的を実現するのに重要な情報である。　この情報が既に仮説等で提唱されているならば、この仮説をパターン認識の手法で追証した事になる。
例）芳香族アミン化合物の化学発癌研究を例にとる。　この研究分野の仮説の一つとして、芳香族アミンの発癌過程ではアミンのパラ位に電子供与基がある事が重要であるという説がある。　通常のパラメータではこの定説を数値データへと変換することは不可能である。　このような情報を扱う時、部分構造パラメータは最適である。
つまり図　に示されるような部分構造を考え、この部分構造より得られる記述子を加え、芳香族アミンの発癌性について解析を行う。　様々な特徴抽出過程を経た後の最終記述子としてこの記述子が残れば、当初の定説を指示する強力なデータが得られた事になる。これらの記述子が特徴抽出過程で落とされれば、この定説以外にも重要な情報が存在する事を暗示していると考えられる。

・具体的部分構造検索キーとして

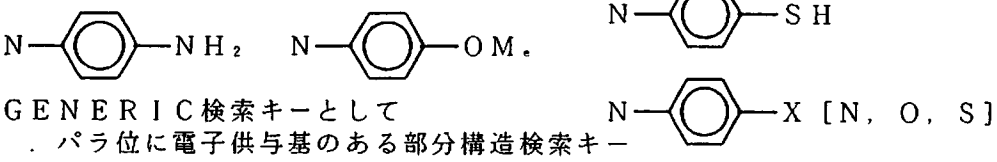
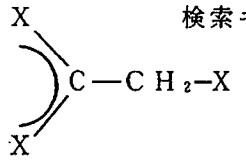
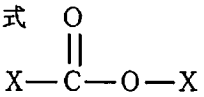
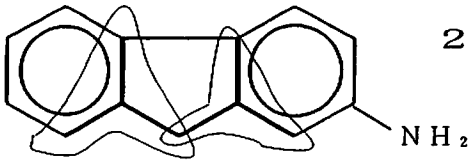
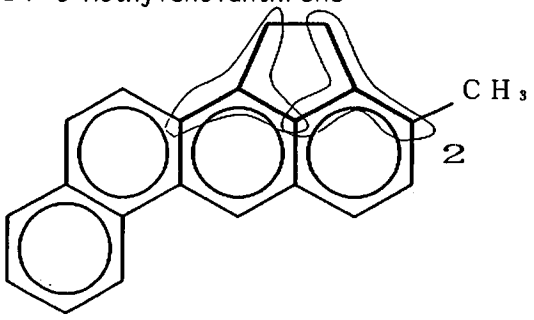
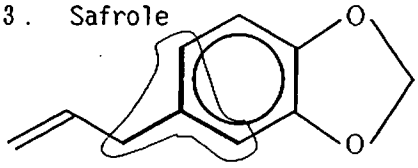
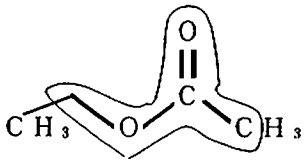


図　．パラ位に電子供与基のある部分構造検索キー

部分構造記述子を創出する為には、化合物中から部分構造を捜し出す部分構造検索の技術が必要である。　この部分構造記述子のヒット部分構造の数は、検索オプションにより様々に変化する。

＊ミニ知識：
部分構造検索は、大きく２タイプに分類できる。　自由度の無い検索と、ある程度の自由度を持たせて検索する２タイプである。　自由度のあるタイプとは検索にある幅をもたせる事を意味し、例えば原子種、結合種、またはこの両方を限定せずに行われる検索であり、一般的にＧＥＮＥＲＩＣ　ＳＥＡＲＣＨと呼ばれている。

□　部分構造パラメータ事例

	検索キー構造式			
				
被検索化合物	部分構造数	ＭＣＩ	部分構造数	ＭＣＩ
1．2-Aminofluorene 	2	3．166	0	0．0
2．3-Methylcholanthrene 	2	3．301	0	0．0
3．Safrole 	1	2．690	0	0．0
4．Ethyl acetate 	0	0．0	1	1．904

＊　部分構造式中にあるＸは水素以外の原子を示す。　つまり、部分構造検索において、この部位の原子種については制限を設けないというＧＥＮＥＲＩＣ検索をおこなっていることになる。

＊　ＭＣＩの値は、部分構造が化合物構造式中に見出されたならばその部分構造を切り出し（Ｘの部分を含む）、その部分構造についてＭＣＩ（分子結合インデックス値）を算出した値を示す。

＊　ＭＣＩの値は切り出された部分構造中、特にＸで示された部分に該当する原子の種類とその結合環境により値が変化する。　これは部分構造パラメータが部分構造廻りの環境を敏感に反映するパラメータになりうる事を意味する。

３．６．　パターン認識による解析時における数値データの取扱について

現在利用されている数値データには様々なものがあるが、解析手続きと解析目的により数値データの使い方が異なって来る。　ここでは、後にのべる特徴抽出との関係について考察する。

①　ある特定の種類の数値データ（記述子）を用いる事が解析前に決まっている時

このケースは入力データと出力データとの相関が明白な時で工学的な分野に多いケースである。　例えばスペクトル解析等の実験を行うといった時はこのケースに入る。　目的とするスペクトル解析に用いられるデータは基本的に種々のスペクトル機器（例Ｃ－１３ＮＭＲ，Ｍａｓｓスペクトル、ＩＲ、その他）より得られたデータを用いる。　これ以外の数値データを用いて解析しても、特別の例を除き、用いた数値データの持つ情報の内容が解析目的と合致する事は少なく、むしろ解析の妨げとなるだけである。

通常の解析では、１つの解析目的に対して１種類のみのスペクトルデータを用いる事が多く、複数種類のスペクトルを用いる事は少ない。　特別に複数種のスペクトルを用いる時は、同時に用いるのでなく、個々の解析ステージに応じて必要な時に使い分けるというパターンを取る。

②　なるべく多くの種類の数値データ（記述子）を用いて解析を行い、特徴抽出によりノイズデータを取り除き、用いたデータセットに対して重要な記述子セットを取り出す事を目的とする解析

この様なケースは、例えば構造活性相関の研究を行う時には重要となる。　一般に化合物の薬理活性を予測／分類する時にはスペクトル解析と異なり、予め使用すべき数値データが特定されている事は少ない。　むしろ、構造活性相関の最終的な目的は考えられる様々な要因から、対象とする化合物の薬理活性を精度高く予測出来るパラメータを如何にして発見するかという事が最重要課題となる。

この様な解析では、解析当初は少しでも解析目的に関与すると思われる要因を考慮すべきであり、このために数多くのパラメータを用いてスタートする。　解析過程で真に必要なパラメータのみを選択しノイズデータを取り除く。　この場合、解析のスタート時点では個々のパラメータ間に何らの重みも無い事が前提条件となる。

しばしばある程度の前提条件が明確になっている時は、この条件を盛り込んだ形で解析をスタートする事がある。　この時は解析の手続きが大幅に軽減される可能性を持つ。しかし、前提条件として設定した情報がデータ解析に対し不適当なものであった時は解析そのものを誤る可能性があるので注意が必要である。

4．　パターン認識／多変量解析におけるノイズ

4．1．　ノイズの種類／定義

ノイズとはこの場合、パターン認識による解析（特にクラス分類／予測）を行う上で邪魔となるもの総てを意味するものとする。　このようなノイズには様々なものが存在するが、主として記述子（数値データ）に関するものと、用いたパターンサンプルに関するものとに大きく分類する事が可能である。　なお、この記述子に関するノイズを取り去る作業を特徴抽出（ＦＥＡＴＵＲＥ　ＳＥＬＥＣＴＩＯＮ）、パターンに関するノイズを取り去る作業をパターン抽出（ＰＡＴＴＥＲＮ　ＳＥＬＥＣＴＩＯＮ）という。

① 記述子に関するノイズ

記述子に関するノイズとしては様々なものが考えられるが、ここでは大きく３種類に分類する。　一つは個々の記述子内部のクラスパターンの重なりに関するノイズであり、２番目は記述子間の情報の重なりに関するノイズ、最後に分類手法の特徴を利用する事でノイズと判定されるもの等がある。　具体的には、これらのノイズはそれぞれフィッシャー比（ＦＩＳＨＥＲ'Ｓ　ＲＡＴＩＯ）、相関係数、及び個々の手法に由来した様々な特徴抽出手法により取り除く事が可能である。

② パターンサンプルに関するノイズ

パターンに関するノイズは主としてそのパターンのロケーションの様子によってノイズパターンと通常のパターンとに分けられる。　一般的に重回帰手法におけるノイズパターンはＯＵＴ　ＬＩＥＲと呼ばれ、パターン認識法／多変量解析でクラス分類を行う時にはＩＮ　ＬＩＥＲと呼ばれている。　この２つのパターンについては４．３．１．で詳しくのべる。　ＩＮ－ＬＩＥＲやＯＵＴ－ＬＩＥＲはパターン認識による分類、解析時にはノイズとなるが、その他の解析手法では、他のパターンと傾向の異なるパターン（例外パターン）であるという観点から情報的には重要なパターンとなる事が多い。　記述子における特徴抽出と異なり、パターンのノイズを取り除く作業はこのような例外パターンと通常パターンとを識別する事となるので、単に使い捨て的にノイズパターンを取り除く事はないように留意する事が必要である。

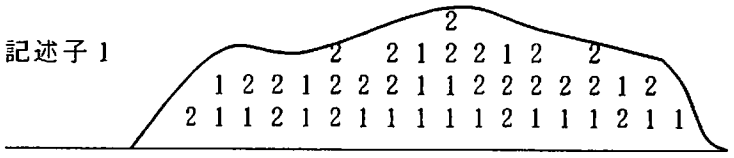
4．2．　記述子に関するノイズ

Ａ．　１個の記述子を対象とする特徴抽出手法に関する考察

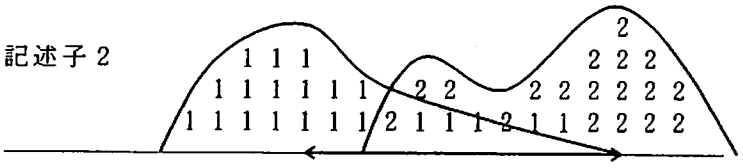
□　データの分布状態及びクラス間の重なりに関するノイズ

パターンは記述子により１次元上に展開されるが、この時クラス毎のパターンの重なりがなく、クラス毎にクラスター（グループ）化され、そのクラスターが狭い領域に集中し、且つクラスター間が大きく離れている記述子がノイズの少ない記述子となる。

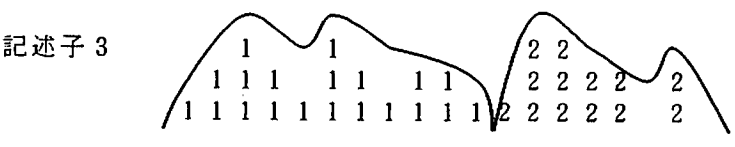
図１には異なるパターン分布をした５種類の記述子のクラスデータが示されている。前記定義により、記述子１～５に変化するにつれてパターン認識の解析に適した記述子（クラス分けに関するノイズの少ない）と変化していることがわかる。



データの重なりが全体に広がっている。



データの重なりが少なくなり、クラス重心の位置が明確になってきている。



データの重なりが無いが、クラス内のデータの広がりが大きい。

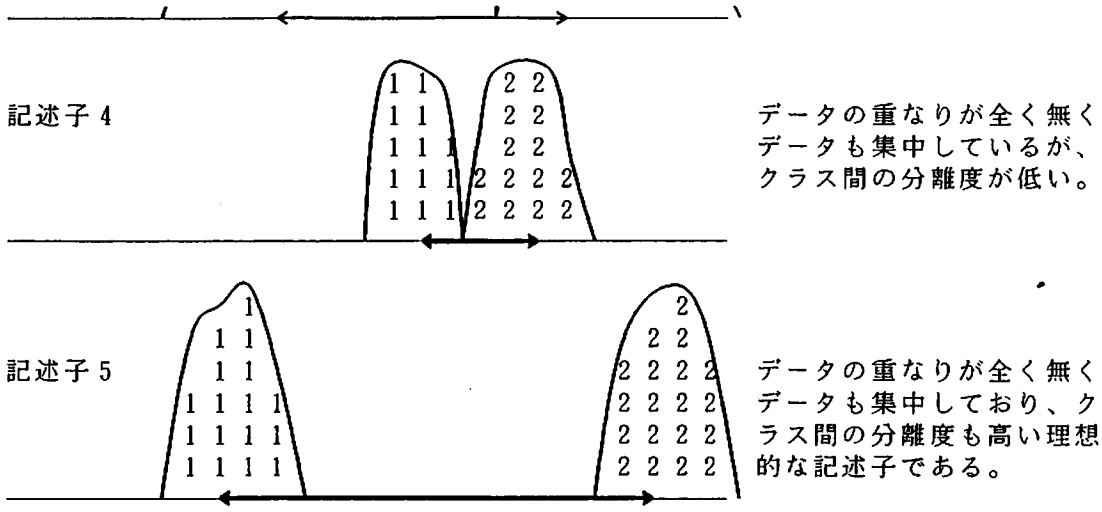


図１．２クラスパターンのクラス分離のパターン

個々の記述子のクラス分類に関する分類能は１次元（１個の記述子）だけで分類を行う事を考えた時、各クラス分布の重なり状態が重要になる事は明確である。　このクラスパターンの重なりを表現するパラメータとしては、クラス間の平均値の差と個々のクラスパターンの分散状態の２つが代表的なものとして考えられる。

① クラス間の平均値の差が大きい程、クラスは分離している
② クラス毎の分散が小さい程、クラスは大きく分離している

この２つの要因を基本としてなんらかのパラメータを求める事で、個々の記述子の分離能に関するモニターを行う事が可能となる。

□ フィッシャー比（ＦＩＳＨＥＲ’Ｓ　ＲＡＴＩＯ）

パターンのクラス毎の分布状態をモニターする特徴抽出手法としてよく利用される手法にＦＩＳＨＥＲ比がある。　この手法は個々の記述子についてクラス間の分布状態（重なり等）に関する情報を取り出すものである。　この手法は、定義式からもわかるように２クラス問題にしか適用できない。

$$F = (\overline{X1} - \overline{X2})^2 / (\sigma 1^2 + \sigma 2^2)$$

$\overline{X1}$：　記述子のクラス１における平均値
$\overline{X2}$：　記述子のクラス２における平均値
$\sigma 1$：　記述子のクラス１における分散
$\sigma 2$：　記述子のクラス２における分散

分子はクラスの平均の差であるのでクラス間の隔たりの情報であり、分母はクラス毎の分散を合わせたものであり、データの分散状態に関する情報である。　従って、クラス間の隔たりが大きく、分散が小さい（狭い場所にパターンが集中している）程ＦＩＳＨＥＲ比は大きくなる。　つまり、パターンの分布形態が理想に近い程ＦＩＳＨＥＲ値が大きくなる事を意味している。

個々の記述子に対し、上記の式に従ってＦＩＳＨＥＲ比を求める。　この値の大小により、対象とする記述子のクラス分類に対する重要性を判断する。

＊Ｆｉｓｈｅｒ比の簡易計算式としてＣｏｏｍａｎｓらによる以下の簡易式が提唱されている。

$$Fc = \frac{| \overline{X1} - \overline{X2} |}{\sqrt{\sigma 1} + \sqrt{\sigma 2}}$$

□　F値の持つ意味

図1にはF値の小さな記述子とF値が大きな記述子とのイメージ図を示す。　この図から分かるように、

F値が小さい時：　クラスパターンの分布は互いに大きく重なっている事を意味する

F値が大きい時：　クラスパターンの分布は互いに重なっておらず、従ってクラス間の分離度が高い事を意味する。

＊　F値が0の時クラス毎のパターンは完全に重なりあっている事になる。

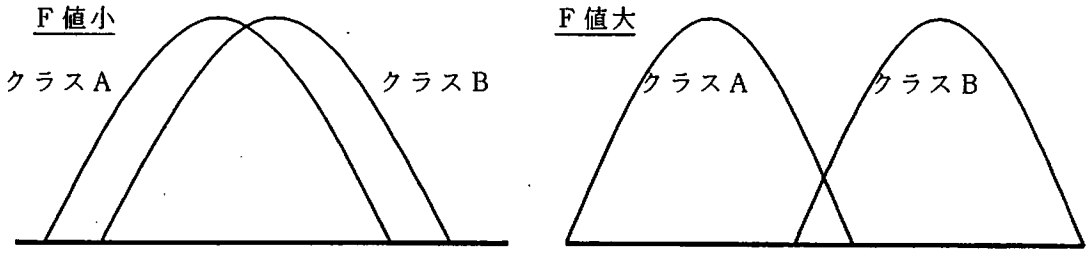


図1．　記述子のクラスA及びクラスBパターンの分布状態の例

ノイズを取り除くという観点から考えた時、F値が小さい記述子はクラス毎のパターンが互いに重なっており、パターンの分類能力は小さいとみなされる。　一方、F値が大きい記述子はクラス毎のパターンの重なりが少なく、パターンの分離能は高い、即ち分類に関しては質の高い情報を持つ記述子という事になる。

特徴抽出ではF値の小さい記述子を取り除く操作が必要となる。　記述子をノイズと判断するか、しないかの基準は解析目的、対象、その他の要因により様々に変化すると考えられる。　後にのべるＡＤＡＰＴ（パターン認識による構造活性相関研究支援システム）ではこの基準は10^{-4}以下の記述子をノイズデータとし、それよりも大きなF値を持つ記述子はノイズではないとしている。

F値　≦　10^{-4}　の記述子　————　ノイズデータ

F値　≧　10^{-4}　の記述子　————　ノイズデータとは見なさない

この値は特に理論的な裏付けは無い。　従って、この値はユーザが独自に設定しても構わないものである。

□　0データの出現回数に従った特徴抽出

個々の記述子の数値データの中には0と0以外の値が含まれる事が多い。　多変量解析では0を数多く含む記述子は解析上あまり望ましいものではない。　特に、0と0以外の値の大きさに大きな差異がある時には問題が多い。　また、0の値自体に意味がなく単に数値データが存在しない時にダミーとして使われる事もある。　このような事から記述子中に含まれる0の値が少ない（ＦＲＥＥ－ＷＩＬＬＳＯＮ解析を除き、通常の解析業務では0の値は存在しない方が望ましい）ものを選びだす。

Ｂ．　複数の記述子同士の関係に関する特徴抽出

□　相関係数による特徴抽出

相関係数は2個の記述子間の情報の重なりをモニターするものである。

$$\text{相関係数} = \frac{\text{2個の記述子の共分散}}{\text{記述子1の標準偏差} \times \text{記述子2の標準偏差}}$$

この相関係数の値が1の時、2つの記述子は全く同じパターンの分散形態をなしている（つまり二個の記述子が有している情報は全く同じ）事を意味する。　また、値が0の時はこの2個の記述子は互いに相関関係（情報の重なり）が全く無い事を意味する。

1　大	↑	2個の記述子の持つ情報は全く同じ
相関係数	↕	
0　小	↓	2個の記述子間に情報の重複はない

ＱＵＩＺ：

① 以下に示す3個の記述子に付いて、ＦＩＳＨＥＲ比と相関係数とを求めよ。

② また、この3個の記述子の1つだけをノイズとして取り出すならばどの記述子を取り出すのがよいか？

	クラス1				クラス2			
記述子1＝（	2	3	4	5	5	7	6	8 ）
記述子2＝（	6	7	8	9	21	25	23	27 ）
記述子3＝（	4	6	2	7	3	5	2	5 ）

＊分散 $= \frac{1}{(n-1)} \sum_{i=1}^{n} (Xi - \overline{X})^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} Xi^2 - \overline{X}^2$

＊標準偏差は分散の平方根である。

C．　解析手法に強く依存した特徴抽出手法

□　線型学習機械法（パーセプトロン）を用いた特徴抽出

(1)　線型学習機械法だけに用いられる特徴抽出手法

線型学習機械法には、この手法の特徴を生かした特徴抽出手法が幾つか存在する。例えば、ウエイトサイン法やバリアンス法などである。　これらの手法は、先にのべたＦＩＳＣＨＥＲ比や相関係数などで特徴抽出を行って選択された（ノイズの取り除かれた）記述子を、さらに線型学習機械法独特の特徴を生かしてさらに厳密に特徴抽出する２次特徴抽出として利用される。

ウエイトサイン法とバリアンス法とを比べた時、バリアンス法の方がより精度の高い特徴抽出を行う事が可能である。　但し、バリアンス法を実行するには線型学習機械法によるクラス分類を数多く実行する事が必要であり、最終的な記述子セットを得るには計算機の計算パワーを大量に消費する事になる。

特徴抽出手法	計算時間	特徴抽出の精度
・ウエイトサイン法	少	低
・バリアンス法	大	高

□　ウエイトサイン法、バリアンス法の基本となるノイズの概念

線型学習機械法では学習により判別式を求め、この判別式を用いてクラス未知パターンのクラス分類を行う。　この分類の時に用いられる判別式は、学習時の初期条件や学習条件[1)]の違いにより、例え同じ母集団を用いても異なる値を持つ判別式となる。　つまり、この判別式を形成している個々の記述子の係数（ウエイトベクトル）が学習毎に異なってくる事を意味する。　一般にこの記述子の係数が初期条件の違いにより大きく異なる値を持つ記述子は不安定な記述子であり、値の変化が小さいものは初期条件の差異に影響されない、むしろ用いた母集団パターンの分類にとり重要な働きをする記述子であると考えられる。

以上の結果を簡単にまとめると、以下のようになる。
即ち、同じ母集団を用いて解析を行った時、初期条件を変えて得られた判別式の係数（ウエイトベクトル）の値の変化が

> ①　大きい時　：　ノイズデータである。
> ②　小さい時　：　その母集団の分類に重要な記述子である。

以下に説明するウエイトサイン法およびバリアンス法はこの判別式のウエイトベクトルの変化に注目した特徴抽出手法である。

[1)]初期条件や学習の条件とは主として、
① 初期ＤＰＯ（ｄ＋１次元目）値の違い。　通常はＤＰＯの初期値として１０００を用いるが、この値を個々の学習（１個の判別式を求める時）毎に変化させる。
② 個々の記述子（次元）の初期値の違い。　通常は個々の記述子の初期値として＋１を用いるが、この値を個々の学習毎に変化させる。
③ 学習時のパターンの参照順を変化させる。　線型学習機械法で最終的に得られる判別式の値は、学習時に参照したパターンの順番によっても変化する。　従って、この参照パターンの順番を変化させる事によっても学習条件を変化させる事は可能である。

等の手法を用いるものであり、上記３つのうちの一つ、または複数同時に採用して線型学習機械法を実行する事により様々な値を持つ判別式を得る。

□　ウエイトサイン法

ウエイトサイン法は同じデータセットを用いて初期条件を変えて学習し、その結果得られた複数の判別式の係数（ウエイトベクトル）の符号（サイン）の変化に注目するものである。

この手法は初期条件を変えて得られた複数の判別式のうち同じ記述子の係数を比較した時、その符号が変わる（＋→－，－→＋）記述子は係数のふらつきが大きく、ノイズとなる記述子であると判定する手法である。　つまり、同じデータセットを分類しているに

もかかわらず、初期条件の違いで得られる判別関数の符号が逆転する程大きく変化する記述子は分類に対し重要な役割を果たしているとは考えられない。

この手法を用いる事で、かなり効果的な特徴抽出を実行する事が可能である。 しかし、実際にはこのウエイトサイン法には内部に幾つかの矛盾を抱えており、このような単純な符号の変化だけで全てのノイズ記述子（判別関数毎の係数の変化量が大きい）を補足出来るとは限らない。

例えば、符号は変化しなくとも係数の絶対的な変化量が大きな記述子が存在し、この様な記述子はウエイトサイン法ではノイズとして補足出来ない。 また、本来ノイズとは見なされない記述子であっても（係数の変化量が小さく、安定している）、その係数の値が小さく、0に近い値を示している時には、極僅かな係数の変化でも符号がかわる為ノイズデータとみなされてしまう。

ウエイトサイン法では以上述べたようにいくつかの矛盾を内包している。 この矛盾を解決する手法としてバリアンス法がある。

□ バリアンス法

バリアンス法では、得られた複数の判別式の係数（ウエイトベクトル）の変化量を求め、この大きなものをノイズデータとして判別式から取り除く手法である。

$$VW_j = \frac{V_j}{\overline{w}_j} \qquad (1)$$

$$V_j^2 = \frac{1}{(n_k - 1)} \sum_{k=1}^{n_k} (W_{jk} - \overline{w}_j)^2 \qquad (2)$$

ここでjは記述子のインデックスであり、kはウエイトベクトルのインデックスである。$\overline{w}_j$は番目のウエイトベクトルの平均値、n_kは用いたウエイトベクトルの数である。

つまり、初期条件を変えて得られた複数の判別式n_k個の中の個々の記述子について1個毎の変化量をもとめる。 この変化量を値の大きい順にならべかえ、値の最も大きな記述子はノイズの最も大きな記述子であり、この値が最も小さな記述子は用いたデータセットの分類に対し最も重要な働きをする記述子である。

n_k個の判別関数を求める作業一回につき、1個から数個の記述子を変化量の大きなものから順に取り除き、残った記述子をもちいて再び同じ事を繰り返す。 どんどん記述子を減らし、これ以上記述子を減少すれば収束しないところまで記述子を減少させる。 この時点での記述子が、用いたデータセットを分類するのに必要な最小単位となる。

□ ＳＩＭＣＡ法による特徴抽出手法

Soft Independent Modelling of Class Analogy
Statistical Isolinear Multicategory Analysis

ＳＩＭＣＡ法から得られる情報としてモデリングパワー（MODELING POWER）とデイスクリミナトリーパワー（DISCRIMINATORY POWER）の二つがある。 これら2つの指標はＳＩＭＣＡ法特有の手法に依存するものであるが、パターン空間における個々のパターンの存在関係に注目した特徴抽出手法として注目に値するものである。 個々のパラメータを求める算出式は既に述べてあるのでここでは特に記載しない。

(a) モデリングパワー（MODELING POWER）

この指標は個々の記述子が母集団のパターンの分散をどの程度説明しているかを表す指標である。 従って、このモデリングパワーが高いという事はパターンの分散をよく説明している記述子という事になる。

(b) デイスクリミナトリーパワー

この指標は対照とするクラスがどの程度分離性良く互いに分散しているかを示す指標である。 従って、この値が大きい記述子は対照とするクラスパターンがそれぞれ互いに離れて／重なりなく分布している事を示す。 従って、特徴抽出に利用出来るものではないが、パターンの分布状態を確認するのに便利なインデイケーターである。

□ その他の手法を基本とした特徴抽出手法

①主成分分析を利用した特徴抽出（因子負荷量の利用）

□ バイプロット図による特徴抽出

(a) バイプロット図により表示される個々の記述子に関する因子負荷量は、直接的には

個々の記述子と主成分軸に対する相関係数として意味づけられるものである。　従って、寄与率の高い、もしくは現在注目している主成分軸に対して低い因子負荷量を持つ記述子は現在注目している主成分軸に関する情報は含まない事になる。　従って、ある特定の主成分軸に対して強い関心を持つ解析を行う時にはこのバイプロット図を用いた特徴抽出は極めて有効である。

(b)　この節の最初の方でも述べたように、分散と分離度は必ずしも一致する物でない事は明らかである。　パターン分類という観点から眺めた時、例え因子負荷量が高くても分類に良好な記述子とは限らない事になる。　従って、バイプロット図による特徴抽出はその利用目的を明確にしながら注意深く行う事が必要である。

②重回帰手法に依存した特徴抽出手法
- Ｆ値
- ｔ値
- 前進選択法
- 後進選択法

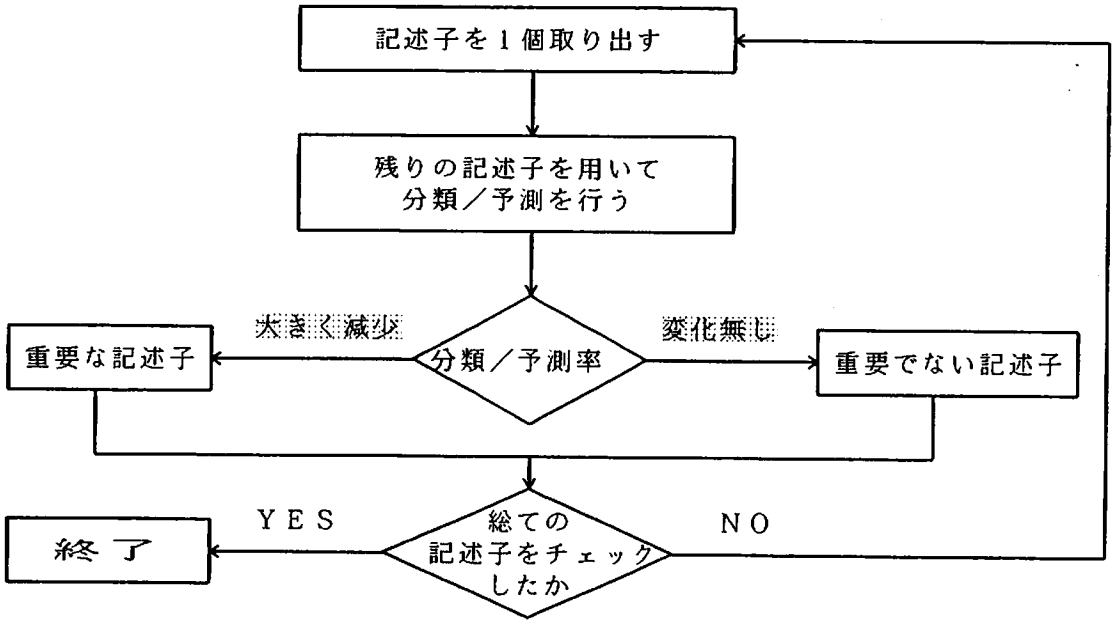
③その他の手法

４．３．　分類／予測率を利用した特徴抽出手法

□　記述子の順次取り出し(ROUND ROBBIN)法による手法

今Ｎ個の記述子が存在する時、１個の記述子を取り除いたＮ－１個の記述子を用いて解析（分類／予測率を求める）を行う。　この時、Ｎ－１個を用いて得られた分類／予測率が大きく減少した時、取り出された記述子は分類に関して重大な情報を含むものと考えられる。　一方、分類／予測率が大きく減少しないならば取り出された記述子は分類に対し重要な記述子ではないという事を利用して特徴抽出を行う手法である。

このアプローチに対する流れ図を以下に示す。



＊　この分類／予測率を求める手法としては、重回帰手法やパターン認識手法のどちらでも構わない。

□　記述子の総当たり法による特徴抽出

記述子の総ての組み合わせに関し分類／予測率を求め、最大の値を示した組み合わせの記述子を最適記述子セットとして選択する。

このアプローチは最も基本的で間違いの無いものであるが、記述子数が増えるに従ってその組み合わせの数は急激に増大してくるという欠陥が存在する。　従って、記述子数が多い時はあまり現実的な手法とはいえない。　このアプローチは他の特徴抽出手法と組み合わせ、それらの手法を１次／２次特徴抽出として記述子を減らし、充分に記述子が減ってきたところで最終的な特徴抽出として行うのが望ましい。

例）記述子数が２０個の時可能な記述子の組み合わせ数

$$2^{20} - 1 = 1048575\text{通り}$$

４．４．　特徴抽出に関する一般的留意事項

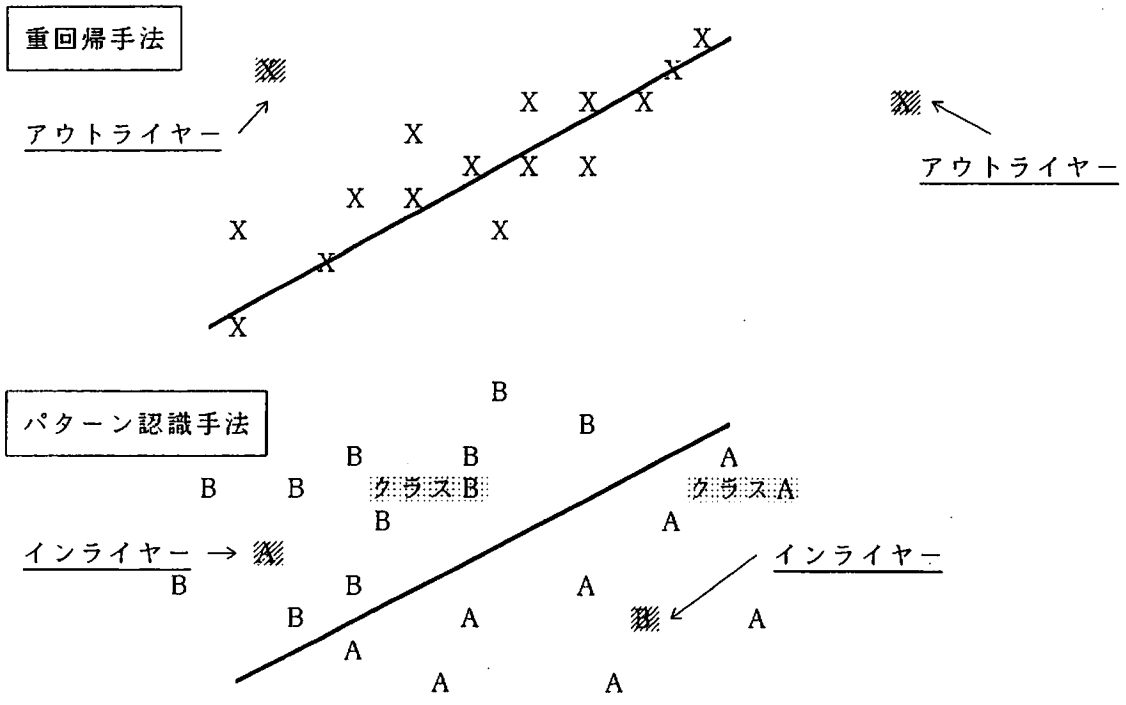
初期記述子数が多い時は特徴抽出過程を２段階にかけて行う事が多い。　第１次特徴抽出で大まかな特徴抽出を行い、第２次特徴抽出でより詳細な特徴抽出を行う。　一般的には第１次特徴抽出ではパターン全体の分散を基準とした統計的手法を主体とした特徴抽出を用い、第２次特徴抽出はより精度の高い且つ解析目的や手法に準じた手法を用いる事が多い。

第１次特徴抽出：Fischer 比、相関係数、零パターン数、その他
第２次特徴抽出：解析手法に強く依存した特徴抽出手法
　　ウエイトサイン法、バリアンスウエイト法、ＳＩＭＣＡ法、その他

４．５．パターンに関するノイズ

□　アウトライヤー（ＯＵＴ－ＬＩＥＲ）／インライヤー（ＩＮ－ＬＩＥＲ）に関する考察

アウトライヤー／インライヤーとはそれぞれ重回帰手法とパターン認識手法とにおけるノイズデータを指した言葉である。　これらのパターンはともに、他の通常のパターンとはその分布状態が異なっているものである。　一般的に、このようなパターンが混在したまま解析を行えば解析自体が完了せず（重回帰の時は高い相関係数が得られない、回帰線が影響を受けてのぞましくない方向に偏る／パターン認識時には学習が収束しない）、また例え解析が完了しても良好な解析結果が得られない等の問題が生じる。



□　ノイズパターンの積極的取り出し、及び利用

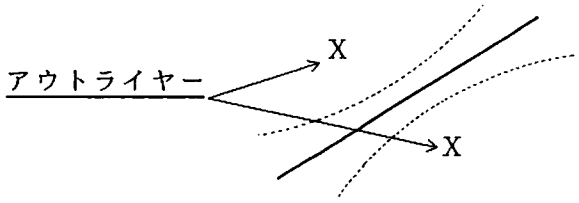
このようなアウトライヤー／インライヤーのパターンは数学的な解析時にはノイズデータとなるが、これらのパターンは従来手法の論理的アプローチによる解析では極めて重要なパターンとなる。　これは、このようなパターンは他の多くのパターンと挙動が異なる例外パターンとなる為、解析時の大きなヒントとなる事が多い為である。　従って、パターン認識、多変量解析を行う時にはこのノイズパターンを取り除き、情報を整理した状態で解析する事が大事であるが、同時にこのノイズパターンをチェックし、その他のパターンと差異を構成する原因を確認する事も重要な作業となる。

ノイズパターン　　取り除いて解析実行　　重要なヒントとなる可能性有り

つまり、パターン認識や多変量解析のパターン抽出機能を活用し、他とは挙動の異なる特殊パターンを積極的に取り出し、それらのパターンについて従来の論理的アプローチを行う事も重要な解析作業となる。

4．6．　信頼区間の考え

重回帰手法においては信頼区間を用いてパターン抽出を行う。　統計的に９０及び９５％信頼区間を設定し、この区間に入らなかったパターンをアウトライヤーとする。　これによりアウトライヤーを取り出す事が可能となる。



□　パターン間の距離を利用したパターン抽出

パターン空間中で他とは異なる特性を有するパターンは他のパターンとは離れた位置に存在するであろう。　この様なパターンはやはり一種のアウトライヤーである。　このようなパターンを取り出すパターン抽出としては、個々のパターンについて他の総てのパターンとの距離を求め、この距離（一般にはユークリッド距離を用いる）が大きい値を持つパターンを取り出す。　このアプローチによりアウトライヤーパターンが取り出される。

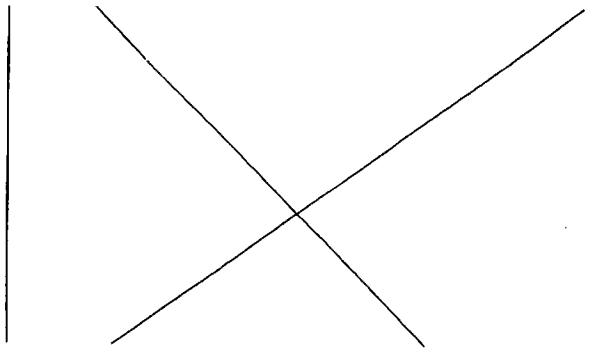
$$Dk = \sqrt{\sum_{j=1}^{m} \sum_{\substack{i=1 \\ i \neq k}}^{n} (Xkj - Yij)^2}$$

ここでｋは現在対象としているパターン、ｍは次元数、ｎはパターン総数を示す。

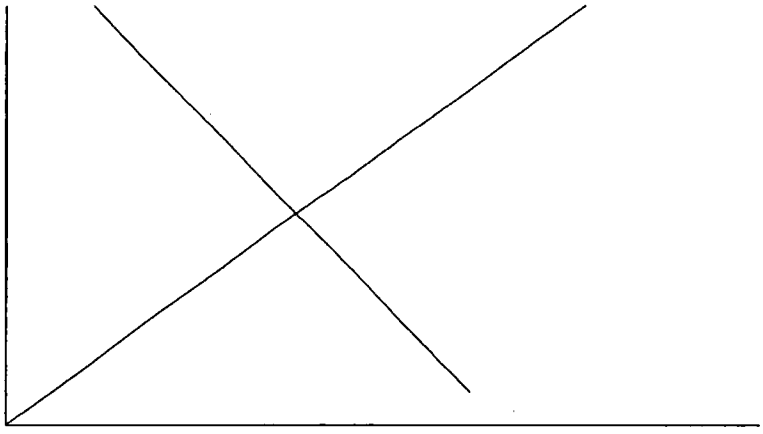
このアプローチは自分の回りに他のパターンが存在しないパターンを取り出すものである。　従って、この手法により取り出されたパターンが必ずしも先に定義したアウトライヤーやインライヤーであるとは限らない事を留意しておく事が必要である。

□　マハラノビスの汎距離によるパターン抽出

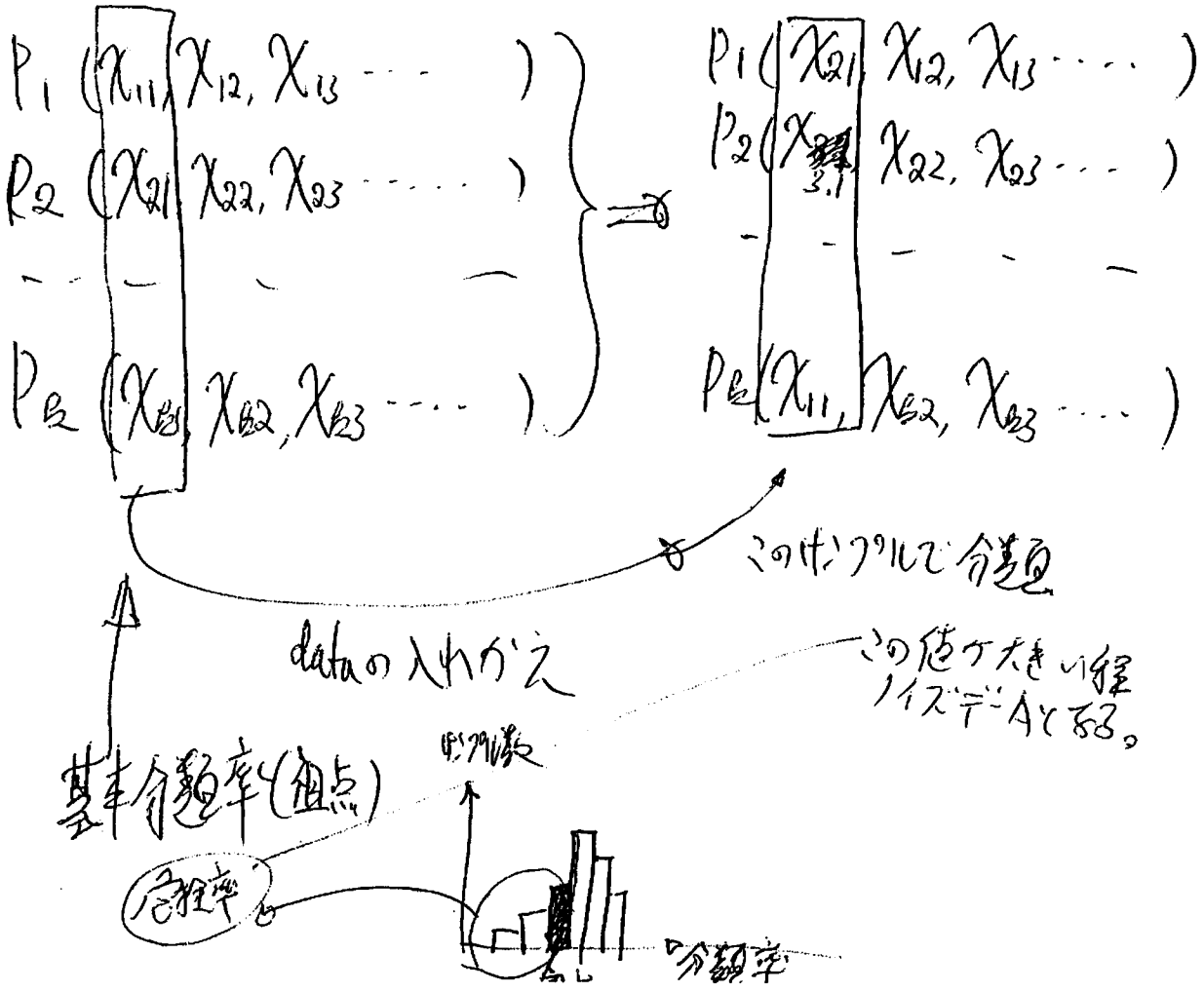
前記のパターン抽出が単なるパターン間の距離だけを指標として行っていたのに対し、このマハラノビスの汎距離を用いた解析では各クラスパターンの分散等も考慮にいれたものとなる。　この距離が大きいものはやはり特殊パターンと見なせる。　但し、パターンとしては特殊であっても、分類という観点からみると理想のパターンである事が多いので解釈には注意が必要である。　どちらかというとパターン間の距離によるパターン抽出と同様に、アウトライヤー的パターンを捜し出すものである。



前記のパターン抽出が単なるパターン間の距離だけを指標として行っていたのに対し、このマハラノビスの汎距離を用いた解析では各クラスパターンの分散等も考慮にいれたものとなる。　この距離が大きいものはやはり特殊パターンと見なせる。　但し、パターンとしては特殊であっても、分類という観点からみると理想のパターンである事が多いので解釈には注意が必要である。　どちらかというとパターン間の距離によるパターン抽出と同様に、アウトライヤー的パターンを捜し出すものである。



○汎用検定法による特徴抽出



5．　分類／予測法について

5．1．　分類率及び予測率について

様々な解析手法が存在しているが、その中でも分類を行う手法が特に多く存在している。　これらの分類手法の利用にあたっては、単に分類を行えば良いというものでなく、その分類結果を正しく評価し、さらなる解析やクラス未知パターンの分類予測につなげてゆく事が必要となる。

この時、最も大事となるのが分類率や予測率である。　即ち、自分が用いたデータセットに対し、行った解析結果の評価がこの分類率や予測率などで示されるからである。つまり、この分類率や予測率の値が低ければ、解析のやり直しを意味する事になるからである。　従って、このように極めて大事な役割をになっている分類率や予測率の信頼性が低ければ、間違った解析結果や判断を解析者に与えてしまう事になる。　さらにこわいのは、例え信頼性が低い結果であっても、単に結果として出てくる数値データをながめただけでは、その数値データに関する信頼性の高低を判断する事が不可能である為である。

このような重大な任務を担っている分類率や予測率というものは、その信頼性を高く保ちつつ、且つ簡単にその値を求められる事が出来るのが理想である。　このような目的に対し、幾つかの手法が提案され、実用化されている。

5．1．1．　分類率

□　分類率の定義

分類率とは、クラス既知パターンについて分類を行った時の分類結果の正答率を示すものである。　従って、分類率は総てのデータ（母集団）を対象として分類を行った時の正答率で示される。

$$\text{分類率（\%）} = \frac{\text{正答したパターンの数}}{\text{解析に用いた全パターンの数}} \times 100\ \%$$

分類率は上記の式で求められるが、この分類率の値の信頼性を高いものに保つ為には解析時に様々な制限事項を満たして行う事が必要である。　この問題は、数多くのパターンを扱う解析作業に常に付きまとう重要な問題である「偶然性（ＣＨＡＮＣＥ　ＣＯＲ－ＲＥＬＡＴＩＯＮ）」の回避という点を考慮する事を意味する。　この問題を無視、或いはクリア出来ない時には、総ての解析結果の信頼性が失われ、折角の解析作業が水泡に帰す事になりかねないし、ある時は全く見当ちがいの結論に導かれる事も少なくない。

この解析結果の信頼性を高度に保つための様々な解析上での制限事項に関しては、第６節で詳しくのべる。

5．1．2．　予測率

□　予測率の定義

予測率とは、クラス未知パターンについてクラス予測を行った時の予測結果の正答率を示すものである。　この予測率を求めるのは、実際にクラス未知パターンのクラスを予測し、その結果が予測と一致しているか否かを確認する事が必要である。　しかし、この様な実験を実際の現場で行う事は極めて困難である。　また、これら分類手法の価値は実際の実験を行う事なくクラス分類予測を高い信頼度で行う事にある。　従って、予測率は既にクラスがわかっているクラス既知パターンを用いて算出し、その結果を用いてクラス未知パターンに適用する事が必要である。

□　予測率を求める為の手法

・　ＬＥＡＶＥ－Ｎ　ＯＵＴ法
　　（ＪＡＣＫ　ＫＮＩＦＥ）　（ＲＯＵＮＤ　ＲＯＢＢＩＮ）

ＬＥＡＶＥ－Ｎ　ＯＵＴ法は予測に対して極一般的に用いられている手法である。　このＬＥＡＶＥ－Ｎ　ＯＵＴ法の内、Ｎが１の時には別名ＪＡＣＫ　ＫＮＩＦＥ法、ＲＯＵＮＤ　ＲＯＢＢＩＮ法とも呼ばれる。

手続的には、全体のデータセット（Ｄパターン）から１個（Ｎ＝１の時）パターンを取

り出し、この取り出されたパターンをクラス未知パターンとし、残りのデータセット（Ｄ－１）を母集団として、この取り出されたパターンのクラスを予測する。　予測が終わったならば、取り出されたパターンを（Ｄ－１）個の母集団に帰してＤ個のパターンに戻す。　再び、先に取り出されたパターンと異なるパターン１個を取り出し、このパターンについてのクラス分類を行う。　この手続きを総てのパターンＤ個について繰り返し、この予測の結果を用いて全体の予測率とする手法である。　図　にこのＬＥＡＶＥ－Ｎ　ＯＵＴ法の流れ図を示す。

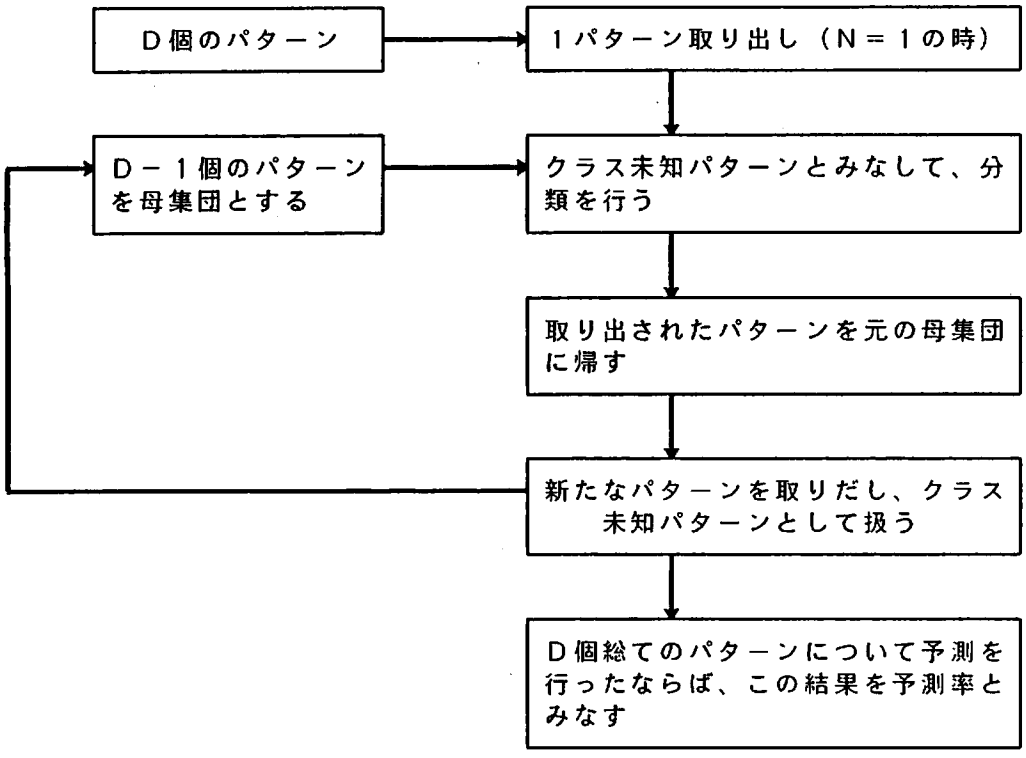


図　．ＬＥＡＶＥ－Ｎ　ＯＵＴ法流れ図

Ｎが２以上の時、取り出されるサンプルの組み合わせは乱数等を用いておこなうのが解析に人為性等の問題を排除する観点からも必要である。　予測率は取り出されるサンプルの組み合わせかたによっても影響をうけることを留意しておく必要がある。

□　ＬＥＡＶＥ－Ｎ　ＯＵＴ法の特徴

この手法の特徴によりＮは１以上の数で実行可能であるが、予測率のデータそのものの信頼性を考えた時、Ｎ＝１の時最も高い信頼度を達成し、Ｎの値が大きくなるにつれて予測に用いられる母集団のパターン数が減少してゆく為、得られる予測率とその信頼度の両方が減少してゆく。　一方、予測率を求める為に必要な予測実験の回数はＮ＝１の時母集団の数と等しくなり、実験回数は最高回数となる。

□　ＬＥＡＶＥ－Ｎ　ＯＵＴ法のＮ値と母集団のサンプル数との関係

ＬＥＡＶＥ－Ｎ　ＯＵＴ法は比較的少数のパターンを扱う時に利用される手法である。特にＮ＝１の時は母集団のサンプルを最大限に利用出来るという点で優れている。　但しサンプル数にも程度があり、極端にサンプル数が少ない時はその予測結果はアウトライヤーパターンがある時は大きくその予測率を下げる結果となる。　この為、少数のパターンを扱う時は注意が必要である。

パターン数がおおくてるにつれ、取り出すパターンの数（Ｎの値）が増大してもその予測率の値はアウトライヤーからの影響を受けにくくなる。

□　ＬＥＡＶＥ－Ｎ　ＯＵＴ法の使用上での注意事項

ＬＥＡＶＥ－Ｎ　ＯＵＴ法は少数のデータを扱う時に貴重な手法であるが、先にも述べたようにその利用サンプル数はある程度の数が必要である。　さらに、予測を行う手法の差異により得られる予測率の値が変化する事がある。　これは特に線型学習機械法を用いる時に起こる。　この時、Ｎ＝１の時よりはＮ＝２、３、と多くなる方が誤分類率は下がる事になる。　この特徴は線型学習機械法で得られる最終判別関数がパターンの接線方向

に落ち着く事に起因するものである。

□　ＬＥＡＶＥ－Ｎ　ＯＵＴ法のＮ値と試行回数及び予測信頼度との関係

解析手法にもよるが、一般的にクラス未知パターンの分類を行う事は計算機の時間がかかる作業である。　従って、パターンの数が少ない時は予測率を求める為の予測実験は、Ｎ＝１として最大数だけ予測実験しても大きな影響は無いが、パターンの数が大きい時はＮ＝１として予測率を求める事は能率が悪い。　そこで、予測の信頼度を保ちつつ予測実験の回数を減らす事が理想であり、取り出されるパターン数（Ｎ）と予測率及びその信頼度との間に最も能率の良い予測率の求め方に関するある種の関係が存在する。

表１には用いたパターンの総数（母集団）が１００個の時のＮの数と、その時の試行回数及び取り出されたＮ個のパターンの予測の為に用いられた母集団の数との関係が示されている。　一般的に、予測率の信頼度はその予測率を得る為に用いた母集団のパターン数に比例するものと仮定する事は自然である。　表１のデータをより分かり易くする為、図で示したものが図１である。

この図より、Ｎが小さい時（Ｎ＝１，２等）は予測率を得る為の予測実験回数が大きく減少するのに対し、母集団パターン数の値はわずかに下がるだけで、予測率とその信頼度という観点からは大きな変化はない。　この事は、予測率と解析信頼度が用いた母集団の数に比例するものと過程するならば、Ｎの値が小さい時にその値を少し大きくすると、予測率と信頼度を殆ど下げる事なく試行回数の大幅な減少を達成する事が可能である事を意味する。

従って、母集団の数が比較的多い時には取り出すパターンの数（Ｎの値）を大きくしても、実験には余り影響がない事がわかる。

表１．　使用パターン数が１００の時のＮの値と試行回数及び母集団の数

Ｎ	試行回数	差	１００－Ｎ	差
１	１００		９９	
		５０		１
２	５０		９８	
		３０		３
５	２０		９５	
		１０		５
１０	１０		９０	
		３		５
１５	７		８５	
		２		５
２０	５		８０	
		１		１０
３０	４		７０	
		２		２０
５０	２		５０	

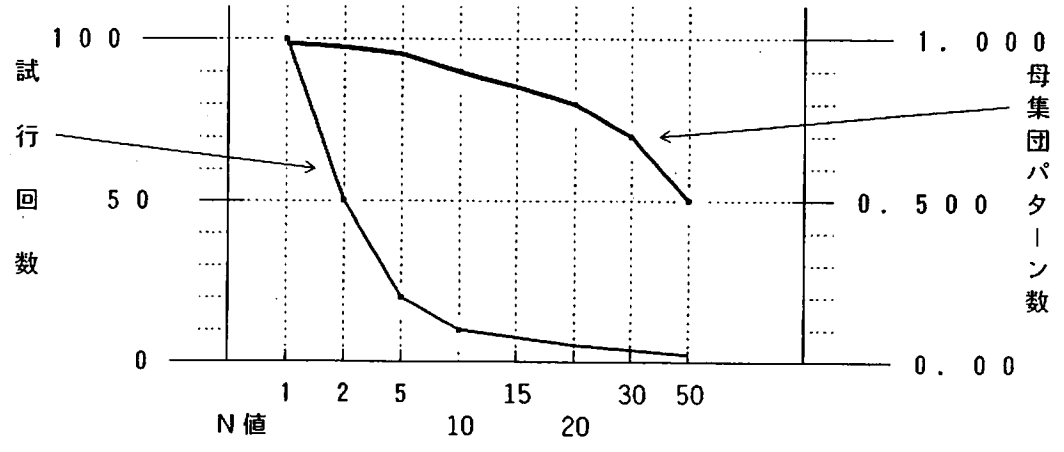


図　．Ｎ値と試行回数との関係

5．1．2．2．　予測率を求めて特徴抽出を行う手法

ここでは予測率を求める手法が、同時に特徴抽出手法にも利用できるアプローチについてのべる。　総ての予測率を求める手法は特徴抽出の総当たり法等と組み合わせることでそのまま特徴抽出にも利用出来る。　このように、予測率と特徴抽出の両方の機能を兼ね備えたものとして現在いくつかの手法が存在している。　このようなアプローチとしては①相互検証法（cross validation）、②ブートストラップ法、③シフト検定法等がある。

これらの手法のうち、②と③の手法は乱数を発生させながらサンプル数を意識的に拡大し、拡大された母集団を用いることでより正確な予測率を求めようとするものである。

ここでのべる特徴抽出は予測率を基本としているので、選択された記述子群は予測に強い、外挿性の高い記述子群が選択されることになる。　従って、予測を重視した解析を行う時はここでのべる特徴抽出を行うことがのぞましい。

□　相互検証法（cross validation）

これはクロスバリデーションと英名そのままで呼ばれることが多い。　手法的にはデータを２分割し、一方を母集団とし、もう一方を予測集団として解析するものである。　従って、Ｎを複数とした時のLEAVE-N OUT 法の変形と考えられる。

本アプローチでは解析母集団と予測集団とを固定してしまうので、解析の信頼性という観点から考えると、サンプル数が少ない時は適用が困難であり、さらに解析母集団と予測集団へのサンプリング過程がもう一つの問題となるであろう。　サンプル群を固定されたグループへと２分割することから、本アプローチは分類問題の時には余り利用されない。線形重回帰といった分類以外の問題を扱う時に利用されることが多い。

□　ブートストラップ法

ブートストラップ法による予測率の算出法についてのべる。　この手続きは以下に示す準で行われる。

①　このＮ個のサンプルについて、乱数を発生させる。
②　この発生した乱数を個々のサンプルに対応させる。
③　乱数の値の大きさに従ってサンプルをならべかえる。
④　並べ変えたサンプルの内、最小のサンプルを取り出しクラス未知サンプルとする。
⑤　取り出された最小のサンプルの所には、２番目のサンプルをコピーして代入する。
⑥　取り出されたサンプルについて予測を行い、正答か否かをチェックする。
⑦　①～⑥を数百回繰り返す。（理論的には３００回程度で良いとされている）
⑧　正答数を試行回数で割り、１００を掛けて最終予測率とする。

$P_1 = (X_{11}, X_{12}, \cdots\cdots, X_{1,N-1}, X_{1N})$
$P_2 = (X_{21}, X_{22}, \cdots\cdots, X_{2,N-1}, X_{2N})$
……
……
$P_{N-1} = (X_{N-1,1}, X_{N-1,2}, \cdots\cdots, X_{N-1,N-1}, X_{N-1,N})$
$P_N = (X_{N1}, X_{N2}, \cdots\cdots, X_{N,N-1}, X_{NN})$

新データセット

予測データ
$P_1 = (X_{11}, X_{12}, \cdots\cdots, X_{1,d-1}, X_{1d})$

解析母集団
$P_2 = (X_{21}, X_{22}, \cdots\cdots, X_{2,d-1}, X_{2d})$
$P_{2'} = (X_{21}, X_{22}, \cdots\cdots, X_{2,d-1}, X_{2d})$　（P_2 のコピー）
……
……
$P_{N-1} = (X_{N-1,1}, X_{N-1,2}, \cdots\cdots, X_{N-1,d-1}, X_{N-1,d})$
$P_N = (X_{N1}, X_{N2}, \cdots\cdots, X_{N,d-1}, X_{Nd})$

□　ブートストラップ法の特徴

本アプローチでサンプリングを行うと、創出しうるデータセット数は最大Ｎ＊（Ｎ－１）となる。　このため、少数のサンプルしかない時でもサンプル数が多い時と同程度の信頼性で予測を行うことが出来る。　この種の問題ではデータセットが大きいということが解析の信頼性向上に結びつく大事な事項である。　本手法では解析母集団の中に同一パターンが存在するという問題はあるが、データセット数を増大させて予測率を求めることが出来る点で、一般的には信頼度の高いアプローチとして評価されている。

視覚的に評価出来る。
各サンプルパターンの正答率がでるので、サンプル単位での予測精度がわかる。
各記述子についての

本アプローチでは取り出されたサンプルの穴の埋め方を工夫することで様々なバリエーションを設けることが可能である。　例えば、単にサンプルをコピーするだけでなく、次元データの一部をさらに他のデータで置き換えたサンプルを持ってきたりすることも可能である。　サンプルを取り出すことなく、単に一部データの入れ換えだけを行うことであればシフト検定法となり、特徴抽出法となる。

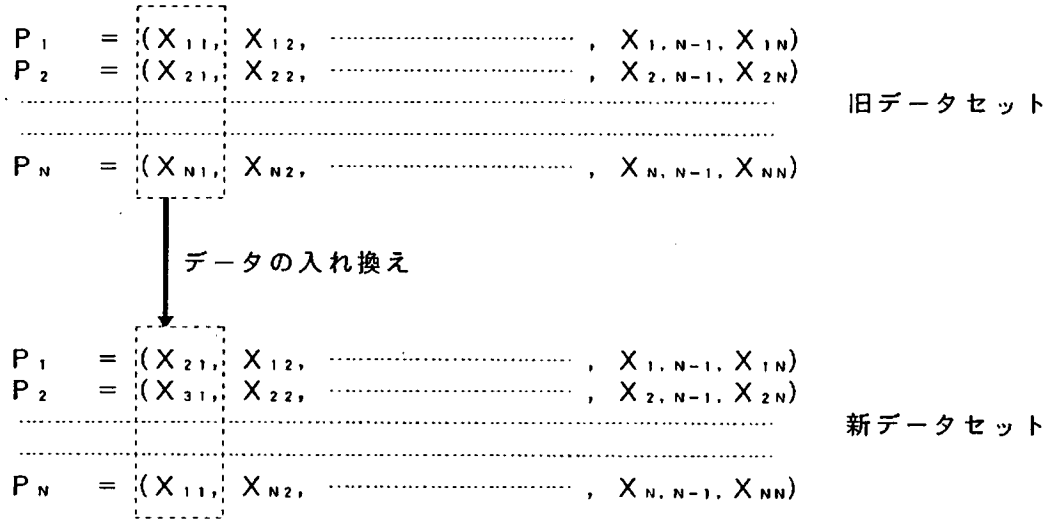


図　．シフト検定法概要

＊ブートストラップ法は一つの矛盾を含んでいる。　これは、同一パターンが解析母集団中に含まれることである。　従って、解析はこの重複パターンにより解析の歪がでるものと考えられる。　しかも、本アプローチが統計の分野で広く発展したものであることを考えるならば、検定という点ではこの点は大きな問題とはならない。　即ち、重複パターンの問題は試行回数を充分に大きく取るということで解決されるのである。　従って、本アプローチが分類や予測を中心とするパターンや多変量解析の分野でなく、統計の分野で発展してきた事を考えれば納得できる。