化学から見た、解析のための
パターン認識概論

パターン認識の化学への適用

パターン認識化学の基礎

〔1〕　パターン認識基本

1．パターン認識とは

1．パターン認識によるアプローチと従来手法によるアプローチとの違い

(1)まじめな人の為の伝統的アプローチ（従来手法）
(2)サボリマンの為のチャッカリアプローチ（パターン認識）

2．パターン認識の活躍はどこで？

(1)様々な利用例
(2)人工知能技術（ＡＩ）との出会い

3．パターン認識の世界へ

(1)パターン認識の歴史
(2) 3－次元空間からＮ－次元空間への招待
(3)パターン認識から画像認識へ
(4)画像認識から化学への展開

4．まとめ

2．パターン認識の様々な手法について

1．パターン認識の定義

(1)人間には解読不可能な情報を計算機に解読（学習）させ、その成果をいただく
(2)学習について
(3)パターン認識の限界及びその打破　（ＭＩＮＳＫＹ　＆　ＰＡＰＰＥＲＴ）
　ＥＸＯＲ問題

2．パターン認識と多変量解析との関係及び分類

(1)教師付き学習（SUPERVISED LEARNING），教師無し学習(UNSUPERVISED LEARNING)
(2)パラメトリック（PARAMETRIC）及び　ノンパラメトリック(NONPARAMETRIC)

3．パターン認識と多変量解析における代表的手法

(1)ニューラルネットワーク
(2)線型／非線型判別分析
(3)主成分分析
(4)因子分析
(5)クラスタリング
(6)マッピング
(7)重回帰手法
(8)その他の手法

4．最適化手法について

(1)シンプレックス法（SIMPLEX OPTIMIZATION)
(2)最小二乗法（LEAST SQUARES METHOD)
(3)ニューラルネットワーク
(4)遺伝的アルゴリズム

3．パターン認識／多変量解析に利用される数値データ（記述子）について

1．Ｎ－次元空間中のパターンはどの様にして表現されるか
(1)パターンと数値データ（記述子）との関係

2．Ｎ－次元パターン空間中のパターン間の距離について
(1)様々な距離基準について

3．数値データの意義及び分類
(1)連続変数／不連続（カテゴリカル）変数

4．数値データ間のスケールや精度について
(1)数値データのスケール／精度
(2)オートスケーリング

5．数値データの詳細
(1)トポロジカル
ＨＯＳＯＹＡ　ＩＮＤＥＸ，　Ｍ．Ｃ．Ｉ．，
(2)トポグラフィカル
ＳＴＥＲＩＭＯＬ　ＰＡＲＡＭＥＴＥＲ
(3)物理化学パラメータ
分子屈折率、分子表面積、その他
(4)その他

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

4．ノイズデータ（記述子／パターン）の選択について

1．ノイズとは？

2．記述子
(1)データの分布　及び　重なりに関するもの
(a)Ｆｉｓｈｅｒ比
(2)複数の記述子同士の関係に関するもの
(a)相関係数
(3)手法に強く依存した特徴抽出法
(a)線型学習機械法
ウエイトサイン法
バリアンスウエイト法
(b)ＳＩＭＣＡ法
DISCRIMINATING POWER
MODELLING POWER
(c)その他の手法
主成分／重回帰／他

(4)分類／予測率を利用した手法
・記述子の順時取り出し(ROUND ROBBIN)法による手法

3．パターン
(1)アウトライヤー／インライヤーの概念
信頼区間、パターン間の距離

(2)マハラノビスの汎距離

5．分類／予測法について

1．分類法

2．予測法
(1) ＬＥＡＶＥ－Ｎ　ＯＵＴ (JACK KNIFE / ROUND ROBBIN) 法
(2) ブートストラップ法
(3) クロスバリデーション法

6．パターン認識適用に当たっての留意事項

1．パターン認識と偶然性との関係

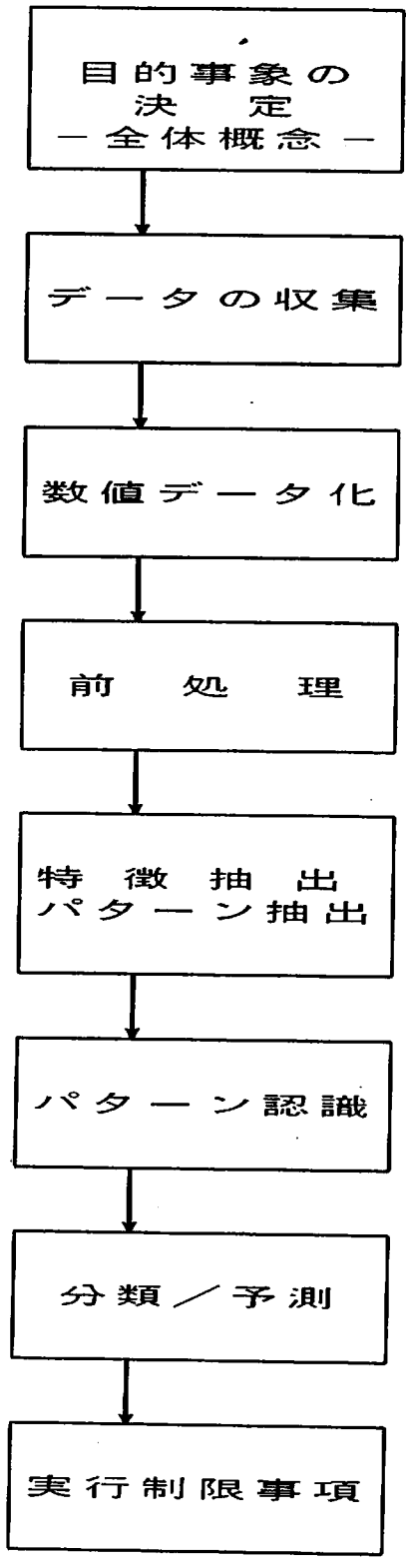
2．適用制限事項（間違った適用をしない為に）
(1)サンプル数と記述子との関係
(2)クラスサンプル数と記述子との関係
(3)使用サンプル数と分類率との関係

3．適用時の留意事項
(1)手法間の情報の差
(2)数値データの種類の差による適用手法の差
　連続変数と不連続変数
(3)情報の重複（相関係数）
(4)クラスデータの分布状態
　クラス間の重なり
(5)ミッシングデータの取扱について

パターン認識の化学への適用

パターン認識の応用

作業解析フロー及び見出し



1．パターン認識とは

3．パターン認識／多変量解析に利用される
数値データ（記述子）について

3．4．数値データ間のスケールや精度について

4．ノイズデータ（記述子／パターン）の選択に
ついて

2．パターン認識の様々な手法について

5．分類／予測法

6．パターン認識適用に当たっての留意事項

〔2〕　パターン認識の化学分野への応用

1．パターン認識の化学への応用
①構造活性相関
②スペクトル解析
・他の一般的アプローチとの比較
③化学反応予測
④バイオテクノロジー分野
⑤その他の分野

2．その他の関連事項

1．ＣＨＥＭＯＭＥＴＲＩＣＳの基本的考え

2．スペクトル解析等への展開

3．パターン認識による化学研究支援システム、ＡＤＡＰＴの機能概要

1．全体機能説明
(1)システムの歴史
(2)システム基本機能／思想

2．単位機能説明
(1)化合物構造式及び薬理データ入力／出力
(2)化合物構造式表示
(3)化合物の０→２次元座標計算
(4)分子力学による化合物の２→３次元座標計算
(5)配座解析機能
(6)化合物重合わせ
(7)記述子創出（化合物構造式→数値データ）
(8)グラフ作成機能
(9)数値データ評価機能（特徴抽出、その他）
(10)パターン認識実行
(11)重回帰実行（ＨＡＮＳＣＨ／ＦＵＪＩＴＡ法）
(12)数値データハンドリング機能

3．ＡＤＡＰＴを用いた実際の解析作業フロー
(1)構造活性相関
(2)構造物性相関
(3)スペクトル解析
(4)その他

4．他システムとの連携
(1)ＡＮＣＨＯＲ
(2)ＣＨＥＭＬＡＢ

4．ＡＤＡＰＴの利用事例

1．構造活性相関
(a)　構造活性相関とは？
・定量／定性的構造活性相関、OPTIMIZATION/GENERATION 他
(b)　定性的構造活性相関の基本的考え
①パターン認識による定性的構造活性相関事例
・定性的構造活性相関への基本的利用事例
・ＬＥＡＤ　ＲＥＴＲＩＥＶＡＬ（リード化合物検索）
・ＬＥＡＤ　ＲＥＣＯＮＳＴＲＵＣＴＩＯＮ（リード化合物再構築）

②ＨＡＮＳＣＨ／ＦＵＪＩＴＡ法による定量的構造活性相関
③化合物３次元座標計算、及び配座解析
(c) 定量的構造活性相関の基本的考え
①重回帰による定量的構造活性相関事例

□ ＡＤＡＰＴによるキノロン系抗菌活性化合物の解析及びそのＬＥＡＤ化合物検索、ＬＥＡＤ化合物再構築への展開
(1) ＡＤＡＰＴによる構造活性相関への一般的解析
(2) ＡＤＡＰＴによるＬＥＡＤ化合物検索
(3) ＡＤＡＰＴによるＬＥＡＤ化合物再構築

２．スペクトル解析事例

(1)マススペクトル

(2)Ｃ－１３　ＮＭＲ

３．その他の適用事例
(1)麝香性化合物の判定

５．その他のパターン認識の化学分野における一般利用事例

(1)燃焼性危険物質
(2)ゲンブ岩
(3)ミルクの判定

〔３〕　その他のパターン認識関連事項

１．パターン認識を主体とした計算機システム

１．総合システム
(1)ＡＤＡＰＴ（化合物の扱いを最優先したパターン認識研究支援システム）
(2)ＡＲＴＨＵＲ（パターン認識の解析力を重視したシステム）

２．単一手法を中心としたシステム
(1)ＳＩＭＣＡ（定性／定量的解析を目的としたＳＩＭＣＡ法を基本としたシステム）
(2)ＡＬＳ（多クラスデータの重回帰的扱いを目指したＡＬＳ法を用いたシステム）

３．統計その他の手法を含むシステム
(1)ＳＡＳ／ＢＭＤ（統計／多変量解析を数多く揃えた汎用解析システム）
(2)ＲＳ／１、Ｓシステム他
(3)ＫＡＭＥＬＯ、ＬＯＴＵＳ他

２．パターン認識のプログラム

１．線型学習機械法

２．Ｋ－ＮＮ（最近隣法）

３．新しい概念に従った、パターン認識手法の展開

１．超ボリューム概念によるパターン認識への新たなアプローチ
(1)パターンに関する新たなアプローチ
(2)超ボリューム概念の導入　（「点」から「空間」へ）

(3)超ボリューム概念に基づいたパターン認識手法
・超球を用いた分類手法、『超球法』
・超球を用いたクラスタリング手法、『超球クラスタリング』
(4)超ボリューム概念に基づいたパターン認識手法による解析例

4．パターン認識における最近の傾向、及び今後の展開

1．新世代のパターン認識、ニューラルネットワーク
(1)ニューラルネットワークの基本
(2)ニューラルネットワークの種類
(3)ニューラルネットワークによる適用事例
(a)最適化問題
(b)分類問題
音声認識
Ｃｈｅｍ　Ｎｅｔによるスペクトル解析の試み（ＤＥＮＤＲＡＬへの挑戦）
反応生成物予測
バイオテクノロジーへの応用
(4)その他のネットワーク構造を有するニューラルネットワーク

2．ファジイ理論の利用
(1)　ファジイとは？
(2)　ファジイによるエキスパートシステム（検討中／実施中）
(3)　曖昧の種類について
(4)　曖昧さの数学的取扱
(5)　従来手法及びファジイ理論に基づいた分類の差異について
(6)　他の解析技術とファジイとの関係
(7)　ファジイの利用例

3．人工知能技術との関係

〔4〕　関連文献

1．ＢＯＯＫ関連

2．ＪＯＵＲＮＡＬ関連

3．その他の資料

〔5〕　用語集

1．パターン認識関連

2．統計その他

パターン認識の化学への適用

その他のパターン認識関連事項

本書はパターン認識に関しなにも知らない人が、化学の幅広い問題にパターン認識を適用する際に利用する事を念頭に書かれた。　内容は御茶ノ水女子大学の大学院生に行った特論『パターン認識の化学への応用』を基本として構成されている。

パターン認識という言葉は知っていても、その実体はよくわからないというのが、大部分の化学者に共通した事であろう。　パターン認識は基本技術である。　その応用は様々な分野におよんでいる。　化学や生物学の分野においてもパターン認識適用の歴史はかなり以前から試みられてきた。　また、書中でものべるがパターン認識の基本となるニューラルネットワークは生体を基本とし、この生体のメカニズムを研究する生理学の分野から発生したものである。　現在、化学の分野では構造－活性・物性・毒性相関等の分野で確固たる地位を築いている。　ケモメトリクス（計量化学）という分野では、パターン認識は主役となっている。　最近注目をあびつつあるバイオテクノロジーの分野でもパターン認識の利用は頻繁になりつつある。　今後、計算機の進歩とともに化学や生物学分野におけるパターン認識の利用は増大するものと考えられる。　パターン認識を身近なものとしておくことが、多くの化学者にとり必要となってゆくであろう。

化学の分野でも様々な機械が導入されるようになってきた。　例え機械をつくることは出来なくとも、使いかたさえ知っていれば目的は達成される。　しかし、どんなに優れた機械であっても使いかたを誤ればとんでもない結果を導き出す。　使い方が優れていれば機械の能力を１００％や２００％にもする事が可能である。
パターン認識も化学の立場から見れば単なる解析道具にしか過ぎない。　パターン認識を知っていれば、研究の幅も深さもひろがる。　しかし、使いかたを誤ればとんでもない結果を導き出す。　パターン認識による解析が恐ろしいのは、誤った使いかたをしても必ずそれらしき結果がでるということである。　パターン認識というオブラートに包まれた解析結果が正しいものであるか、そうでないかを判断するのは結果をみただけでは困難である。　正しい結果を得る為には、自分の責任範囲でパターン認識を正しく使う事が必要である。

著者がパターン認識に関する解析業務を行うに際し留意してきた事は、“パターン認識は解析の道具である”という事を認識した上で間違いの無い運用を行う事である。　従って、ここではパターン認識を利用する立場からみた時に最低限度必要となるパターン認識の基礎知識、パターン認識を正しく自分の目的に適用する為の留意点等を中心として解説する。　最終的に①パターン認識の概要がつかめる事、②パターン認識を自分のテーマに適用出来る事、③間違った運用はしないの３点を理解されれば本書の目的は達成されたものと思う。

パターン認識を基礎の理論から解説した本は多数出版されているが、パターン認識を使うという立場から書かれた本は少ない。　理論と応用との間には大きなギャップが存在する。　パターン認識の専門家はパターン認識の各分野への応用に関しては得意でなく積極的でもないことが多い。　特にパターン認識の研究者と化学の研究者との間には相当なギャップが存在する事は事実である。　私自身は化学者であるが、パターン認識の研究者との間には、パターン認識に対する考え方、見方、扱い方等様々な点で差異が存在することを感じている。　この原因はパターン認識に対する見方であり、基本理論を重視するかその応用面を重視するかで大きな差異が生じる。
著者は数式は苦手である。　数式が苦手であるがゆえに積極的に薬学部に進学したという経緯もある。　自分の癖として、常に感覚にたよりながら判断し、最終的にイメージとして把握ることが出来ないと納得しないということがある。　従って本著では数式の利用は最低限度の使用に止まっている。　寧ろ、定性的な言葉で解説するところが多くなるであろう。　数式に親しんだ方や、よりパターン認識の原点に近づき体と思う方に取っては物足りない点があるであろう。　そのような時は他の専門書を参照される事を勧める。

1. パターン認識と従来手法との様々な観点からの比較

1. パターン認識とは

読者の中に『パターン認識』という言葉を知らない人は殆どいないであろう。 それほどパターン認識という言葉は日常生活に溶け込んできている。 最近では『ニューロ』や『ファジイ』といった言葉も頻繁に利用されるようになり、これは日常生活を支える基本技術としてパターン認識がますます重要になりつつある証拠とも言えよう。

しかし、パターン認識という言葉は知っていても、

・パターン認識とは一体どんなものなのだろうか？
・パターン認識はどのような問題の解決に利用されるのであろうか？
・パターン認識を自分がかかえている諸問題に適用できるのだろうか？

という問いに即座に答えられる人は少ないであろう。

本書を読むことでパターン認識は基本技術であり、様々な分野での利用が可能である事がわかっていただけるものと思う。 実際、化学上の問題でも分析関連では“ケモメトリクス（化学に関連する問題を統計やパターン認識といった計量学的手法で解決する）”という分野を支える基本技術となっている。 このほか、構造－活性・物性相関分野では定性／定量的な解決手段を与える技術として定着しつつあり、最近ではバイオテクノロジー分野での利用も急速に進みつつある。

□ パターン認識が対象とする問題

パターン認識は厳密な理論で組立てられているが、このパターン認識が対象とする問題は“論理的に不明確”なものを対象とする。 但し、この‘不明確’とは本質的に不明確なものでなく、人間が問題解決をする事が困難であるという意味での不明確さである。通常のアプローチではこの不明確な部分（原因と結果間の因果関係）を解明し、その結果えられる何らかの数式やルールを用いて問題解決（分類／予測）をすることになる。従って、この不明確な部分が明確なものにならない限り仕事は進まない。 一方、パターン認識ではこの不明確な部分をブラックボックスとしたままパターン認識を適用する事が可能で、当面必要とする解を得る事が可能である。 これがパターン認識による問題解決の大きな特徴である。

工学的な分野におけるパターン認識の利用はこのブラックボックスに手をつけないままで解析するのが一般的である。 つまり、画像認識や音声認識分野では認識率をいかにして向上するかということが最大の問題であり、原因と結果との因果関係を追求する目的は存在しない。 これに対し、理学や薬学／農学といった分野におけるパターン認識の利用は、単に解析結果だけでなく、同時に何故そのような結果になったのかという因果関係についても考察する事が求められる。 この因果関係を求める基本となる要因解析はパターン認識の様々な手法を駆使し、各手法から得られる視点の異なる情報を総合的に解析することで行える。

パターン認識の最大の利点は対象となる問題をブラックボックスにしたまま、当面必要となる結果（分類／予測等）をえることが出来るという点につきる。 この特徴は従来手法には無い特徴で、パターン認識の最大の魅力でもある。 このようにブラックボックスを残した状態で結果を得ることが出来るのは、複雑な要因を多数含み数式化やルール化が困難な化学上の諸問題にたいし強力な解析ツールとなりうる。 先ず問題解決を先行させ、その過程を通じて新しい情報を取り出して次の展開に結びつけるというアプローチをとることが可能となる。

2. パターン認識と従来手法による問題解決の差

パターン認識を利用するためには、パターン認識はどのような問題に適用できるかという問いに正しく答えられることが必要である。 すなわち従来手法によるアプローチとパターン認識によるアプローチとの差異が正しく認識されていることが必要である。

パターン認識の適用は対象とする問題により変化するが、数式の適用性および問題の不明確さを基準として判断するのが最も確実な判断基準となる。 数式の適用が可能で、ルール化可能な問題は従来手法によるアプローチで解決出来る。 一方、数式の適用性が低い（困難）問題は、①複雑な要因が多数関係し単純化が困難、②正常パターンと異常パターンの区別がつきにくい、③事象が起こっている総ての要因を手持ちのデータで説明出来ない、といった様々な問題を抱えている。 このような問題の解決にはパターン認識によるアプローチが適している。

このような関係を理解する為、以下に簡単な問題について考察を試みる。

QUIZ 1

以下に示された５つの問題について、問題解決における理論／数式の適用性と不明確さ（理論／数式の適用が困難な程高い）の程度の２つの観点からそれぞれ高い順に番号を付けよ。

設問	1	2	3	4	5
理論／数式の適用性					
問題の不明確さ					

設問：

1．ハレー彗星が次回現れる日時を予測する。

条件：彗星に関する軌道／位置／時間の情報が数多く存在し、算出式もある。

2．未知のスペクトルチャートが与えられた時、そのスペクトルの化合物の構造式を推定する。（スペクトルチャートの種類は問わない）

条件：化学構造の分かっている同一種のスペクトルチャートが数多く存在する。

例）
Ａスペクトル ⟶ 化合物構造式Ａ
Ｂスペクトル ⟶ 化合物構造式Ｂ
Ｃスペクトル ⟶ 化合物構造式Ｃ
……

3．活性未知の化合物が与えられた時、その化合物に生物活性があるか否かについて予測する。

条件：生物活性と構造式の分かっている化合物が多数存在する。

例）
Ａ化合物 ⟶ 生物活性 有り
Ｂ化合物 ⟶ 有り
Ｃ化合物 ⟶ 無し
…… ……

4．コーヒーの味覚に関する情報を与えられた時、その味覚を実現するコーヒーのブレンド比を解として与える。

条件：与えられた味覚に関する情報を実現する為に必要なコーヒーの種類とそのブレンド比の実例が数多く存在する。

例）
苦さ＝3 ⟶ ブルーマウンテン＝０．３
香り＝1 ⟶ モカ ＝０．１
…… ⟶ キリマンジャロ ＝０．４
…… ……

5．名曲、あるいはスタンダードとよばれる曲を作曲する。

条件：名曲、スタンダード、その他の曲がたくさんある。

□ ＱＵＩＺ１に対する理論／数式の適用性と曖昧性／感覚性についての考察

設問１～５は仕事の内容がそれぞれ異なっている。 数式の適用やルール化が容易なものから困難なものまで様々である。 数式化／法則化、その対照としての不明確さという点から考慮すると前記の設問はそれぞれ以下のように考えられる。

設問１ 全ての事象がある特定の法則に従って支配されており、数式化が成功すれば問題は完全に解決可能である。 その法則の数も少ない。（解答は１つ）

設問２ 単純な問題に対してはある程度の数式化が可能である。 しかし、全ての事象を数式で説明できる所までは到っていない。 複雑な化合物は、数式だけ

で進めるよりも、比較的数多く存在する経験則を中心として解析する事が望ましい。　（解答は一つ）

設問3　数式化／法則化はあまり期待出来ない。　少数例を除き、数学的手法は単なる解析手法として利用されるにしか過ぎない。（完全解答はない）

設問4　数式化／法則化は殆ど不可能。　データ自体に、人間の感覚的なものが混在し、絶対的な数値データ化は極めて困難である。　また、解答が必ずしも1つとは限らず、その答えは時と場合により変化する事もある。

設問5　曲を作成する為のルールは数多く存在するが、名曲、スタンダードである為の基準はないに等しい。　現在の技術では捕らえがたい芸術の世界である。名曲は音楽の教育を受けない人、音譜を読めない人でも感覚的に感じる事は可能である。　つまり、ルールを知らずともその良し悪しを判定する事が可能という事である。

以上の結果を簡単にまとめると表1のようになる。　数式の適用性は設問1から5になるにつれ困難性が増大する。　現実的には4及び5の数式化は不可能である。　一方、不明確さという点では5から1に向かってその傾向が減少して行く。

表1．設問に対する数式の適用性と不明確さの程度

	可能		中程度		困難	
数式の適用性	1		2	3	4 5	設問番号
	大		中		小	
問題の不明確さ	5	4	3	2	1	設問番号

□　理論／数式の適用性及び不明確さとパターン認識との関係

この数式の適用性と問題の不明確さとの関係はパターン認識の適用領域問題に関係する。

パターン認識は従来のアプローチ（数式の適用による問題解決）では解決が困難な問題領域をカバーする事が可能である。　従って、先に掲げた1～5の問題中不明確な要因を多く含む問題こそパターン認識適用の可能性が高い事になる。

設問中、実際にパターン認識が適用された事例が存在するのは2～4番である。　2と3番の設問に関しては理論や数式を用いたアルゴリズミックな従来型アプローチも試みられている。　しかしこの場合、化合物の種類やスペクトルの範囲等を限定する等の制限付きである事が多く、このような制限を設けないで解析する事は困難である。　4番の設問は現時点において、アルゴリズミックなアプローチは不可能であり、パターン認識的アプローチか、人工知能（ARTIFICIAL INTELLIGENCE)的アプローチにたよるしかない。

設問5はパターン認識手法でも試みられていない問題である。　しかし、現時点で解決へのアプローチがなされていないというだけで、作曲でなく、単に名曲か否かという程度の判定ならばパターン認識でも解決可能かもしれない。　興味ある方は一度試みられる事をお勧めする。

3．従来手法によるアプローチの考察

□　問題解決に対する従来手法（論理思考型）による一般的アプローチ

従来から行われている、問題解決に対する一般的なアプローチとはどのようなものであろうか？　ここで、簡単に振り返ってみる。

従来のアプローチにおいては、なんらかの手段を用いてある定まった事象例のグループに共通する普遍的なルールを見出し、このルールを当面の問題に適用するという段階を経る。　この一般的な解析フロー中、常に律速段階となるのが図中太い四角で囲まれた「問題解決の為のルール発見」部分である。　このルールの発見が最も困難な過程であり、このルールが見出されればその問題は解決したことになる。

図1に問題解決についての一般的（従来手法）な解析フローが示されている。

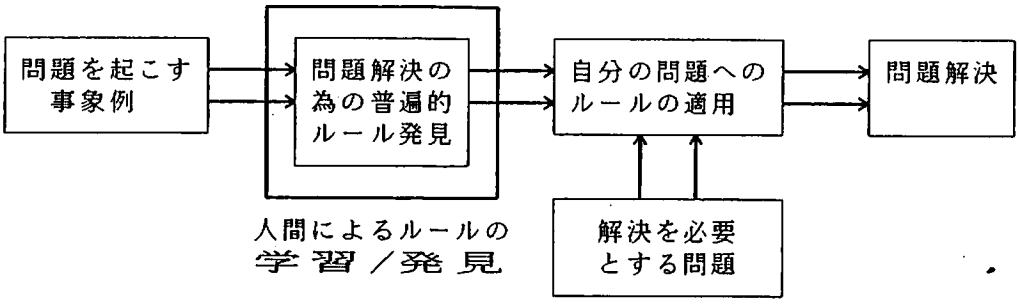


図１．　従来型アプローチによる、問題解決の為の一般的解析フロー図

□　ＱＵＩＺ１との関係

このアプローチは前記設問中、設問１がこのタイプの典型となる。　即ち、彗星の周回に関する数式が得られれば（この場合、図１中太線で囲まれた「問題解決の為のルール発見」に該当する）、その式にデータをあてはめる事により簡単に事象の完全予測が可能である。　天体に関する多くのの法則は厳密な数式で表現可能である。　この数式は、少しの誤差（例えば統計の様なアプローチでは局部的に見るならば誤差が存在する）も許さない厳格なものである。

この例に対し、設問２では単純な化合物であるならば数式を用いてスペクトル予測を行う事は可能である。　しかし、化合物の構造が複雑になると数式のみで総てのピークを予測する事は困難となる。　また複雑なスペクトルチャートを扱う時は、経験者（エキスパート）の第六感が重要な働きをする事になる。

□　従来型アプローチ（論理思考型）の特徴

このアプローチは、問題解決の為の法則（方程式）が発見される限りにおいては理想的なアプローチである。　しかしこれは問題解決の為の法則発見に多大な努力を必要とし、地道な研究活動が要求される。

世の中の多くの事象は多くの研究者の努力にもかかわらず、問題解決の為の法則を簡単に見出すことが困難である事の方が多い。　このような問題に対しては、従来からの伝統的なアプローチは無力である。

例えば、最も簡単な例として人間が自然に行う他人の識別という点について考える。この時、人は一度会った人を識別するのに、識別のルールを明確にしてから行ってはいない。

むしろ、明確なルールはなくとも、全体的なパターン情報を用いてモヤモヤとした中から総合的に判断を行っている。　このように明確なルールを設定する事は困難であるが、全体的な情報から法則なく解決する事が可能という問題は多く、このような問題こそ後に述べるパターン認識が最も得意とするものである。

４．パターン認識によるアプローチの考察

□　問題解決に対するパターン認識（無ルール型）によるアプローチ

従来手法によるアプローチでは解決困難だった様々な問題（データが複雑で統一性／法則性を見出す事が困難な問題、曖昧性及び感覚性が主体となる問題）に対するアプローチとしてパターン認識が考えられる。　このパターン認識による問題解決の流れは図１と基本的に同じである（図２）。　この図２と図１との大きな違いは、問題解決の為のルール発見部分を人間がやるのか（従来手法）、計算機がやる（パターン認識による問題解決）かの違いである。　パターン認識によるアプローチでは、この最も手のかかる問題解決の為のルール発見部分は計算機が担当する。

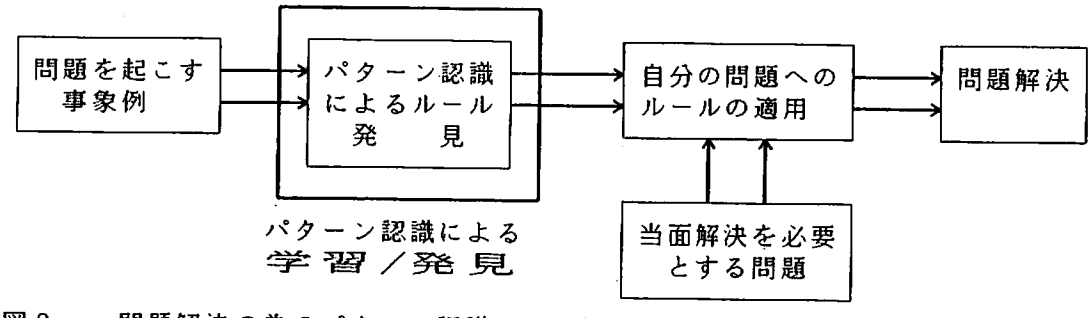


図２．　問題解決の為のパターン認識による解析フロー図

□　ＱＵＩＺ１との関係

設問２のスペクトル解析問題では、数式抱けを用いてすべてのスペクトルを解読し、その化合物構造式まで予測する事は殆ど不可能である。　この不足分を補うものとして、定性的な解析ルールや経験者のノウハウ等がある。　また、パターン認識による解析も可能である。

設問３では構造と活性を直接結びつける相関式を作成する事は殆ど不可能である。　ＨＡＮＳＣＨ／ＦＵＪＩＴＡ法等重回帰式を用いる手法があるが、これは法則というよりは経験式というものである。　その適用に当たっては化合物構造式等での制限が多いが、現在最も厳格に薬理活性を予測出来る手法である事は間違い無い。　薬理活性の予測は余りにも考慮すべき要因が多すぎて、そのすべての要因を考慮した解析を行う事は不可能である。　このような分野では法則化すべき部分をブラックボックスとしたパターン認識による解析も一つの大きなアプローチとして試みられている。

設問４ではコーヒーの味覚という人間の感覚に関する問題を扱う事が必要となる。　一般にこのような感覚に関する事象は数式に展開するどころか、その感覚を値として表現する事すら困難である。　また、データの信頼性という点からも問題が多い。　仮に甘味を強／中／弱と３段階にわけても、そのデータを入力する人により同じ化合物に対する値そのものが変わりうる。　このような問題に対してはパターン認識／多変量解析のアプローチが有効となる。

設問５では、あまりにも不確定要因が多く、またそれらの要因を明確な形で取り出す事すらも困難であり、このような問題に対し現在可能となるアプローチは存在しない。　人間の感覚的な作業にたよるしかない。

□　パターン認識によるアプローチの特徴

パターン認識によるアプローチでは、計算機内部で発見された問題解決の為の法則は計算機が使う為のものであり、人間が直接理解出来る形にはなっていない。　例えそのような機能が存在したとしても、多次元情報をより見易い形で提供する等の機能にしかすぎない。　従って、計算機が内部に構築した問題解決法則をそのまま用いて、より一般的な数式やルールを導くことは困難である。　しかし、パターン認識から得られる様々な情報をまとめることで問題解決の助となる情報を見出し、新たな展開に導くことも可能である。

このようにパターン認識を利用する事で従来のアプローチでは解決不可能だった当面の問題解決を実現し、同時に、何故問題解決出来るのかという問題に関する様々な情報を取り出すことが可能である。　表２にはパターン認識による問題解決と従来手法による問題解決における基本的な差異をまとめた。

表２．パターン認識と従来手法による問題解決の違い

パターン認識	ブラックボックス　従来手法によるアプローチ
要因が複雑で数式化困難	←　要因　→　要因が明確で数式化可能
問題が本質的に多次元	←　次元　→　少数次元の問題
データ量は豊富	←　データ　→　データ量は少なく

５．パターン認識の実体は？　その活躍はどこで行われているか？

□　パターン認識を構成する２種類のアプローチ

現在、パターン認識を行うのに利用される手法は大きく２種類存在する。　ひとつは、最近ニューロ等の言葉で一般的なものとなりつつあるニューラルネットワークである。残る一つはいわゆる多変量解析と呼ばれる解析手法である。

①　ニューラルネットワーク：　人間の神経細胞の動きをシミュレートする事から出発した学問であり、人間が得意とする画像／音声認識を計算機で行わせようとする。

②　多変量解析：　多次元データを扱う事を目的として展開された数学上の一分野。

このパターン認識に対する２種類のアプローチのうち、画像／音声等の認識問題は伝統的にニューラルネットワークによるアプローチがとられており、一般的な多次元データを多角的に扱う時は多変量解析が適用される事が多い。　現在では、両方のアプローチを混在して利用する事も多く、一般的に化学の分野におけるパターン認識という時は両方のアプローチを代表した言葉として用いられる事が多い。

化学分野以外でパターン認識が利用されるのは画像認識で代表される「認識」作業分野が多い。　この認識作業はその対象とするものの種類によりさらに細かく分類される。

例えば、認識対象が文字の時は文字認識、音声の時は音声認識、その他波形認識、図形認識、３次元形状認識等がある。

□　化学以外の分野における様々な利用例

パターン認識は様々な分野で応用が進んでいる。　ここでは化学以外の伝統的（パターン認識にとり）な分野における応用について概観する。

以下には画像／文字認識と音声認識について実際にパターン認識が利用されている事例をいくつかリストアップする。　このパターン認識の技術が日常生活に溶け込んできている事がおわかりかと思う。

①　画像／文字認識技術としてのパターン認識

・郵便番号の判定
・指紋の判定
・農産物製品のグレード（等級）判定

②　音声認識技術としてのパターン認識

・言葉の判定
・声紋による喉頭癌の判定
・音波による潜水艦の国籍判定

③　その他の利用事例

・顔料混合時のカラー予測
・パターン（スペクトル／センテンス／他）の高速検索

最近ではこのような単純な認識（RECOGNITION)問題“パターン認識”に止まらず、認識で獲得した情報を用いて理解(UNDERSTAND)するという、“パターン理解”を行なう方向に研究が進みつつある。　例えば、ある実験室の内部を見て（パターン認識）、その実験室では何の実験が行われているか（パターン理解）ということまで認識するもので、より高度な判断が要求されるものである。　このようなレベルの研究では単なるパターン認識だけでは不可能で、人工知能やその他の技術を取り込んだ形での展開が必要となる。

6．パターン認識の世界へ

パターン認識を行う為のアプローチとしては第５節でも述べたように大きく２種類存在する。　この２種類のアプローチは互いにその起源も歴史も異なるが、単に多次元データを扱うという点では同じ土俵上にある。

ここでは当初よりパターン認識を行う為の手法として開発され、最近注目を浴びつつあるニューラルネットワークを中心として述べる。　パターン認識を支える第２の手法である多変量解析は次節で詳しく述べる。

□　パターン認識（ニューラルネットワーク）の歴史

（パターン認識の歴史は計算機の実現よりも古い）

表１に１９４３年のマッカロ(MuCulloch)とピッツ(Pitts)らによる論理ニューロンモデルの導入とウイナー（N.Wiener）によるサイバネティクス（CYBERNETICS)の提唱から、現在にいたるまでのニューラルネットワークに関する歴史的経過を年代順に示す。　この表によると、計算機が実用化される前にパターン認識に関する学問が始まっている。　この事実は、パターン認識を実際に実行するにはどうしても計算機の助けが必要であるという現実から想像すると、意外に感じられるかもしれない。

表１．ニューロンモデルからパターン認識及びニューラルネットワークへの歴史的展開

１９４３．	論理ニューロンモデル及びサイバネティクスの提唱
（１９４６．	世界初の実用計算機ＥＮＩＡＣが完成）
１９４９．	Ｈｅｂｂの学習則発表
１９５８．	Ｒｏｓｅｎｂｌａｔｔによるパーセプトロンの提唱
１９６８．	中野らによるアソシアトロンの提唱
１９６９．	Ｍｉｎｓｋｙ　＆　Ｐａｐｅｒｔらによるパーセプトロン限界説発表
……	
……	パターン認識（特にパーセプトロン）暗黒の時代
……	
１９８３．	Ｈｉｎｔｏｎ　＆　Ｓｅｊｎｏｗｓｋｙによるボルツマンマシン発表、ニューラルネットワークの世界に突入
１９８５．	Ｈｉｎｔｏｎ　＆　Ｒｕｍｅｌｈａｒｔによる一般化デルタルールの発表
１９８９．	ニューラルネットワークの化学上の問題（反応成績体の比率予測）に対する適用事例がＣ＆ＥＮｅｗｓに掲載される

ニューラルネットワークの研究は、Ｈｅｂｂの学習則の提唱及びその学習則を用いたパーセプトロンの発表により実用段階に達し、パターン認識の全盛時代を迎える。　人間が行う最も高尚な“学習”過程をシミュレートする事で、従来の技術では解決不可能な“認識”に関する問題が解決可能となるという事実は、当時華々しい成果を挙げつつあった計算機とリンクされる事で過剰とも思われる期待がかけられた。

しかし、１９６９年にＭｉｎｓｋｙ＆Ｐａｐｅｒｔらによるパーセプトロン限界説（パーセプトロンでは２次元上の最も簡単な分類問題すら解決不可能である、という事実の証明）が発表されてから“パターン認識暗黒の時代”がはじまる。　この発表により、パターン認識に寄せられていた過剰期待の失望感から、多数の研究者がパターン認識から去っていった。　しかし、新たなニューラルネットワークが発表されるまでの約１０年間、理論的には限界があったとしても実用的レベルでは様々な分野で応用が進み、地味ではあるが数多くの実績が築きあげられてきた。

１９８３年のボルツマンマシンの発表により旧来手法のニューラルネットワークが抱える限界を越す事が可能となり、第２世代とも呼べるパターン認識の時代、即ち新たなニューラルネットワークの時代に突入する。

その後、１９８５年のＨｉｎｔｏｎ＆Ｒｕｍｅｌｈａｒｔによる一般化デルタルールの発表により、最もポピュラーなネットワーク構造を持つ階層型ニューラルネットワークが実用化され本格的な普及が始まり、第２のブームともいえる状態になっている。　本書では都合上１９８３年のボルツマンマシンに至るまでに開発されたニューラルネットワークを第一世代ニューラルネットワーク、ボルツマンマシン以降のニューラルネットワークを第二世代ニューラルネットワークと呼ぶ。

パターンとは？　パターン空間とは？　その数値データ化

「パターン認識」という言葉の「パターン」とは一体何を意味するのであろうか？　その外にも「パターン空間」という言葉も頻繁に使われる。　これらの言葉や概念は、今後実際にパターン認識を実行するにあたり、常に付きまとってくる重要な概念であるので簡単に考察してみる。

一般的に言われる“パターン認識”とは、全体的な形状や状態を認識する問題に関する学問として理解されていると思う。　この時、“パターン認識”という言葉の中のパターンとは形状／状態を意味するものといえる。　通常、この形状（パターン情報）は多数の数値データへと変換する事が可能である。　従って、パターンとは数学的には多数の数値データ（多次元）から表現される、最小限度の単位を意味するものと考えられる。

つまり、解析対象となる母集団を構成する個々のサンプルがパターンであり、このサンプル（パターン）が存在する多次元空間をパターン空間と称する。　パターン認識とは、このパターン空間（パターンがｎ個の数値データで表現されている時はｎ次元空間）に浮かんでいるパターン同志の存在関係に関する情報を用いて、分類その他の解析を行うアプローチを意味するものといえる。　また、このパターンそのものはｎ個の数値データで表現されている。　この多次元データを扱うという観点で見ると、多変量解析手法の適用もパターン認識に適用可能である事がわかる。

パターン認識の“パターン”はその適用分野により様々に変化する。　例えば、画像認識分野では画像そのものであり、写真、文字、その他がパターンとなり、音声認識分野では音声スペクトルがパターンとなる。　その他、構造－活性相関分野では化合物の構造式がパターンとなり、スペクトル解析では様々なスペクトルチャートがパターンとなる。

7．2／3－次元パターン空間からＮ－次元パターン空間（多次元空間）への展開

□　パターン認識を行う為の第１歩（パターン情報の数値データ化）

パターン認識を適用する為には対象となる事象がなんらかの形で数値データ化されている事が必要である。　即ち、パターンと称される解析対象が数学及び計算機による取扱が可能となるように数値データへと変換されている事が必要である。

パターン認識による解析に利用される数値データの内容／種類はそのパターン認識が適用される分野により異なるが、パターン認識に用いられるデータの基本は一個のパターンが複数の数値データで表現される多次元データである事である。

□　個々のパターンから、Ｎ個の数値データ（Ｎ次元）への導入

Ｎ次元空間中にＭ個のパターンが存在する時、個々のパターンＰはそれぞれＮ個の数値データX_1，X_2，……，X_{N-1}，X_Nを用いて表現される。　従って、ｉ番目のパターンP_iは以下のような形で示され、これはパターンベクトルと呼ぶ。

$$P_i = (X_{i1}, X_{i2}, \cdots\cdots, X_{i,N-1}, X_{iN})$$

われわれが生活している３次元空間上のパターンＰは３次元データである（X_1，X_2，X_3）で表現される。この時、（X_1，X_2，X_3）がそれぞれＸ、Ｙ、Ｚの座標軸上の数値であるならば、ｉ番目のパターンP_iの座標は３次元空間中で

$$P_i = (X_i, Y_i, Z_i)$$

で表される。　この３次元空間とＮ－次元空間との関係が図３に示されている。

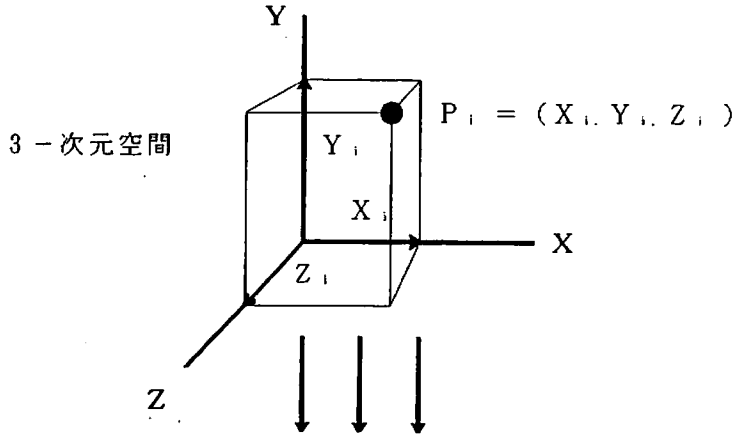


図３．　３－次元空間からＮ－次元空間への拡張

Ｎ－次元空間中のパターンの分布状態を、われわれの感知できる２、３－次元空間上で完全に図示する事は不可能である（近似的には可能、第　章参考）。　このように人間には３次元より大きな次元をイメージするのは困難であるが、思考上、或いは計算上では３次元がＮ次元になっても、パターンを表現する為の情報が３個からＮ個に増えたというだけで、特に問題になる事はない。

パターン認識では、このような多次元空間（パターン空間）上に存在するパターンの位置関係に関する情報を様々な観点から解析を行う事が総ての基本となっている。

＊　統計等の分野では、すべての基本は確立密度分布に起因し、この確立密度関数の様々な取扱が、そのまま統計手法のバリエーションにつながっている。

８．画像認識における多次元データ事例（文字認識を例として）

□　始めに数値データありき（文字情報のサンプル（学習用事例）収集）

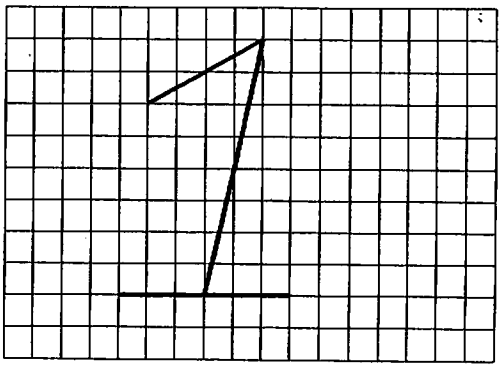
パターン認識解析の基本となる多次元データは、実用上ではどのような形で利用されているかについて分かり易い例を取って説明する。　例は、現在パターン認識分野で古くから行われてきた文字（数字）認識（実用例としては郵便番号の自動読み取りが有名である）を例にとって説明する。

ここでは文字情報の数値データへの変換手法として、数値が描かれている領域をメッシュで区切り、数値情報を０／１の多次元数値データへと変換する手法について言及する。文字パターンの数値データへの変換方法はこの手法以外にも様々なアプローチが取られている。　個々の目的に合わせ、自分の解析目的に最も相応しいと考えられる変換方法を工夫し、採用する事がパターン認識による解析を成功させる近道である。

□　文字情報の数値データ（多次元）への変換

①　文字情報の細分化

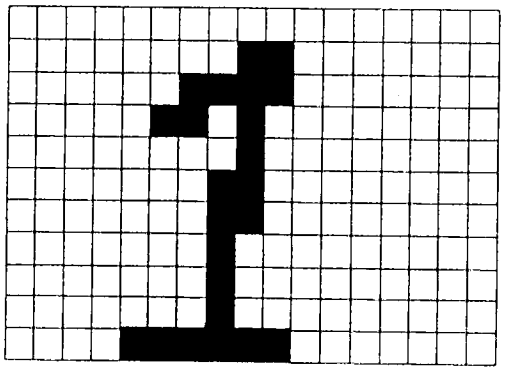
数字が書かれる範囲を１７×１１のブロック（総数１８７個）に分割する。
このブロック上に数字の１を描き、一つの文字情報とする。



1 2 3 4 5 6 7 …… 151617

②　文字情報の数値データ化

ブロック上に描かれた１という文字が、ブロック上を走っている総てのブロックを塗りつぶすと右図のようになる。
この時、塗りつぶされたブロックを１、何も塗られていないブロックを０と定義する。
この結果、文字情報１は０と１から構成される１７×１１＝１８７次元のデータへと変換された事となる。



1 2 3 4 5 6 7 …… 151617

文字データ１＝
数値データ（００００…… １０１１０ …… ００００）
次元　（１２３４ …… …… …… 218, 219, 220, 221）