6．「超ボリューム概念」のパターン認識への導入

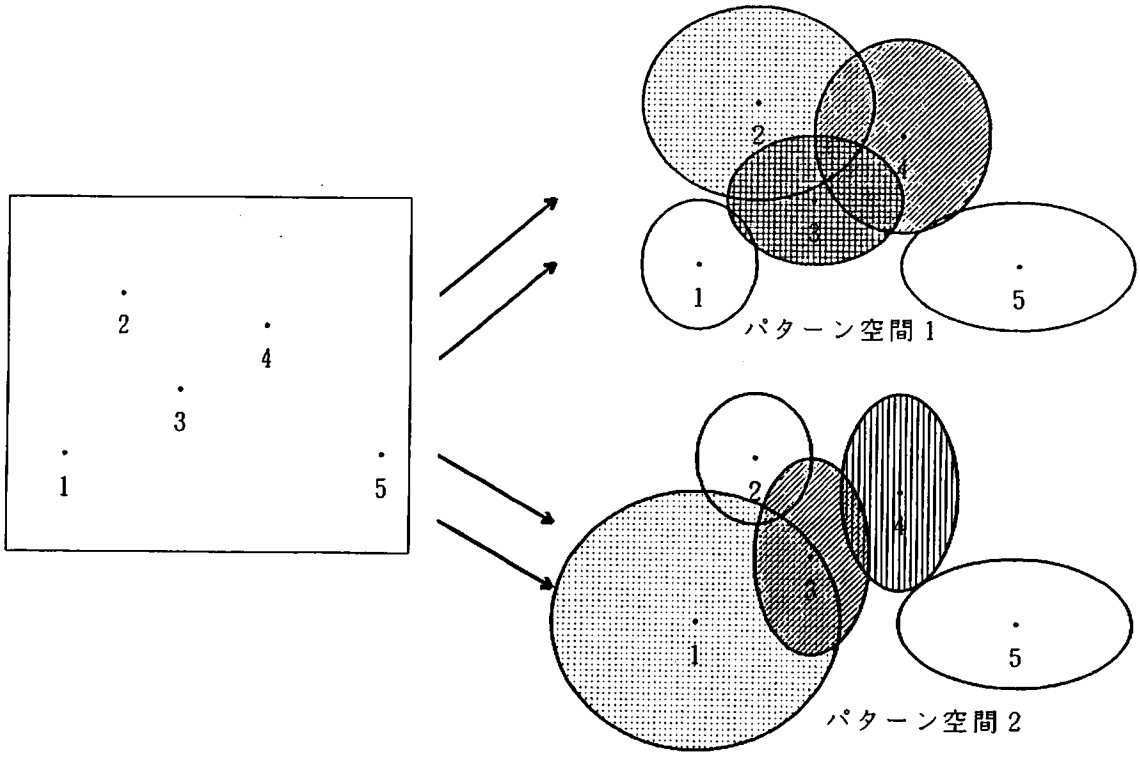
本節ではパターン認識／多変量解析で扱う個々のパターンに対する基本概念を“点”から“ボリューム”へと拡張する事によってもたらされる新たな展開に関する議論を行う。この展開は著者が行っているものであるが、パターン認識や多変量解析の新たな解析手法の導入、従来手法の改良／強化等を実現する事を可能とするものとなる。

6．1．「超ボリューム概念」の基本的考え

パターン認識に限らず統計／多変量解析において、解析対象とされるパターンは総てボリュームの無い“点”として扱われてきた。　この理由はパターンを点で代表する事で数学的な扱いが簡単になる事が大きな理由であった。　しかし、対象パターンを点で代表させるよりもある一定の“ボリューム”を持つ空間で代表させる方が様々な解析を行う上でより自然と考えられる事が多い。　概念的には空間で代表される超ボリューム概念がより一般的で、各パターンを点で表す従来の概念は超ボリューム概念の特殊例であると考えることが出来る。

> 「超ボリューム概念」では、“点”として扱われてきたＮ次元空間中の対象パターンを、ある点を中心とした特定の領域（ボリューム）として扱う。

図１には個々のパターンが点として取り扱かわれるパターン空間から、超ボリュームとして扱われるパターン空間へと拡大された様子を示す。　個々のパターンはある重心を持つ超ボリュームとして表現されており、この重心点だけをみる時従来手法と超ボリューム概念に従った空間は全く同じものである。　しかし、超ボリュームの形は様々であり、且つその内容にも差がある。　この結果、超ボリュームを用いた表現では例え同じ座標データを用いたパターン空間を用いたとしても、超ボリュームの形態により様々な超ボリューム空間を実現する事が可能となる。　この事は、従来手法と比較して超ボリューム概念の方が様々な情報を多様な形で表現出来る事を表している。



従来手法による
Ｎ次元パターン空間

$P = (X1, X2, \cdots\cdots, Xn)$

超ボリューム概念による
Ｎ次元パターン空間

$P = ((X1, X2, \cdots\cdots, Xn), (Y1, Y2, \cdots\cdots, Yk), (Z1, Z2, \cdots\cdots, Z1))$

図１．従来手法と超ボリューム概念によるパターン空間

このようにパターンを点から超ボリュームへと拡張する事で、従来手法には無い特徴が

生じてくる。 つまり、従来手法によるパターン空間ではパターンベクトルを構成する要素（X1, X2, ……, Xn)は総てパターン空間上の座標を表す座標データとみなされ、これらの要素はデータ的に総て等価なものとして扱われてきた。 一方、超ボリューム概念に従ったパターン空間では、個々のパターンはパターンの位置を示す座標ベクトル（X1, X2, ……, Xn)、超ボリュームの形状（図1中各楕円系の形）を表す形状ベクトル（Y1, Y2, ……, Yk)、及び超ボリュームの特性（図1中各楕円中の違い）を表す特性ベクトル（Z1, Z2, ……, Zl)の3種類に分類される。 この3種類のデータを目的に応じて特定の意味／情報を定義する事が可能である。 またこれら3種類のデータを目的に応じて使い分ける事で様々な解析手法の展開へと導く事が可能である。 これが超ボリューム概念の最大の特徴である。

「超ボリューム概念」でのパターン空間および超ボリュームの持つ意味

先ず個々のパターンを代表する超ボリュームが存在する空間（以後超ボリューム空間と呼ぶ）は、パターン間の相対的位置関係を意味し、この空間の持つ意味は従来手法でのパターン空間と何ら変わる所はない。 従って、パターン空間における超ボリュームの重心位置だけを対象として議論すれば、従来手法の議論と全く同じものになる。

一方、超ボリュームに関してはこの ①超ボリューム相互の関係（例えば超ボリューム同士の重なり等）と ②超ボリュームそのものの特性に関する問題とに分けて考える事が必要である。 この①は、その超ボリュームを形成するパターンが及ぼす（又は及ぼされる）影響の範囲や状態、即ち個々のパターン独自（ローカル）の環境に関する情報を盛り込んだ解析と考えられる。 ②の超ボリュームそのものに関する情報、即ち「モデル」（サイズ、密度、カラー、他）は個々のパターン固有の情報を盛り込んだ解析として考えられる。 勿論このモデルの持つ意味は、超ボリューム概念が適用される分野毎に異なってくる。 以上の関係を簡単にまとめると、以下のようになる。

(1) パターン空間 ―――――――――― パターン間の相対的位置関係

(2) 超ボリューム

① 超ボリュームの重なり等で代表される特性空間同士の相互関係

―――――――――― パターン固有の環境情報

② 超ボリュームそのものの特性（カラー、密度、他） ―― パターン固有の情報

例えばこの超ボリュームの形が球状（多次元空間では超球となる）の時、超球中心の座標は $x_1, x_2, \cdots\cdots, x_n$ の座標成分で示され、形状ベクトルとして超球の半径がrで、更に超球の特性に関する特性ベクトルの様々な情報が、例えば超球の重さwやカラーc等である時、以上3種類のベクトル成分をあわせた新ベクトルとして表される。

$$P = [x_1, x_2, \cdots\cdots, x_n, r, w, c] \quad (3)$$

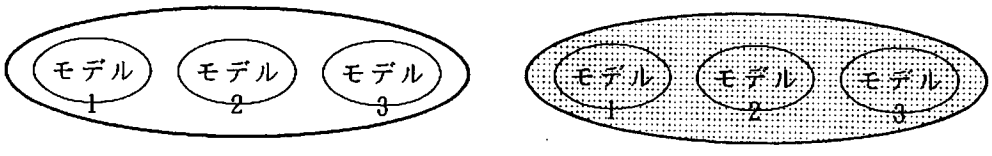
超ボリューム概念導入による「手法」と「モデル」の概念について

この「超ボリューム概念」の導入を行うことで、新たなパターン認識や統計手法を導き出す事が出来る。 この時、新手法への展開はこの“超ボリューム”の取扱の差異に従って2つのタイプに分類される。 この2つは、前節で扱った(1)と(2)とに該当するもので、

①“超ボリューム”の多次元パターン空間中における分布状態の解析 →→→「手法」

②“超ボリューム”そのものに関する様々な特性（形状、サイズ、重さ、密度、カラー、他）を考慮しつつ展開される解析 →→→「モデル」

である。



手法1　　手法2

図2．「手法」と「モデル」との関係

つまり、「手法」と「モデル」という2つのタイプである。 この2つの関係は、①に従った展開は「・・手法」とし、解析手法としては大きな分類に該当する。 また②の展開は「・・モデル」とし、「手法」の中での展開となり小さな変化／バリエーションを「手法」にもたらすものである。 以下にこの「手法」と「モデル」について述べる。

「手法」について

「手法」とはその定義上、パターン空間上に散在する超ボリューム相互の分布状態に関する解析を行うものである。 この解析アプローチは超ボリュームの重心を扱う事で簡略化すれば、大筋において従来の“点”を基本とするアプローチと一致する。 例えば「分

類」、「クラスタリング」、「重回帰」、その他のアプローチで代表されるようなものである。　但し、解析過程で“点”を対象として展開するか、“ボリューム”を対象として展開するかという点が従来手法との大きな差異である。　この展開は超ボリューム同士の分布状態（相対的位置関係というよりは超ボリューム同士の重なり状態）のデータを用いて行われる。　従ってこの「手法」に関する解析では、超ボリュームの形状や重なりが重要な要素となって展開される。

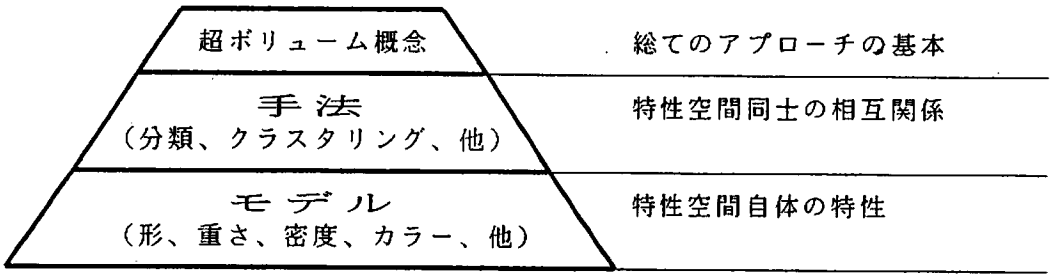


図２．「超ボリューム概念」、「手法」、「モデル」の階層関係

「モデル」について

「モデル」とは超ボリュームそのものの特性に関する情報を扱うものである。　この特性とは、例えば超ボリュームの重さ、カラー、密度、硬さ等で示される情報であり、個々のパターンに特有なものである。　この為パターン相互の関係は薄く、独立性の高い情報となっている。　この「モデル」の特性を考慮する事で、先の「手法」に関し様々なバリエーションを持たせた展開が可能となる。

３．　「超ボリューム概念」のパターン認識における分類問題への適用

個々のパターンをボリュームのある特性空間として扱う超ボリューム概念は、パターン認識や統計等の分野の基本概念として様々な形で利用出来る。　この概念を基本とする事で、従来手法にない特徴を持つ新たな手法を導く事が容易となる。　ここでは超ボリューム概念適用の一事例として、新たに開発された分類手法「超球法」について述べる。

３．１　「超ボリューム概念」に基づいた新分類手法、「超球法」の開発

この分類「手法」は、超ボリューム概念に従った特性空間の形状を最も単純な「超球」とし、パターンのクラス決定はこの超球同士の重なりを検討して行うものである。

即ち、「超球法」でのクラス決定はクラス未知パターンＸの中心が落ち込んでいる（重なっている）超球を形成する全てのパターンのうち、最大多数を占めるパターンと同じクラスに帰属して行われる。

尚、これ以降の式では、ｎはパターン数、ｍはクラスの数、Ｄｉｓｔ（Ａ、Ｂ）はパターンＡとＢとの距離を示すものとする。

まず前提条件として、クラス既知パターンα_iがｎ個（ｉ＝１～ｎ）あり、且つクラスがｍクラス存在し、分類に用いられる超球の半径ｒは全てのパターンについて同じであるとする。　この時、クラス未知パターンＸの中心とクラス既知パターンα_iの中心との距離をＤｉｓｔ（α_i，ｘ）とし、この距離が超球の半径ｒよりも小さいパターンα_iのみを取り出し、このパターンを分類の為の基準データとして用いる。　つまり、取り出されたクラス既知パターンの総数をｋとすると、ｋは

$$k = P_1 + P_2 + \cdots + P_j + \cdots + P_m \quad \text{——————（４）}$$

で表される。　この式で、P_jは取り出されたクラス既知パターンのうち、クラスが同じｊであるパターンの数を示す。

３．１．１　クラス未知パターンＸの中心が他の超球と重なる時

クラス未知パターンＸの最終的なクラス決定は、このP_j（ｊ＝１～ｍ）中最大の値を示すものと同じクラスｊに帰属される。

$$Class(X) = Class_P\{\max_{j=1}^{m}(P_j)\} \quad \text{——————（５）}$$

ここで、Ｃｌａｓｓ（Ｘ）はパターンＸの所属するクラスを、$\max_{j=1}^{m}(P_j)$はP_jがｊ＝１～ｍの中で最大値を持つP_jをとる事を示す。　また、$Class_P$は取り出されたP_jのクラスの値ｊを取り出す事を意味する。

この時、最大値を示すP_jが複数あるならば、パターンＸの分類は「保留」とする。

３．１．２　クラス未知パターンＸの中心が他の超球と重ならない時

パターンＸの中心が他の超球　α_i（ｉ＝１～ｎ）と重ならない時、このパターンＸのクラス決定は「不可能」とする。　即ち、ｒ＜Ｄｉｓｔ（α_i，ｘ）　の時、

Class（X）　＝　分類不可能　　――――――――――――――――――――――（6）

従って超球法による分類では、3．1．1の分類「保留」と3．1．2の分類「不可能」とが存在する。　この2つの結果は分類率に反映されないが、分類問題における「保留」と「不可能」のデータの持つ意味は異なる。　また、個々のパターンについて「保留」か「不可能」かの認識が行われる事で、従来手法よりも詳細な解析が可能となる。

尚、ここで展開された超球法は特性空間モデルとして最も単純な「ソリッドモデル（超球内部が均密なもの）」を用いて展開されている。　従って、分類時に必要となる超球同士の重なりチェックは、単に超球の重なりの有無だけで判定している。

3．2　従来手法、K－NN法について

従来手法のうち、この超球法にアルゴリズム的に最も類似した分類手法として「K－NN（最近隣）法」がある。

このK－NN法ではクラス未知パターンXとn個のクラス既知パターン　α_i（i＝1～n）との距離を求め、距離が小さいものから順にクラス既知パターンをK（1、3、5、・・（奇数））の数だけ取り出す。　クラス分類は、Kが1の時には取り出されたクラス既知パターンと同じクラスに帰属する。　Kが3以上の時は、取り出されたK個のパターンが所属するクラスのうち最大多数を占めるクラスと同じクラスに未知パターンのクラスが帰属される（多数決による決定：多数決を行う為、Kは常に奇数となる）。

以下では簡単の為K＝1に限定して議論を展開するが、ここでの展開はクラス決定が多数決の原理に基づく以外はK＝3以上の時も全く同じである。

先ず、クラス未知パターンXのクラスは（7）式により決定される。

$$\mathrm{Class}(X)=\mathrm{Class}\left[M_{i=1}^{n}\{\mathrm{Dist}(\alpha_i, x)\}\right] \qquad \text{――――――――――――――――――――――（7）}$$

但しClass（X）はパターンXの所属するクラスを、nはクラス既知パターンの総数を、Dist（α_i，x）はi番目のパターンα_iとパターンXとの距離を、又

$M_{i=1}^{n}\{\mathrm{Dist}(\alpha_i, x)\}$はn個のパターン$\alpha_i$とXとの距離のうち、最短距離を持つパターン$\alpha_i$を取り出す事を示す。

3．3　「超球法」と従来手法「K－NN法」との比較

超球法とK－NN法との主な差異はクラス決定の時利用されるパターン数の差と、パターン間の距離制限の有無との2点である。　つまり、K－NN法ではクラス決定に用いる隣接パターン数K（Kは奇数1、3、5・・・）は予め設定されており、パターン間の距離制限は存在しない。　一方第3章第1節でのべたように、超球法ではクラス決定に利用されるパターン数はパターン毎に異なり（超球の重なり状態は個々のパターン毎に変化する）、パターン間の距離は超球の半径が事実上の制限距離となっている。　即ち、超球法は個々のパターンの環境に従い、クラス分類に必要な隣接パターン数を超球の重なり情報に従ってダイナミックに選択しつつ分類する手法である。　従って、最初から解析に必要となるパターン数が固定されているK－NN法と異なり、パターン毎の分布状態や環境に従った詳細な分類が可能である。

手法的にはK－NN法に多少の修正を加える事で、超球法と同じ結果を得る事が可能となる。　しかし、単なる分類手続きや分類結果だけで基本概念の異なる2手法を同列のものとして扱うことは出来ない。　例えば、分類結果は同じでも超球法での特性空間の持つ意味はK－NN法には代用となるものが無い。　更に、超球法では超球そのものの特性を考慮する事で新たな手法への展開が容易となるが、K－NN法では更なる展開への余裕は小さい。　ここでは分類手続きや分類結果の細かな評価は行わず、手法としての本質に関する問題について詳しく検討する。　具体的には、パターン空間中での個々のパターンの分布状態に関する考察から出発し、K－NN法における手法上の矛盾を指摘し、次にこの矛盾が超球法及び次章で述べる「傾斜超球法」とで解決出来る事を示す。

3．3．1　K－NN法における分類上の矛盾

図3はN次元パターン空間中におけるクラス未知パターンX1、X2及びクラス既知パターンA、Bとの相対的位置関係を示す。　この図を用いて、K－NN法が抱えているパターン分類上での2つの矛盾を以下に示す。

（1）　矛盾例　1（粗悪なデータに基づいた強引な分類）

『クラス決定時に利用されるデータ（パターン間の距離）に大きな差異が有っても、分類上等価なデータとして扱われる』

第3図中クラス未知パターンX1とクラス既知パターンA（クラスA）及びB（クラスB）との間の距離は

Dist(B，X1)　＞　Dist(A，X1)

であり、パターンX1はK－NN法の定義に基づき近い距離にあるパターンAの属するクラスAに分類される。　ここで、Dist(B，X1)はパターンBとパターンX1との距離を示し、他も同様である。　また、クラス未知パターンX2に関しては

Dist(B，X2)＞　Dist(A，X2)

であり、パターンX1と同様、近接パターンAの属するクラスAに分類される。　この時

、Ｘ１とＸ２のクラスを決定した情報はＤist(Ａ，Ｘ１）とＤist(Ａ，Ｘ２）とであるが、この２つの距離の大小関係はＤist(Ａ，Ｘ１）≫Ｄist(Ａ，Ｘ２）であり、２つの距離には大きな差異がある。　従ってこの場合、Ｘ１とＸ２とのクラス決定に用いたデータの質に際立った差異があるにもかかわらず両パターンは同じクラスに分類されている。　この事実はクラス決定がパターン間の絶対的な距離でなく、相対的な距離を基準として分類が行われている事を意味する。　従って、解析時にパターン間の絶対的な距離を重視する問題にＫ－ＮＮ法を使用すれば、分類率や分類結果の信頼性の低下がおこる。

（２）　矛盾例　２　（良質なデータの切り捨て）

『近接パターンが複数ある時、クラス分類上重要な（近い所に存在する）パターンであっても分類に利用されない時がある。』

図３中、クラス未知パターンＸ２について考えると、Ｄist(Ｂ，Ｘ２）＞Ｄist(Ａ，Ｘ２）である。　従ってパターンＸ２はクラスＡに帰属されるが、図からもわかるようにＸ２のパターンＡ及びＢに対する絶対的距離は殆ど等しい（Ｄist(Ｂ，Ｘ２）≒Ｄist(Ａ，Ｘ２））。　しかし、Ｋ－ＮＮ（Ｋ＝１）法の定義からクラス決定はＤist(Ａ，Ｘ２）の情報のみを基準として決定され、例えパターンＢとＡがＸ２の近くに存在するパターンであっても、Ｄist(Ｂ，Ｘ２）の情報、即ちパターンＢのＸ２に対する影響は全く無視されて決定される。　この事実は粗悪な情報を用いて無理に分類する矛盾例１と反対の関係にあり、これは良質の情報を無視して分類が行われる事を意味し、これも矛盾例１と同様に分類結果や信頼性の低下を導く大きな原因となるものである。

３．３．２　「超球法」によるＫ－ＮＮ法の矛盾の解決

ここでは「超ボリューム概念」を基本とした「超球法」を用いて第３章３．１節と同じデータを解析すると、前記の矛盾が解決される事を示す。

（１）　矛盾例１に対する「超球法」による解

先ずクラス既知パターンＡ及びＢを中心として超球を形成する。（図４）　この時、パターンＡとＢの超球とクラス未知パターンＸ１、Ｘ２との距離関係は以下のようになる。

Ｄist(Ａ，Ｘ１）、Ｄist(Ｂ，Ｘ１）＞　ｒ（超球の半径）

ｒ　＞　Ｄist(Ａ，Ｘ２）、Ｄist(Ｂ，Ｘ２）

図４でもわかるようにクラス未知パターンＸ１は超球の外部に、又パターンＸ２は超球Ａと超球Ｂの重なり部分に位置している。　この時個々の超球は第２章で説明したように、一般的には個々のパターンの影響の及ぶ範囲を示している。　つまり、超球内部に落ち込んだパターンは超球中心のパターンと相互に影響している、または類似しているものと考えられる。　従ってこのパターンは超球中心のパターンと同じクラスに分類されるのが自然である。　この定義からすれば、パターンＡ、Ｂいずれの超球にも重ならないパターンＸ１はパターンＡ及びＢとの関係が殆ど無い／似ていない事を意味している。　このような時、パターンＸ１のクラスを決定する事は不可能であり、実際超球法ではパターンＸ１のクラス分類は行われない（（６）式、分類「不可能」とされる）。　この事はパターン間の距離に大きな差があっても強引に分類するＫ－ＮＮ法との大きな差異の一つである。

この様に、超球法では他の超球と重ならないパターンの分類は行われないが、これはパターンＸ１がその他のパターンと異なる特徴を有するものとして明確に区別された事を意味する。　このようなパターンを積極的に分類する事も、ある特定分野における解析では重要な要因となる。

（２）　矛盾例２に対する「超球法」による解

図４中、超球Ａ及びＢに重なるパターンＸ２はＡとＢに近接しており、両パターンの影響を強く受けている（又はＡとＢに近い特性を持つ）。　従って分類時にはこのような影響の強いパターンＡとＢ両方の情報を考慮してクラス分類を行なう事が必要である。

超球法ではクラス未知パターンが重なる超球全てについての情報を用いてクラス分類を行っており（多数決によるクラス決定、但しクラスポピュレーションが同数となった時クラス分類は行わない。　分類には反映されないが、この事実は分類「保留」という情報として認識される）、第３章第２節の矛盾例２のように、単なる手法上の定義から情報が意味なく無視される事はない。　この為超球法による分類結果は、たとえ従来手法のＫ－ＮＮ法と同じ分類率であったとしても、その結果の信頼性は飛躍的に高くなっている。

４．「超球法」の“ソリッドモデル”から“傾斜（ファジイ）モデル”への展開

４．１．“ソリッドモデル”の限界

前章では個々のパターンをボリュームを持つ超球として扱う事で、Ｋ－ＮＮ法に内在する様々な分類上の矛盾を解決出来る事を示した。　この展開は「手法」のレベルに相当するものであり、この時用いた「モデル」は最も単純な“ソリッドモデル”であった。　この“ソリッドモデル”とは、クラス未知パターンＸの落ち込んだ超球内部の位置に関係なく、超球の内部にクラス未知パターンＸの中心が存在するか否かの基準だけでクラス分類を行う単純なモデルである。　しかしこのような単純な超球、即ち“ソリッドモデル”の

導入だけでは解決出来ない問題も存在する。　このような問題は、例えば超球がＮ次元パターン空間中、ある特定の形で分布している時に発生する。

このケースは図５のように超球同士（超球Ａ及びＢ）が近接（もしくは接する）し、且つそれぞれの超球に属するクラス未知パターン（Ｘ１及びＸ２）も近い距離にある時に発生する。　即ち図５中Ｘ１とＸ２はそれぞれＡ及びＢの超球内部に位置し、クラスＡとＢとに分類される。　この時、Ｘ１とＸ２のパターン空間中の距離Ｄｉｓｔ(Ｘ１，Ｘ２）は極めて小さい。　従ってパターン間の類似／相関関係を考えた時、このＸ１とＸ２は互いに類似した特性を持つものと考えられる。　しかしながら実際にはＸ１及びＸ２は、より距離の遠いパターンＡ及びＢを基準としてクラス決定され、互いに異なったクラスに帰属されている。　この結果はパターン空間中での距離関係を基準とする分類問題での基本概念、「パターン空間中近い位置にあるパターンは同じクラスに分類される」に反する事実である。　この問題を解決するには、「パターンＸはクラスＡではあるが、クラスＡでもない」という柔軟な解析が出来るように超球法を改良する事が必要となる。

４．２　“傾斜（ファジイ）モデル”の導入による「傾斜超球法」への展開

前節の矛盾は、未知パターンが超球の内部にあるか外部にあるかという２値情報だけで分類している事に原因がある。　従って、このような矛盾を解決する為には、分類基準となるデータを１、０のディジタルデータから、連続量を表現できるアナログなデータに変換する事が必要である。　そこで、超球内部の密度（クラス決定の為の基準）が全て均一であるソリッドモデル（超球法）の代わりに、この密度に傾斜がある“傾斜モデル”の導入を行った。　この傾斜モデルを用いた分類には、今後の展開の可能性を高くし、且つ適用分野の変化に柔軟に対応出来るという観点からファジイ理論（Ｚａｄｅｈ，１９６５）を導入した。（“傾斜モデル”→→→“傾斜（ファジイ）モデル”）

この“傾斜（ファジイ）モデル”を用いた「傾斜超球法」はクラス決定の為の基準となる密度の傾斜をメンバーシップ関数で表現し、クラス分類についてはファジイ代数和を求め、最も高いメンバーシップ値を示したクラスに未知パターンのクラスを帰属させるものである。　この結果、クラス決定時に‘クラスらしさ’の程度を表すパラメータ（Ｍ＝０～１）を用いる事が出来、柔軟な分類が可能となる。　尚、クラス決定に用いるメンバーシップ関数は、その適用する分野により多種多様なものが利用できる。　更に、複数のメンバーシップ関数を同時に設定する事で、より複雑な問題にも対応可能となる。（例：似ている、少し似ている、かなり似ている、似ていない、殆ど似ていない、他）

図６には超球法で用いられたソリッドモデルと傾斜超球法で用いられた傾斜（ファジイ）モデル（ＧＡＵＳＳ関数を用いた時）のクラス決定の為の関数が示されている。　ソリッドモデルでは超球中心から外周部まで総て密度Ｍ＝１である。　従って、超球内部のどの部分にクラス未知パターンの中心が落ち込んでも、その未知パターンは超球中心のパターンと類似度（‘～らしさ’）が１（即ち全く同じ特性を持つ）になる。

一方、ＧＡＵＳＳ関数を用いた傾斜（ファジイ）モデルでは超球中心の密度がＭ＝１で、超球外部に移動するにつれて密度が小さくなり外周部でＭ＝０となっている。　これはクラス未知パターンの中心が他の超球中心に落ち込んだ（完全に重なった）時のみ超球中心パターンとの類似度が１となり、超球の外部に行くに従い類似度が減少し、外周部で完全に類似性がなくなる（共通特性が全く無い）事を示している。　この関数を用いた「モデル」のほうが、超球に落ち込んだパターンが超球内部のどの部位にあっても超球中心部のパターンと全く同一の特性（Ｍ＝１）を持つとして分類する“ソリッドモデル”よりも柔軟な解析が可能であり、現実を反映していると考えられる。

４．３　「傾斜超球法」によるクラス分類

傾斜超球法はＮ次元パターン空間中に超球を設定し、クラス未知化合物が落ち込んでいる超球に関するメンバーシップ値を検討しつつクラス分けを行う手法である。　以下には傾斜超球法のクラス決定ルールを示す。

４．３．１　傾斜超球法におけるクラス決定ルール

傾斜超球法ではクラス未知パターンＸが落ち込んでいる（重なった）超球パターン全てについて、クラス別にメンバーシップ値（αカット値）を求め、この中で最も大きなメンバーシップ値を持つクラスに帰属する。

（１）　超球の内部にクラス未知パターンＸが存在する時　（図７中Ｘ１、Ｘ２）

クラス未知パターンＸが落ち込んでいる（関与する）パターンの総数をｋとすると、ｋはクラス毎のパターンの数Ｐ」との間に、式（４）の関係が成立している。

ケース１：　超球の重なり部分以外にクラス未知パターンが存在する時（図７、Ｘ１）

帰属法；クラス未知パターンが落ち込んでいる超球を形成するクラス既知パターンと同じクラスに帰属する。　この時、クラス分類に対してはメンバーシップ値の大小は考慮しない。　但し、クラス未知パターンＸの周辺環境を知る為メンバーシップ値を求める。

Ｃｌａｓｓ（　Ｘ　）＝　Ｃｌａｓｓ（　Ａ　）　――――――――――――――――（８）

但し、Xはクラス未知パターン、AはXが落ち込んでいる超球を形成するクラス既知パターンであり、Class（X）及びClass（A）はそれぞれパターンX及びAが属するクラスを示す。

ケース2： 超球の重なり部分にクラス未知パターンXが存在する時（図7、X2）

帰属法；クラス未知パターンXが関与する（落ち込んでいる）超球のメンバーシップ値を求め、クラス毎にファジイ代数和を求める。 得られたメンバーシップ値のうち最大値のものが属するクラスに帰属する。

$$Class(\ X\ ) = Class_M\ [\ M_{j=1}^{m}\ (\ P_{i=1}^{p}\ \{\ M_{ji}\ )\ ]\quad \text{——————} \quad (9)$$

但し、mはクラスの数を、p_jはクラス未知パターンXが落ち込んでいる超球k個のうちクラスが同じjである超球の数を、M_{j1}はこのk個のパターンのうちクラスjの超球で1番目のもののメンバーシップ値を、$P_{i=1}\{\ M_{ji}$はM_{j1}を1が1からp_jまでp_j個ファジイ代数和したもので$M_{j1} \| M_{j2} \| \cdots\cdots \| M_{j,p_j-1} \| M_{j,p_j}$を意味している。 また$M_{j=1}^{m}\ (\ P_{k=1}\{\ M_{jk}\ )$はクラス毎に求められたメンバーシップ値m個のうち、最大のものを取り出す事を示す。 $Class_M$（M）はメンバーシップ値Mが与えられたパターンが帰属しているクラスを意味する。

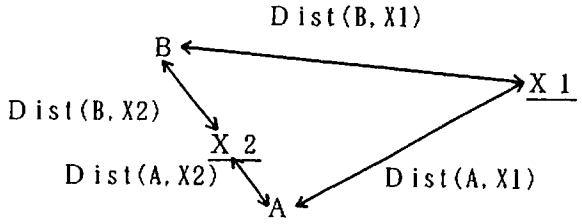
（2） 超球の外部に存在する時 （図7、X3）

帰属法； クラス未知パターンのクラスを決定する事は「不可能」とする。

$$Class(\ X\ ) = \text{分類不可能} \quad \text{——————} \quad (10)$$

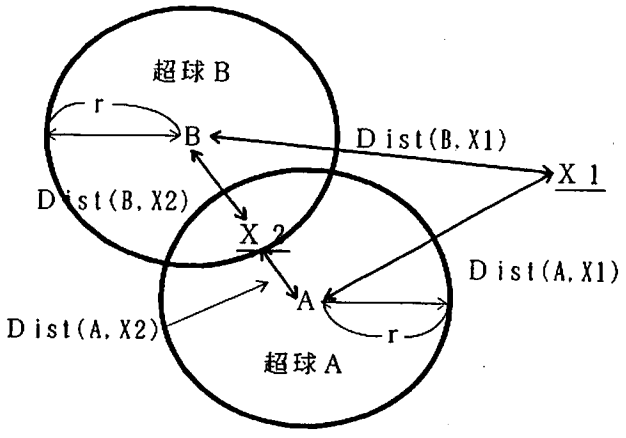
4．3．2 「超球法」と「傾斜超球法」との違い

超球法での分類は基本的にパターン数の大小を基準として行われている為、不連続データを用いている。 従って、詳細な分類は不可能であり、分類「保留」がしばしば起こる。 一方、傾斜超球法では連続データを用いている為、より細かなレベルでの分類が可能であり、超球法のように分類「保留」は生じない。 また、メンバーシップ値の大小を検討する事で、クラス未知パターンXの他の超球に落ち込んでいる状態を詳しくモニターする事が出来る。 さらに、メンバーシップ関数を変える事で様々な問題に対応可能であり、超球法よりも柔軟性の高い分類手法となっている。



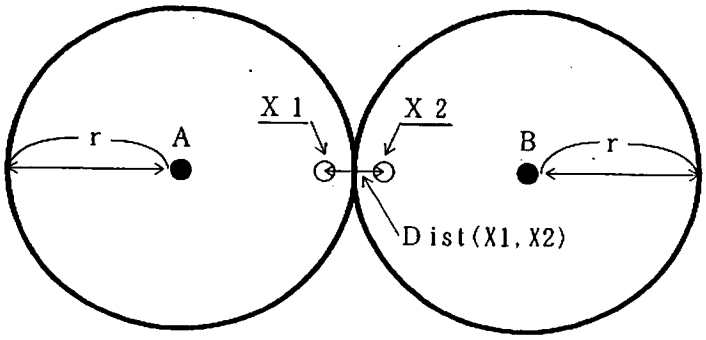
$$Dist(B,X1) > Dist(A,X1)$$
$$Dist(B,X2) > Dist(A,X2)$$

図3． K－NN（K＝1）法によるクラス未知パターンX1、X2のクラス分類



$$r < Dist(B,X1),\ Dist(A,X1)$$
$$Dist(A,X2) \fallingdotseq Dist(B,X2)$$

図4．「超球法」によるクラス未知パターンX1、X2のクラス分類



r ≫ Dist(X1, X2)

Class（X 1）＝Class（A），Class（X 2）＝Class（B）

図５．超球が互いに隣接し、クラス未知パターンＸ１，Ｘ２が近い位置にある時のクラス分類

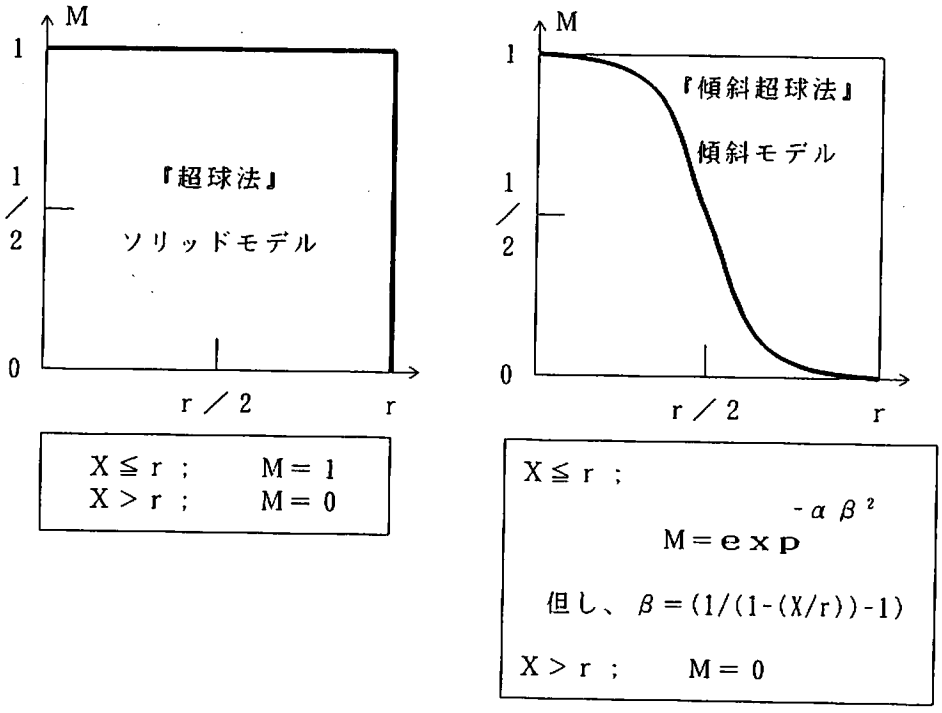


図６．ソリッドモデルと傾斜（ファジイ）モデルとの相違
傾斜（ファジイ）モデルに使用したメンバーシップ関数はＧＡＵＳＳ関数を用いた

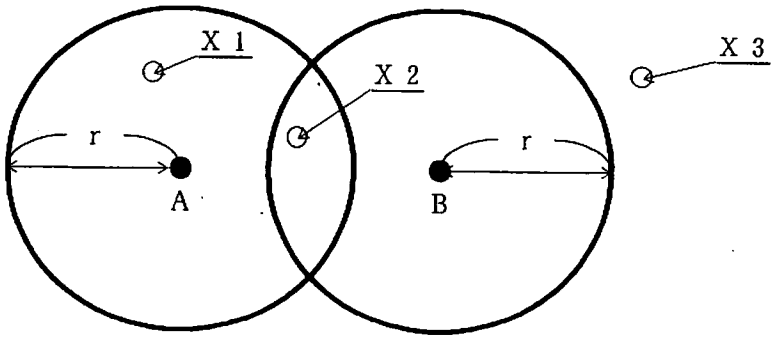


図７．　クラス未知パターンＸ１，Ｘ２，Ｘ３　及びクラス既知パターン
Ａ，Ｂとの関係

2．　構造活性相関分野への「超球法」の適用

2．1　構造活性相関の基本原理

現在行われている様々な形式での構造活性相関に対するアプローチを表1に示す。　この中には、統計的アプローチ（パターン認識を含む）を筆頭とし、理論化学計算によるもの、グラフィックディスプレイ（3次元）を多用したもの等数多く存在する。　このように数多くのアプローチがあるにもかかわらず、これらアプローチの根底に潜んでいる基本原理は単純且つ限られている。　この限られた原理中最も代表的なものが『似た化合物は似た活性を示す』（＊）というものであり、構造活性相関分野ではほぼ公理といえる程重要なものである。

表1．　構造活性相関に対する計算機によるアプローチ

分　類	使　用　手　法	使用基本理論	使用パラメータ
統　計 パターン認識	定量的構造活性相関 ・Hansch-Fujita 法 ・Free-Wilson 法 定性的構造活性相関 パターン認識による	各　種	物理化学的 パラメータ使用 特に限定しない
理論化学計算を 用いたアプローチ	分子軌道 分子力学 配座解析	ドラグレセプター	電子密度、 分極率、 その他 歪みエネルギー その他
グラフィック ディスプレイヲ 多用したもの	化合物表示技術 ・スケルトン ・空間充塡図 ・ドット図／他	ドラグレセプター	電子密度、 分子間相互作用、 水素結合力、
その他の手法	・化合物の重ね合わせ ・共通部分構造探索 ・人工知能技術の利用	各　種	各　種

例えば、理論化学計算やグラフィックディスプレイの多用で代表される「ドラグレセプター（Key & Rock）理論」は最終的には「同一レセプターサイトにフィットする似た化合物を探す」という言葉に置き換える事が可能である。　統計やパターン認識を用いたアプローチや、その他のアプローチも基本的には「似た」化合物を探す点では同じである。

2．2　「似た」概念への従来手法によるアプローチ

この「似た」という概念を数値データで扱うアプローチとしては、従来からいくつかの試みが行われてきた。　例えば様々な類似関数を考え、この値の比較を行うのが最も単純なものである。　（＊）　しかしながら、構造活性相関分野では比較的単純な類似関数を用いて活性を予測出来る例は極めて少数に限られている。

より一般的には、N次元パターン空間中のパターン相互の距離を「似た」という概念を表す一つの指標として利用する事が考えられる。　これが現在行われているパターン認識によるアプローチの基本原理である。　確かにパターン空間中近い位置に有るものは遠い位置にあるパターンよりも似た活性を示す可能性は高いが、先にも述べたようにパターン空間は薬理活性の相対的な強度を代表する空間でもある。　従って、この薬理活性と「似た」という概念がパターン空間上で一致しなければ『似た化合物は似た活性を示す』というルールの適用は意味がない。　現在行われているパターン認識による構造活性相関研究では、手法論が先に立ち、この薬理活性と『似た』という概念の一致や相互関係についての本格的な議論なくして解析が行われている。　構造活性相関上どこまでの距離がパターン空間上で実際に「似た」といえるのか、この基本を明確にしない限り統計やパターン認識による『類似』を基本とする構造活性相関解析は意味がなくなる。　しかしながら、現在利用可能な手法ではこの点（特にパターン間の相互関係）に関する考慮を行う事が不可能であり、この事が現在行われているアプローチの限界でもある。

2．3　構造活性相関の基本原理と「超球法」との結合

前節で述べた問題点を解決する事を目的とし、先に展開された超ボリューム概念を基本

とする超球法及び傾斜超球法を構造活性相関の問題に適用する事を試みた。　適用にあたり、予め超球法に用いられる超球の持つ構造活性相関上での意味を明確にしておく事が必要である。　これは先の『似た化合物は似た活性を示す』という事実が超球法ではどのような言葉で表現されるかという問題に置き換えられる。　即ち、

『超球体の内部はその超球体を形成する化合物と「似ている」、即ち活性が似ている（同じ）領域を意味する』　（図１）

ものと定義する。この結果、「似ている」という概念はパターン空間上での距離から、“超球”という絶対的な大きさを持つ有限な領域として明確化される。　従って、従来手法のように単なる相対的な距離関係だけで類似関係を議論するものと異なり、解析上での目的が明確、且つ分類結果の信頼性も飛躍的に向上させる事が可能である。

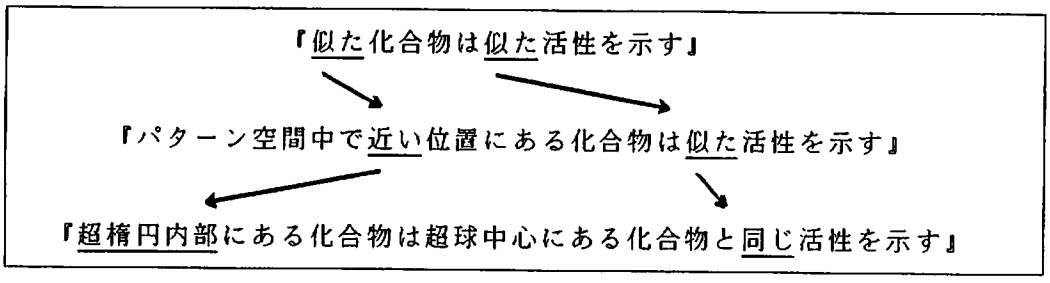
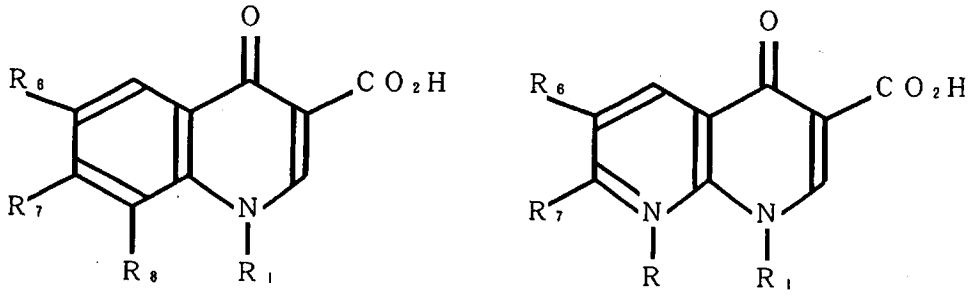


図１．構造活性相関と「超球法」との相互関係

３．　実験：Ｑｕｉｎｏｌｏｎｅ系抗菌活性化合物

Quinolone 基本骨格を有する化合物は１９６０年にピリドンカルボン酸系化合物に抗菌活性が見出されて以来、新種の抗菌活性化合物として注目されてきた。　特にグラム陰性菌に対する活性が強いため今後の開発が期待されている化合物である。

今回解析に用いたデータは１９８６年Ｄomagla（１９８６）らが行った薬理試験のデータを用いた。　化合物はQuinolone 及び1,8-Naphthyridine基本骨格を有する５７化合物である。　図２にはその代表的な基本骨格を示している。　化合物は大腸菌に関する最小阻止濃度（ＭＩＣ）の値に応じて２クラスに分類し、６．３μｇ以下の化合物を活性化合物、１２．５μｇ以上の化合物を不活性化合物とした。　それぞれ活性化合物として２７化合物、不活性化合物として３０化合物、総数５７個を用いて解析を行った。



Ｑｕｉｎｏｌｏｎｅ誘導体　　１，８－Ｎａｐｈｔｈｙｒｉｄｉｎｅ誘導体

図２．使用化合物誘導体構造式

解析に用いる数値データは構造活性相関支援システムＡＤＡＰＴ（Automated Data Analysis using Pattern recognition Toolkits）（Ｓtuper et al., １９７９）の化学構造式から数値データへの変換機能を用いて、約八十種類の数値データへと変換した。

続いて、様々な特徴抽出手法により、今回用いるデータセットに対し最適な記述子５種が選びだされた（湯田他、１９８８）（表２）。　超球法、傾斜超球法及びその他のパターン認識による解析は、この５種の数値データを用いて行なわれた。

表２．　解析に用いた数値データ

1．　パスの数／原子数
2．　化合物中の窒素数
3．　化合物中の基本リングの数
4．　求核スーパーデローカライザビリテイ
5．　トーション歪みエネルギー

今回用いる超球法及び傾斜超球法のプログラムはＡＤＡＰＴに組み込まれ、構造式入力～数値データへの変換～特徴抽出～パターン認識の実行等全ての作業をＡＤＡＰＴ上で行った。　図３には今回行ったＡＤＡＰＴ上での作業フローを示す。　尚、傾斜超球法に用いるメンバーシップ関数としては２次関数、１次関数及びステップ関数の３種類用いた（図４）。　又従来手法との比較を行う目的で、今回の超球法及び傾斜超球法と分類法上でのアプローチが基本的に類似しているＫ－ＮＮ（最近隣法）法も実行した。

化合物構造式 薬理データ　入力	Ｑｕｉｎｏｌｏｎｅ／１,８-Ｎａｐｈｔｈｙｒｉｄｉｎｅ Ｅ．Ｃｏｌｉ，ＭＩＣ（最小阻止濃度）μｇ
記述子創出	約８０種（２次元／３次元／物理化学的パラメータ）
特徴抽出実行	バリアンスウエイト法、相関係数、他
「超球法」及び 「傾斜超球法」実行	クラス分類及びメンバーシップ値による 分布状態のチェック

図３．　ＡＤＡＰＴによる解析作業流れ図

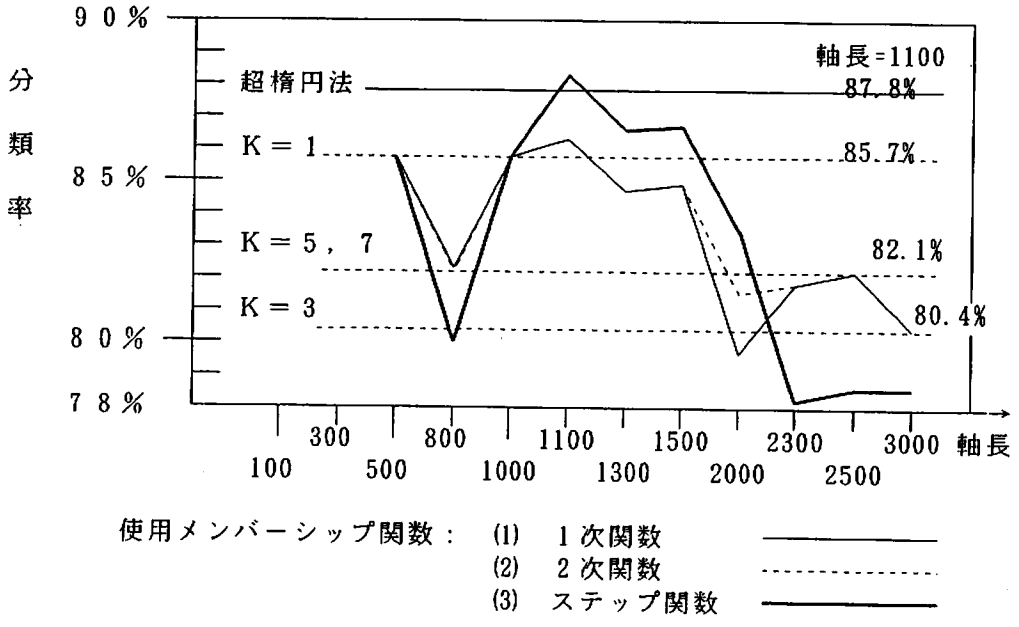


図４．超楕円法、傾斜超楕円法による分類結果と用いた軸長サイズとの関係。
傾斜超楕円法に用いたメンバーシップ関数は３種類である。尚、参考

４．　実験結果及び考察

表３に超球法、傾斜超球法及びＫ－ＮＮ法とによる分類結果が示されている。　超球半径が１１００（無名数）の時、超球法及び傾斜超球法（ステップ関数使用）ともにＫ－Ｎ

N法よりも僅かではあるが高い値を示している。　超球法及び傾斜超球法において超球の内部に存在しなかった化合物（アウトライヤー）は活性クラスに3化合物、不活性クラスに2化合物の総数5化合物存在し、全化合物数の約一割弱であった。　尚、超球法のソリッドモデルではこれらアウトライヤーの他にクラス決定が出来なかったもの（活性及び不活性クラスとで化合物数が等しい）が活性及び不活性化合物それぞれ1化合物ずつ出ている。

表3．　超楕円法、傾斜超楕円法及びK－NN法による分類結果

	活性化合物	不活性化合物	総　合
「超楕円法」 軸長；1100	91．3	84．6	87．8
ｸﾗｽ未決定化合物数	1	1	2
「傾斜超楕円法」 ステップ関数 軸長；1100	91．7	85．2	88．2
アウトライヤー数	3	2	5

「K－NN法」	活性化合物	不活性化合物	総　合
K＝1	85．2	86．2	85．7
K＝3	81．5	79．3	80．4

図4は傾斜超球法で用いた超球の半径とその超球を用いた時の分類率との関係を示す。用いた関数は3種類（2次関数、1次関数、ステップ関数）（図5）である。　超球半径と分類率との関係には1100をピークとする山形を形成している事がわかる。　用いた3種類のメンバーシップ関数は何れも同じ傾向を示しており、この中で今回はステップ関数による結果が最も高い分類率を示している。

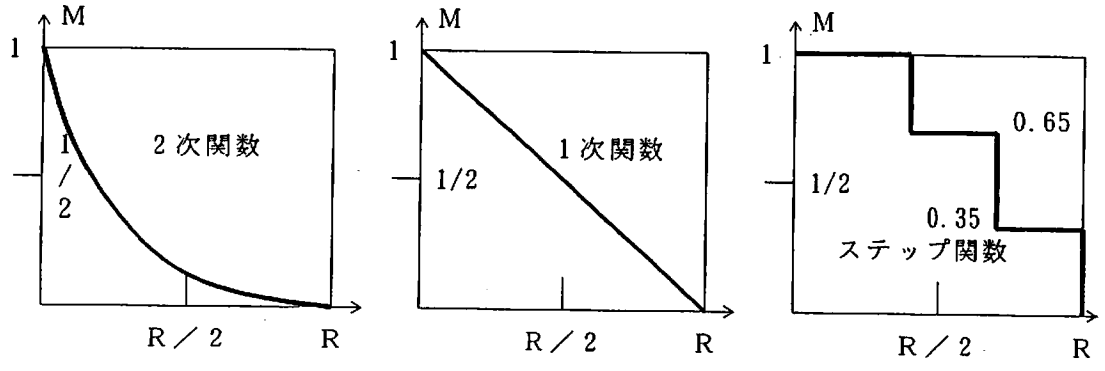


図5．傾斜超楕円法に使用したメンバーシップ関数（3種類）

図6に明らかなように、分類率は半径が800以下になると再び上昇するが、この時は分類不可能なアウトライヤーの数も急激に上昇してくる（800で11個、500で21個）為分類に利用されるパターンの数が少なくなり、分類率の信頼性が急速に低下するため今回は半径が800以下の結果は検討の対象としていない。

60

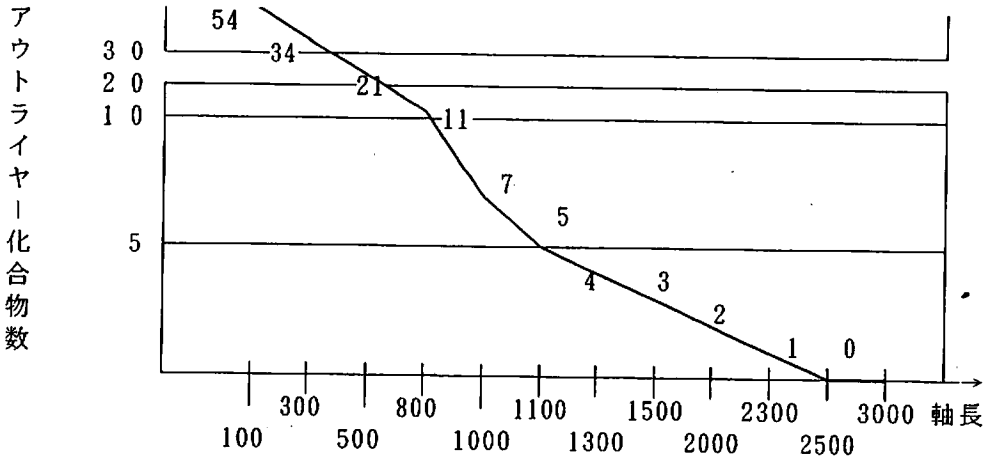


図６．超楕円及び傾斜超楕円法に用いた軸長サイズとアウトライヤー化合物数

採用すべき超球半径の値としては、分類率が最高に成った時の半径を採用することとする。　従って今回用いたデータに対しては、３種類のメンバーシップ関数総てに最高の分類率を示す１１００を超球の半径として採用した。

採用された１１００のサイズの超球を用いた分類の結果アウトライヤーとして認識された５個の化合物に共通する特徴としては、化合物を構成する環の数が他の化合物と大きく異なっている点である。　即ち、他の大部分の化合物が２～３環性化合物であるのに対し、取り出されたアウトライヤー化合物は単環性化合物と４、５環性化合物であり、残りの２化合物が３環性化合物である。　即ち他の化合物とは基本構造が大きく異なる化合物が分離されている事がわかる。　アウトライヤーの残る２化合物は３環性化合物であるが、他の要因（残る４記述子）との総合的な影響の為と考えられるが、明記出来る程の構造上の差異は見られない。

又、超球のサイズとアウトライヤーの数との関係では、軸長が２５００になった時点で０となる、即ち全てのパターンが超球の内部に取り込まれた事を示している。　又半径が１０００以下では急激にアウトライヤーの数が増えるのに対し、１０００以上ではその数の減少程度が小さくなっている事がわかる。　この事実は、全パターンは空間上で互いに最大距離２５００以内の空間に分散している事、及び大部分のパターンは互いに１０００以下の距離で存在している事を示すものである。

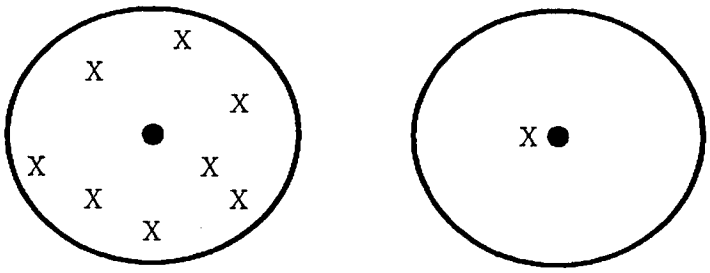
今回用いたアプローチでは単に分類率のみならず、個々のパターン毎に算出されるメンバーシップ値を求める事により、そのパターン周辺の環境についての情報が得られる。この結果、同じクラスに分類された化合物同士のランキングや誤分類の原因追求等が容易になり、構造活性相関上より高度な解析が可能となる。

表４．　メンバーシップ値が殆ど同じ２化合物の比較

化合物ＩＤ	重なった超楕円体のメンバーシップ値	化合物の総合メンバーシップ値
化合物２９	0.3640 0.1457 0.0949 0.0846 0.0785 0.0627 0.0368 0.0192 0.0185	0.63951
化合物４３	0.5831	0.5831

例えば、同じメンバーシップ関数値を持っていても、超球内部に落ち込んだ化合物の数が大きく異なる場合（表４）。　表中、化合物２９と４３とはメンバーシップ値が０．６３９５と０．５８３１と似た値を持っている。　しかし、超球内部に落ち込んだ化合物の数は２９と４３とで大きく異なっている。　即ち、化合物２９は超球中に９個パターンが落ち込んでいるのに対し、化合物４３は１個しか落ち込んでいない。　つまり、化合物２９と４３とで分類の基準となるメンバーシップ値の値は殆ど同じであるが、これら２化合物周りの環境は大きく異なっている事を示す。　この２化合物周りの環境をイメージ化したのが図７である。　つまり、化合物２９が関与する超球は何れも外周部に落ち込んでお

り（９化合物）、一方化合物４３は極めて近い位置に１個の化合物が存在している事になる。　この時、２９、４３どちらのパターンがより重要かどうかについての判断は超球法が適用される分野により異なってくる。　この他にもメンバーシップ値の検討により構造活性相関上貴重な情報が多数得られているが、ここでは省略する。



化合物２９　　化合物４３

図７．メンバーシップ値による化合物２９、４３の環境イメージ
Ｘは隣接パターンの中心位置を示す

５．　結　論

個々のパターンをボリュームのある超球として扱うという超ボリューム概念を基本として展開された超球法、及びその応用系である傾斜超球法を用いて構造活性相関分野の分類問題に適用を試みた。　構造活性相関分野においては最も基本的な公理として『似た化合物は似た活性を示す』という事実がある。　この事実は構造活性相関における解析においてはパターン相互の関係を重視しつつ解析する事が必要であるという事に等しい。　従来手法がこのようなパターン間の相互関係を無視して解析を行っていたのに対し、今回開発された手法は、このようなパターン間の相互関係に関する情報を十二分に取り込んだ形での分類が可能である。　以下に超球法及び傾斜超球法の構造活性相関解析における特徴を簡単にまとめる。

（１）　解析手法が構造活性相関の基本公理、『似た化合物は似た活性を示す』を忠実に再現する形で展開されている為、解析手法及びその結果と構造活性相関との関係が明白であり、情報の解析が容易になる。

（２）　分類結果の信頼性が従来手法（Ｋ－ＮＮ）よりも高い。　これは超球法及び傾斜超球法ではＫ－ＮＮ法にあった分類に関する様々な矛盾点が解決されている為である。

（３）　構造活性相関解析では重要な“アウトライヤー”の識別が容易である。

（４）　メンバーシップ値の検討により、個々のパターン周りの環境に関する様々な情報が容易に得られる。

（５）　多クラス分類が可能である。

ここで示した特徴は、Ｋ－ＮＮ手法と共通である（５）番目の特徴を除けば総て超球法及び傾斜超球法独自の特徴である。　従来の化合物（パターン）を“点”として扱い、単に点同士の距離関係だけを用いて分類する手法に比べ、ボリュームを持った超球として表現した超球法は構造活性相関の分野に適用するならば、従来の手法と比べ数多くの長所を有する事が実際の薬理データを用いた解析により示された。

尚、この超球法における今後の課題としては、解析時に用いられる超球の半径の決定であり、この半径決定の為のアルゃリズムの開発がある。　又、超球法自体の解析精度を向上する観点から今回は一種類の半径のみを用いたが、これを個々のパターン毎に異なる半径を用いた解析が行えるようにする事が必要である。

又、今回は傾斜超球法においては３種類のメンバーシップ関数を用いたが、これらは何れも「活性が強い」という事を前提に考えられたものである。　今後はこれらの他に、「活性が有る／小さい／殆ど無い」等の場合にマッチしたメンバーシップ関数を設定し、より詳細な分類が可能となるように展開を行う予定である。

〔参考文献〕

Domagra, J.M., Hanna, L.D., Heifetz, C.L., Hutt, M.P., Mich, T.F., Sanchez, J.P. & Solomon, M., (1986). New Structure-Activity Relationships of the Quinolone Antibacterials Using the Target Enzyme. The Development and Application of a DNA Gyrase Assay. J.Med.Chem., 29, 394-404.