## *VII-3.　クラスター分析*

### *VII-3-1. 機械学習としてのクラスター分析*

何らかの尺度で表現されたデータを使って、集団をいくつかのクラスに分けることをクラスター分析と言います。人は日常生活の中で、集団を共通した特徴を持ついくつかのグループ（クラス）に分けることを感覚的にやっています。水族館の水槽の中にいくつかの種類の魚がいるとき、複数の種類の魚がいるとなんとなく人は感じます。これは感覚的なクラスター分析です。この能力は人によって違っていて、細かい違い気が付いて似たような魚を違う種類の魚だと瞬時にして認識する人もいれば、違いに気が付かずに同種だと認識してしまう人もいます。統計学的なクラスター分析も同じで、分析結果として納得できるクラス分けをすることもあれば、よく意味の解らないクラス分けをする場合もあります。

クラス分けが妥当だったりうまくいかなかったりするのは、「似ている」こと、「同じ仲間と認識する」ことの定義が、人によってあるいは分析方法によって違うからです。「似ている」ことや、「同じ仲間である」ことは哲学的に難しく面倒くさい問題です。感覚的認識でも統計数学的分析でも同じで、それは物の認識に伴う本質的な問題です。類似度（似ていること）について考えてみます。個々のデータ間の類似度については、変数の数値が近いこと、つまり、空間的な距離の近さを類似度とすることが考えられます。しかし、数値をそのまま類似度の計算を使うことが不適切な場合もあります。プードルにはスタンダードプードル、ミニチュアプードルなど大きさの違ういくつかの種類がありますが、皆同じプードルの形をしていて、プードルだと認識されています。もし、数値を比較すると、中型日本犬とプードルを同じ中ぐらいのサイズの犬と認識し、トイプードルとチワワを小さい犬と同じ仲間として認識してしまうかもしれません、こういう場合は、形態的な変数の数値の比が似ているということを類似度にした方が良いでしょう。社会の特徴や人の行動の類似性についても比を取った方が良い場合が多いでしょう。このような比の類似性は、変数のベクトルの方向の一致の程度として捉えることが出来ます。内積を二つのベクトルの距離の積で割って、ベクトルの角度のコサインを取ればー１から１の数値になります。これを類似度とすれば、１の時にすべての比が完全に一致していることになります。「同じ仲間であること」の判定の仕方も様々なやり方が考えられるのです。数値の近さを類似として扱える場合にも、個々の変数の分散に違いがあったり、変数間に相関があったりしますから、変数を互いに独立だと考えて、そのままユークリッド距離をとるのか、変数間の相関を補正したマハラノビスの距離をとるのかということも問題になります。また、距離を数値化するにしても、グループの中心を仮定して個々のデータとの距離を考えるのか、中心を仮定せずに個々の期待感の距離を考えてより近いもの同士をつなげていくのか、遠いものを取り分けていくのか、グループ間の距離をどのように定義するか（グループの重心間の距離、最も近いもの個体間の距離、最も遠い個体間の距離等々）様々なことを考えなくてはなりません。これらの方法のどれが良いかは結局、得られた結果が納得できか、結果から何らかの意味を読み取ることが出来るかで判断するしかありません。普通、データには様々なグループから得られた情報が入って

いますから、それらを仕分けるために、クラスター分析をするのはやむを得ない自然なことですが、結果は意外に頼りなくて、いくつかの方法を試してみるしかないのです。
クラスター分析は代表的な多変量解析の一つで、主成分分析や対応分析などとともに、多くの教科書でその手法が解説されています。この解説でも最初は VI 章にクラスター分析の解説も入れようと思っていたのですが、そもそも似ている（類似度）とは何かとか、似ていれば同じグループにして良いのか（他人の空似）とか考え始めると、そちらの説明が長くなってしまいます。VI 章は線形代数学的な説明の流れを意識して全体を構成しています。クラスター分析でも、類似度をどのように考えるか、相関がある場合の距離の問題等々、線形代数学的な説明も必要なのですが、より大事なのは類似度やクラスの属することの考え方の説明だと考えました。生まれたばかりの赤ん坊は識別性の悪い漠然とした感覚の中で様々な刺激を受け取り、その刺激を何となく仕分けしながら、お母さんとか、風とか、母乳とかの違いを認識し、「お母さん」、「風」、「母乳」というクラスの違いを発見し、やがてそれに言葉を与えていきます。その過程はおそらく試行錯誤的です。いろいろな仕組みでクラスター分析して、その結果から納得できる妥当なクラス分けをして、そのクラスター構造とそれぞれのクラスの意味を納得していくという行為をしているのです。こういう学習を「教師なし学習」と言います。おそらくこれが私たちの知のベースです。クラスター分析には類似度の決め方、クラスの連結法などに様々な考え方があって、それらを紹介すると、一体、何をどのように選択すればよいのかという質問が来ます。データに使われている尺度も違えばデータ分布の特性も違います。初めから正解があるわけではありません。試行錯誤的に結果として与えられるいくつかのクラスター構造に納得できるものを探すしかないのです。こういう作業では、コンピュータは便利な道具です。大量の繰り返し的なプロセスを効率よく行うことが出来るからです。そこで、機械学習のところで、クラスター分析のいくつかの方法をまとめて紹介して、それぞれの方法で、何を類似度としているのか、同じクラスに属することをどのように定義しているのかを比較しながら解説すれば、結果として与えられたクラスター構造を理解して妥当な構造を選択することを助けると考えました。そのために、まず、クラスター構造の作り方の違いが理解しやすい階層的クラスター分析を説明し、次に、非階層的クラスター解析として K-means 法および、混合ガウスモデルによる確率的クラスタリングの３つを紹介します。