1-2 樹形図ができるまで

食い倒れの街大阪の大阪市役所から港町神戸の神戸市役所までの直線距離は、約28㎞。この数字は、高校で学ぶ座標平面上の2点間の距離を計算する方法で求められます。ところが、クラスタ分析を行うときは、2つのデータ間の距離に加えて、すでに作られているクラスタとデータの距離、さらに、クラスタとクラスタの距離を考える必要が出てきます。これらの距離をどのように定めるかがクラスタリングの方法の違いとなり、どのクラスタリングの方法を選ぶかによって、同じ変数の組でも樹形図が変わることがあります。

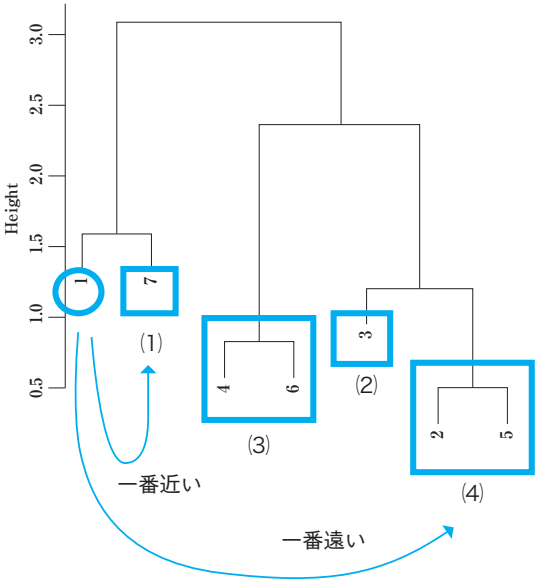
1 クラスタの構成

「すでに説明したとおり、クラスタ分析は、似たもの同士のデータをまとめてクラスタを作ってデータを分類し、多変量データのおよその全体像をつかむ手がかりにする方法だ。そして、似たもの同士を寄せ集めるときの基準が、距離の考え方だ」

「距離が長いか短いかで、似ているか似ていないかを決めるのですか？」

「そう。樹形図で説明しよう」

図1-9　樹形図でみるデータ間の距離
（図1-3の再掲載）



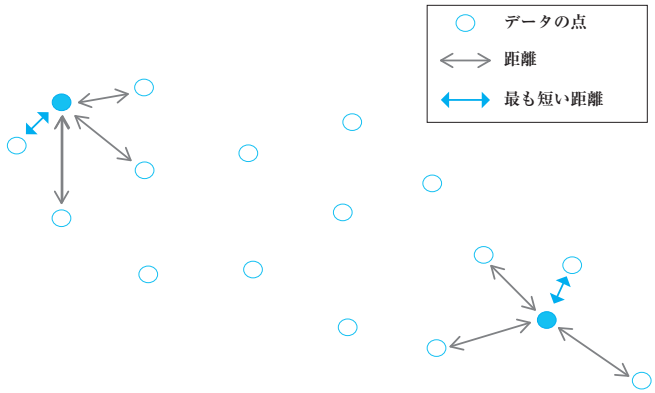
Catfood01のデータを使って描いた図1-3を再度使って説明します。図1-3に必要事項を書き込んだものが図1-9です。データ間の距離は、樹形図では枝の長さで表されます。図1-9で一番左にあるデータ番号1の葉を基準にして考えます。

1. データ番号1の葉から最寄りの分岐までもどってたどり着けるのは、データ番号7
   ⇒ データ番号1との距離が最も短いのはデータ番号7
2. データ番号1の葉から根の分岐まで戻って反対側の枝をたどったとき、一番近いところにある葉はデータ番号3
   ⇒ 2番目に距離が短いのはデータ番号3
3. データ番号1の葉から最も長く枝をたどらなければ行き着くことができない葉は、データ番号2と5
   ⇒ 最も距離が長いのは、データ番号2と5

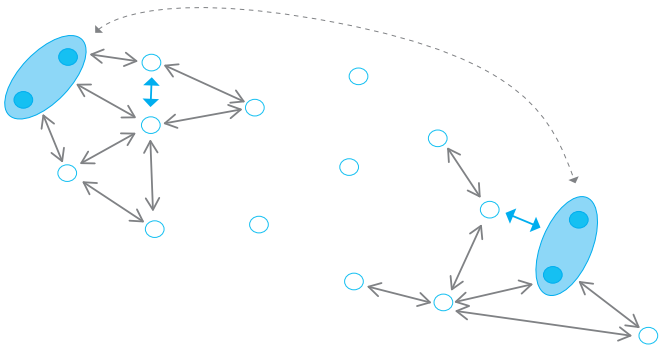
「このようにしてデータ同士の距離が樹形図に反映されて、距離が近いもの同士がまとめられて、クラスタが作られていくのですね」

「距離を使ってクラスタが作られていく手順を、模式図を使って説明しよう」

図1-10　クラスタの構成手順の模式図



1. ある1つの点に着目し、他のすべての点との距離を測る
2. すべての点について1の作業を行う
3. 最も近い点と最初のクラスタを作る



4. クラスタ同士の間、クラスタとデータの間、データ同士の間の距離を測る