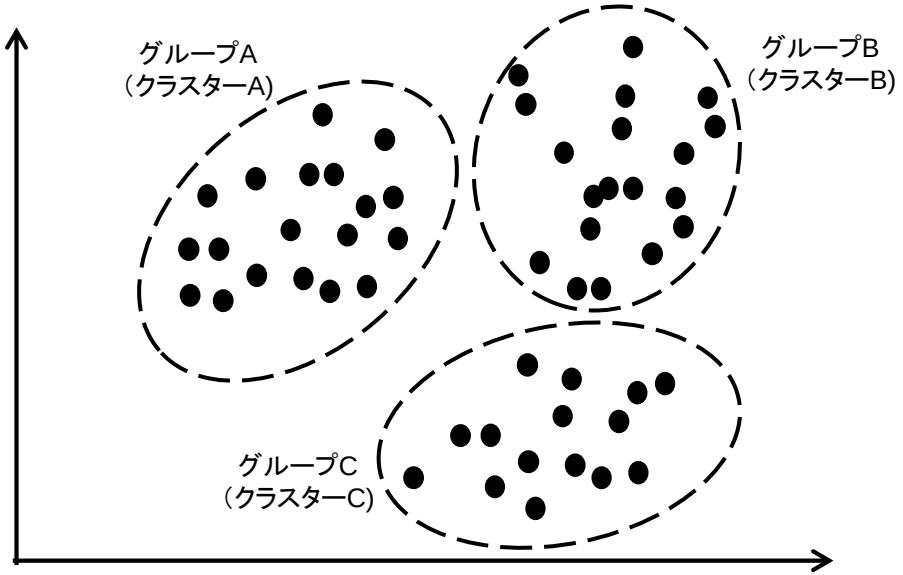
ホームページ：　分かりやすくまとめた会社実務のための基礎知識と事例
URL　http://www.bisquerhino12.sakura.ne.jp/

クラスター分析の概要

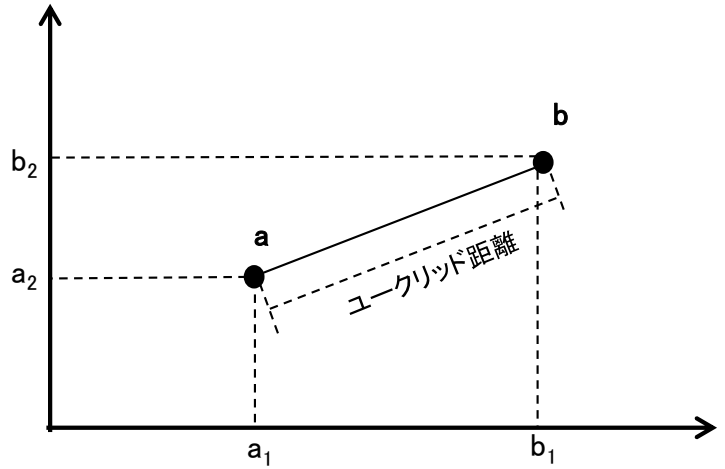
2019年　ver. 1.0
倉谷　隆博

6－1. クラスタリング

クラスタリングとは、特徴の似たデータ同士をクラスター（塊り）にすること、グループ化することである（図1）。
例えば、コンビニエンスストアは、その立地などによって売れ筋商品が異なる。その違いなどによってコンビニエンスストアをグループ化しておき、グループごとに異なるセールスプロモーションを実施することなどが考えられる。また例えば、各種の潤滑剤や脱臭剤などをB to Bで販売する会社では、商品数が多い場合、商品の機能（特徴）ごとにクラスタリングによってグルーピング化しておき、そのデータを技術営業に活用していくことが考えられる。



（図1） クラスタリングの考え方の模式図



（図2） ユークリッド距離

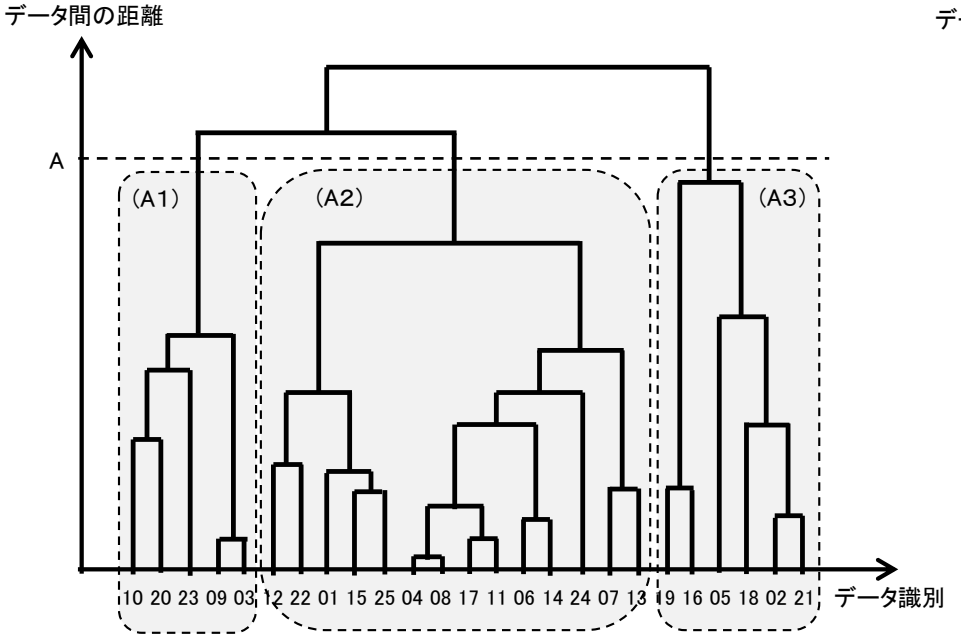
代表的なクラスタリングの方法は、樹形図（デンドログラム）を作る方法とk-meansクラスタリングである。
クラスタリングでは、数多くのデータのなかでデータ同士の類似度を測定する（どの程度似ているのかの判断）には、データ間の距離が使用される。距離としては、多くの場合、ユークリッド距離が使用される。

2つのデータが、$\mathbf{a}=(a_1,a_2,\cdots,a_n)$ と $\mathbf{b}=(b_1,b_2,\cdots,b_n)$ であるとき、
2つのデータのユークリッド距離は、$[(a_1-b_1)^2+(a_2-b_2)^2+\cdots+(a_n-b_n)^2]^{1/2}$ である。
（図2）は、$n=2$ のときのユークリッド距離を示す。

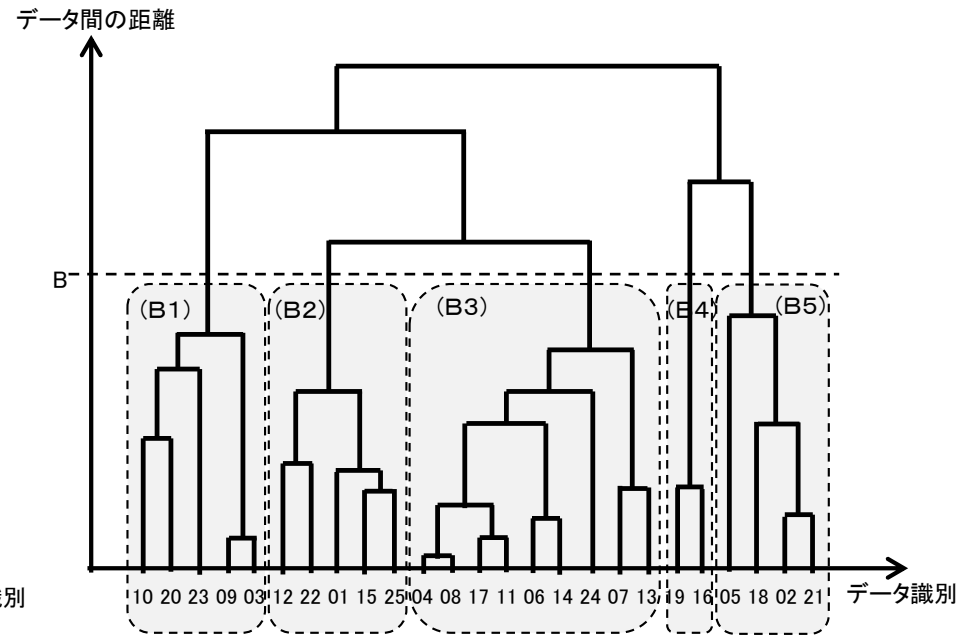
６－２．樹形図（デンドログラム）

デンドログラムは、距離の近いデータ同士で、つまり特徴が似ているデータ同士で1つずつ順番に塊り（クラスター）を形成していく方法である。データ同士の類似度を可視化できる。図の横軸は個々のデータの識別（データ番号など）、縦軸はデータ間の距離（類似度）である。よって例えば、それぞれのデータを関連付ける縦の棒の長さが短いほどデータ間の類似度は高いことになる。

（図２－１）と（図２－２）のデンドログラムは同じデータのものである。
（図２－１）のデンドログラムは、グループ間の距離をAに設定した場合のクラスター分割の様子を示す。（A1）、（A2）、（A3）の3つのクラスターに分割できる。一方、（図２－２）は、グループ間の距離を、（図２－１）のAより短いBに設定した場合のクラスター分割の様子を示す。（B1）、（B2）、（B3）、（B4）、（B5）の5つのクラスターに分割できることを示している。
クラスター分割するときのクラスター間の距離設定を変えることによって、分割されるクラスターの数が変わる。



（図２－１） 樹形図の例　クラスター数3で分類



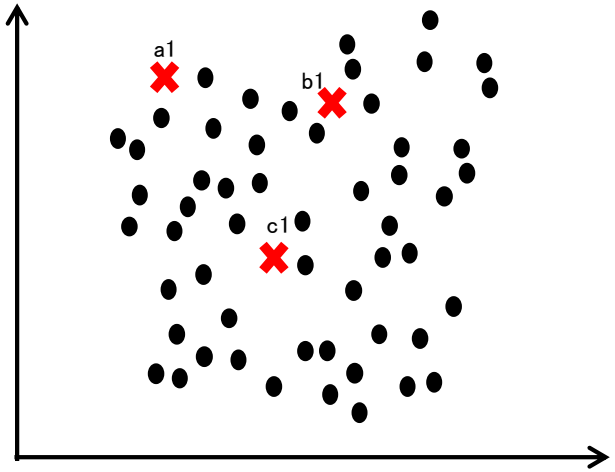
（図２－２） 樹形図の例　クラスター数5で分類

６－３．k-meansクラスタリング

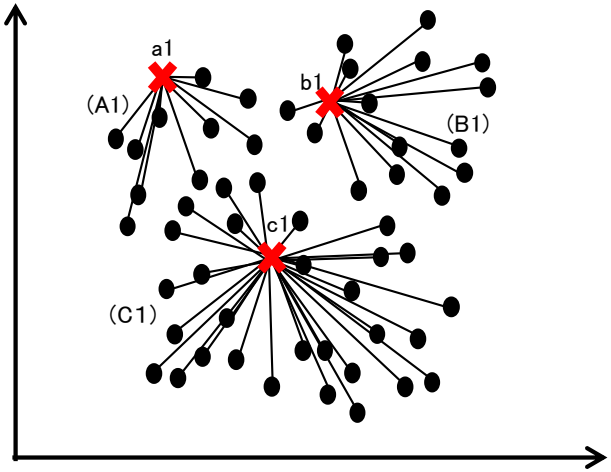
k-means法は、試行錯誤を繰り返して似通ったグループの塊り（クラスター）を形成していく方法である。

（手順）

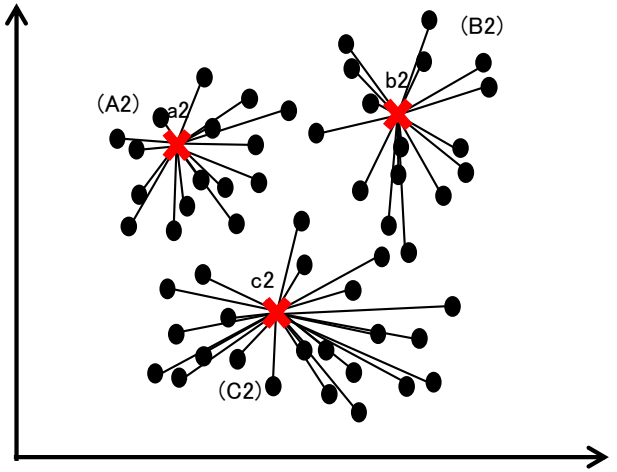
1）分割したいクラスターの数に合わせて、クラスターの仮の中心点を適当に配置する。（図３－１）のa1、b1、c1の３点の相当する。
2）（図３－２）に示すように、それぞれのデータごとにデータからの距離が最も近い中心点をa1、b1、c1のなかから選んで、データの属するクラスター（A1）、（B1）、（C1）を作り上げる。
3）次いで、それぞれのクラスターの重心（平均値） a2、b2、c2を求め、この重心からの距離が近いデータを選び直して、データの属するクラスター（A2）、（B2）、（C2）を作り上げる。
4）それぞれのクラスターに属するデータの割り振りが変動しなくなるまで、3）の操作を繰り返す。



（図３－１） クラスタリングの考え方の模式図
（中心点の仮配置）



（図３－２） クラスタリングの考え方の模式図
（仮の中心点を使った分割）



（図３－３） クラスタリングの考え方の模式図
（重心を使った分割）

k-means法を使用する上で注意することは、最初に配置するクラスターの仮の中心点の場所によって、クラスターの分割結果が変わってしまうことがある、ということである。最初に配置する中心点の位置をいろいろ変えてみて、クラスター分割の結果がどのように変わるか、確認する必要がある。
また、クラスターの分割数は最初に配置する中心点の数によって決まってしまうので、分割数をいくつに設定するのがいいのかも、結果を見て検討する必要がある。

なお、k-means法という名前は、k個のクラスター分割するに際して、クラスタの重心、つまり平均値（mean）を使用することに由来する。

参考にした書籍

河本薫：「会社を変える分析の力」、 講談社現代新書
小山昇：「数字は人格」、 ダイヤモンド社
永野裕之：「ビジネス×数学＝最強」、 すばる舎
竹内薫：「数学×思考＝ざっくりといかにして問題をとくか」、 丸善出版
中西達夫：「統計データをすぐに分析できる本」、 アニモ出版
中西達夫：「すぐれた判断が統計データ分析から生まれる」、 実務教育出版
豊田裕貴：「マンガでわかる ビジネスを成功に導くデータ分析」、 ナツメ社
向後千春、冨永敦子：「統計学がわかる」、 技術評論社
石井俊全：「意味がわかる統計学」、 ベレ出版
涌井良幸、涌井貞美：「中学数学でわかる統計の授業」、 日本実業出版
涌井良幸、涌井貞美：「統計学の図鑑」、 技術評論社
西内啓：「統計学が最強の学問である」、 ダイヤモンド社
西内啓：「統計学が最強の学問である（実践編）」、 ダイヤモンド社
西内啓：「統計学が最強の学問である（ビジネス編）」、 ダイヤモンド社
森岡毅、今西聖貴：「確率思考の戦略論」、 角川書店
デビッド・マクアダムス：「世界一流企業はゲーム理論で決めている」、 ダイヤモンド社
河村真一ほか：「本物のデータ分析力が身に付く本」、 日経BPムック
末吉正成、末吉美貴：「Excel ビジネス統計分析 この分析できますか？」、 翔泳社
谷岡一郎：「社会調査のウソ リサーチ・リテラシーのすすめ」、 文藝春秋
林知己夫：「調査の科学」、 ちくま学芸文庫
八谷大岳：「データ解析」シリーズ 全15回、（例）データ解析 第1回 ベクトルの復習 – YouTube