

# クラスター分析

1997/10/31

阿部圭司

E-mail: abek@tcue.ac.jp

## 目的

証券市場で売買される株式は、所属産業で分類されている。こうした分類は証券投資にはよく言及されることも多い。当然、それらの銘柄の値動きは似たようなものである。が、分類には時間的な安定性が弱く、産業分類の方法自体にも問題点は存在する。より明確な基準を用いて分類を行うことを研究者が意図したとき、定性的な分類を離れ、定量的な分類を試みることは価値があると思われる。定量的な分類を行う多変量解析の方法がクラスター分析である。

## クラスター分析

### 1.概要

#### 1.1.一般的な目的

クラスター分析とは、異質なものが混じっている対象（ケース、変数）の中から、似ているものを凝集し、グループ（クラスター、cluster）分けを行うためのアルゴリズムの総称を指す。すべての研究領域で各研究者が直面している一般的な問題は、観測されたデータをいかに意味のある体系に組織立てるか、すなわち、いかに分類を行うかにある。分析手法は大きく3つのカテゴリーに分かれている。つまり、凝集法（ツリークラスタリング）、Two-way 法（ブロッククラスタリング）、K-means 法である。

#### 1.2.統計的有意性検定

クラスター分析ではクラスタリングアルゴリズムを重視し、統計的有意性検定については触れられていない。実際、クラスター分析は、オブジェクトをクラスターに分類する様々なアルゴリズムの“集まり”であって典型的な統計的検定は含まれない。他の多くの統計的なプロシジャーと違って、クラスター分析法は、事前の仮説が無く、探索的な立場で分析を行っている場合に使用される。ある意味では、クラスター分析は“最も有意となる可能性のある解”を見つけ出す。よって、統計的有意性検定は、（K-means 法のように）p 値が出力される場合でも、この場合はあまり適切でない。

## 2.凝集法（ツリークラスタリング）

### 2.1.一般的な考え方

概要で示された例は、凝集的またはツリークラスタリングアルゴリズムの目標を提示している。このアルゴリズムの目的はオブジェクトをある類似性や距離の測度を利用して、逐次的に大きくなるクラスターと一緒に結びつけていくことにある。クラスタリングのこのタイプの典型的な結果が階層的なツリー（樹状図：デンドログラム）である。

### 2.2.樹状図

今、水平樹状図を考える。このプロットの左側は、各オブジェクトが1つのクラスとなっている。そこで、何と何が同じで、何が異なるという基準をゆるめることを考える。すなわち、2つまたはそれ以上のオブジェクトが同じクラスターのメンバーであると宣言するための基準を下げる。結果として、オブジェクトをどんどん結びつけ似てない構成要素を増やしながらかラスターを大きくしていく。最終的に、すべてのオブジェクトが1つに集まるようになる。これらのプロットでは、水平の軸は結合距離を表す（垂直樹状図では垂直な軸が結合距離を表す）。グラフの各ノード（新しいクラスターが形成される所）で、各エレメントが一緒になり新しい1つのクラスターとなる基準の距離を読みとることができる。データに、互いに似ているオブジェクトのクラスターの形でのはっきりとした“構造”が含まれているときは、この構造が樹状図にはっきりと現れる。凝集法による成功した分析の結果として、クラスター（枝分かれ）を見つけることができ、その枝分かれの解釈ができる。

### 2.3.距離測度

凝集的またはツリークラスタリング法では、クラスターを形成する際に、オブジェクト間の非類似度や距離を使用する。これらの距離は1次元または多次元を基にしている。たとえば、インスタント食品のクラスタリングを行いたいとする。このとき、食品に含まれるカロリー、価格、味などを考慮する必要がある。多次元空間内のオブジェクト間の距離を計算する最も端的な方法は、ユークリッド距離を計算することである。2または3次元の空間の場合、この測度は空間内のオブジェクト間の実際の幾何的距離と一致する。しかし、凝集アルゴリズムは、そこで使用された距離が実際の距離、または、研究者にとってより意味のある距離であるかどうかについては何の保証もない。特定の応用に対して正しい方法を選択することは研究者の責任となる。

A.ユークリッド距離：これは，多分最もよく使われる距離であり．これは多次元空間の単なる幾何的な距離である，次のようにして計算される：

$$D(x, y) = \left( \sum_{k=1}^T (x_k - y_k)^2 \right)^{\frac{1}{2}}$$

B.標準化ユークリッド距離：オブジェクトの分散の大小による距離の影響を避けるために，各変数の分散を1に標準化した値を考える．この距離は次のように計算される：

$$D(x, y) = \left( \sum_{k=1}^T (z_{xk} - z_{yk})^2 \right)^{\frac{1}{2}}$$

ここで， $z_x = \frac{x - \bar{x}}{S_x}$  だったりする．

C.ユークリッド距離の2乗：より離れているオブジェクトに大きな重みを与えたい場合は，ユークリッド距離の2乗を考える．この距離は次のように計算される：

$$D(x, y) = \sum_{k=1}^T (x_k - y_k)^2$$

D.マハラノビス距離：一般に，マハラノビス距離は2つまたはそれ以上の相関がある変数で定義される空間内での2点間距離の測度である．2つの変数に相関がない場合，点(ケース)は標準的な2次元散布図にプロットすることができる．この場合，2点間のマハラノビス距離はユークリッド距離と一致する．3つの無相関の変数があった場合，(3次元プロット内で)物差しで単純に距離が測れる．4つ以上の変数があった場合，プロット内の距離を表すことができない(当然か)．また，変数に相関がある場合，単純なユークリッド距離は適切な測度ではなくなる．一方，マハラノビス距離は相関も十分に考慮された形で情報を提供できる．この距離は以下の式で表される<sup>1</sup>．

---

<sup>1</sup> ここでのマハラノビス距離は判別分析で用いられるものと少し違うように見える．判別分析で用いられるp変量(サンプル数n)のマハラノビス距離は，

$$D^2 = AS^{-1}A'$$

で与えられる．ここで，Aは $[x_1 - \bar{x}_1, \dots, x_p - \bar{x}_p]$ を要素とするベクトル．Sは $x_p$ の分散共分散行列である．

$$D(x, y) = \sum_{j=1}^T \sum_{k=1}^T s^{jk} (x_j - y_j)(x_k - y_k)$$

$s^{jk}$  : 分散共分散行列 ( $s_{jk}$ ) の逆行列の (j,k) の要素

E.市街地 (マンハッタン) 距離 : この距離は単純な各次元の差である . ほとんどの場合 , この距離はユークリッド距離と似た結果になる . しかし , ( 2 乗してないので ) 1 つの大きい距離 ( 外れ値 ) の影響は抑えられる . 市街地距離は次のように定義される :

$$D(x, y) = \sum_{k=1}^T |x_k - y_k|$$

F.チェビシェフの距離 : この距離は , 1 つの次元でも違っているものは違っていると定義したい場合に適切なものとなる . チェビシェフ距離は次のように定義される :

$$D(x, y) = \max |x_k - y_k|$$

G.べき乗距離 (ミンコフスキー距離) : 非常に離れた距離に対する重みを増やしたり , 減らしたりすることを考える . これはべき乗距離で可能である . べき乗距離は次のように定義される :

$$D(x, y) = \left( \sum_{k=1}^T |x_k - y_k|^p \right)^{\frac{1}{r}}$$

ここで , r と p が指定できるパラメータとなる . 2 , 3 の例を計算すれば , この距離がどのように “ 振る舞う ” のか分かる . パラメータ p は個々の次元の違いに与える重みを調整し , パラメータ r はオブジェクト間の大きな差に与える重みを調整する . r と p が共に 2 であれば , この距離はユークリッド距離になる . 1 であれば市街地距離になる .

H.不一致割合 : この測度は , 分析に含まれる次元に対するデータがカテゴリ型であるときに特に有用である . この距離は次のように定義される .

$$D(x, y) = \frac{x_k \neq y_k \text{ の数}}{\text{変数の数}}$$

#### 2.4.凝集または結合ルール

最初の段階では、各オブジェクトが各クラスターを表しており、オブジェクト間の距離が選択された距離測度によって定義される。しかし、一度いくつかのオブジェクトが一緒に結合された後、新しいクラスター間の距離をどのように定義するのだろうか？すなわち、2つのクラスターが結合されるべく十分に似ていることを決定するための結合または凝集ルールが必要になる。これにはいろいろな可能性がある。たとえば、2つのクラスター内の任意の2つのオブジェクトが、結合距離より互いに近い場合に、その2つのクラスターを結合する。また、クラスター間の距離を定義するためにクラスター間の“最近隣距離”を使う；この方法は最近隣法と呼ばれる。このルールは“糸を引くような”タイプのクラスター、すなわち、偶然結合されたようなたった1つのオブジェクトで“一緒に結合されている”クラスターを作る。その他の方法として、クラスター間の互いに最も遠いオブジェクト間の距離を利用する方法がある。これは最遠隣法と呼ばれる。他にもいろいろな結合ルールがある。

A.最近隣法：2つのクラスター間の距離はそれぞれのクラスター内の最も近いオブジェクト間の距離として定義される。このルールは、ある意味では、オブジェクトを一緒につなげていきクラスターを作成する。よって、結果として得られるクラスターは長い“チェーン”になりがちである。クラスター p とクラスター q を統合し、クラスター t を作成した場合、任意のクラスター r との距離  $s_{tr}$  は、次の式で与えられる。

$$s_{tr} = \min(s_{pr}, s_{qr})$$

B.最遠隣法：クラスター間の距離は、それぞれのクラスター内の任意の2つのオブジェクトの距離の最大値として定義される。この方法は、オブジェクトが実際にはっきりとした“集団”を形成している場合には、常に非常によく機能する。クラスターが長くなったり、“チェーン状”になる場合は、この方法は適切ではない。距離は以下の式で与えられる。

$$s_{tr} = \max(s_{pr}, s_{qr})$$

C.群平均法：2つのクラスター間の距離は、2つのクラスター内のすべてのオブジェクトのペアの距離の平均として定義される。この方法も、オブジェクトが実際にはっきりとした“集団”を形成している場合には非常に有効である。しかし、長くなったりチェーン状のクラスターでも同じくらいの機能を発揮する。各クラスターのオブジェクト数をそれぞれ  $n_p, n_q, n_r$  とする

と,

$$s_{ir} = \frac{n_p s_{pr} + n_q s_{qr}}{n_p + n_q}$$

D.重み付き群平均法：計算において各クラスターの大きさを重みとして使用する点を除けば，群平均法と同じである．この方法は，クラスターサイズが非常にアンバランスであると思われる場合に使用される．具体的算式は不明．

E.重心法：クラスターの重心は次元によって定義される多次元空間内の平均的点となる．それは，ある意味で，各クラスターに対する重力の中心である．この方法では，2つのクラスター間の距離を重心間の距離として定義する．

$$s_{ir} = \frac{n_p}{n_p + n_q} s_{pr} + \frac{n_q}{n_p + n_q} s_{qr} - \frac{n_p n_q}{(n_p + n_q)^2} s_{pq}$$

F.重み付き重心法：計算において各クラスターの大きさの違いを考慮する点を除けば，重心法と同じである．クラスターのサイズがかなり違う場合（またはその疑いがある場合）には，この方法の法が好ましい．具体的算式は不明．

G.ワード法：この方法は，クラスター間の距離を分散分析的アプローチで評価しているため，他の方法とはかなり異なる．簡単にいえば，この方法は，各ステップで形成される任意の2つのクラスターの平方和を最小にするよう試みる．一般にこの方法は，非常に有効だが，小さいクラスターを作りやすい傾向にある．

$$s_{ir} = \frac{n_p + n_r}{n_t + n_r} s_{pr} + \frac{n_q + n_r}{n_t + n_r} s_{qr} - \frac{n_r}{n_t + n_r} s_{pq}$$

### 3. Two-way 法

#### 3.1.概要

先に，この方法についてクラスタリングされるべき“オブジェクト”という言葉を使って説明した（凝集法（ツリークラスタリング））．研究者が興味を持っている問題は常に，ケース（観測値）または変数という言葉で表現される．だが，どちらを用いてもクラスタリングが意味のある結果もつ場合がある．たとえば，標本内の企業（ケース）の健康状態についての様々

な測定値（変数）を収集しているとする．同じような症状の企業のクラスターを見つけるためにケース（企業）をクラスタリングすると同時に，変数（健康状態に関する測定値）をクラスタリングして，同じような健康能力を示すクラスターを作ることありうる．

### 3.2. Two-way 法

上の段落で，ケースまたは変数どちらかをクラスタリングすることについて説明したが，両方同時にクラスタリングしたい場合もある．これについては Two-way 法がある．Two-way 法はケースと変数両方が同時に意味のあるクラスターを形成している場合に役に立つ．結果として得られる構造（クラスター）は本来均質ではない．これはやや混乱を引き起こすかもしれない．実際，他のクラスタリングの方法（凝集法（ツリークラスタリング）と K-means 法参照）に比べ，Two-way 法はあまり使われない．しかし，研究者の中には，この方法は非常に強力な探索的なデータ解析のツールであると主張する人もいる．

## 4.K-means 法

### 4.1.概要

#### A.一般理論

このクラスタリングの方法は，凝集法（ツリークラスタリング）や Two-way 法とは非常に異なる．いま，ケースまたは変数のクラスターの数に関する仮説が既にあるとする．ここでたとえば，コンピュータに向かってできるだけはっきりとした3つのクラスターを作るように指示することを考える．これは K-means 法によって解決できる問題である．一般に，K-means 法は最もはっきりとした K 個のクラスターを作成する．

#### B.計算

計算上，この方法は分散分析の逆だと考えることができる．まず k 個のランダムなクラスターからスタートする．次に，(1)クラスター内での変動が小さくなり，(2)クラスター間の変動が大きくなるよう，クラスター間でオブジェクトを移動させる．グループ内の平均が互いに異なるという仮説の検定を行うときに，分散分析の有意性検定がグループ内の変動に対してグループ間の変動を評価しているという意味で，これは“分散分析の逆”だといえる．K-means 法では，プログラムは，分散分析の結果が最も有意になるように，オブジェクト（ケースなど）をグループ（クラスター）から出し入れする．（K-means 法の標準的な出力の一部は分散分析の結果となる．）

### C.結果の解釈

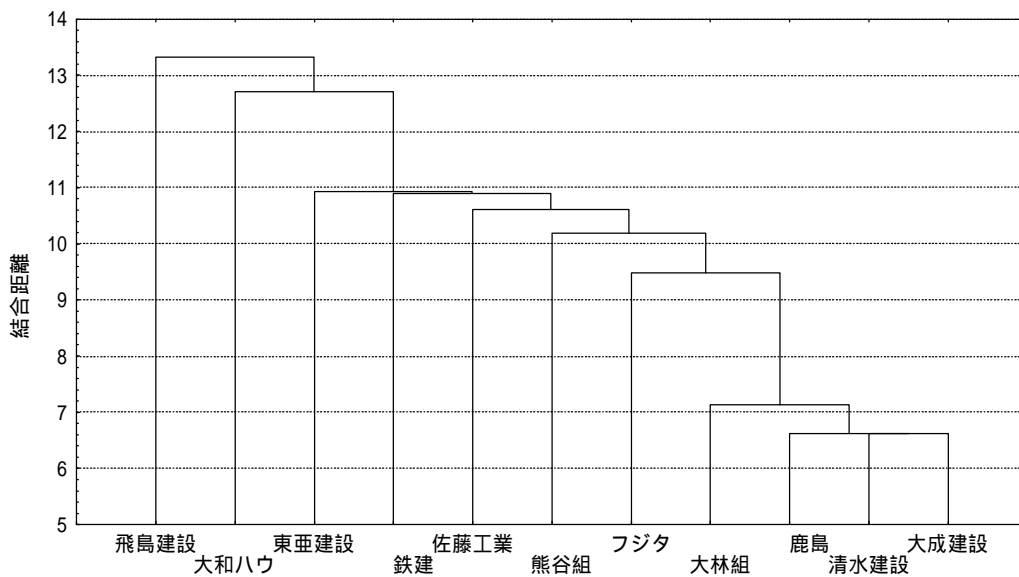
K-means 法クラスター分析の結果として、k 個のクラスターがどの程度はっきりしたものであるかを測るために、各次元の各クラスターに対する平均を調べる。理想的には、分析している次元ほとんどで平均が異なることが望ましい。各次元に対して行われた分散分析の F 値の値は、クラスター間の判別を各次元がどの程度行っているかを示す尺度となる。

### 5.個別銘柄のグループ分類

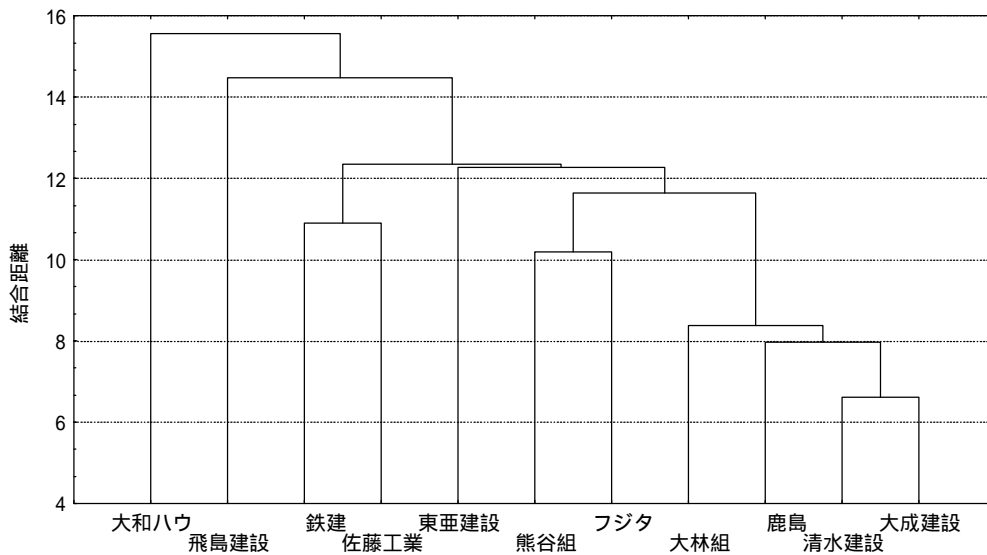
以下のデンドログラムは 1977 年 7 月～1993 年 6 月までの 192 ヶ月間における日経 225 採用銘柄のうち、建設業（11 銘柄，92 年 1 月時点）によるクラスター分析の結果である。それぞれ距離尺度としてユークリッド距離を用い、最短距離法，最遠距離法，群平均法，ウォード法を用いてクラスタリングした。また，収益率はあらかじめ基準化したものを用いる。

目立った違いとしては(1)最短距離法ではチェーン状にクラスターが形成されたこと，(2)ウォード法では早期に幾つかのクラスターが形成されやすいこと，(3)すべての方法において，鹿島，大林，清水，大成は類似した動きをしていること，(4)大和ハウス，飛鳥建設は他とは異なる動きをしていることなどが指摘できる。

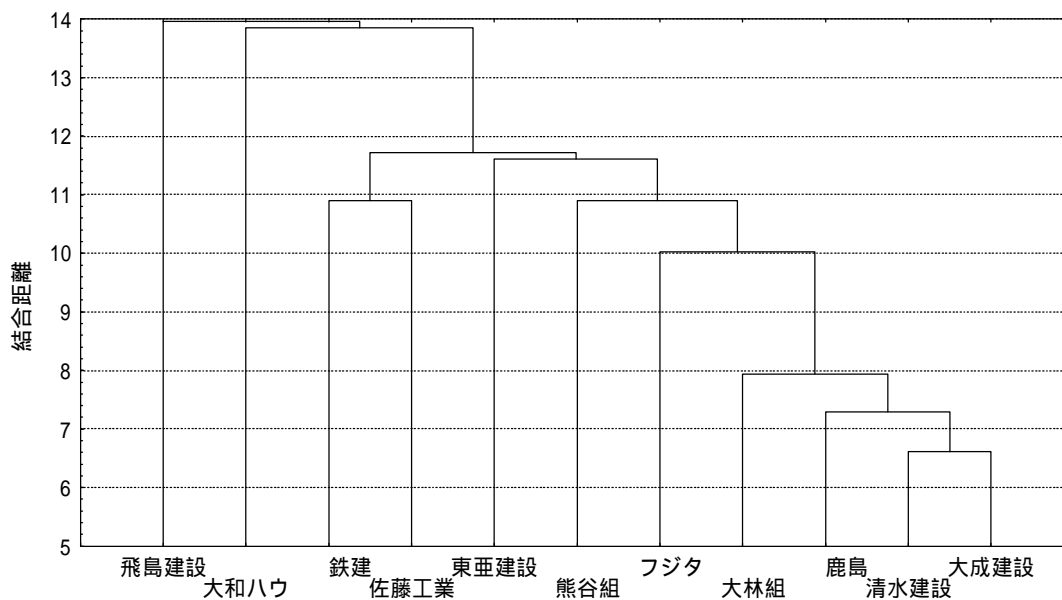
樹状図 11 変数  
最近隣法  
ユークリッド距離



樹状図 11 変数  
最遠隣法  
ユークリッド距離



樹状図 11 変数  
群平均法(UPGMA)  
ユークリッド距離



樹状図 11 変数  
ワード法  
ユークリッド距離

